

1st Annual Conference of Computational Literary Studies | Darmstadt 2022

CCLS2022 Conference Reader

Session 1

- Judith Brottrager, Annina Stahl, Arda Arslan, Ulrik Brandes, Thomas Weitin: [Modeling and Predicting Literary Reception – A Data-Rich Approach to Literary Historical Reception](#)
- Marijn Koolen, Julia Neugarten, Peter Boot: [»This book makes me happy and sad and I love it« – A Rule-based Model for Extracting Reading Impact from English Book Reviews](#)

Session 2

- Leonard Konle, Anton Ehrmanntraut, Thora Hagen, Fotis Jannidis, Simone Winko, Merten Kröncke: [Modeling and Measuring Short Text Similarities – On the Multi-Dimensional Differences between German Poetry of Realism and Modernism](#)
- Chiara Palladino, Farnoosh Shamsian, Tariq Yousef: [Using Parallel Corpora to Evaluate Translations of Ancient Greek Literary Texts](#)
- Keli Du, Julia Dudar, Christof Schöch: [Evaluation of Measures of Distinctiveness – Classification of Literary Texts on the Basis of Distinctive Words](#)

Session 3

- Almas Abdibayev, Daniel Rockmore, Yohei Igarashi, Allen Riddell: [Limericks and Computational Poetics: The Minimal Pairs Framework – Computational Challenges for Poetic Analysis and Synthesis](#)
- Melanie Andresen, Benjamin Krautter, Janis Pagel, Nils Reiter: [Who Knows What in German Dramas? A Composite Annotation Scheme for Knowledge Transfer – Annotation, Evaluation, and Analysis](#)

Session 4

- Anna Mareike Weimer, Florian Barth, Tillmann Dönicke, Luisa Gödeke, Hanna Varachkina, Anke Holler, Caroline Sporleder, Benjamin Gittel: [The \(In-\)Consistency of Literary Concepts – Formalising, Annotating and Detecting Literary Comment](#)
- Julian Schröter, Keli Du: [Validating Topic Modeling as a Method of Analyzing Sujet and Theme](#)
- Evelyn Gius, Michael Vauth: [Towards an Event Based Plot Model – A Computational Narratology Approach](#)

Session 5

- Heejoung Shin, Ted Underwood: [Analyzing the Positive Sentiment Towards the Term “Queer” in Virginia Woolf through a Computational Approach and Close Reading](#)
- Yvonne Völkl, Sanja Saric, Martina Scholger: [Topic Modeling for the Identification of Gender-specific Knowledge – Virtues and Vices in French and Spanish 18th Century Periodicals](#)

Judith Brottrager  ¹

Annina Stahl  ²

Arda Arslan ²

Ulrik Brandes  ²

Thomas Weitin  ¹

1. LitLab, Technical University Darmstadt, Darmstadt.

2. Social Networks Lab, ETH Zurich, Zurich.

Keywords:

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

Licenses:

This article is licensed under:



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 catalogues, 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

1

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, pp. 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, Stahl, and Arslan 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 temporarily as well as 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 light fiction and high literature in German literary history (Heydebrand and Winko 1996, pp. 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 judgements and circulating library catalogues as proxies for audiences’ interests. 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. In the discussion (Section 7), we will illustrate how the theoretical framework of historical evaluation is reflected in our results.

2. Previous Work

59

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 (Underwood and Sellers 2016; Porter 2018; Underwood 2019; Algee-Hewitt, Allison, et al. 2016), where these categories reflect to some degree reception-related aspects: In their paper “The Longue Durée of 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, pp. 323–325, see also Underwood 2019, pp. 68–110). Algee-Hewitt et al. (2016) similarly determine their investigated texts’ 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, which they model as the combination of the number of reprints and translations (2016, p. 3). Capturing modern readers’ responses, Porter (2018) constructs a score representing the popularity of authors by combining metrics taken from Goodreads (the number of ratings, the number of reviews, and the author’s average rating). Analogous to Algee-Hewitt et al. (2016), prestige is determined by counting MLA entries (2018, pp. 3–4).

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, p. 11).

When discussing the historical specificity of literary reviews and their implicitly marked registers (2019, p. 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, p. 62), a verbal value judgement can be defined as an illocutionary act of utterance through which an object is ascribed an attributive value. This attributive value in turn links

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

back to a defined value system. 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, pp. 111–131, 134–162).

In addition to verbal value judgements, 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 all levels of the literary system, from a publisher's acceptance of a manuscript to a reader's individual buying decision (1996, p. 79). Selective decisions by literary critics² 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, p. 99).

Similar to our previous work on the issue of canonization (Brottrager, Stahl, and Arslan 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; Schröter et al. 2021; Pichler and Reiter 2021). This work is also in line with a turn towards creating and publishing datasets and corpora, either as full text repositories (e.g. Odebrecht, Burnard, and Schöch 2021) or as deduced text formats (e.g. Schöch et al. 2020).

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, p. 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, p. 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

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

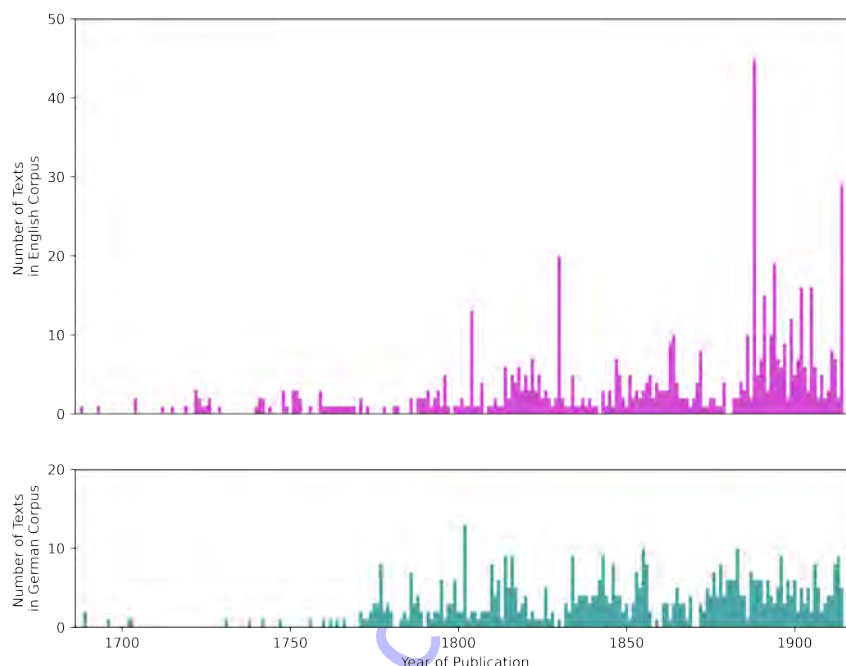


Figure 1: Temporal distribution of texts in our corpora

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, p. 4) of already digitized material with a “made” list (2015, p. 15) of, if not commissioned, but still purposely gathered texts. To capture the essence of normative best-of lists, we used exclusive narrative literary histories and anthologies. Lists of popular literature and marginalized literary texts were reconstructed by including specialized sources (e.g. companions to literature by women authors and literature from geographical peripheries, sources on light fiction and popular genres) 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.³ For texts not already available as digitized full-text, we looked for high-quality scans or scanned and retro-digitized them ourselves. As Algee-Hewitt et al. (2016, p. 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.

3. 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)

The compilation resulted in an English corpus of 605, and a German corpus of 547 texts. 155
The temporal distribution of publication dates in both corpora is shown in Figure 1. 156
In both corpora, the number of texts increases around 1770, which corresponds to 157
historically informed expectations linked to the rise of the novel in both English and 158
German literary history. Later spikes in the English corpus are primarily caused by the 159
inclusion of collections of (short) stories, which are incorporated as individual texts. 160

4. Complementary Data

161

To be able to model literary evaluation as described by Heydebrand and Winko (1996), 162
we expanded our dataset to include representations of verbal value judgements and read- 163
ers' selective choices. While verbal value judgements are directly preserved in historical 164
reviews, the reconstruction of readers' choices is not as straightforward. Transferring 165
Heydebrand and Winko's idea of the buying decision to the time frame in question 166
seems impractical because particularly for earlier time periods covered by our corpora, 167
reliable sales numbers are not available. Additionally, we wanted to introduce a measure 168
that explicitly encapsulate a text's popularity with lay audiences in contrast to expert 169
opinions recorded in reviews, and historically, buying books was simply not the way the 170
majority of readers accessed their reading materials. Here, entries in circulating library 171
catalogues seem to be a better suited proxy: Circulating libraries relied heavily on the 172
popularity of the items they advertised and had to adapt to audiences' preferences in 173
order to remain profitable (E. H. Jacobs 2003), which makes the existence of catalogue 174
entries a suitable representation of a text's popularity. 175

4.1. Reviews

176

In both the English and German-speaking Europe, the rise of literary periodicals coin- 177
cides with the commercialization of the literary market (see Italia 2012), which lead to 178
an exponential growth of available reading material and a resulting need for selection. 179
As a consequence, literary periodicals can be seen as structuring devices (Plachta 2019) 180
that place the reviewed texts along a gradient from well to poorly received, but also 181
distinguish between texts that were interesting enough to be reviewed and the remain- 182
ing mass of texts that were published at the same time. In addition to reviews being 183
written by professional readers, numerous influential publications were directly linked 184
to central figures of the literary sphere: Authors such as August Friedrich Kotzebue and 185
Tobias Smollett, for example, acted as founders and editors of the *Blätter für literarische* 186
Unterhaltung and *The Critical Review*, respectively. This direct involvement of authors as 187
professional reviewers (see Heydebrand and Winko 1996, pp. 188–210) further accentu- 188
ate the difference between evaluations by (peer) experts and popularity with broader 189
audiences, as it is recorded in circulating library catalogues described below. 190

Due to the sheer number of literary journals published in the time span covered by our 191
corpora, the selection of representative journals is based on considerations of influence 192
and outreach, but also availability. For the English dataset, we were able to rely on 193

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 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, pp. 1–134). By lending books to people who, as Gamer puts it, “would never have considered buying fiction” (2000, p. 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, pp. 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*,

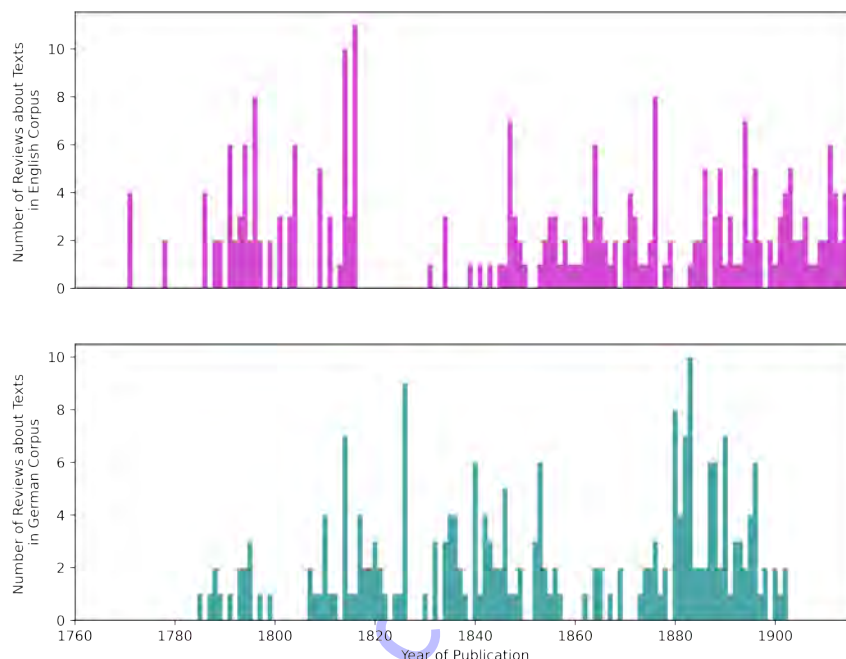


Figure 2: Temporal distribution of reviewed texts

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

Analogously to our approach to literary journals, the selection of specific catalogues 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 catalogues (Martino 1990, pp. 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 catalogues 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 catalogue of *W. Storry’s General Circulating Library* (York), the 1829 catalogue of *Hookham’s Library* (London), and two catalogues (1873 and 1907) for *Mudie’s Select Library* (London). Due to the municipal library of Vienna’s digital research focus on library catalogues, the German dataset is heavily skewed towards Viennese libraries and includes the 1790 and 1812 catalogues of rentable books at *Johann Georg Binz’s* bookstore, *Carl Armbruster’s* 1813 catalogue, *J. August Bachmann’s* 1851 catalogue, *Friedrich Gerold’s* 1850 catalogue, and the 1901 catalogue of the *Literatur-Institut Ludwig und Albert Last*. Linking our corpus texts with entries in these catalogues 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 catalogue.

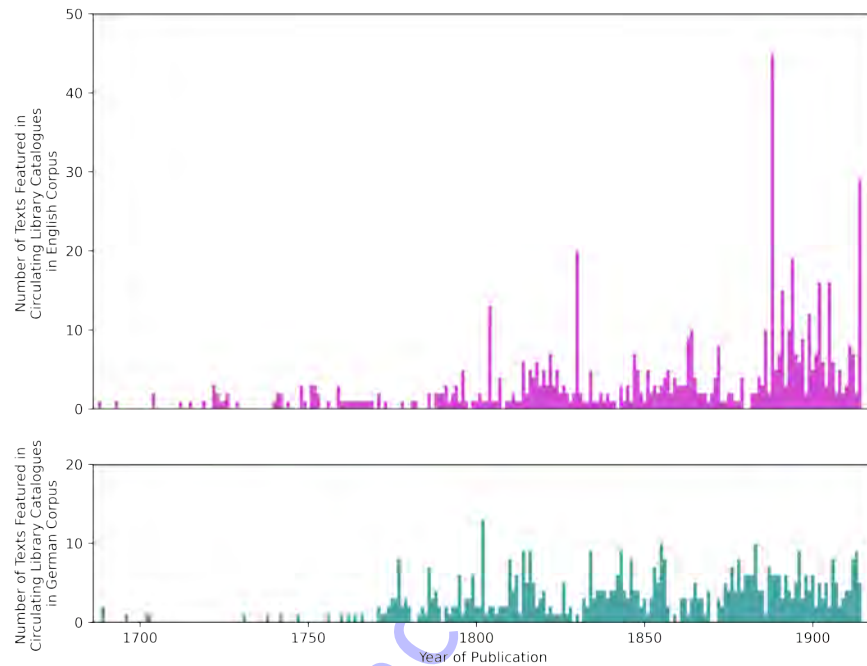


Figure 3: Temporal distribution of entries in circulating library catalogues

Of all 1,153 novels and narratives in our corpora, 763 were referenced in at least one of the catalogues 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 catalogue. 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 catalogues 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 catalogue 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 catalogue 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 by using sentiment analysis. We employed both an established and a custom sentiment analysis tool and assigned a sentiment score to each review. Then, we identified and extracted textual

features from our corpus texts that represent a text’s lexical and syntactic complexity 273
and its distinctiveness within the corpus. Based on these features and the reception data, 274
we trained a regression model to predict the sentiment scores of reviews, and classifiers 275
to predict the popularity with both reviewers and lay audiences. 276

5.1. Evaluative Language in Reviews 277

As described above, a basic sentiment analysis alone often fails to detect differences 278
between historical reviews (Du and Mellmann 2019). This is partially due to the tools 279
being designed for modern language usage, but also due to specificities of evaluative 280
language in literary reviews. When examining the collected reviews, it becomes apparent 281
that especially negative reviews are often quite vague in their criticism and balance 282
out criticism by mentioning minor positive aspects. Additionally, the reviews differ 283
significantly in length—some of them consist of only a few sentences, while others 284
span over several pages, featuring detailed plot synopses. Unsurprisingly, tools such 285
as TextBlob (Loria 2018) and its extension for German, textblob-de (Killer 2019), are 286
often not able to detect these subtleties. In a preliminary experiment with a test set of 287
15 positive and negative reviews for each dataset, TextBlob correctly identified all 15 288
positive English reviews and 13 positive German reviews, but only 8 negative English 289
and 6 negative German reviews. With precision rates of 68.18 and 59.09 percent, we 290
decided to implement an alternative approach using word embeddings to define the 291
positive and negative poles of evaluative language in the specific context of historical 292
reviews. 293

From a linguistic point of view, the evaluative language to be detected is an instance of 294
appraisal (Halliday and Matthiessen 2014, Martin and White 2007). To be able to include 295
not only explicit evaluative expressions on the word level (e.g. ‘this is an excellent novel’) 296
but also more implicit forms of appraisal (e.g. the positive connotation of ‘Gestalt’ and 297
negative connotation of ‘Geschöpf’ described by Du and Mellmann 2019, p. 13) we 298
ascribe words a value that represents their similarity to explicit evaluative expressions 299
by calculating their distances in word embeddings. 300

Adapting an approach to sentiment analysis suggested by Jacobs (2019), we define the 301
reference points by using what Jacobs calls “label words”. However, in contrast to Jacobs 302
who uses a theoretically and empirically tested set of emotion words, we use manually 303
compiled lists of evaluative words that stood out as especially positive or negative in a 304
close reading of a sample set of reviews.⁴ 305

4. 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, dürrig, seltsam, schädlich, unfertig, frech, enttäuschung, schwäche, tadel, simpel, übertrieben, überflüssig, fehler, niedrig, grauenhaft, umständlich, oberflächlich, mittelmäßig, unnatürlich*

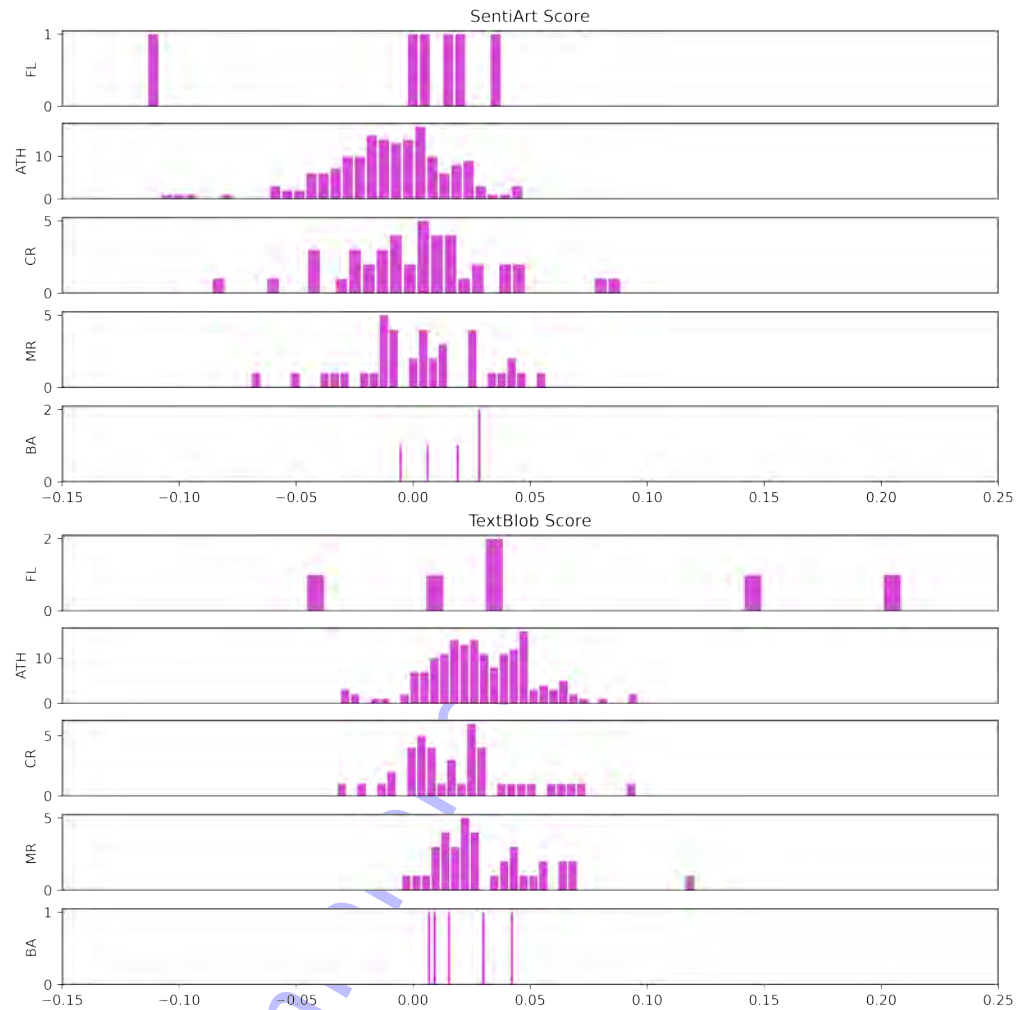


Figure 4: Distribution of sentiment scores across reviews in English journals

We then generated word2vec embeddings (Mikolov et al. 2013) for both languages, 306
 using the corpora and reviews as textual basis. For each manually determined label 307
 word, we added the words that were the most similar in the word embeddings⁵ to the 308
 respective lists of positive or negative label words. Then, we filtered both the label words 309
 and newly added similar words according to the following criteria: With our approach, 310
 a focus on evaluative language on the word level seems most practicable, which is why 311
 we excluded all word classes but adjectives and nouns. As an additional prerequisite, we 312
 only included words whose relative frequency in the reviews is higher than their relative 313
 frequency in our corpora. By doing so, we model the particular register of reviews and 314
 thus exclude words used in the plot descriptions. Finally, to ensure some degree of 315
 generalizability, we only included words that belong to the 10,000 most frequent nouns 316
 and adjectives in all reviews. 317

After applying these limitations to the lists, we performed an affinity propagation 318

5. 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.

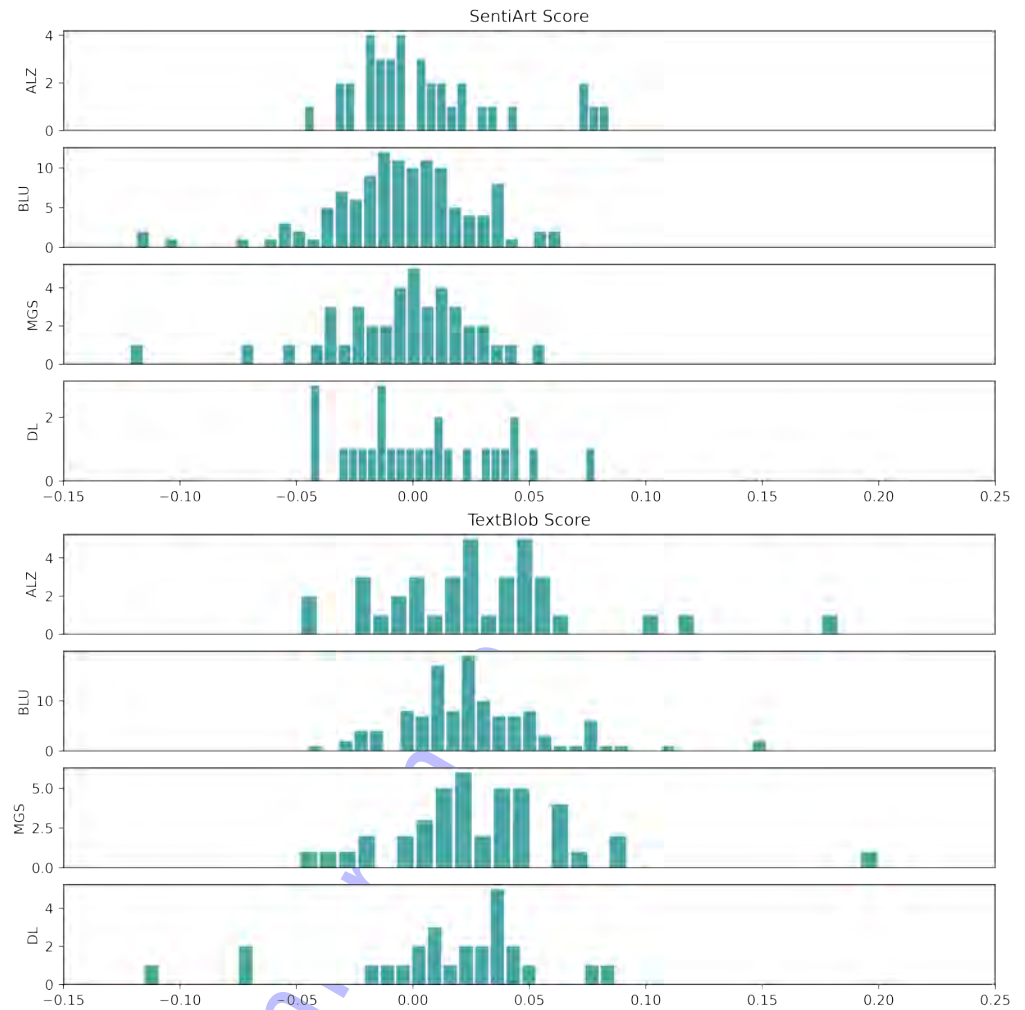


Figure 5: Distribution of sentiment scores across reviews in German journals

clustering algorithm for both positive and negative evaluative words. This is necessary 319
because all evaluative words are relatively close to each other in the word embeddings 320
and combining positive and negative words helps to identify stable and unambiguous 321
clusters. Then, we manually chose the most representative clusters to define the positive 322
and negative poles of evaluation, represented by the centroid of each of these clusters. 323
Based on the centroids, we calculated the cosine similarities between the positive and 324
negative clusters and each word that belongs to the 10,000 most frequent adjectives or 325
nouns more typically used in reviews. By subtracting the normalized sum of the negative 326
similarities from the normalized sum of the positive similarities, we then determine 327
whether a specific word is closer to the positive or negative cluster centroids. 328

Compared to TextBlob, our *ad hoc* SentiArt approach performs better at recognizing 329
negative reviews: 12 out of 15 English and 10 out of 15 German negative reviews in the 330
test set were attributed correctly. However, the SentiArt implementation performs worse 331
for positive reviews, correctly identifying 9 positive English and 8 positive German 332
reviews. The distributions across reviews from different journals in Figure 4 and 5 333

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 that surveyed the use of 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. 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 lead to some loss of data, since we did not include 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 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 a text as one 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 the two datasets in two ways: For the document plus averaged chunks dataset, the document-based dataset was left unchanged and combined with an average across the chunks of a text in the chunk-based dataset. In the other dataset, called the chunks plus copied document dataset, the document-type features were added to the chunk-based dataset of each respective text.

On the level of characters, we included the ratio of various special signs (punctuation marks, whitespaces, digits, 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-, bi-, and trigram entropy.⁶

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

6. Entropy is a measure of the information content of a sequence of symbols (Baeza-Yates and Ribeiro-Neto 1999; Bentz et al. 2017). If one symbol makes up the majority of the sequence and the other symbols have a very low frequency, 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.

Table 1: Text Features

	Chunk-type	Document-type
Character	Character frequency Ratio of punctuation marks Ratio of whitespace Ratio of digits Ratio of exclamation marks Ratio of question marks Ratio of commas Ratio of uppercase letters	
Lexical	Type-token ratio n -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	
Semantic		Intra-textual variance Stepwise distance Outlier score Overlap score
Syntactic	Tag distribution Production rule distribution Tag unigrams Tag bigrams Tag trigrams	
Text Length	Average number of words per sentence Max. number of words per sentence Average word length Average paragraph length Chunk text length	
Other	Flesch reading ease score	

story—a n -gram had to occur in at least two corpus texts to contribute to the distance. 371
 We also added a second version of the unigram corpus distance, where a word had to 372
 occur in at least 5 percent but no more than 50 percent of the documents, with the goal of 373
 finding words that are particular to selective writing styles. To account for the semantic 374
 complexity of a text, we used four measures introduced by Cranenburgh, Dalen-Oskam, 375
 and Zundert for computing different concepts of distance between the chunks of a 376
 text (2019). We calculated each of them with both document embeddings (Le and 377
 Mikolov 2014) and sentence BERT (SBERT) embeddings (Reimers and Gurevych 2019). 378
 Intra-textual variance measures how similar the individual chunks are to the average of 379
 all chunks making up a document, the centroid, while stepwise distance is a measure 380
 of the distance between successive chunks. The outlier and overlap scores look at the 381
 similarity to other works in the corpus. The former is the smallest distance between 382
 the centroid and another document’s centroid, while the latter is the share of chunks 383
 belonging to other documents among the k chunks that are nearest to the centroid, with 384
 k being the number of chunks in the text. 385

We also included features on the syntactic text level. Using the natural language process- 386
 ing library spaCy for Python, we tagged the words in each text with their part-of-speech 387
 (POS) and counted the number of single tags as well as the number of two or three tags 388
 occurring subsequently, here called the tag bigrams and tag trigrams. “ADJ-NOUN- 389
 VERB”, for example, is such a tag trigram, which means that an adjective followed by a 390
 noun, which is then followed by a verb, occurs in the text. Due to the number of possible 391
 combinations, we included only the frequency of the 100 most common tag n -grams. 392
 The production rule distribution served as another syntactic feature, but is available 393
 only for the English texts. A production rule is the pattern according to which one 394
 grammatical part of a sentence is followed by another part. We used NLTK, a different 395
 Python NLP library, and included the frequency of the 100 most common production 396
 rules. 397

The average word length, the average and maximum number of words in a sentence, the 398
 average length of a paragraph and the text length of a chunk are measures for the general 399
 complexity of the text. The Flesch reading ease score accounts for how challenging it is 400
 to read a text (Flesch 1948). Previous research has found a negative correlation between 401
 readability and literary success (Ashok, Feng, and Choi 2013). 402

5.3. Prediction 403

To test if the review sentiment is dependent on text features, we ran a regression pre- 404
 dicting review sentiment. Further, we trained two classifiers: The first one predicted 405
 whether a review to a work had been written or not, the second one determined if the 406
 review sentiment was positive, neutral, or negative. Finally, we ran a classifier predicting 407
 if a text had been added to a library catalogue. 408

5.3.1. Cross-validation

409

Since we had the choice of different models, features, and model parameters, we ran a cross-validation to find the combinations of options that achieved the highest performance for each of the four prediction tasks.

For the classifications, we implemented two different classifiers, XGBoost and SVM, further detailed in Section 5.3.3. Then, we tested whether the document-based dataset, the chunk-based dataset, or one of the combinations of the two performed best in combination with the models. To avoid 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 these features amounted to 100 columns in the dataset. While we aggregated the scores from TextBlob and our modified SentiArt approach for classification, for the regression we tested each score individually and a combination of the two. For the SVM classifiers, we also tested different options for the regularization parameter C .

We implemented a 10-fold cross-validation for regression, meaning that we split the data into 10 folds of approximately equal size, and trained the models 10 times on 9 of the datasets combined, leaving out a different dataset each time and using it for evaluating the model. All works written by an individual author were put into the same fold to avoid overfitting to an author's writing style instead of learning the textual features that might be connected with the positive reception of the text. We only used 5 folds for the classifications, because the number of negatively reviewed texts in our dataset was too small to be spread over more folds. Instead, we implemented a stratified cross-validation where each fold had approximately the same number of texts from each class, so that all classes were represented in both the training and test set.

5.3.2. Regression

433

We ran separate regressions for the TextBlob and SentiArt-generated scores. If a text had multiple reviews, we assigned 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 next section (Section 5.3.3), 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 used XGBoost, a Python machine learning library that is based on decision trees, as the prediction model, and tested it with different combinations of features and feature levels as described in Section 5.3.1. For evaluating the performance of the model, we calculated the correlation between the true and the predicted labels with Pearson's r . The Python library SciPy automatically calculates the p -value along with the correlation coefficient. The p -value of each model tested in the cross-validation was then calculated by taking the harmonic mean of the p -values of the individual folds (Wilson 2019).

5.3.3. Classification

449

Besides predicting the review sentiment from the texts, we also 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 using a binary variable indicating if a text had generated a review or not, the size of the dataset increased, because we could also include texts that had not been reviewed. 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. Finally, we ran another two-class classification that predicted if a work had appeared in a circulating library catalogue.

We tested two classifiers, XGBoost adapted for classification and SVM from the Python machine learning library scikit-learn, for both two- and multi-class classification. SVMs are algorithms that try to fit a hyperplane that separates the data points belonging to different classes. We included the choice of the optimal regularization parameter C of the SVM in the cross-validation, testing values between 0.1 and $10'000$. We only used the document-based dataset and the 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 chunk-level classification would be even more difficult to interpret for multi-class classification, 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 labelled as either positive or negative and all others as neutral⁷ (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, which was the case for 6 texts in the English corpus and for 3 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. The oldest reviewed texts in the corpus were published in 1771 for English and in 1785 for German. We only included texts published during and after the respective years so that texts that had no chance of being reviewed because they were published too early did not distort the classification. We also only included works that were younger than the first works that were part of a circulating library catalogue for the same reason.

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 multi-class classification and negatively reviewed texts were especially underrepresented. The number of reviewed texts in each class after

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

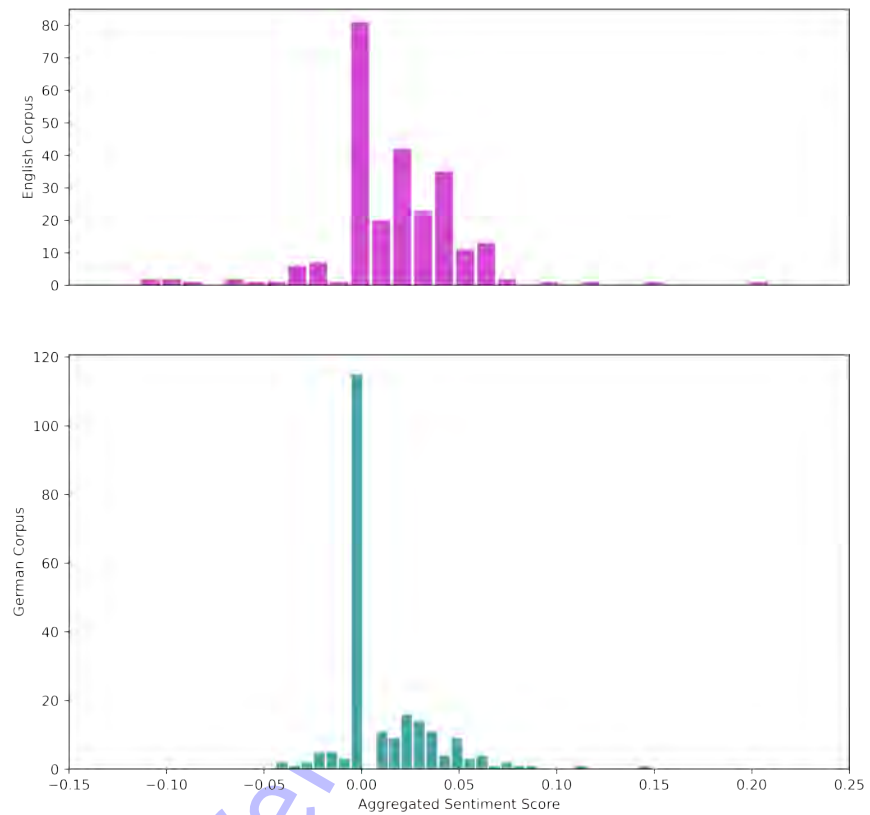


Figure 6: Distribution of aggregated sentiment scores for both corpora

filtering for publication years is shown in Table 2. The majority of English texts were included in a library catalogue, which is why this dataset is also imbalanced (see Table 3, again filtered for publication year). As mentioned in Section 5.3.1, we used a stratified cross-validation to make sure that each training and test set contained all classes, and adapted the evaluation metrics to account for class imbalance.

The two two-class classifications were evaluated based on accuracy, which is the number of instances where the predicted label is correct, divided by the total number of samples. Using balanced accuracy instead to account for class imbalance did not improve the result. The evaluation metric used for multi-class classification was the F1 score, the harmonic mean between precision and recall. Scikit-learn’s implementation of the F1 score has several options for averaging over the F1 scores of each class to calculate the final F1 score. Because of the class imbalance we used the ‘macro’ option, which gives equal weights to each class.

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 catalogues

	English	German
Not Featured	146	240
Featured	457	306

6. Results

501

6.1. Regression

502

The highest significant correlation coefficient from the cross-validation, or the highest coefficient if none was significant, are reported in Table 4. In Figure 7, the true and the predicted scores are plotted against each other. Running a regression with the scores from each of the two tools separately delivered small but significant correlation coefficients. The sentiment scores from the SentiArt approach can be predicted from text features to a small extent, while the correlation coefficients for the TextBlob scores were around 0. Running the regression using the combined scores led to correlation coefficients of around 0 that were not significant.

510

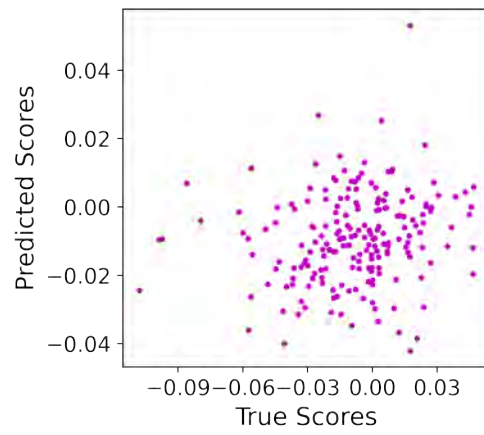
Table 4: Regression results

	English	German
SentiArt	0.233**	0.198**
TextBlob	-0.01*	0.049**
Combined	0.131	0.074

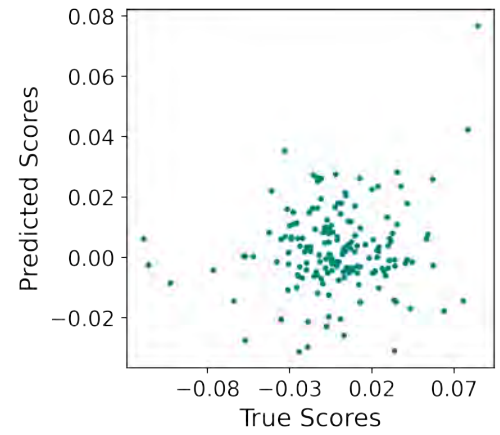
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

The cross-validation showed that using the document-based dataset and dropping the POS features was the best choice for both languages when working with the SentiArt scores, while using the chunk-based dataset for English and the document plus averaged chunks dataset for German was better for TextBlob.

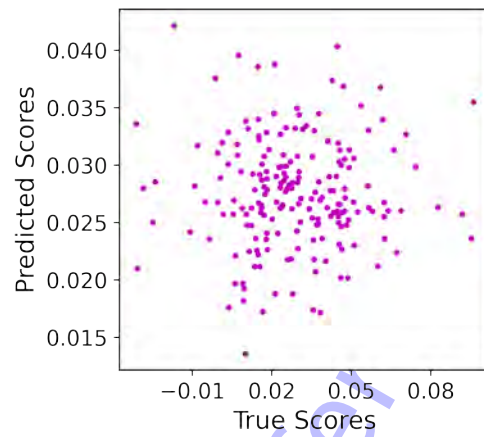
514



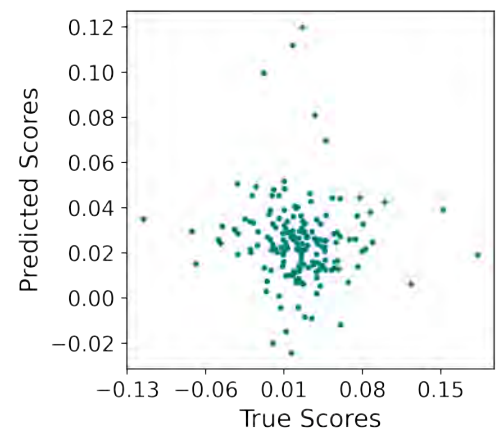
(a) English, SentiArt



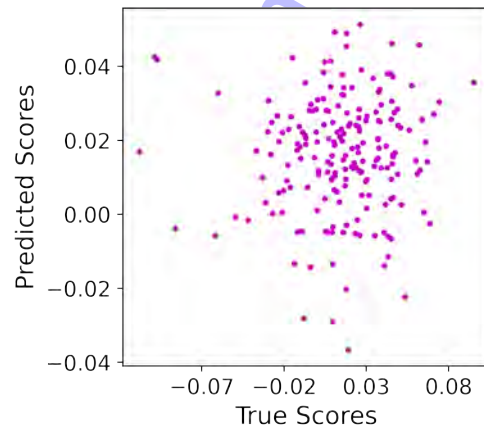
(b) German, SentiArt



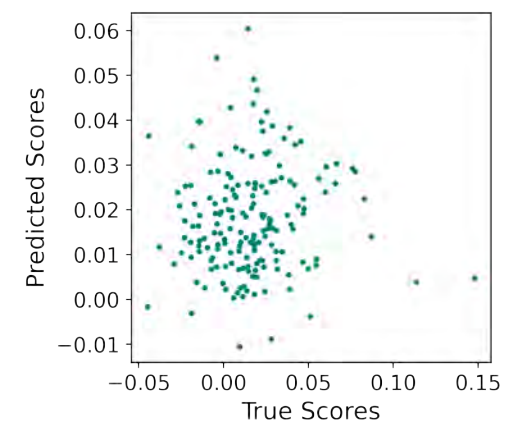
(c) English, TextBlob



(d) German, TextBlob



(e) German, Combined



(f) German, Combined

Figure 7: Sentiment scores and predicted scores

6.2. Classification

515

6.2.1. Reviewed/not reviewed

516

Using XGBoost as the classifier and dropping the POS features was the best choice for both languages when classifying texts according to whether they had been reviewed or not. In combination with the document plus averaged chunks dataset, the model achieved an accuracy of 0.715 on the English texts, and an accuracy of 0.638 on the German texts when using the document-based dataset.

The crosstabs below show how many texts from each class were predicted to be a specific class.

Table 5: Crosstab for reviewed/not reviewed classification, English

		Predicted		Total
		Not reviewed	Reviewed	
True	Not reviewed	239	126	365
	Reviewed	56	135	191
	Total	295	261	556

Table 6: Crosstab for reviewed/not reviewed classification, German

		Predicted		Total
		Not reviewed	Reviewed	
True	Not reviewed	251	79	330
	Reviewed	103	70	173
	Total	354	149	503

6.2.2. Multi-class Classification

524

For multi-class classification, Xgboost was the best choice of model for both languages, achieving a F1 score of 0.390 for English by using the document-based dataset, and by dropping the POS features. For German, a F1 score of 0.305 was reached by using the document-based dataset without the tag distribution feature.

Table 7: Crosstab for multi-class classification, English

		Predicted			
		Not reviewed	Negative	Neutral	Positive
True	Not reviewed	238	5	41	81
	Negative	6	1	4	4
	Neutral	15	4	22	22
	Positive	29	6	34	44
	Total	288	16	101	151

Table 8: Crosstab for multi-class classification, German

		Predicted			
		Not reviewed	Negative	Neutral	Positive
True	Not reviewed	239	8	47	36
	Negative	7	1	1	1
	Neutral	57	0	16	13
	Positive	35	2	18	22
	Total	338	11	82	72

6.2.3. Library Catalogues Classification

529

For classifying if an English text had been added to a library catalogue, Xgboost was the best choice of classifier, along with using the document plus averaged chunks dataset and dropping the parts-of-speech features. This combination achieved an accuracy of 0.676. However, using SVM with regularization parameter $C = 10'000$ performed better than XGBoost on the German texts. The best choice of features was the document-based dataset, and dropping either the POS or no features resulted in the exact same crosstab and accuracy of 0.590.

530

531

532

533

534

535

536

Table 9: Crosstab for library catalogues classification, English

		Predicted	
		Not featured	Featured
True	Not featured	55	91
	Featured	115	342
	Total	170	433

Table 10: Crosstab for library catalogues classification, German

		Predicted		
		Not featured	Featured	Total
True	Not featured	150	90	240
	Featured	134	172	306
	Total	284	262	546

7. Discussion

537

The low correlation coefficients of the best model with 0.233 for English and 0.198 for German texts show that even with our adapted SentiArt approach, there is only a weak correlation between the measured sentiment in reviews and textual markers of the reviewed texts. We did not further analyze the contribution of individual features due to this weak effect. 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.

545

With the highest correlations between text features and sentiment scores being achieved by using our own *ad hoc* sentiment analysis approach instead of the established TextBlob tool, we conclude that by taking into account typical characteristics of historical literary reviews, as, for example, the implicitness and vagueness of negative comments and by constructing a register more commonly used in reviews than in narratives and novels, our approach was more adept at identifying the particularities of evaluative language in reviews. This finding demonstrates that established methods, including but not limited to sentiment analysis, have to be adapted to the time period and the peculiarities of the source material.

554

Despite the imbalanced data, the models differentiated between texts with reviews and texts without reviews with an accuracy of over 0.7 for English and 0.6 for German without predicting the majority class for all labels. 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, p. 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.

562

The connection between popularity and text characteristics seems to be similar, since the accuracy scores for differentiating between texts featured in circulating library catalogues and others are close to those for predicting if a text had been reviewed or not, even though the class imbalance is even bigger for the English dataset. The fact that a text had been added to a library might be viewed as a similar indicator of interest by a

567

broad audience. There seem to be detectable textual qualities that spark interest among the public in the first place.

The higher accuracy for English library catalogues can be linked back 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 catalogues to include contemporary texts, which can be assumed to be stylistically more homogeneous. Moreover, the last two catalogues surveyed are from the same library, Mudie's, whose owner Charles Edward Mudie has been claimed to only advertise books that satisfied his personal moral and literary standards (Katz 2017, Roberts 2006). Assuming that these factors lead to more quantitatively detectable similarities within the set of texts advertised in library catalogues, a higher accuracy seems plausible.

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. The exemplified workflow has proven to be productive. 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.

Based on this data-rich literary history dataset, predicting review sentiment from texts alone proved to be successful only to a limited extent. Historical literary data is scarce, and a larger dataset might have led to different results. However, the small but significant correlation between the sentiment scores calculated with our SentiArt-inspired approach and the scores predicted by our models show that a text's rating by reviewers can be explained to some extent by the texts themselves. 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. Similarly, the classification of texts featured in circulating library catalogues proved to be comparatively accurate, suggesting that popular texts share certain textual qualities.

In our future work, we plan to include additional data in order to produce more reliable and generalizable results. This means on the one hand that we will add additional journals and circulating library catalogues to our dataset, but will also work on alternative operationalizations of a text's popularity and proliferation.

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 with the same textual features.

Therefore, it seems also reasonable to add a time component to our SentiArt approach to evaluative language, for example, by extracting period-specific evaluation words and computing period-specific evaluation scores. As the SentiArt approach has proven to be useful, we will work on fine-tuning the word embeddings to increase the approach's accuracy in the detection of positive reviews.

For the analysis of the corpus texts, we plan to include text representations with embeddings as separate features and not just as the basis of the already included semantic features. The high number of dimensions of the dataset due to the POS and production rule distribution features, as well as the planned embeddings, will be addressed with suitable dimensionality reduction.

So far, we excluded texts that had contradicting positive and negative reviews from the classification, which led to the underrepresented class of negatively reviewed text being even sparser. In a next step, we will consider all texts that had any negative reviews as negatively reviewed without considering the number of positive and neutral reviews. This step is also justified due to the general tendency of reviewers to give positive reviews and to attenuate negative criticism.

9. Data and Code Availability

The scripts are available at https://github.com/sta-a/jcls_reception; corpora, reviews, metadata, trained word embeddings, and sentiment scores can be accessed via <https://figshare.com/s/98d85345c50d0594bb59>.

10. Data availability 627

<https://figshare.com/s/98d85345c50d0594bb59> 628

11. Software availability 629

https://github.com/sta-a/jcls_reception 630

12. Acknowledgements 631

This work is part of “Relating the Unread. Network Models in Literary History”, a 632
project supported by the German Research Foundation (DFG) through the priority 633
programme SPP 2207 Computational Literary Studies (CLS). Special thanks to Joël Doat 634
for his advice on the formal aspects of word embeddings. 635

13. Author contributions 636

Judith Brottrager: Formal Analysis, Data curation, Writing – original draft 637

Annina Stahl: Formal Analysis, Writing – original draft 638

Arda Arslan: Formal Analysis 639

Ulrik Brandes: Supervision, Writing – review & editing 640

Thomas Weitin: Supervision, Writing – review & editing 641

References 642

- Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, 643
and Hannah Walser (2016). “Canon/Archive. Large-scale Dynamics in the Literary 644
Field”. In: *Pamphlets of the Stanford Literary Lab* 11. URL: [https://litlab.stanford](https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf) 645
[.edu/LiteraryLabPamphlet11.pdf](https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf). 646
- Algee-Hewitt, Mark and Mark McGurl (2015). “Between Canon and Corpus: Six Per- 647
spectives on 20th-Century Novels”. In: *Pamphlets of the Stanford Literary Lab* 8. URL: 648
<http://litlab.stanford.edu/LiteraryLabPamphlet8.pdf>. 649
- Alvarado, Rafael C. (2019). “Digital Humanities and the Great Project: Why We Should 650
Operationalize Everything - and Study Those Who Are Doing So Now”. In: *Debates in* 651
the Digital Humanities 2019. Ed. by Matthew K. Gold and Lauren F. Klein. Minneapolis, 652
MN: University of Minnesota Press, pp. 75–82. doi: [10.5749/j.ctvg251hk](https://doi.org/10.5749/j.ctvg251hk). 653
- Ashok, Vikas, Song Feng, and Yejin Choi (2013). “Success with Style: Using Writing 654
Style to Predict the Success of Novels”. In: *Proceedings of the 2013 Conference on Empir-* 655
ical Methods in Natural Language Processing. Ed. by Association for Computational 656
Linguistics, pp. 1753–1764. 657

- Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (1999). *Modern Information Retrieval – The Concepts and Technologies Behind Search*. Harlow: Addison Wesley. 658
659
- 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, p. 275. DOI: [10.3390/e19060275](https://doi.org/10.3390/e19060275). 660
661
662
- Bode, Katherine (2018). *A World of Fiction: Digital Collections and the Future of Literary History*. Ann Arbor, MI: University of Michigan Press. 663
664
- 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*. CHR 2021. 665
666
667
Amsterdam: CEUR Workshop Proceedings, pp. 195–205. URL: [http://ceur-ws.org](http://ceur-ws.org/Vol-2989/short_paper21.pdf) 668
[/Vol-2989/short_paper21.pdf](http://ceur-ws.org/Vol-2989/short_paper21.pdf). 669
- City University London (2001). *The Athenaeum Projects*. URL: <https://athenaeum.city.ac.uk/> (visited on 04/22/2022). 670
671
- 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, pp. 625–650. DOI: [10.1007/s10579-018-09442-4](https://doi.org/10.1007/s10579-018-09442-4). 672
673
674
- Du, Keli and Katja Mellmann (2019). “Sentimentanalyse als Instrument literaturgeschichtlicher Rezeptionsforschung”. In: *DARIAH-DE Working Papers* 32. URL: <http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2019-32.pdf>. 675
676
677
- Flesch, Rudolph (1948). “A New Readability Yardstick”. In: *The Journal of Applied Psychology* 32 (3), pp. 221–33. 678
679
- Gamer, Michael (2000). *Romanticism and the Gothic: Genre, Reception, and Canon Formation*. 40. Cambridge; New York: Cambridge University Press. 680
681
- Garside, Peter (2011). *British Fiction 1800-1829: A Database of Production, Circulation & Reception*. British Fiction 1800-1829. URL: <http://www.british-fiction.cf.ac.uk/> (visited on 04/22/2022). 682
683
684
- 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 [u.a.]: Oxford Univ. Press. 685
686
687
- GJZ18 (2021). *Gelehrte Journale und Zeitungen der Aufklärung*. GJZ18. URL: <https://gelehrte-journale.de> (visited on 04/22/2022). 688
689
- Halliday, M. A. K. and Christian M. I. M. Matthiessen (2014). *Halliday's Introduction to Functional Grammar*. Milton Park, Abingdon, Oxon: Routledge. 690
691
- Heydebrand, Renate von and Simone Winko (1996). *Einführung in die Wertung von Literatur*. Paderborn, München, Wien, Zürich: Schöningh. 692
693
- Italia, Iona (2012). *The Rise of Literary Journalism in the Eighteenth Century: Anxious Employment*. 3. London: Routledge. 694
695
- 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, p. 53. DOI: [10.3389/frobt.2019.00053](https://doi.org/10.3389/frobt.2019.00053). 696
697
698
- Jacobs, Edward H. (2003). “Eighteenth-Century British Circulating Libraries and Cultural Book History”. In: *Book History* 6.1, pp. 1–22. DOI: [10.1353/bh.2004.0010](https://doi.org/10.1353/bh.2004.0010). 699
700

- 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*. Bd. 20. Wiesbaden: Harrassowitz. 701-704
- Jockers, Matthew Lee (2013). *Macroanalysis: Digital Methods and Literary History*. Topics in the digital humanities. Urbana: University of Illinois Press. 705-706
- Katz, Peter J. (2017). "Redefining the Republic of Letters: The Literary Public and Mudie's Circulating Library". In: *Journal of Victorian culture : JVC*. 22.3, pp. 399-417. 707-708
- Killer, Markus (2019). *textblob-de*. Version 0.4.4a1. URL: <https://textblob-de.readthedocs.io/en/latest/index.html>. 709-710
- 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)*, pp. 184-195. doi: [10.23919/FRUCT48121.2019.8981504](https://doi.org/10.23919/FRUCT48121.2019.8981504). 711-714
- Le, Quoc and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on Machine Learning*. 715-716
- Loria, Steven (2018). "textblob Documentation". In: *Release 0.15.2*. 717
- Martin, J. R. and Peter Robert Rupert White (2007). *The Language of Evaluation: Appraisal in English*. Basingstoke, Hampshire: Palgrave Macmillan. 718-719
- Martino, Alberto (1990). *Die deutsche Leihbibliothek: Geschichte einer literarischen Institution (1756-1914)*. Wiesbaden: Harrassowitz. 720-721
- Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *Computer Science*. 722-723
- Moretti, Franco (2013). *Distant Reading*. Verso. 724
- Odebrecht, Carolin, Lou Burnard, and Christof Schöch, eds. (2021). *European Literary Text Collection (ELTeC)*. COST Action Distant Reading for European Literary History. doi: [10.5281/zenodo.4662444](https://doi.org/10.5281/zenodo.4662444). 725-727
- Pichler, Axel and Nils Reiter (2021). "Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse. Eine Annäherung über Norbert Altenhofers hermeneutische Modellinterpretation von Kleists *Das Erdbeben in Chili*". In: *Journal of Literary Theory* 15.1, pp. 1-29. doi: [10.1515/jlt-2021-2008](https://doi.org/10.1515/jlt-2021-2008). 728-731
- Plachta, Bodo (2019). "Literaturzeitschriften". In: *Grundthemen der Literaturwissenschaft: literarische institutionen*. Ed. by Norbert Otto Eke and Stefan Elit. Boston, MA: De Gruyter, pp. 345-356. 732-734
- Porter, J.D. (2018). "Popularity/Prestige". In: *Pamphlets of the Stanford Literary Lab* 17. URL: <https://litlab.stanford.edu/LiteraryLabPamphlet17.pdf>. 735-736
- 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 [u.a.]: Oxford Univ. Press. 737-739
- 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, pp. 982-992. doi: [10.18653/v1/D19-1410](https://doi.org/10.18653/v1/D19-1410). 740-744

- 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)*. Valletta, Malta: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/490_Paper.pdf.
- Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, 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, p. 4853. doi: [10.3390/a9224853](https://doi.org/10.3390/a9224853).
- Roberts, Lewis (2006). "Trafficking Literary Authority: Mudie's Select Library and the Commodification of the Victorian Novel". In: *Victorian Literature and Culture* 34.1, pp. 1–25. doi: [10.1017/S1060150306051023](https://doi.org/10.1017/S1060150306051023).
- Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke (2020). "Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen". In: *ZfdG* 5. doi: [10.17175/2020_006](https://doi.org/10.17175/2020_006).
- 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, pp. 81–108. doi: [10.1515/jlt-2021-2011](https://doi.org/10.1515/jlt-2021-2011).
- Stamatatos, Efstathios (2009). "A Survey of Modern Authorship Attribution Methods". In: 60.3, pp. 538–556. doi: [10.1002/asi.21001](https://doi.org/10.1002/asi.21001).
- Underwood, Ted (2019). *Distant Horizons: Digital Evidence and Literary Change*. Chicago: The University of Chicago Press.
- Underwood, Ted and Jordan Sellers (2016). "The Longue Durée of Literary Prestige". In: *Modern Language Quarterly* 77 (3). doi: [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, pp. 1195–1200. doi: [10.1073/pnas.1814092116](https://doi.org/10.1073/pnas.1814092116).

A Rule-based Model for Extracting Reading Impact from English Book Reviews

Marijn Koolen ¹

Julia Neugarten ²

Peter Boot ²

1. DHLab, KNAW Humanities Cluster, Amsterdam.

2. Literary Studies, Huygens ING, Amsterdam.

Keywords:

reading impact, online book reviews, Goodreads, impact model

Licenses:

This article is licensed under:



Abstract.

Being able to identify and analyse reading impact expressed in online book reviews allows us to investigate how people read books and how books affect their readers. In this paper we investigate the feasibility of creating an English translation of a rule-based reading impact model for Dutch book reviews. We extend the model with additional rules and categories to measure reading impact in terms of positive and negative feeling, narrative and stylistic impact, humor, surprise, attention and reflection. We created ground truth annotations to evaluate the model and find that the translated rules and new impact categories are effective in identifying reading impact expressed in English book reviews. Additional rules are needed to improve recall and some impact aspects are hard to extract with our type of rules. When applying the model to a large set of reviews, lists of the top-scoring books in the categories show the model's prima-facie validity. Correlations among the categories include some that make sense and others that require further research. Overall, the evidence suggests this is a suitable approach for investigating the impact of books.

1. Introduction

1

Online book reviews are an important source of data for analysing how people read books and how they describe reading experiences (Holur et al. 2021). This paper builds on our earlier work (Boot and Koolen 2020) in detecting the impact of reading fiction as it is expressed in online book reviews. That paper presented a rule-based model for measuring four categories of reading impact (affective, narrative, stylistic and reflective) in Dutch-language book reviews. As these rules are language-specific, the model cannot be used on the huge numbers of English-language reviews available online. In that article, we also mentioned potential types of reading impact that the model did not capture, such as suspense, humor and surprise. In this paper, we present a model for measuring reading impact expressed in English-language book reviews. We created this

2

3

4

5

6

7

8

9

10

11

model by translating the Dutch model and adding rules for four new categories of impact: attention, humor, surprise and negative impact. To account for these new categories and refine the Dutch model, we re-categorised some rules and added more rules based on manual analysis of modes of expression in a corpus of Goodreads reviews. We analyze and validate the English model using crowdsourced ground truth annotations.

We formulate two research questions:

1. How effective is our adaptation of the Dutch model?
 - (a) Can the new impact categories we add to the model be captured in a rule-based model? Can these new categories be meaningfully identified by human annotators?
 - (b) Is adapting an existing rule-based model for use in another language a productive approach? Is our method of translating and changing rules an effective way to do this? What are the challenges and advantages of transferring knowledge or tools from Dutch to English through translation and adaptation?
2. Is a rule-based model a productive tool for assessing the impact of fiction as expressed in online book reviews? What are the advantages of a rule-based model compared to other approaches, such as machine learning?

We first discuss the impact model and explain our selection of new impact categories in Section 2. Then, we describe how we created the rules that make up the English-language mode by adapting the Dutch model in Section 3. We evaluate these adapted rules using the ground truth annotations and do an error analysis in Section 4. Human annotators recognize and distinguish categories of impact with some consistency, resulting in acceptable Inter-Rater Agreement. For several impact categories the rule-based model attains good performance in terms of precision and recall, but more ground truth data is needed to reliably validate some other categories, and for some categories more rules are needed to cover the various ways impact can be expressed. To assess the quality of our results, we use the model to detect reading impact in a large set of Goodreads reviews for a set of popular novels in Section 5. We observe, aggregated over many reviews per novel, that the results mostly meet expectations. We conclude in Section 6 with suggestions for how to improve the model, and argue that taking a rule-based approach to assessing reading impact is a productive approach that may, in future work, be supplemented with other methods and tools.

Both the annotations and the rule-set used in the current paper are publicly available.

2. Impact model and New Categories

2.1. Book Reviews

Online book reviews are increasingly used to gauge reader response to books (Rebora et al. 2019; Spiteri and Pecoskie 2016). Using online reviews for this purpose has its

problems: the reviews are not necessarily representative of all readers, they do not necessarily reflect readers' 'true' opinions and they may be fraudulent. We discuss these issues briefly in Boot and Koolen (2020). Prompted by the epistemological issues raised by one of this paper's reviewers, we want to make clear that we do not argue here that online reviews directly reflect the reading experience or necessarily provide insight into the general public's reading experiences. Instead, we posit that creating rule-based models for the detection of impact in book reviews can generate insights into what types of impact are expressed. This can give insight into the differences that exist among reviews or among reviewers, and we hypothesize that such differences also reflect differences between reading experiences that reviewers choose to report in reviews. By examining these differences, we aim to increase our understanding of how reading affects readers more generally. Not all readers are reviewers, but studying the reviewers can show aspects of reading impact that also apply to readers more generally. In that sense, reviews provide a complementary source of evidence from reader response captured through interviewing readers (G. Sabine and P. Sabine 1983; Ross 1999), or through using questionnaire data from readers on reading selected short stories and passages in a controlled setting (Nell 1988; Miall and Kuiken 2002; Koopman and Hakemulder 2015; Koopman 2016). As online reviews are a more public form of reader response than interviews and questionnaires, we will have to remain aware that differences between reviews can also be attributable to social factors.

Nonetheless, using online book reviews as data also has advantages: the texts are accessible online in a digital format and they are primarily produced by groups of readers overlooked in much traditional literary scholarship. The writers of reviews on platforms like Goodreads are around 75% female (Thelwall and Kousha 2017), and users of Goodreads represent various nationalities and ethnicities (Champagne 2020). Thus, these reviews offer diverse perspectives that much of the field of literary studies lacks. We therefore consider them a useful source of information for literary scholarship in general and reception studies in particular.

Given the brevity of most online book reviews, we do not expect our model to perfectly identify all impact expressed in individual reviews. Instead, our aim is to develop a model that can identify relationships between aggregates of reviews grouped together by features like length, book genre or author gender, and the kinds of reading experiences described in reviews. In other words, we are producing a tool that enables literary scholars to assess the impact of books or collections of books on groups of readers by comparatively analyzing the way these books are reviewed online. Even though the representation of reading experience in reviews is nowhere near exhaustive, differences between these representations can nonetheless lead to insights into the impact of reading on reviewers. Questions that we eventually hope to be able to answer include: How does the impact of the *Harry Potter* books change over the course of the series? How do readers differ in their responses, for instance by age, gender, or reading preferences? What patterns can we discern in the impact of specific genres or authors? Do reviewers review books differently depending on author gender or book popularity? Are there discernible patterns in how reviewers develop as readers?

We define impact as any effect a book has on its reader, large or small, permanent or fleeting. Following Boot and Koolen (2020) we investigate the following four categories of impact: *Reflection*, *Positive affect* and its two subcategories *Narrative feeling* and *Stylistic feeling*, as well as a number of new categories.

2.2. Existing and New Impact Categories

In the final section of Boot and Koolen (2020) we express the expectation that smaller and more clearly defined impact categories might be better suited for validation in a survey. We added four categories to our English version of the model: *Humor* as an additional subcategory of *Positive affect*, and three independent categories: *Attention*, *Surprise* and *Negative feeling*. We chose to add these categories for the following reasons:

Attention is one of the dimensions of Story World Absorption (M. M. Kuijpers et al. 2014), defined as ‘a deep concentration of the reader that feels effortless to them. As a consequence the reader can lose awareness of themselves, their surroundings and the elapse of time.’ Green and Brock (2000, p. 702) hypothesize that this feeling of absorption relates to changing beliefs and attitudes in readers. We chose *attention* as a category rather than *suspense*, although we consider the two closely related, because textual manifestations of attention can be distinguished more clearly than those of *suspense*. *Attention* is predicted in our model by terms such as ‘immersed’, ‘absorbed’ and ‘engrossed.’

Humor, perceiving events or language as humorous, is a distinctive form of appreciation, related to but separate from stylistic or narrative feeling. Defining it as a separate category might make the categories of stylistic and narrative feeling more homogeneous. Humor is also relevant for its role in introducing young people to reading (Shannon 1993).

We added *Negative feeling*, such as being bored or disappointed by a book, to help differentiate between positive and negative expressions of impact. Although some research examines the negative effects of reading (Schmitt-Matzen 2020) and a negative response to prescribed reading (Poletti et al. 2016), previous research has overwhelmingly focussed on trying to validate the hypothesis that reading is good for personal development and social behaviour (Koopman and Hakemulder 2015), while negative feelings towards reading are often overlooked.

Surprise shows engagement with a story, because surprises are unexpected story elements. Thus, experiencing surprise requires one to have expectations of a book which are subsequently defied, and these expectations are a sign of engagement. We therefore considered including *Surprise* in *Narrative feeling*. On the other hand, surprise shows cognitive processing (Tobin 2018) and could be considered part of *Reflection*. It is also possible to conceptualize *Surprise*, which can incorporate elements of ‘violence and enlightenment, physical attack and aesthetic pleasure’ (Miller 2015) as a separate impact type. We chose to try to measure *Surprise* by itself. Correlations with other categories could help us theorize the nature of *Surprise* further.

2.3. Definitions 132

These considerations led to the following definitions for eight categories: 133

- **Attention:** the reader's feeling of concentration or focus on their reading. 134
- **Positive affect:** any positive emotional response to the book during or after reading. 135
A feeling is positive if it contributes to a positive reading experience, so even sad 136
or awful story-events can contribute to a positive affective response. 137
 - **Narrative feeling:** a subcategory of positive affect, specifically response to a 138
book's narrative properties, including feelings about storylines, characters, 139
scenes or elements of the story world. 140
 - **Stylistic feeling:** a subcategory of positive affect, specifically response to a 141
text's stylistic properties such as feelings of admiration or defamiliarization 142
about its tone, choice of words, use of metaphor or the way the sentences 143
flow. 144
 - **Humor:** a subcategory of positive affect, specifically a response of laughter, 145
smiling or amusement; the effect of any type of humor in the text. 146
- **Surprise:** a feeling of surprise at some element of the book, such as a plot devel- 147
opment, part of the story world or a stylistic feature. 148
- **Negative feeling:** feelings of dislike or disapproval towards any element of the 149
book. This could mean a dislike for a storyline or character or a feeling of boredom 150
or frustration with the book as a whole. A feeling is negative if it contributes to a 151
negative reading experience, so unsympathetic characters or dark story elements 152
that a reviewer appreciates as part of a story do not fit within this category. 153
- **Reflection:** any response to a reading experience that makes the reader reflect on 154
something from the book, such as a theme or topic, or on something in the real 155
world. 156

3. Methods 157

This section introduces how the impact model works and explains the method of its 158
validation. 159

3.1. Model Development 160

Our model uses a set of rules to identify different types of impact expressed in individual 161
sentences of reviews, similar to the setup used by Boot and Koolen (2020). Each rule 162
belongs to a category and consist of an impact term, an impact term type and in some 163
cases a condition. For each combination of sentence and rule the software checks 164
whether the impact term is present in the sentence and, if there is a condition, whether 165
that condition is met. If so, it outputs a rule match with the associated impact type. 166

Impact terms can be lemmas or phrases. If they are lemmas, their impact term types include a POS-tag. For example, if the impact term is ‘mesmerize’, and the type is ‘verb’ the software will check for each word in the sentence whether it is a verb form with that lemma. POS-tags can also be ‘noun’, ‘adjective’ or ‘other’. In phrasal impact terms, no lemma or POS information is used, and terms can contain wildcards (*), so ‘redeeming qualitit*’ finds both ‘redeeming quality’ and ‘redeeming qualities’. Phrases consist of groups that are matched to tokens in the input sentences. A group can be a single word or a set of alternatives, such as ‘(hard|difficult)’. A phrase can be continuous or discontinuous. In a continuous phrase the groups must match a set of contiguous tokens. In a discontinuous phrase each group must match a token in the same sentence in the same order as in the phrase, but they need not be adjacent. For examples, see Table 1.

Conditions can also have different types. Most common is a reference to one of six groups of book aspect terms: *plot*, *character*, *style*, *topic*, *reader* and *general*. For example, aspect terms in the *reader* group are words referring to the reader, such as ‘I’, ‘you’, ‘the reader’ and the *general* group includes words like ‘book’ and ‘novel’. The implied condition is that one of the words from the aspect group must occur in the same sentence as the impact term. Thus, a rule linking the impact term ‘great’ to the aspect group *style* results in a hit when the word ‘great’ is present in combination with ‘writing’, ‘language’, ‘prose’ or other words in the *style* category. Conditions can also be groups of individually named words, such as ‘(part|series|sequel)’. It is also possible to negate a condition. In that case the impact term may not be combined with words from the condition. For example: ‘engage’ is an impact term related to *Narrative feeling*, unless it is combined with ‘to’ because ‘engaged to’ is more likely to refer to marriage than to narration.

To create the rules, we began by translating the 275 rules of Boot and Koolen (2020). To account for the new impact categories, we reassigned some rules to different categories. We also created new rules by manually examining a large collection of Goodreads reviews to find terms related to impact that online reviewers use. In total, the English model has 1427 rules. The growth of the set of rules has three main reasons. Firstly, the addition of four categories required adding many rules. Secondly, there are many possible translations or equivalents for the words and expressions used in the Dutch model. For example, some words relating to emotional investment in the Dutch model led to eight new rules in the English model containing various verbs combined with the noun ‘heart’ (‘break’, ‘steal’, ‘touch’, ‘rip’ and others). Thirdly, there are many more

type	Impact		Condition	
	term	term type	aspect	negate
Attention	on the edge of (my your) seat	phrase-continuous	-	-
Positive affect	makes (me you reader) sad	phrase-discontinuous	-	-
Narrative feeling	enamoured	lemma-adj	reader	-
Stylistic feeling	elegant	lemma-adj	-	-
Narrative feeling	engage	verb-adj	‘to’	y

Table 1: Example rules from the English reading impact model.

reviewers writing in English. Many have their own national variety of English and many of them are not native speakers. The range of expressions used can be assumed to be larger than in Dutch and we added many idioms based on manual analysis of a corpus of reviews from Goodreads. As we found (Boot and Koolen 2020) that human annotators often detected impact that their Dutch impact model overlooked, we expect that adding more rules will lead to a better model.

Our choice to follow the rule-based approach needs to be considered next to alternatives approaches, such as creating ground truth annotations and using Machine Learning (ML) techniques to train a generalised model. Our main reason to use rules instead of ML is that we expect ML to require many more ground truth annotations to train and test a stable and effective model that can capture subtle expressions of impact. Our model was developed ahead of gathering ground truth annotations to evaluate it (as discussed in the next sections). An ML model only learns from the annotated examples, while our rules potentially also cover cases not seen in the ground truth. If the evaluation shows that our model captures the different impact categories well, then we have reason to assume that the rule generation process achieved its aim and that the approach generalises well. With ML this is not necessarily so, although the recent advances with context-sensitive token-based word embeddings and fine-tuning of large pre-trained transformer models like BERT (Devlin et al. 2018) allow such approaches to better capture latent meanings (Yile Wang, Cui, and Zhang 2019; B. Wang et al. 2019; Ehrmanntraut et al. 2021) and generalise beyond the surface forms of the annotated impact expressions. We will discuss this further in Section 6.

3.2. Ground Truth Annotations

The rules we formulated determine how the impact model defines the various categories of impact. Next, we needed to verify that the rules we had formulated correctly operationalized the intended categories of impact. After all, the definitions implicitly created through the formulation of our impact rules might not agree with a common-sense idea of how these categories of impact are expressed. To validate our impact rules, we surveyed recipients of relevant mailing lists, students and conference attendees. We asked the participants to annotate sentences from reviews on the presence of the eight impact types. The sentences were sampled from a collection of 15 million English-language Goodreads reviews, crawled by Wan and McAuley (2018) and Wan, Misra, et al. (2019), and parsed using spaCy.¹ We manually removed sentences that we considered impossible to annotate, such as sentences containing only punctuation or incorrectly split (partial) sentences. Each sentence was annotated by at least three different annotators. After reading an explanation, each annotator was presented with ten sentences to annotate. Each annotator could annotate as many sentences as they wanted. The questions were presented to them as shown in Figure 1.

Aside from rating the presence of all eight categories of impact on a five-point scale,

1. The sentences were from a held-out set of reviews, not used to create the impact rules. We used spaCy version 2.3, <https://spacy.io>

Sentence 1: *Just wish it was longer!*

☐ For this sentence these questions are impossible to answer. ⓘ

☐ This sentence expresses no reading impact at all. ⓘ

Rate the impact on the reviewer based on this sentence (0 = Not or doubtful, 4 = Clearly or very strongly).

	0	1	2	3	4
Does this sentence express the reviewer's attention while reading? ⓘ	☐	☐	☐	☐	☐
Based on this sentence, did the reviewer experience positive emotional impact? ⓘ	☐	☐	☐	☐	☐
Did these feelings relate specifically to the narrative? ⓘ	☐	☐	☐	☐	☐
Did these feelings relate specifically to the style? ⓘ	☐	☐	☐	☐	☐
Did these feelings relate specifically to humor? ⓘ	☐	☐	☐	☐	☐
Based on this sentence, did the reviewer experience negative emotional impact? ⓘ	☐	☐	☐	☐	☐
Based on this sentence, did the reviewer experience surprise? ⓘ	☐	☐	☐	☐	☐
Does this sentence show reflection on the part of the reviewer? ⓘ	☐	☐	☐	☐	☐

Figure 1: Questions in the survey.

participants could choose to indicate that the questions were impossible to answer, 241
 such as if the text only contained gibberish or required more context to interpret, or 242
 that a sentence expressed no reading impact at all such as if it contained only a factual 243
 statement about a book. We ran the survey from October 2020 until April 2021. 244

4. Evaluation 245

In this section we assess agreement among the annotators and between the annotators 246
 and our model, and analyze which impact categories our model can meaningfully 247
 identify. 248

The survey resulted in 266 sentences that were annotated by at least three annotators, 249
 with ratings by 79 different annotators. This number excludes sentences judged to be 250
 impossible to annotate. The majority of annotators rated 10 sentences, some stopped 251
 after only a few sentences, and others annotated multiples of 10 (up to 80). We asked 252
 annotators to rate sentences on the presence of impact-types on a five-point scale from 253
 0 (not or doubtful) to 4 (clearly or strongly) for each impact type. The distribution 254
 of ratings per impact type is shown in Figure 2. On the left, only the zero ratings are 255
 shown. *Positive affect* has the fewest ratings of 0, with just over 40%, while *Stylistic feeling* 256
 and *Surprise* have around 70% 0 ratings and *Humor* has more than 80%. On the right, 257
 the distribution of ratings 1–4 are shown, also with distinct differences between types. 258
Positive affect and *Narrative feeling* tend to get high ratings (3 or 4), while *Attention* and 259
Surprise get mostly low ratings (1 or 2). 260

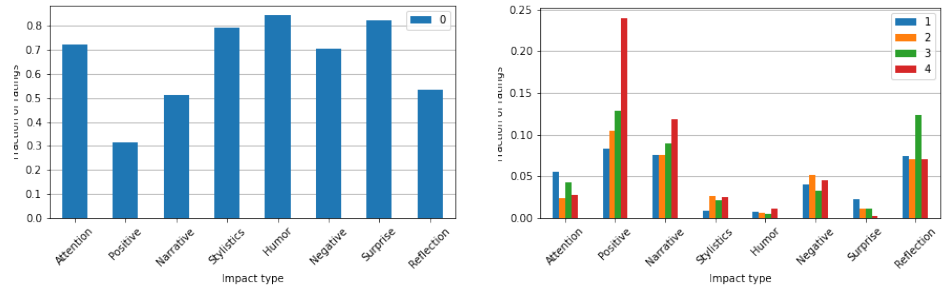


Figure 2: Fraction of 0-ratings among all ratings (left) and fraction of positive ratings (1, 2, 3 or 4) among all ratings (right).

4.1. Inter-Annotator Agreement

261

In (Boot and Koolen 2020) we calculated inter-annotator agreement using the Inter-Rater Agreement (IRA) statistic $r_{wg}^* = 1 - \frac{S_X^2}{\sigma^2}$, where S_X^2 is the variance of the ratings for a sentence and σ^2 is the expected variance based on a chosen theoretical null-distribution (Lindell and Brandt 1997). We used the same r_{wg}^* measure, but with a uniform null-distribution instead of an inverse triangular one (which assumes annotators tend to pick ratings at the two extremes), given that we observe a more uniform distribution of positive ratings when combining ratings across all categories and a larger fraction of zero ratings (so the overall variance is closer to a uniform distribution than to an inverse triangular distribution). In addition, we report Fleiss’ Kappa (κ) on binarized ratings where any rating above 0 is mapped to 1, as is more commonly reported in sentence annotation tasks for e.g. sentiment analysis (Alm and Sproat 2005; Sprugnoli et al. 2016; Schmidt, Burghardt, and Dennerlein 2018). Finally, we also report the number of sentences rated zero on a particular impact category by all three annotators, to get insight into how commonly each impact category is observed.

Category	% all zero	r_{wg}^*	κ
Attention	0.37	0.58	0.27
Positive affect	0.26	0.71	0.57
Narrative	0.36	0.55	0.40
Style	0.49	0.72	0.29
Humor	0.72	0.91	0.19
Negative	0.56	0.79	0.52
Surprise	0.50	0.74	0.25
Reflection	0.39	0.60	0.19

Table 2: Inter-Annotator Agreement per impact category averaged over 266 sentences. Agreement measures are r_{wg}^* and Fleiss’ Kappa.

Agreement is moderate (0.51-0.70) to very strong (0.91-1.00) according to r_{wg}^* (column three in Table 2), but the κ scores are much lower, in the range of 0.20 – 0.50 (column four). Scores in this range are common for related tasks like sentiment annotation (Alm and Sproat 2005; Sprugnoli et al. 2016; Schmidt, Burghardt, and Dennerlein 2018; Klenner et al. 2020). The low κ of the more commonly observed categories should not

be interpreted as low agreement, because in the original five-point scale, the difference between 0 and 1 is small while in the binarized version it is counted as disagreement.

To understand how the differences between r_{wg}^* and κ should be interpreted, we look at the number of sentences for which all three annotators agreed on a rating of zero. Since the majority of the ratings (69%) is zero, this can easily lead to a high r_{wg}^* , especially for categories that are rarely rated above zero. If a category is rarely observed, it is easy for annotators to agree on the many sentences where it is clearly not present, but they might disagree on the few sentences where at least one annotator thinks it is present. Only 26% of all sentences are rated zero on *Positive affect* by all three annotators, so its high r_{wg}^* is not caused by being rarely observed. In contrast, for *Humor*, 72% of the sentences are rated zero by all annotators, meaning it is rarely observed. For this category, a high r_{wg}^* could be caused by agreement that the category is rare, thus masking disagreement on which sentences do express impact of humor. The κ score of 0.19 (below the conventional 0.2 threshold for weak agreement) signals that agreement is lacking. For *Reflection*, only 39% of sentences are rated zero by all annotators, so this category is not uncommon, but the κ score of 0.19 also suggests a lack of agreement. We stress again that a low κ does not necessarily mean lack of agreement, as the binarization removes information from the five-point rating scale, but for *Humor* and *Reflection* these combined measures strongly suggest that either these categories are difficult to identify with our current definitions, or that reliable annotation of these categories requires more training than of the other categories.

The disagreement among annotators signals that this task is difficult and that some types of impact are more subjectively interpreted than others. This could indicate that we need to discard the categories with really low agreement. However, several recent papers suggest that disagreement between annotators is not necessarily a problem and should not be removed from the published annotation dataset (e.g. Gordon et al. 2021), but should either be retained in the form of an opinion distribution (Basile 2020; Klenner et al. 2020) or a special class label *Complicated* (Kenyon-Dean et al. 2018). Since our data is based on a rating scale, it makes sense to distribute the annotated sentence data with the full rating distributions. In the following sections, we discuss whether all impact categories should be retained in the ground truth data and the rule-based model.

4.2. Evaluating the Model

To compare our model against the ratings of the human annotators, we select the median of the three ratings per sentence and impact category as the ground truth rating and compare that to whether our model finds at least one matching impact rule for that category in the sentence. If the model works well, then it should find matching rules for an impact category in sentences that received a high median rating from human annotators.

We measure recall, precision and F_1 of our model's performance on the annotated sentences, using two different binarizations. As we have a 5-point rating scale, we want to know if our model finds impact in sentences that clearly express impact, that is, where

Impact	Model # Sent.	$r_{median} \geq 1$				$r_{median} \geq 3$			
		# Sent.	Prec.	Rec.	F_1	# Sent.	Prec.	Rec.	F_1
Attention	9	83	0.78	0.08	0.15	44	0.78	0.16	0.26
Positive	90	148	0.82	0.50	0.62	102	0.59	0.52	0.55
Narrative	39	101	0.72	0.28	0.40	60	0.51	0.33	0.40
Stylistic	8	59	0.50	0.07	0.12	26	0.50	0.15	0.24
Humor	7	18	1.00	0.39	0.56	4	0.57	1.00	0.73
Negative	15	68	0.73	0.16	0.27	40	0.60	0.23	0.33
Surprise	2	51	0.00	0.00	0.00	17	0.00	0.00	0.00
Reflection	19	68	0.53	0.15	0.23	19	0.26	0.26	0.26

Table 3: Model evaluation per impact category on 266 sentences, with number of sentences for which the model identifies impact (column 2), and precision and recall of our model for binarization of ratings based on median rating $r_{median} \geq 1$ and $r_{median} \geq 3$

the median rating is high, i.e. 3 or 4, but also for sentences that express any impact at all, i.e. those with ratings of 1 or higher. The results are shown in Table 3, with the number of sentences that have a binary rating of 1 for each binarization (columns 3 and 6). The model scores above 0.7 precision on five of the eight categories for binarization $r_{median} \geq 1$: *Attention*, *Positive affect*, *Narrative feeling*, *Humor* and *Negative feeling*. In the majority of cases, the matching rules for these aspects correspond to the type of impact identified by the median annotator, and therefore at least two of the three annotators. For *Stylistic feeling* and *Reflection* it scores around 0.5 precision, so in half of the cases, the matching rules incorrectly signal impact. For *Surprise* the model completely fails. It only finds *Surprise* in two sentences—both of which are incorrect according to the ground truth—while there are 51 sentences with a median rating of at least 1. For binarization $r_{median} \geq 3$, precision is mostly lower, showing that the model regularly predicts impact where human annotators consider it doubtful. *Humor* is rarely observed by the annotators, with low agreement, and our model also rarely finds matching rules, but with high precision for $r_{median} \geq 1$ and high recall for $r_{median} \geq 3$. When annotators agree that *Humor* is clearly expressed, our model detects it (in the few cases in this ground truth dataset), and when our model detects *Humor*, it is in places where annotators perceive *Humor* to some extent. Two examples where annotators and our model clearly agree demonstrate this. For the sentence ‘I loved Blaire’s personality she was sassy, funny, extremely witty, I laughed out loud frequently, much to my embarrassment.’ our model has three matching rules, *funny*, *witty* and *laugh out loud*, and the annotators gave an average rating of 3. For the sentence ‘We actually bought a copy for our music history teacher who would appreciate the humor in this book (he was Jewish, sarcastic, clever.’ our model has two matching rules, *humor* and *sarcastic* (which in this sentence does not refer to impact of the book) and annotators gave an average rating of 3.33. This sheds further light on the low Fleiss’ Kappa scores for *Humor*. There are clear cases where annotators agree that humor is expressed, so the low agreement seems to come from doubtful cases where some annotators are not sure and give a low rating of 1 or 2 and others say it is not expressed. The binarization we used to compute Fleiss’ Kappa creates a complete disagreement in such doubtful cases, where the original five-point

ratings signal only slight disagreement. The model performance suggests that, although we need more ground truth annotations and perhaps a better definition to improve agreement, this is a viable category to include.

The generally low recall scores show that our model misses many expressions of reading impact. This suggests that our set of impact rules is incomplete. Overall, the precision and recall scores suggest that our approach of translating and extending our Dutch impact model is viable for most of the categories and that with additional rules and some improvements to the existing rules, the model can capture enough of the expressed reading impact in individual reviews to derive a reliable overall estimate of a book's impact, at least for books with more than a handful of reviews. The only clear exceptions are *Surprise*, where the model fails completely, and *Stylistic feeling* and *Reflection* where the models not only misses many expressions of impact, but also makes many mistakes.

4.3. Error Analysis

Annotators found impact in many instances where the model failed to detect it. For example, the model scored a 0 in the positive emotion category for the sentence "I was born to love this book," which received a rating of 4 from all annotators. This suggests we should add rules to increase the sensitivity of the model. We should also revise the way that the model processes impact terms to nuance the model. Currently, the model marks the presence of negative terms like 'skim' as negative impact, but it turns out that this is not always accurate: "I didn't skim at all" actually indicates positive impact. The negation of the negative impact term 'skim' should flip the predicted impact to positive. To improve performance on sentences with such negations of typical sentiment words, we could adopt the sentiment flipping technique used in the VADER sentiment analyzer (Hutto and Gilbert 2014). This technique looks for negations in the word tri-gram preceding a sentiment term, which captures almost 90% of the negated sentiments in their ground truth data. However, negation should not always flip the valence from positive to negative or vice versa (Dadvar, Hauff, and De Jong 2011; Socher et al. 2013). When a reviewer says that a book is 'not terrible' they probably don't mean to say it is good.

The responses to the survey showed that annotators struggled to understand some categories and regularly disagreed over them, albeit to a different degree for different categories. For instance, the sentence "And then there was Jacob O'Connor," which we feel expresses no impact, was rated by annotators with a score of 3.5 in the surprise-category. Annotators also found *Attention* difficult to distinguish from *Positive affect* and *Narrative feeling*. They also struggled with negative story elements that can add to a positive reading experience, such as a 'creepy' character. Respondents tend to annotate such sentences as negative impact, while that is often impossible to judge without context. In another example, annotators judged the sentence "My soul is beautifully crushed" to indicate negative impact, but in our view a reviewer who writes this is expressing positive impact. These differences between annotator-ratings and our own conceptions of impact categories point towards one of the complexities of developing

computational models for literary studies: while defining categories of impact and formulating rules for our model, our own subjective understanding and academic knowledge of impact categories and the impact of reading became part of the model we produced. These conceptions may not necessarily align with the conceptions of other people. To resolve issues of annotator agreement, we could consider recruiting annotators with a background in reception studies or literary studies for future research, since they will presumably have a shared understanding of these impact categories based on the scholarly literature. Therefore, these annotators would probably be better-equipped to distinguish and detect our eight impact categories, but it is also possible that they would skew results with their pre-existing definitions of the categories. Another option would be resolving disagreement between annotators using the method outlined by Oortwijn, Ossenkoppele, and Betti (2021), or recruiting annotators from within the community of people actively writing English-language reviews on Goodreads. This way, we could validate the model using conceptions from within the community we are studying. While we tried to do this by contacting the moderators of various Goodreads groups, we received little response. In the end, developing a flawless model to measure how reading impact is expressed in online reviews may be impossible, because of the subjectivity and fluidity of the categories such a model tries to measure. In the act of operationalizing impact categories through rulesets, some of their polysemic meanings are inevitably lost. Nonetheless, we believe that our current imperfect model has pointed us towards some interesting insights into the impact of reading expressed in our corpus of reviews. We discuss these insights in Section 5.

5. Analyzing Reading Impact of Novels

In this section, we analyze the impact identified by our model by applying it to a collection of 1,313,863 reviews of 402 well-known books, from the Goodreads crawl introduced in Section 3.2. As the results from the previous section cast doubt on the viability of measuring some of the categories of impact, in this section we ignore *Surprise* and *Reflection*. We selected books with at least 10 reviews in both Dutch and English so that, in future research, we may compare the current and future versions of the English-language model against the Dutch model.

5.1. Impactful Books

Our model generated a rating for each of the 402 books in each of the model's categories. This rating gives an indication of how often a specific type of impact was mentioned in a specific review. After normalizing the scores for the length of the reviews we computed which books scored highest and lowest in each category. Table 4 lists the books scoring highest on *Stylistic feeling* and *Humor*. The left column contains mostly literary classics that received high critical acclaim; we would expect those novels to score high on *Stylistic feeling*. The right column contains mostly books that are well-known for their comic appeal. Similar lists for other categories are not always easy to evaluate, for example because lesser-known novels appear in the list or because there is no canon of

Title	Author	Title	Author
Monsieur Linh and his Child	Philippe Claudel	Weird Things Customers Say in Bookshops	Jen Campbell
Lolita	Vladimir Nabokov	Look Who's Back	Timur Vermes
All the Light We Cannot See	Anthony Doerr	The Hitchhiker's Guide to the Galaxy	Douglas Adams
Stoner	John Williams	The Secret Diary of Hendrik Groen, 83¼ Years Old	Hendrik Groen
The Sense of an Ending	Julian Barnes	The Hundred-Year-Old Man Who Climbed Out of the ...	Jonas Jonasson
The Discovery of Heaven	Harry Mulisch	The Girl Who Saved the King of Sweden	Jonas Jonasson
HHhH	Laurent Binet	Me and Earl and the Dying Girl	Jesse Andrews
The Vanishing	Tim Krabbe	The Rosie Project	Graeme Simsion
A Visit from the Goon Squad	Jennifer Egan	A Totally Awkward Love Story	Tom Ellen
The Book Thief	Markus Zusak	Geek Girl	Holly Smale

Table 4: Top ten titles on Stylistic impact (left) and Humor (right)

narratively engaging novels, the way there is one for literary novels. Still, some results suggest that our rules are pointing in the good direction. For example, one would expect that non-fiction titles score low on *Narrative feeling*. Indeed, the four worst-performing titles in terms of *Narrative feeling* are non-fiction titles, including Marie Kondo's *The Life-Changing Magic of Tidying Up*. These results provide prima facie evidence for the validity of the rules that we use to define these impact categories.

5.2. Correlations between Impact Types

In this section, we analyze the correlation between impact types and the correlation between impact types and the average rating of reviews, when aggregated per novel for the same set of 402 novels. For this analysis, we computed impact score per category based on the recommendation of Koolen, Boot, and van Zundert (2020), where we suggest weighing the number of impact rule matches per review by the log-length of the review in number of words. This weighing should account for the fact that long reviews potentially have more impact matches without actually indicating stronger impact. The Pearson correlations are shown in Figure 3, with levels of correlation above 0.2 highlighted in green. Unsurprisingly, *Positive affect* is positively correlated with its components *Narrative feeling*, *Stylistic feeling* and *Humor*. We discuss the correlations of *Attention* and *Negative feeling* with the other impact factors and the correlations with reviewer rating.

5.2.1. Correlations of Attention

The most important correlation (.60) for *Attention* is with *Narrative feeling*. This suggests that *Narrative feeling* draws readers in and leads to a sense of absorption and immersion. Attention-related questions are also an important part of the Story World Absorption Scale (M. M. Kuijpers et al. 2014). That there is no correlation between *Attention* and *Stylistic feeling* similarly suggests that stylistic appreciation is not that important for absorption. *Attention* is weakly negatively correlated with *Humor*. Knoop et al. (2016), in their analysis of evaluative terms, distinguish between emotionally charged terms

	Att	Pos	Nar	Sty	Hum	Neg	Rating
Att	1.00	0.40	0.60	0.14	-0.26	0.39	-0.13
Pos	0.40	1.00	0.77	0.40	0.31	0.33	0.02
Nar	0.60	0.77	1.00	0.17	-0.14	0.42	-0.03
Sty	0.14	0.40	0.17	1.00	-0.09	0.04	-0.10
Hum	-0.26	0.31	-0.14	-0.09	1.00	-0.01	-0.01
Neg	0.39	0.33	0.42	0.04	-0.01	1.00	-0.61
Rating	-0.13	0.02	-0.03	-0.10	-0.01	-0.61	1.00

Figure 3: Pearson correlation coefficients between impact types and rating of reviews aggregated per novel.

such as ‘sad’ and ‘beautiful’ and more cognitive terms such as ‘funny’ or ‘humorous’. The relationship between cognitive and emotional impact is an area of further research for the refinement of our model.

5.2.2. Correlations of Negative feeling

It is surprising that *Negative feeling* is weakly to moderately positively correlated with *Attention*, *Positive affect* and *Narrative feeling*. As this is not just a book-level effect (*Positive* and *Negative feeling* are also correlated within individual reviews), we speculate that these correlations occur because negative terms are often used concessively, as in ‘the plot may be a bit unrealistic but the characters are lovely’. But more research on these correlations is needed.

5.2.3. Correlations of Impact Categories and Reviewer Rating

On Goodreads, reviewers have the option of rating a book on a five-star scale in addition to, or instead of, providing a written review. Only one impact category shows a correlation with reviewer rating: *Negative feeling*. The moderate negative correlation suggests that negative terms are not just used concessively but often do express a lack of appreciation.

The lack of correlation between rating and the other impact categories is surprising. Positive feeling, as measured by sentiment analysis tools, is known to predict rating (De Smedt and Daelemans 2012). We would also expect *Attention*, which is closely related to enjoyment (M. M. Kuijpers et al. 2014), to correlate positively with rating. This lack of correlations could indicate that the impact model succeeds in extracting new information, independent from rating, from the review text.

The correlations among impact types, or lack thereof, as well as those between impact types and rating, call for further analysis of the nature of their relation. Reader characteristics may also influence this relation. For instance, we found a negative correlation between impact in the *Reflection* category and reviewer ratings (not shown in Figure 3). This could mean that reviewers are less appreciative of books that encourage reflection. But it could also mean that readers who engage in more reflection generally give more

moderate ratings. More generally, how does rating behavior relate to reading preference 488
and other reader characteristics? 489

6. Discussion 490

Our findings allow us to address our two main research questions and to indicate a 491
number of areas for future research into the impact of fiction and the usefulness of 492
measuring and analyzing that impact computationally in online book reviews. In future 493
research, we will build on the findings presented in the current paper. 494

6.1. Conclusions 495

1. *How effective is our adaptation of the Dutch model?* 496

Based on the results from the English impact model so far, the model is effective 497
in some categories but not all of them. For several impact categories the rule- 498
based model attains good performance in terms of precision and recall, but more 499
ground truth data is needed to reliably validate some other categories, and for 500
some categories more rules are needed to cover the various ways impact can be 501
expressed. When ranking books by scores in individual impact categories, the 502
model appears to do a good job. In future work, we intend to compare the English 503
impact model presented in this paper with the existing Dutch model. 504

(a) *Can the new impact categories we add to the model be captured in a rule-based model?* 505

Can these new categories be meaningfully identified by human annotators? 506

We added four new impact categories to the impact model described in 507
Boot and Koolen (2020), in the hope that adding more categories would 508
lead to a more fine-grained and accurate model. Some of these newly added 509
categories proved difficult for annotators to identify consistently. For example, 510
annotators frequently seemed to confuse *Attention* and *Narrative feeling*. For 511
example, according to the annotators "Lots of twists and turns and good 512
characters" indicated *Attention* as well as *Narrative feeling* while we, and 513
the rules of our model, see this sentence as indicating only *Narrative feeling*. 514
Conversely, annotators labelled the sentence "The third book of the trilogy is 515
just as compelling as the other two" as both *Narrative feeling* and *Attention*, 516
while our model would only see it as *Attention*. Such overlap, disagreement 517
or confusion between categories shows that, similar to the original categories, 518
identifying which sentences express a specific type of impact remains a 519
difficult and subjective task. 520

One way of approaching this issue might be to compare the correlations 521
between the impact categories as established by our model and those between 522
the impact categories as rated by the annotators. That could provide us with a 523
sense of how the annotators' conceptualisation of the impact categories differs 524
from our model's conceptualisation. However, that some of the new impact 525
categories can be meaningfully identified by a rule-based model is already 526

clear from the combination of inter-annotator agreement analysis, evaluation of the model based on the ground truth annotations, and comparing the reading impact of novels identified in sets of reviews.

(b) *Is adapting an existing rule-based model for use in another language a productive approach? Is our method of translating and changing rules an effective way to do this? What are the challenges and advantages of transferring knowledge or tools from Dutch to English through translation and adaptation?*

Our results indicate that the translation of the rules, in combination with adding new rules specific to English, is a viable approach to building a reading impact model for English-language reviews and expanding on the existing Dutch model. However, since human annotators detect impact in many words and phrases that the model disregards, it seems that adding still more rules may be necessary. Also, adapting the model was a labor-intensive process. This is a definite drawback of taking a translation-approach to a rule-based model. On the other hand, the advantage of translating the model from Dutch to English is that it makes the impact model accessible to a wider user base of researchers.

2. *Is a rule-based model a productive tool for assessing the impact of fiction as expressed in online book reviews? What are the advantages of a rule-based model compared to other approaches, such as machine learning?*

A rule-based model has advantages and drawbacks when compared to other approaches, like Machine Learning (ML). Rule-based approaches are more transparent than trained ML models because users can inspect each and every rule and understand how the model arrived at a specific decision. With ML models, especially neural network-based models, the knowledge is distributed over and represented by a large number of weights between the network nodes. As in a rule-based model researchers can add or translate impact-rules, they can adapt the tool to specific research questions and language domains without requiring large amounts ground truth annotations to train a ML model. Moreover, for fine-grained annotation in specific domains, like identifying expressions of different types of reading impact, it can be difficult to attain good performance with ML, as ML models need to be trained on domain-specific data to adapt to the domain-specific terminology and nuances (Thelwall, Buckley, et al. 2010; Wu and Huang 2016; Mishev et al. 2020), which requires large amounts of training data. For instance, for the simpler task of sentiment polarity classification, many thousands or tens of thousands of annotated examples are needed (Mishev et al. 2020; Yao and Yan Wang 2020). At the same time, formulating and validating rules is also a labor-intensive process and our model did not attain great results for every category. However, the impact model presented in this paper could potentially be used to gather such data. Thus, the best approach for future research may be to combine rule-based and machine learning methods.

6.2. Directions for Future Research

568

As discussed in 2, online book reviews are not necessarily representative of the un- 569
mediated reading experience, let alone of the spectrum of reading experiences that a 570
book may evoke in different readers. Given the increasing amount of work that uses 571
online book response in the study of reading, research that bridges these gaps seems 572
particularly urgent, for ourselves as well as for the wider field of research in literary 573
reading. 574

Another ambitious next step in studying reading impact is the possibility of connecting 575
the impact reported in online book reviews to specific features of individual books. This 576
is the aim of the *Impact and Fiction* project (<https://impactandfiction.huygens.knaw.nl/>), where we will develop new metrics to computationally identify high-level 578
features of literary texts such as mood, style and narrative structure, in order to examine 579
the relationship between these book-intrinsic features and the impact of books expressed 580
in online reviews. Additionally, we will differentiate between groups of readers to take 581
into account that different (groups of) readers may respond differently to these book 582
features (Van den Hoven et al. 2016). The research presented in this paper serves as a 583
first step towards answering such questions. 584

7. Data availability 585

Data can be found here: <https://zenodo.org/record/5798598> 586

8. Software availability 587

Software can be found here: <https://github.com/marijnkoolen/reading-impact-model/> 588
589

9. Acknowledgements 590

10. Author contributions 591

Marijn Koolen: Formal Analysis, Conceptualization, Software, Writing – original draft 592

Julia Neugarten: Formal Analysis, Conceptualization, Writing – original draft 593

Peter Boot: Formal Analysis, Conceptualization, Writing – original draft 594

References 595

- Alm, Cecilia Ovesdotter and Richard Sproat (2005). “Emotional sequencing and development in fairy tales”. In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp. 668–674. URL: https://link.springer.com/chapter/10.1007/11573548_86. 596
597
598
599
- Basile, Valerio (2020). “It’s the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks”. In: *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP*. Vol. 2776. CEUR-WS, pp. 31–40. URL: <http://ceur-ws.org/Vol-2776/paper-4.pdf>. 600
601
602
603
- Boot, Peter and Marijn Koolen (2020). “Captivating, splendid or instructive? Assessing the impact of reading in online book reviews”. In: *Scientific Study of Literature* 10.1, pp. 66–93. DOI: [10.1075/ssol.20003.bo](https://doi.org/10.1075/ssol.20003.bo). 604
605
606
- Champagne, Ashley (2020). “What Is A Reader? The Radical Potentiality of Goodreads to Disrupt the Literary Canon”. In: *Digital Humanities 2020*. 607
608
- Dadvar, Maral, Claudia Hauff, and Franciska De Jong (2011). “Scope of negation detection in sentiment analysis”. In: *Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR 2011)*. Citeseer, pp. 16–20. URL: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.998.6048&rep=rep1&type=pdf>. 609
610
611
612
- De Smedt, Tom and Walter Daelemans (2012). ““Vreselijk mooi!” (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives.” In: *LREC*, pp. 3568–3572. 613
614
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*. 615
616
617

- Ehrmanntraut, Anton, Thora Hagen, Leonard Konle, and Fotis Jannidis (2021). "Type- and Token-based Word Embeddings in the Digital Humanities". In: *CHR 2021, Proceedings of the Conference on Computational Humanities Research 2021*. CEUR-WS, pp. 16–38. URL: http://ceur-ws.org/Vol-2989/long_paper35.pdf.
- Gordon, Mitchell L, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein (2021). "The disagreement deconvolution: Bringing machine learning performance metrics in line with reality". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. URL: <https://dl.acm.org/doi/abs/10.1145/3411764.3445423>.
- Green, Melanie C and Timothy C Brock (2000). "The role of transportation in the persuasiveness of public narratives." In: *Journal of personality and social psychology* 79.5, p. 701. URL: <https://psycnet.apa.org/journals/psp/79/5/701.html?uid=2000-00920-003>.
- Holur, Pavan, Shadi Shahsavari, Ehsan Ebrahimzadeh, Timothy R Tangherlini, and Vwani Roychowdhury (2021). "Modeling Social Readers: Novel Tools for Addressing Reception from Online Book Reviews". In: *arXiv preprint arXiv:2105.01150*. URL: <http://arxiv.org/abs/2105.01150>.
- Hutto, Clayton and Eric Gilbert (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. URL: <https://ojs.aaai.org/index.php/ICWSM/article/download/14550/14399>.
- Kenyon-Dean, Kian, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhandari, Robert Belfer, Nirmal Kanagasabai, et al. (2018). "Sentiment analysis: It's complicated!" In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1886–1895. URL: <https://www.aclweb.org/anthology/N18-1171.pdf>.
- Klenner, Manfred, Anne Göhring, Michael Amsler, Sarah Ebling, Don Tuggener, Manuela Hürlimann, and Martin Volk (2020). "Harmonization sometimes harms". In: *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. swisstext-and-konvens-2020. URL: <https://www.zora.uzh.ch/id/eprint/197961/1/Konvens2020.pdf>.
- Knoop, Christine A, Valentin Wagner, Thomas Jacobsen, and Winfried Menninghaus (2016). "Mapping the aesthetic space of literature "from below"". In: *Poetics* 56, pp. 35–49. DOI: <https://doi.org/10.1016/j.poetic.2016.02.001>.
- Koolen, Marijn, Peter Boot, and Joris J. van Zundert (2020). "Online Book Reviews and the Computational Modelling of Reading Impact". In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18-20, 2020*. Ed. by Folgert Karsdorp, Barbara McGillivray, Adina Nerghes, and Melvin Wevers. Vol. 2723. CEUR Workshop Proceedings. CEUR-WS.org, pp. 149–169. URL: <http://ceur-ws.org/Vol-2723/long13.pdf>.
- Koopman, Eva Maria Emy (2016). "Effects of "Literariness" on Emotions and on Empathy and Reflection after Reading". In: *Psychology of Aesthetics, Creativity, and the Arts* 10.1, p. 82.

- Koopman, Eva Maria Emy and Frank Hakemulder (2015). "Effects of literature on empathy and self-reflection: A theoretical-empirical framework". In: *Journal of Literary Theory* 9.1, pp. 79–111. URL: <https://www.degruyter.com/document/doi/10.1515/jlt-2015-0005/html>.
- Kuijpers, Moniek M, Frank Hakemulder, Ed S Tan, and Miruna M Doicaru (2014). "Exploring absorbing reading experiences." In: *Scientific Study of Literature* 4.1. URL: <https://www.jbe-platform.com/content/journals/10.1075/ssol.4.1.05kui>.
- Lindell, Michael K and Christina J Brandt (1997). "Measuring interrater agreement for ratings of a single target". In: *Applied Psychological Measurement* 21.3, pp. 271–278. URL: <https://journals.sagepub.com/doi/abs/10.1177/01466216970213006>.
- Miall, David S and Don Kuiken (2002). "A Feeling for Fiction: Becoming What We Behold". In: *Poetics* 30.4, pp. 221–241.
- Miller, Christopher R (2015). *Surprise: The poetics of the unexpected from Milton to Austen*. Cornell University Press. DOI: <https://doi.org/10.7591/9780801455780>.
- Mishev, Kostadin, Ana Gjorgjevikj, Irena Vodenska, Lubomir T Chitkushev, and Dimitar Trajanov (2020). "Evaluation of sentiment analysis in finance: from lexicons to transformers". In: *IEEE Access* 8, pp. 131662–131682.
- Nell, Victor (1988). *Lost in a Book: The Psychology of Reading for Pleasure*. Yale University Press.
- Oortwijn, Yvette, Thijs Ossenkoppele, and Arianna Betti (2021). "Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks". In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 131–141.
- Poletti, Anna, Judith Seaboyer, Rosanne Kennedy, Tully Barnett, and Kate Douglas (2016). "The affects of not reading: Hating characters, being bored, feeling stupid". In: *Arts and Humanities in Higher Education* 15.2, pp. 231–247. DOI: <https://doi.org/10.1177/1474022214556898>.
- Rebora, Simone, Peter Boot, Federico Pianzola, Brigitte Gasser, J Berenike Herrmann, Maria Kraxenberger, Moniek Kuijpers, Gerhard Lauer, Piroška Lendvai, Thomas C Messerli, et al. (2019). *Digital humanities and digital social reading*. OSF Preprints. URL: <https://osf.io/preprints/mf4nj/>.
- Ross, Catherine Sheldrick (1999). "Finding without Seeking: the Information Encounter in the Context of Reading for Pleasure". In: *Information Processing & Management* 35.6, pp. 783–799.
- Sabine, Gordon and Patricia Sabine (1983). *Books That Made the Difference: What People Told Us*. ERIC.
- Schmidt, Thomas, Manuel Burghardt, and Katrin Dennerlein (2018). "Sentiment annotation of historic german plays: An empirical study on annotation behavior". In: *Proceedings of the Workshop on Annotation in Digital Humanities 2018 (annDH 2018)*. RWTH Aachen. URL: <https://epub.uni-regensburg.de/43701/1/Schmidt%20et%20al.%20-%20SentimentAnnotation.pdf>.
- Schmitt-Matzen, Cassie D (2020). "Adult Retrospectives on Unhealthy Adolescent Responses to Reading Fiction". PhD thesis. Tennessee Technological University. URL:

- <https://www.proquest.com/openview/87797daf04873815817102580ef87d25/1> 706
707
- Shannon, Donna M (1993). "Children's responses to humor in fiction". PhD thesis. The 708
University of North Carolina at Chapel Hill. URL: <https://www.proquest.com/openview/0d8c439362b9954efe1c4ad141fef083/1>. 709
710
- Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, An- 711
drew Y Ng, and Christopher Potts (2013). "Recursive deep models for semantic 712
compositionality over a sentiment treebank". In: *Proceedings of the 2013 conference on 713
empirical methods in natural language processing*, pp. 1631–1642. URL: <https://aclanthology.org/D13-1170.pdf>. 714
715
- Spiteri, Louise F and Jen Pecoskie (2016). "Affective taxonomies of the reading expe- 716
rience: Using user-generated reviews for readers' advisory". In: *Proceedings of the 717
Association for Information Science and Technology* 53.1, pp. 1–9. doi: <https://doi.org/10.1002/pra2.2016.14505301032>. 718
719
- Sprugnoli, Rachele, Sara Tonelli, Alessandro Marchetti, and Giovanni Moretti (2016). 720
"Towards sentiment analysis for historical texts". In: *Digital Scholarship in the Humani- 721
ties* 31.4, pp. 762–772. URL: <https://academic.oup.com/dsh/article-abstract/31/4/762/2748263>. 722
723
- Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas (2010). 724
"Sentiment strength detection in short informal text". In: *Journal of the American society 725
for information science and technology* 61.12, pp. 2544–2558. 726
- Thelwall, Mike and Kayvan Kousha (2017). "Goodreads: A social network site for 727
book readers". In: *Journal of the Association for Information Science and Technology* 728
68.4, pp. 972–983. doi: <https://doi.org/10.1002/asi.23733>. eprint: <https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23733>. URL: 729
<https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23733>. 730
731
- Tobin, Vera (2018). *Elements of surprise: Our mental limits and the satisfactions of plot*. 732
Harvard University Press. doi: <https://psycnet.apa.org/doi/10.4159/9780674919570>. 733
734
- Van den Hoven, Emiel, Franziska Hartung, Michael Burke, and Roel M Willems (2016). 735
"Individual differences in sensitivity to style during literary reading: Insights from 736
eye-tracking". In: *Collabra* 2.1. 737
- Wan, Mengting and Julian J. McAuley (2018). "Item recommendation on monotonic 738
behavior chains". In: *Proceedings of the 12th ACM Conference on Recommender Systems, 739
RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*. Ed. by Sole Pera, Michael D. 740
Ekstrand, Xavier Amatriain, and John O'Donovan. ACM, pp. 86–94. doi: [10.1145/3240323.3240369](https://doi.org/10.1145/3240323.3240369). URL: <https://doi.org/10.1145/3240323.3240369>. 741
742
- Wan, Mengting, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley (2019). "Fine- 743
Grained Spoiler Detection from Large-Scale Review Corpora". In: *Proceedings of the 744
57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, 745
July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, 746
and Lluís Màrquez. Association for Computational Linguistics, pp. 2605–2610. doi: [10.18653/v1/p19-1248](https://doi.org/10.18653/v1/p19-1248). URL: <https://doi.org/10.18653/v1/p19-1248>. 747
748

- Wang, Bin, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo (2019). "Evaluating word embedding models: Methods and experimental results". In: *APSIPA transactions on signal and information processing* 8. 749-751
- Wang, Yile, Leyang Cui, and Yue Zhang (2019). "How Can BERT Help Lexical Semantics Tasks?" In: *arXiv preprint arXiv:1911.02929*. 752-753
- Wu, Fangzhao and Yongfeng Huang (2016). "Sentiment domain adaptation with multiple sources". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 301-310. 754-756
- Yao, Fang and Yan Wang (2020). "Domain-specific sentiment analysis for tweets during hurricanes (DSSA-H): A domain-adversarial neural-network-based approach". In: *Computers, Environment and Urban Systems* 83, p. 101522. 757-759

Conference

Ancient Greek Literary Texts

An application of Text Alignment for Digital Philology Research

Chiara Palladino  ¹

Farnoosh Shamsian  ²

Tariq Yousef  ²

1. Classics, Furman University, USA.

2. Leipzig University, Germany.

Keywords:

Translation Alignment, Translation
Analysis, Philology, Critical
Translation Studies, NLP, Annotation

Licenses:

This article is licensed under:



Abstract.

This paper presents a workflow to analytically compare translations of Ancient Greek texts into English and Persian through the analysis of parallel corpora aligned manually at word level using UGARIT translation alignment editor. We extracted the translation pairs, measured word intersections, match ratios, and part of speech data, in order to observe how close the translations were to each other and to the original text. The corpus we propose includes the *Iliad*, the *Hippolytus*, and *Against Neaira*. In addition to the direct translations, we have included and analyzed some indirect translations in the Greek-Persian corpus where French has been used as the mediating language.

1. Introduction

1

In this study, we propose an application of translation alignment for the study of translations of Ancient Greek texts in English and Persian. We introduce the general principles of translation alignment and its challenges in the domain of historical languages, and examine how the alignment of parallel texts at word level can support a comparative analysis and the individuation of certain translation phenomena.

2

3

4

5

6

Translation alignment is defined as the operation of aligning parallel texts, i.e. two or more texts in different languages. It is an essential task of Natural Language Processing, the main purpose of which is to define which parts of a source text correspond to which parts of a second text. The result is often a list of pairs of items (words, sentences, or larger chunks of text like paragraphs or documents) (Kay and Röscheisen 1993). A collection of parallel texts aligned at some level is also defined as a parallel corpus.

7

8

9

10

11

12

Translation Alignment is a task that can be performed automatically, semi-automatically or manually, through the establishment of translation pairs. The most important current

13

14

methods for automatic translation alignment belong to two categories: statistical or 15
 neural models. The first statistical lexical models for automatic word alignment, known 16
 as IBM models, were introduced in the 1990s (Brown et al. 1993); more recently, Giza++ 17
 was introduced to perform automatic alignment based on similar principles, and it was 18
 long considered the state of the art for automatic alignment with statistical methods 19
 (Och and Ney 2003). However, statistical methods require an enormous amount of 20
 training data, and tend to perform poorly in the absence of large corpora. Recently, 21
 neural models have been developed as an alternative, exploiting static or contextualized 22
 word embeddings extracted from multilingual language models and semantic similarity 23
 matrices, to create accurate alignments even without training data: for example, recent 24
 tools like AWESOME aligner (Dou and Neubig 2021) and SimAlign (Jalili Sabet et al. 25
 2020) fall into this category. Moreover, multilingual contextualized language models 26
 such as mBERT and XML-R can be fine-tuned on monolingual and bilingual datasets 27
 used in a supervised and unsupervised manner, to predict word-level alignments for 28
 under-resourced languages (Yousef, Palladino, Wright, et al. 2022; Yousef, Palladino, 29
 Shamsian, Ferreira, et al. 2022). 30

Despite the popularity of automatic models, manually aligned parallel corpora remain 31
 an essential resource used in a variety of fields, especially if aligned according to specific 32
 guidelines and at high levels of granularity (word and sentence level). Primarily, they 33
 provide training data for statistical methods, or gold standards against which neural 34
 models can be tested. However, they are also used in many other contexts, including 35
 text mining, pedagogy, and text reuse (Dagan, Church, and Gale 1999; Graça et al. 2008; 36
 Véronis 2000). Parallel corpora are also used in the analysis of languages and translations, 37
 in lines of research such as Corpus-Based Translation Studies (CTS), which analyze 38
 parallel corpora, coupled with information like part of speech and morphosyntax, to 39
 provide a better understanding of translational dynamics or textual traditions (Baker, 40
 Francis, and Tognini-Bonelli 1993; Laviosa 2008). 41

For these reasons, many tools are designed to facilitate a user-based creation of parallel 42
 texts at word and sentence level. A first category includes tools that offer an annotation 43
 interface to generate translation equivalents without improving the visualization of the 44
 performed alignments: these include the Blinker Project (Melamed 1998), which was 45
 used to align different versions of the Bible in French and English; the LDC Word Aligner, 46
 for the alignment of Arabic-English and Chinese-English broadcast texts (Grimes et 47
 al. 2010); TagAlign, which allows users to annotate texts with a pre-defined tagset 48
 (Caseli, Feltrim, and Nunes 2002) for Portuguese and English. A second category of 49
 tools empowers various kinds of methods to visualize and query the annotated texts: 50
 Yawat (Germann 2008), Alpheios (Almas and Beaulieu 2013), SWIFT Aligner (Gilmanov, 51
 Scrivner, and Kübler 2014), and CLUE-Aligner (Barreiro, Raposo, and Luís 2016), enable 52
 users to create alignments manually and offer various options for visualizing them, such 53
 as side-by-side view, interlinear text view, and alignment matrices (Yousef, Palladino, 54
 Shamsian, and Foradi 2022a). 55

1.1. Introducing Ugarit: A Tool for Translation Alignment of Low-resourced Languages 56 57

Typically, the methods and tools developed around parallel corpora are conceived for 58
modern languages. One important exception is Alpheios, which was designed with 59
Ancient Greek and Latin texts in mind, but it does not allow the open publication of the 60
alignments on the web (Almas and Beaulieu 2013). 61

Historical, indigenous, and generally under-resourced languages lack the necessary 62
infrastructure to successfully apply automated methods for annotation or alignment, 63
and manually annotated data are often the only resource available to create and analyze 64
parallel corpora. Most of these languages are ancient or minority languages, for which 65
the descriptive and historical study of the textual and linguistic tradition is of unques- 66
tionable importance, but their status determines a set of peculiar issues. As a general 67
principle, the more distant two languages are by typology (e.g. analytical vs. synthetic 68
languages), the more difficult it is to establish exact translation correspondences, be- 69
cause of the structural variations in morphology and word order. For ancient languages, 70
there is the added difficulty of having to deal with a tradition and a culture that are 71
radically distant, that cannot be verified with native speakers, and that typically require 72
a completely different approach to reading and comprehension (Crane 2019). Typically, 73
ancient texts will also have a long and complicated translation history (Nergaard 1993; 74
Bettini 2012), with translations derived indirectly from other modern languages, textual 75
corruptions, and several manipulations (Lefevre 1992). 76

The tool used for this study, Ugarit, is a web-based Translation Alignment editor de- 77
signed with ancient or low-resourced languages in mind (<http://ugarit.ialigner.com/>). 78
It is a crowd-sourcing project that enables users to align up to three parallel texts at 79
sentence or word level, specifically focusing on texts less represented in translation 80
alignment. 81



Figure 1: The home page of Ugarit.

The workflow is very simple: the user uploads the desired texts or imports them from 82

the Perseus Digital Library, and clicks on the words to align, which are then stored in the database as translation pairs. A progress bar allows the users to see how much of a text has been aligned. Users can create translation pairs aligning one word to another word (1-1), one word to many words (1-N), many words to one (N-1), and many to many (N-N). By default, the alignments are published on the platform in the 'New Alignments' panel, although users may opt out by simply selecting a different visibility option. The translation pairs can be further examined using the Alignment Statistics chart provided by Ugarit, which counts the frequency of the types of pairs created, or by downloading the whole dataset in XML or tabular format. It is also possible to analytically inspect published alignments by hovering with the mouse on each token: aligned words and expressions are highlighted in both texts. An additional service of transliteration of non-Latin alphabets is also provided for most of the languages currently aligned.



Figure 2: Public view of an alignment, showing the transliteration feature.

At the database level, Ugarit creates translation graphs, which can be used for dynamic lexica induction. Further, Ugarit allows users to inspect how other people aligned a specific word using the translation pairs search functions, which provides a contextualized visualization of an aligned pair (Yousef, Palladino, Shamsian, and Foradi 2022a).

Because of these powerful supporting features, Ugarit has been variously used for research, machine translation development, and language learning (Foradi 2019; Yousef, Palladino, Shamsian, and Foradi 2022b; Shukhoskvili 2017; Yousef, Palladino, Shamsian, Ferreira, et al. 2022; Yousef, Palladino, Wright, et al. 2022). The user pool currently counts 581 users, and more than 40 different languages including Ancient Greek, Persian, Latin, Egyptian, Coptic, Georgian, and Arabic, while more than 250,000 texts have been aligned by scholars, teachers, students, and non-experts.

2. Methodology

Ugarit demonstrates the potential of translation alignment in analytical tasks on texts and languages, based on the reflective evaluation of correspondences between words (Palladino, Foradi, and Yousef 2021): for this reason, it can also be used for the systematic comparison of translations of ancient texts (Shukhoskvili 2017).

The analytical study of translations through alignment is an operation of philology

and close reading (Berti 2019; Eve 2019). While it involves a certain amount of distant reading and NLP operations, such as POS tagging, lemmatization, and various kinds of queries, it also supports fine-grained research questions that require control on the data in a way that automatic methods alone do not allow (see below in our Conclusions). For example, it enables researchers to establish phrasal correspondences based on the peculiarities of the texts: Homeric texts, for instance, will have standard formulas that may be reflected in translations in various ways, and that a researcher may want to query to individuate particular trends. Moreover, it enables the study of languages currently not supported by effective NLP pipelines, such as Persian, as presented in this paper.

We used Ugarit to conduct the collection of translation pairs (TPs) used for this study. We used aligned translations of texts in Ancient Greek, selecting samples from Greek tragedy (Euripides' *Hippolytus*) and Homeric epic (*Iliad*). The texts were aligned against competing translations in English (Euripides) and in Persian (*Iliad*). The same annotator completed each category of alignments (Ancient Greek vs. English and Ancient Greek vs. Persian), ensuring that a consistent strategy was adopted in the establishment of translation pairs. Each annotator followed a set of guidelines designed for that particular language pair to provide more homogeneous alignments and reduce the chance of mistakes to a minimum¹.

Although Ugarit provides a local option to download the translation pairs and to visualize alignment statistics, we extracted all translation pairs directly from the database, so that we did not have to repeat the process multiple times with each individual alignment. Then, we analysed the following variables: 1) rate of non-aligned words in both languages; 2) word intersections, to investigate the rate of semantic overlap across the translations being compared; 3) TP ratios, measuring how many times one word in the original matched against one word in the translation (1-1) or against more than one word (1-N) and vice versa (N-1), and how many times groups of words were aligned in both texts (N-N); 3) for Ancient Greek and English, we were also able to measure intersections across parts of speech, to investigate how close the grammatical structures used in the translations were to the original text.

In this article, we present the results of the analysis on a small sample, as a showcase for our methodology: we plan to expand the study to a much larger dataset in future iterations.

2.1. Texts selection and rationale

The selection of the texts was limited chiefly by some contingent factors: first, we needed several adequately digitized translations of Ancient Greek works, available on the web, preferably covering a wide timespan and with some variation in audience and

1. This strategy was effective. By the end of the study, only two relevant mistakes had been detected in the whole corpus. This shows the importance of guidelines to reduce the chance of error, and reinforces the idea that guidelines and supervision need to be established especially when various annotators are at work.

destination (translations conceived for critical editions vs. translations addressed to a more general public or for performance, for example)².

Euripides' *Hippolytus*, a tragedy written in 428 BCE on the basis of a previous version now lost, is a text well attested on the web with plenty of digitized translations to choose from. The notorious character of the play, dealing with subjects such as incest and misogyny, makes it a frequent choice for both scholarly and more popular translations, while its longstanding tradition in Ancient Greek literature ensures the existence of early translations; moreover, the fact that this was a text conceived for theatrical performance opened more possibilities in terms of variety. The selection of the texts to align was very easy for anyone who knows the play: the prolog of Aphrodite, where the goddess introduces the main character Hippolytus (vv. 1-20), and the misogynistic monologue by Hippolytus himself (vv. 616-638).

For Persian, the scarcity of direct translations from Ancient Greek is the main challenge, as most translations are indirect and derived from mediating translation(s). Although most Ancient Greek texts have not only one, but multiple indirect translations in Persian, we wanted to include at least one direct translation, which limited the range of choices. The *Iliad* is one of the very few texts that, in addition to two indirect translations, also has a direct translation in Persian. Using translation alignment for a comparison between indirect and direct translations gives us practical information for evaluating accuracy and reliability. Considering that indirect translation are the main method for transmittion of the Ancient Greek texts to Persian, the question of their accuracy is of great significance. Moreover, the three translations of the *Iliad* come from different backgrounds and therefore show sufficient variation for testing our methodology.

3. Alignment of Euripides, *Hippolytus*

The user compared four competing translations of the Greek tragedy *Hippolytus*.³:

- D. Kovacs, 1995 (Euripides. *Children of Heracles. Hippolytus. Andromache. Ecuba*).⁴. Alignment of 1-20⁵, and 616-638⁶. This translation was selected as a specimen of a recent scholarly edition and previously praised for its programmatical faithfulness to the original (Gibert 2022).
- G. Theodoridis, 2010 (Euripides, Volume Three. *Medea, Herakleidae, Herakles, Hippolytus*).⁷. Alignment of 1-20⁸ and 616-638: ⁹. This translation is the only one

2. The only exception was D. Kovacs's English translation from Loeb, where individual passages were selected by a Ugarit user from the edition in print. The edition was chosen because it was a particularly appropriate example of a standard scholarly work on the *Hippolytus*.

3. In this early version of the paper, we selected the vv. 1-20 and 616-638. More are going to be added in the final version.

4. Edited and translated by D. Kovacs. Loeb Classical Library, Cambridge: Harvard University Press

5. <http://ugarit.ialigner.com/text.php?id=31127>

6. <http://ugarit.ialigner.com/text.php?id=31118>

7. Made available on the web for noncommercial use at <https://bacchicstage.wordpress.com/euripides/hippolytus/>, Accessed on 24 November, 2021

8. <http://ugarit.ialigner.com/text.php?id=31124>

9. <http://ugarit.ialigner.com/text.php?id=31119>

of the corpus that was written with a theatrical performance in mind (although not necessarily for a particular representation), rather than for reading.

- Ian Johnston, first edition 2016 (*Euripides, Hippolytus*¹⁰). Alignment of 1-20 and 616-638¹². A translation written specifically for a general public, including teachers and students of the tragedy, and the only one in poetry.
- E.P. Coleridge, 1910 (*The Plays of Euripides*¹³), Available on WikiSource. Alignment of 1-20¹⁴ and 616-638¹⁵. Commissioned as a prose translation by the publisher, it was delivered by the translator with the intent of being "an accurate rendering of the Greek text with some elegance of expression" (preface, p. 11). Evidently, the language is very distant from the three modern translations selected for this study.

The translations were aligned against the original text by the same user and with a consistent alignment method. The baseline was provided by already existing guidelines for the alignment of Ancient Greek and English¹⁶. However, these guidelines were conceived for the creation of alignment gold standards for the improvement of machine-actionable translation alignment: therefore, they prioritized linguistic principles and a rigid approach to translation units, regularly privileging word-to-word alignments, which are more useful to train automatic methods to an array of extremely diverse texts and authors. Therefore, a slightly revised version was used for this study, with the main goal of increasing tolerance towards author- and text-specific constructs, and consequently the number of phrase-to-phrase alignments, which are less useful for the implementation of automatic methods but are functional to the retainment (and query) of features that a translation scholar may deem important for analysis.

3.1. Discrepancies: analysis of non-aligned words

The visualization of the alignments on Ugarit provides a nice overview on the most visible characteristics of each translation, alongside a quick glance on the percentage of aligned and not aligned tokens between the compared texts.

10. Translated by I. Johnston, Vancouver island University. Nanaimo, British Columbia. URL: <http://johnstoniatexts.x10host.com/euripides/hippolytushtml.html>, Accessed on 24 November, 2021
 11. <http://ugarit.ialigner.com/text.php?id=31128>
 12. <http://ugarit.ialigner.com/text.php?id=31120>
 13. Translated into English Prose from the Text of Paley by Edward P. Coleridge. G. Bell and Sons, London
 14. <http://ugarit.ialigner.com/text.php?id=31126>
 15. <http://ugarit.ialigner.com/text.php?id=31121>
 16. https://github.com/UgaritAlignment/Alignment-Gold-Standards/blob/main/grc-eng/guidelines_grc-eng.pdf

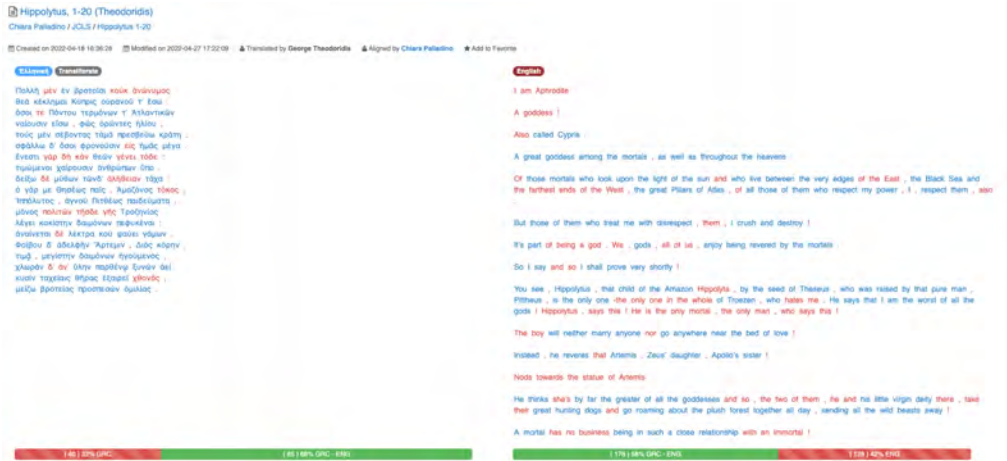


Figure 3: A screenshot of alignment ID 31124 (Theodoridis’s translation of vv. 1-20). Aligned tokens are in blue, while non-aligned tokens are in red. The progress bars below each text indicate an estimate of the non-aligned tokens in percentage, including punctuation.

Visualized alignments also make it easier to individuate overarching tendencies in non-aligned words. While Ugarit provides percentages that include punctuation, we have excluded punctuation from this analysis and provide exclusively numbers for individual words.

Translation	NA in Greek			
Hipp. 1-20	NA in English			
(126 words)				
Hipp. 1-20	NA in Greek			
(126 words)				
Hipp. 616-638	NA in English			
(143 words)				
Hipp. 616-638				
(143 words)				
Kovacs	13	10	6	18
Theodoridis	20	82	17	102
Coleridge	11	14	6	5
Johnston	19	31	8	30

Table 1: Calculation of non-aligned words in the translation and in the original in two sections of *Hippolytus*.

The lack of alignment in Ugarit indicates that the user, following the guidelines provided, did not find an acceptable correspondence for an individual word or group of words in the other text. Wherever the number of non-aligned tokens in one of the two texts is significantly lower or higher than the other, it suggests that something in the original was omitted or overlooked, or that there is a tendency to expansion and paraphrasing in the translation. Obviously, an analysis of the non-aligned words is required: we extracted the list of non-aligned tokens from the Ugarit database, both in the source and target text, for further investigation. As it is to be expected, the POS and intersection data reveal that most translators omit functional words, conjunctions, and particles in the Ancient Greek text, and add some on their own that are not in the original: therefore,

words like μέν, τε, καί, δέ, δή, ἄν, γάρ, ἦ are frequently omitted by all translators, while 222
English words like 'and', 'even', 'but', 'then', are also frequent additions. 223

In the translation by Theodoridis, the number of non-aligned tokens presents the most 224
staggering ratio between original and translation (overall 37 non-aligned words in Greek, 225
and 184 non-aligned words in English!). This measure suggests that, while a substantial 226
part of the original was left out, there is a very visible counter-tendency to expand on 227
the original, as a stylistic choice going beyond what a translator would normally do to 228
explain a word or expression to their audience. This is confirmed by the analysis on the 229
individual non-matching words, which include typologies way beyond stopwords and 230
particles (Greek constructs: *κοῦκ ἀνώνυμος*; nouns and concepts: *βάρος, γυνή, κακόν,* 231
καλόν, χαίρων, γένει, χθονός, πολιτῶν, ἀλήθειαν, etc.). 232

As it is perhaps to be expected, Kovacs shows a balance in the number of non-aligned tokens in both languages (19 words in Greek, 28 in English), with little more than stopwords and particles being omitted from the original, and some additional explanatory words in the translation, particularly in vv. 616-638. Moreover, the ratio between non-aligned tokens in Ancient Greek and English shows that there is no strong tendency towards expansion in the translation, as the number of non-matching English tokens is not remarkably higher than the Greek ones.

Somehow more surprisingly, the rate is much more skewed for the other modern trans- 240
lator, Johnston, where the number of non-aligned English tokens is significantly higher 241
than the number of non-aligned Greek tokens. Moreover, while the Greek mostly in- 242
cludes stopwords, pronouns, and particles (e.g. γάρ, μέν, καί, τε, εις, δή), the English 243
clearly shows a tendency towards expansion, with the addition of significant words and 244
concepts that tend to be explanatory of the Greek (e.g. the would-be husband, wife, 245
worthy family, disparage, bestow, Hippolyta, lad, women, god, time, etc.). 246

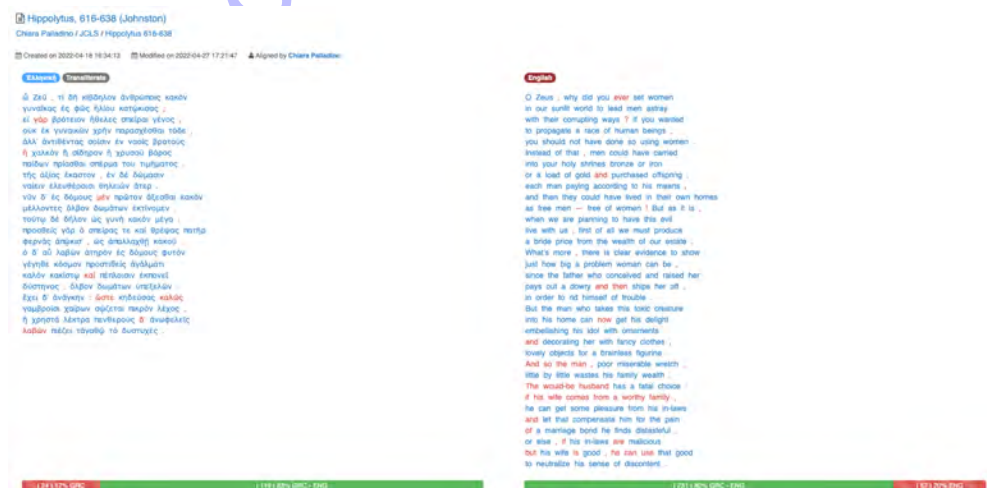


Figure 4: A screenshot of alignment ID 31120 (Johnston’s translation of vv. 616-638). Aligned tokens are in blue, while non-aligned tokens are in red. The progress bars below each text indicate an estimate of the non-aligned tokens in percentage, including punctuation.

Finally, perhaps the most surprising of all is Coleridge, which has the lowest and most uniform number of non-aligned words across the board (17 in Greek, 19 in English). In only two cases the omission of the Greek is particularly relevant, both in vv. 1-20: the expression *κοῦκ ἄνώνυμος* referred to Aphrodite (lit. "not anonymous", famous), which is paraphrased and incorporated in the rest of the verse as "wide o'er man my realm extends, and proud the name", and the verb *ψάυει* (lit. "scorns", scil. marriage), which is replaced in context with "will (have) none of it". In the remaining cases, most of the words omitted are stopwords or redundancies (e.g. the word *μύθων*, 'words', is omitted from the expression 'and the truth of this (i.e. these words)').



Figure 5: A screenshot of alignment ID 31126 (Coleridge's translation of vv. 1-20). Aligned tokens are in blue, while non-aligned tokens are in red. The progress bars below each text indicate an estimate of the non-aligned tokens in percentage, including punctuation.

3.2. Similarities: analysis of intersection data

We extracted intersection data from all four alignments, then compared intersections across all translations and across any combination of them. We observed that the intersection between each pair of translations is *always* minimal. Overall, among all four translations in *Hipp.* 1-20 (125 Greek words) the intersection was for 8 translation pairs (TPs) after capitalization (but before lemmatization): 'ἡμᾶς' - 'me', 'δ' - 'but', 'τ' - 'and', 'τάμὰ' - 'my', 'κράτη' - 'power', 'Ἀρτεμιν' - 'Artemis', 'κόρην' - 'daughter', 'ἀδελφήν' - 'sister'.

In *Hipp.* 616-638 (143 Greek words) the overall intersection was for 7 translation pairs (TPs): 'ἢ' - 'or', 'ἐν' - 'in', 'σίδηρον' - 'iron', 'τί δὴ' - 'why', 'εἰ' - 'if', 'γυναικῶν' - 'women', 'τε καὶ' - 'and'.

Overall, the intersection is not only minimal, but relatively insignificant as to the typologies of overlapping words, which include in the majority adpositions, such as *ἐν*, particles such as *τε* or *δέ*, and conjunctions such as *καί* and *εἰ*, which have a limited array of options for translation.

Some more intersections could be added by including minor changes due to editorial choices (e.g. presence of determiners or different capitalization), and focusing exclusively on semantic similarity after lemmatization, i.e. only considering how a lemma was translated regardless of how its inflected form was rendered. These additions allow

us to expand the list a little more, but they omit important contextual information re- 275
 garding syntactic choices of each translator, and should be taken with caution.. With this 276
 increased level of tolerance, we may include the following pairs: 'θεά' - 'goddess', 'Ζεύς' 277
 - 'Zeus', 'Τροζηνία' - 'Troizen', 'Πιτθεύς' - 'Pittheus', 'Ἱππόλυτος' - 'Hippolytus', 'ὄλβος' 278
 - 'wealth', 'χρυσός' - 'gold', 'Ζεύς' - 'Zeus', 'γυνή' - 'woman', 'γένος' - 'race', 'λέκτρον' - 279
 'wife', 'φερνή' - 'dowry' (but note 'dower' in Coleridge). Finally, we may add cases where 280
 there is a 3/4 overlap and where the fourth translation is only minimally different. These 281
 include: 'ἕκαστος' - 'each man', 'ἐλεύθερος' - 'free', 'χαλκός' - 'bronze', 'βροτός' - 'man', 282
 'χρηστός' - 'good', 'Ἀμαζών' - 'Amazon' (but note Kovacs, 'the Amazon woman'), 'ἐγώ' - 283
 'I', 'θεός' - 'god'. 284

Overall, we observe more regular overlap in the following categories: 285

- Proper nouns: 'Ἄρτεμις' - 'Artemis', 'Ζεύς' - 'Zeus', 'Ἀμαζών' - 'Amazon', 'Θησεύς' - 286
 'Theseus', 'Πιτθεύς' - 'Pittheus', 'Ἱππόλυτος' - 'Hippolytus'. . 287
- Functional words, like prepositions, adverbs, conjunctions, and pronouns 'καί' - 288
 'and', 'ἐν' - 'in', 'εἰ' - 'if', 'ἐς' - 'into', 'ἐν' - 'among', 'ὅσοι' - 'all', 'γάρ' - 'for'. 289
- Common or very common words that have a standardized meaning in a given con- 290
 text, especially including words related to the religious or family sphere: 'ὄλβος' 291
 - 'wealth/family wealth', 'δῶμα' - 'home/house/estate', 'λέκτρον' - 'wife', 'θρέπω' 292
 - 'to raise', 'φερνή' - 'dowry', 'γυνή' - 'woman', 'βρότειος/βροτεία/βροτός' - 'mor- 293
 tal/mortal kin/mortal man', 'ναός' - 'temple/shrine', 'οὐρανός' - 'heaven', 'σέβω' 294
 - 'to respect (a deity)', 'πατήρ' - 'father', 'πενθερός' - 'in-law', 'τιμάω' - 'to receive 295
 honors/be revered'. 296
- Technical or rare words that have few established meanings, with little left to 297
 stylistic choice: 'σίδηρος' - 'iron', 'χαλκός' - 'bronze', 'χρυσός' - 'gold'. 298

Even in these cases, however, some words belonging to the same categories do not 299
 overlap across translations. For example, certain proper nouns are sometimes translated 300
 differently. The most prominent case is 'Πόντος', a name that simply stands for 'sea', 301
 but is conventionally referred to the Black Sea when capitalized. Kovacs and Johnston 302
 both opted for the Latinized name Euxine Sea (Pontus Euxinus), more traditional 303
 for scholarly translations, Theodoridis preferred the more modern version 'the Black 304
 Sea', while Coleridge simply translated 'the sea' (non-capitalized). An interesting case 305
 occurs with the genitive 'Θησεώς', which is always translated as "of Theseus", except in 306
 Theodoridis, who expands the translation adding contextual meaning "by the seed of 307
 Theseus". 308

Additional patterns of overlap can be seen across each pair of translations: 309

For the two translations that have the highest intersection (Kovacs and Johnston), there 310
 is a high level of literal overlap in all the categories described above: we observed a 311
 total of 22 functional words (prepositions, conjunctions, adverbs), 6 proper names, 312
 and 11 family or religious names. However, there are also more correspondences in 313
 cases where conscious stylistic choices were made: both authors translated 'Κύπρις' as 314

Translations (126 words) (143 words)	<i>Hippolytus</i> 1-20 <i>Hippolytus</i> 616-638	
Kovacs-Johnston	22 (17.46%)	29 (20.27%)
Kovacs-Coleridge	20 (15.87%)	25 (17.48%)
Theodoridis-Johnston	12 (9.5%)	13 (9.09%)
Theodoridis-Coleridge	11 (8.7%)	13 (9.09%)
Theodoridis-Kovacs	15 (11.9%)	15 (10.04%)
Johnston-Coleridge	13 (10.31%)	14 (9.07%)
Average values	15 (11.9%)	18 (12.58%)
Total intersection	7 (5.5%)	7 (4.89%)

Table 2: Word intersections across translations of the *Hippolytus*.

‘Aphrodite’, ‘Πόντος’ as “the Euxine Sea”, and the genitive ‘Ἀτλαντικῶν’ as ‘(the Pillars) of Atlas’. Moreover, certain common words were translated in the same way: ‘χθονός’ - ‘land’, ‘ταχύς’ - ‘swift’, the gen. ‘οὐρανοῦ’ was translated by both authors as a locative (‘in heaven’), and the word ‘φυτόν’ was translated by both with the neutral ‘creature’, as opposed to both Theodoridis and Coleridge, who chose derogatory words (‘beast’ and, remarkably, ‘weed’).

Coleridge’s much older translation does not fare as bad as we would expect, although still below the average in most cases. The type of overlap, however, is limited to functional words and a few of the common categories we have observed above, with very little that may be connected to conscious stylistic choices. On the other hand, Coleridge distinguished himself in a few cases, where a more liberal translation was chosen: e.g. ‘οὐρανός’ - ‘heaven’s courts’, as opposed to the prevalent ‘heaven’; the acc. of motion ‘χλωρὰν ὕλην’ (lit. through the green forest) - ‘through the Greenwood’ (capitalized); ‘λέκτρον’ (lit. the marriage bed) - ‘Love’ (capitalized); ‘ἥλιος’ (the sun) - ‘the sun-god’; ‘δῶμα’ (house) - ‘independence’; ‘λέχος’ (lit. marriage bond) - ‘wife’.

Theodoridis’s translation, on the other hand, regularly scores low intersections. This may be explained by the very different destination of this translation, which was conceived for the scene, rather than for reading. One relevant exception is the overlap with Kovacs in the peculiar translation of the word ‘λέκτρον’ as ‘the bed of love’, which is unique in our dataset. This phenomenon is coupled with very distinctive choices in cases where there is strong semantic overlap in the other three: ‘παρθένος’ (lit. virgin) - ‘his little virgin deity’, the gen. ‘Θησέως’ - ‘by the seed of Theseus’, ‘ναίω’ (to dwell) - ‘live out their lives’, ‘χαλκός’ - ‘some piece of bronze’, ‘σπείρω’ - ‘to sow the seeds’, ‘ἐκπονεῖ’ (lit. he works out, finishes off) - ‘he begins the little game of cajoling’ (sic!), ‘λαμβάνω’ - ‘to bring’, as opposed to ‘to take’ in the other three.

As much as similarities matter, there are also parts of the text that are regularly translated in completely different ways. This was most commonly observed in fixed expressions and idiomatic constructs, which were addressed very differently by each translator. For example, the expression ‘φρονοῦσιν μέγα’ (lit. they think great things) was only translated literally by Kovacs (‘think proud thoughts’), but it was bound together and

paraphrased by the other three: 'treat with disrespect' (Theodoridis), 'stuffed with pride' (Johnston) and 'vaunt themselves' (Coleridge). More conventional expressions that would be recognized by any student of Ancient Greek are always translated in a different way. For example, the popular tragic expression ἔχει δ' ἀνάγκην (lit. it is necessary): 'there is a fatal necessity' (Kovacs), 'And then come the unavoidable choices of his constraints' (Theodoridis), '(he) has a fatal choice' (Johnston), 'For he is in this dilemma' (Coleridge). Other fixed, recurring expressions such as νῦν δέ' (lit. but now): 'But as matters stand' (Kovacs), 'As it is now' (Theodoridis), 'But as it is' (Johnston), 'But now' (Coleridge), and τοῦτω δὲ δῆλον (it is clear from this): 'The clear proof is this' (Kovacs), 'Here's the clear proof of it' (Theodoridis), 'What's more there is clear evidence to show' (Theodoridis), 'Tis clear from this' (Coleridge).

We observed the widest and most consistent disagreement in the translation of one single word: the neutral adjective κακόν, which appears repeatedly in vv. 616-638 as a signpost for 'woman'. Each translator used a noticeable variety of synonym, including 'curse', 'evil', 'bane', 'problem', 'plague', 'trouble', 'unbearable burden', 'mischief', 'worthless' and 'brainless figurine' (sic!), with remarkable variety in the space of about twenty verses.

3.3. Translation pair ratios

The charts below show the ratios of Translation Pairs (TPs) across the four translations: 1-1 indicates a match of one word in the original to one word in the translation, 1-N one word in the original with more than one word in the translation, N-1 more than one in the original with one word in the translation, and N-N indicating many words to many words in both.

In general, the ratios are consistent across the group, with about a third of 1-1 TPs, regularly higher in Coleridge and in Kovacs, and lower in Theodoridis and Johnston. 1-N TPs are also regularly between 50 and 60% of the total in all four translations. This phenomenon can be partly explained with the fact that Ancient Greek is an inflected language, where meaning is added by means of changing the ending part of words, while English is very marginally inflected, and it tends to use more words to convey the same ideas; moreover, English and Ancient Greek make a very different use of determiners (e.g. definite articles), and English tends to use them much more often, effectively duplicating the number of words used.

Part of these trends, however, can be explained as the result of conscious translation choices. The rates of 1-N and N-N TPs are particularly high for Theodoridis and Johnston. In the case of the former, this further substantiates the impression that his translation has a tendency towards expansion, as observed above. Johnston, however, is a close second. In fact, despite the fact that his translations are often semantically similar to the rest of the group, and to Kovacs in particular (see above), the high 1-N ratio shows a considerable tendency towards expansion. Compare cases like ἐκτίνομεν - 'we must produce a bride price from' (as opposed to 'we pay out, we bring to the ground'), the dative of disadvantage ἀνθρώποις (lit. against men) is translated emphatically with 'to

lead men astray', 'φῶς ἡλίου' (lit. the sunlight) - 'our sunlit world', the dative 'κακίστῳ' (lit. for the worst) - 'for a brainless figurine' (sic!).

To sum up our observations so far, the combination of non-aligned words, semantic overlap across translations, and TP ratios can be used to reveal features of translations that have tendencies to build upon, or explain, the original. Johnston is an interesting case in point: while the intersection data reveals a nice semantic overlap with the other modern and scholarly translation by Kovacs, the combination of TP ratios and non-aligned words suggest that Johnston expands and explains more broadly and more freely, and omits more substantially as well. His translation, in fact, is not necessarily created to be read alongside the original, as in the case of Kovacs (the Loeb edition of Euripides features the Ancient Greek text to the side), but is conceived for an extended public of teachers, students, and general readers interested in the ancient world but not necessarily familiar with Ancient Greek.

Kovacs may be expected to be the most consistent translator in the ratio of translation pairs. Quite surprisingly, however, Coleridge's translation shows very similar overall scores, if not even superior (e.g. the higher percentage of 1-1 TPs). So, while the intersection data suggest the peculiarity of the language and translation choices of Coleridge, the non-aligned words and the TP ratios tell a different story, showing how he is still very adherent to the original text, with a very high degree of word-to-word correspondence, and very little tendency towards expansion or omission.

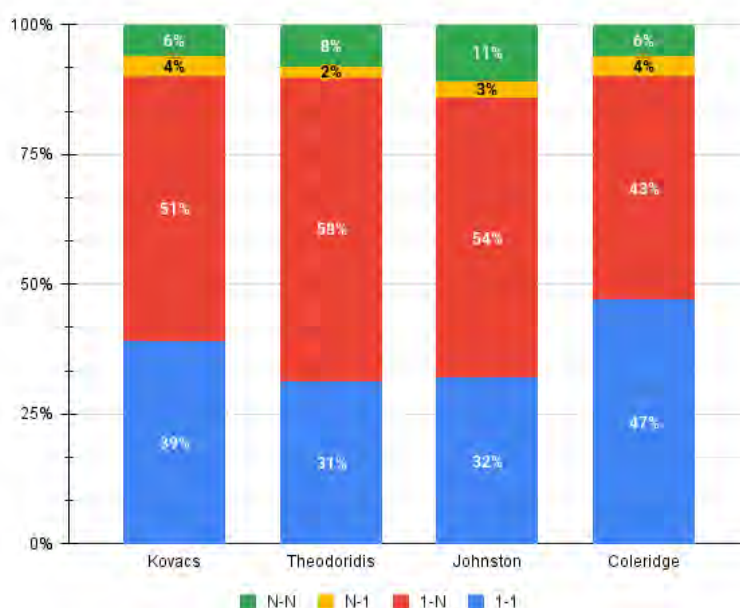


Figure 6: Translation Pair ratios across the four translations of *Hippolytus*, vv. 1-20.

	Euripides, Hippolytus, vv. 1-20				Euripides, Hippolytus, vv. 616-638			
	1-1	1-N	N-1	N-N	1-1	1-N	N-1	N-N
Kovacs	38 (43%)	48 (54%)	1 (1%)	2 (2%)	43 (39%)	56 (51%)	4 (4%)	6 (6%)
Theodoridis	27 (34%)	49 (61%)	1 (1%)	3 (4%)	30 (31%)	56 (58%)	2 (2%)	8 (8%)
Johnston	30 (37%)	46 (57%)	2 (2%)	3 (4%)	33 (32%)	56 (54%)	3 (3%)	11 (11%)
Coleridge	44 (48%)	44 (48%)	0 (0%)	3 (3%)	51 (47%)	46 (43%)	4 (4%)	7 (6%)

Table 3: Translation Pair ratios across the four translations of *Hippolytus*, with percentages.

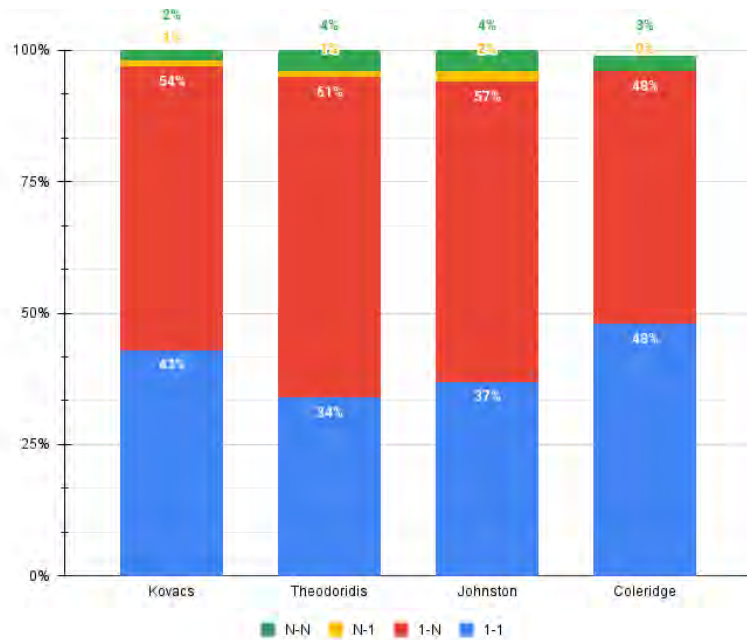


Figure 7: Translation Pair ratios across the four translations of *Hippolytus*, vv. 616-638.

3.4. POS Tags

406

We also examined parts of speech in the Greek and compared them against each translation, to investigate whether translators would tend to use similar grammatical structures to what they found in the original. We used UDPipe¹⁷ trained on an Ancient Greek dataset to extract POS data (Straka and Straková 2017, Celano, Crane, and Majidi 2016), and revised the results manually to increase accuracy¹⁸.

411

The categories where variation was less common were nouns and adjectives: often nouns would be translated with other nouns, adjectives with other adjectives, and proper nouns with other proper nouns. Often, however, English would add words not present in the original, such as determiners, e.g. 'φῶς' - 'the light', 'θεά' - 'a/the goddess', 'κακὸν' - 'a bane' or 'this plague', 'φύτον' - 'this creature' or 'the weed', 'ὄλβος' - 'the

416

17. <https://lindat.mff.cuni.cz/services/udpipe/>

18. Overall, UDPipe trained on the Perseus model for Ancient Greek gave remarkably good results, which only needed minimal revision. This is encouraging for future study, where the amount of manual correction will necessarily reduced, as the size of the dataset increases. See further in our Conclusions.

wealth'; or possessive adjectives, e.g. 'ἄγαλμα' - 'his idol', 'γαμβρός' - 'his in-laws', 'δῶμα' - 'their/his/their own house. Notably, in a few cases nouns were translated with a pair adjective+noun, a phenomenon mostly recurring in Johnston: e.g. 'βροτός' - 'mortal man', 'ὁμιλία' - 'a close relationship', 'παιδεύμα' - 'a student trained'.

Verbal forms displayed the highest degree of variation across all four translations: very often, translators would alter tense, person, mood, or voice of a given form in the original. This happened frequently, but not exclusively, with Greek participles. E.g. 'σέβοντας' - 'reverence', 'τιμώμενοι' - 'from those honours', 'ὑπεξελών' 'there goes bit by little bit', 'κηδεύσας' - 'a man makes', 'λαβών' - 'takes', 'μέλλοντες' - 'we would'.

Other variations were due to linguistic factors: while in Greek the subject of a verb can be left implicit and rendered only with the verb's personal ending, in English the subject has to be explicit, e.g. 'ἀναίνεται' - 'He shuns' / 'he refuses', 'ἐξαίρει' - 'he clears of', 'τιμᾷ' - 'he honors', 'ἤθελες' - 'you wanted'.

Overall, full overlap in POS was not frequent except for functional words and particles, even considering some of the categories described above as partial matches: interestingly, the highest rate of overlap was in Kovacs and in Coleridge, while Johnston and Theodoridis, predictably, had the lowest score. The situation was not substantially changed by considering both full and partial matches. This reinforces the observations made above, that Coleridge, although being stylistically distinct from the rest of the translations, seems indeed to be more "faithful" to the original even in morphology, while the one translation designed with a substantially different destination in mind is also the most distant from the original in every way.

Translations	Matching POS			
<i>Hippolytus</i> 1-20 (126 words)	Non Matching POS			
<i>Hippolytus</i> 1-20 (126 words)	Matching POS			
<i>Hippolytus</i> 616-638 (143 words)	Non Matching POS			
<i>Hippolytus</i> 616-638 (143 words)				
Kovacs	24	65	34	75
Theodoridis	16	64	23	73
Coleridge	22	69	36	72
Johnston	14	67	24	79

Table 4: Matching and non-matching part-of-speech tags across *Hippolytus*.

4. *Iliad* 1-67: Comparing Ancient Greek and Persian translations 439

The third subset includes alignments of *Iliad* 1.1-67 with three Persian translations: 440

- Saeed Nafisi, 1958 (Nafisi 1958), Derived from the French translation.¹⁹ 441

19. Alignment: <http://www.ugarit.ialigner.com/text.php?id=28503>

- Mir Jaleddin Kazzazi, 1998. (Kazzazi 1998), Derived from the French translation.²⁰
- Farnoosh Shamsian, 2020, translated directly from Greek²¹.

All three translations were aligned by the same annotator following the same guidelines. Two out of the three translations are indirect, using French translations as mediating texts. Unlike Greek to English, direct translations from Greek to Persian are rare; however, indirect translation is a common practice and most major texts even have multiple indirect translations, usually from English, French or German. Although indirect translation might be less efficient for translation alignment, it is still the main medium for the transfer of Greek texts to Persian and consequently, has a significant impact on the reception of Greek culture among Persian speakers.

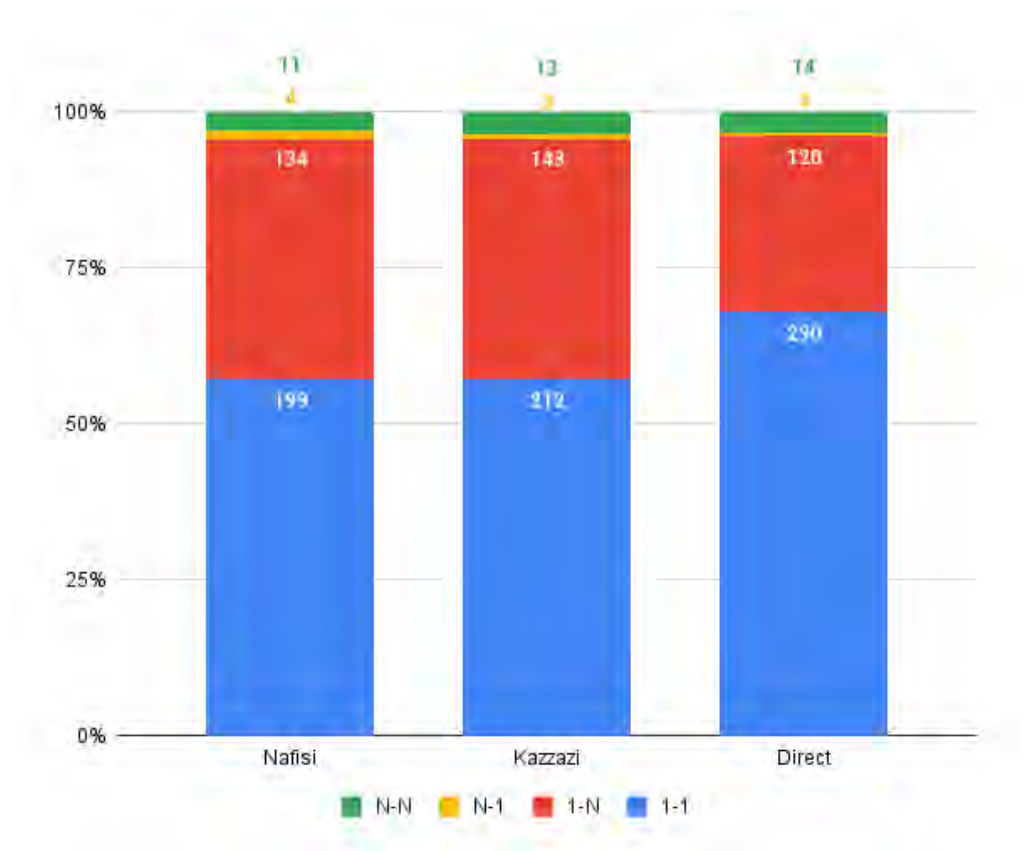


Figure 8: Translation pair ratios across all translations of the *Iliad*

There are similar trends between the two indirect translations in comparison with the direct one. Both indirect translations have a lower number of 1-1 pairs and higher number of 1-N pairs in comparison to the direct translation. This is mainly caused by phrasal translation of certain Greek words, particularly of epithets. For instance,

20. Alignment: <http://www.ugarit.ialigner.com/text.php?id=28502>

21. Available on <https://github.com/farnoosh-shamsian/Iliad> Alignment: <http://www.ugarit.ialigner.com/text.php?id=28504>

translation of the word "ἐὐκνήμιδες" has 4 tokens both in Kazzazi and in Nafisi, but only 1 token in Shamsian. The ratio of N-N or N-1 pairs doesn't show considerable differences between the three translations.

However, the most substantial difference is not in the ratio of the pairs, but in the number of non-aligned tokens. The indirect translations are generally longer, Kazzazi's translation with 901 and Nafisi's with 742 token in comparison to 542 of Shamsian, and have a much higher number of non-aligned tokens, Nafisi with 233 and Kazzazi with 337, while the non-aligned tokens in the direct translation are minimal.

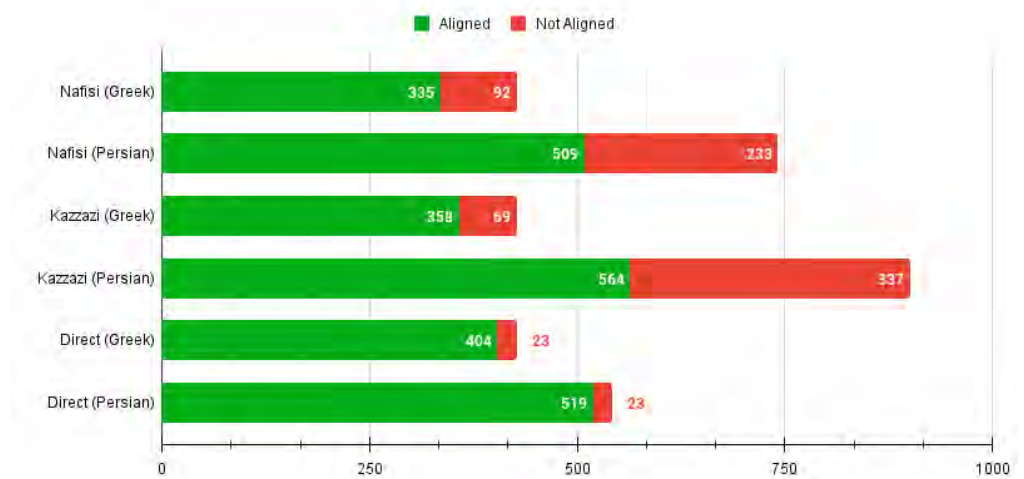


Figure 9: Ratio of aligned and non-aligned tokens across all translations of the *Iliad*

One reason for the higher number of non-aligned tokens in the indirect translations might be that they tend to be more descriptive and use multiple synonyms which have no equivalent in the Greek text but correspond with the mediating text. Both indirect translations are derived from the French edition by Eugène Lasserre (Homer 1965) while consulting other translations such as Mazon (Homer 1962) and Leconte de Lisle (Homer 1867). The differences might be better demonstrated on a sentence level. For instance, both indirect translations of the *Iliad* 1.25 have 16 tokens and the direct translation has 12 tokens (Except in graphs, all Persian texts have been transcribed for easier formatting).

Greek	ἀλλὰ κακῶς ἀφίει , κρατερὸν δ' ἐπὶ μῦθον ἔτελλε (Hom. Il. 1.25)	
English Translation	<i>Murray:</i>	He sent him away harshly, and laid upon him a stern command
French Translation	<i>Lasserre:</i>	Méchamment il renvoya Chrysès, sur cet ordre rude
	<i>Mazon:</i>	Brutalement il congédie Chrysès, avec rudesse il ordonne
	<i>Leconte de Lisle:</i>	Il le chassa outrageusement, et il lui dit cette parole violente
Nafisi Translation (indirect)	از سر خشم و کبر کریزس را روانه کرد و به سختی و خشونت فرمان داد az sar-e xašm va kebr kerizes rā rawāne kard va be saxti va xošunāt farmān dād [from] [base of] [anger] [and] [pride] [kerizes] [rā*] [going] [made] [and] [with] [harshness] [and] [violence] [order] [gave]	
Kazzazi Translation (indirect)	سخت و درشت ، شریسه را بازگردانید و خروشان و آتشی خوی ، در ستیزه با او ، فرمود saxt va dorošt, šerise rā bāzgrdānīd va xorušān o ātašin-xuy, dar setize bā 'u farmud [harsh] [and] [coarse] [šerise] [rā*] [sent back] [and] [roaring] [and] [with fiery disposition] [in] [dispute] [with][him/her][ordered]	
Shamsian Translation (direct)	دگر ظالمانه [خروس را] معزول میداشت و با کلام قاطع فرمان میداد degar zālemāne (xoruses rā) ma'zul midāšt va bā kalām-e qāte' farmān midād [rather] [ruthlessly] [xoruses] [rā*] [discharged] [was making] [and] [with] [speech] [forceful] [order] [was giving]	
*Postposition, commonly used as a marker of the object		

Figure 10: Translations of Hom. Il. 1.25 with transcription and glosses.

Since the guideline prioritizes 1-1 alignment, only one of the multiple synonymous 473
equivalents was aligned with the Greek, leaving other synonyms unaligned. In the 474
example of *Iliad* 1.25, the Greek word κακῶς is translated to [zālemāne] in the direct 475
translation, and to [az sar-e xašm va kebr] and [saxt va dorošt] in the indirect translations. 476
The word κρατερὸν in the same line is translated to [qāte'] in the direct translation and to 477
[be saxti va xošunāt] and [xorušān va ātašin-xuy, dar setize ba 'u] in the indirect translations. 478

It should be considered that a change of approach in the guideline could significantly 479
affect the ratio of translation pair. For instance, according to our guidelines, when a 480
word in the Ancient Greek is translated to two or more synonymous word, only one 481
of the equivalents should be aligned. A different approach in the guidelines could 482
have resulted in mutiple 1-N pairs by including the synonymous equivalents instead of 483
leaving them unaligned. 484

Non-aligned tokens of the Greek text Not surprisingly, the number of non-aligned tokens 485
in the Greek text is higher in the indirect translations, 92 in Nafisi, 69 in Kazzazi compared 486
to 23 in Shamsian. Most Greek words without equivalents in the direct translation are 487
particles, often δε and τε, with 8 and 6 incidences respectively out of the 23. On the 488
other hand, the non-aligned tokens in the indirect translations also include nouns, verbs 489
and even phrases, caused by semantic variation through the mediating texts. 490

Greek	ἰστὸν ἔποιχόμενῃν (<i>Hom. Il. 1.31</i>)	
English Translation	<i>Murray:</i>	as she walks to and fro before the loom
French Translation	<i>Lasserre:</i>	tissant la toile
	<i>Mazon:</i>	allant et venant devant le métier
	<i>Leconte de Lisle:</i>	tissant la toile
Nafisi Translation (indirect)	Rajzanān	رجزنان بر دار [bāfandegi] [knitting/weaving] [on] [loom of] [weaving]
Kazzazi Translation (indirect)	āngāh ke piše'i rā mivarzad	آنگاه که پیشه‌ای را می‌ورزد [then] [that] [a profession] [rā*] [practise]
Shamsian Translation (direct)	ānjā dar barābar kārgāh-e man dar raft o āmad xāhad bud	آنجا در برابر کارگاه من در رفت و آمد خواهد بود [there] [in] [front of] [workshop of] [me] [in] [going] [and] [coming] [will] [be]

*Postposition, commonly used as a marker of the object

Figure 11: Translations of Hom. Il. 1.31 with transcription and glosses.

While the indirect translations do not correspond with the Greek text, they can be aligned 491
with the French translation, particularly with Mazon. For instance, the alignment of 492
Nafisi's translation of Hom.II.1.31 with Mazon would produce the following pairs, 493
leaving only two tokens unaligned in Persian, [ānjā] and [man]: 494

'allant et venant' - [dar raft-o-āmad xāhad bud], 'devant' - [dar barābar-e], 'métier' - [kārgāh] 495

Intersections The intersection data extracted from all three translations indicates a high 496
degree of variance and there seems to be no significant difference between direct and 497
indirect translation: 498

Translations	Intersection data
Iliad 1-67	
Nafisi-Kazzazi	70
Nafisi-Shamsian	71
Kazzazi-Shamsian	75
All	70

Table 5: Word intersections across translations of *Iliad*

Most of the intersection consists of pronouns and certain particles. Some examples 499
are 'εἴ' - [agar] meaning 'if', 'ἀλλ' - [amā] meaning 'but', 'ἡμῖν' - [mā] meaning 'we', or 500
'ἐνὶ' - [dar] meaning 'in'. There are also instances of some common words that have 501
a standardized translation, such as 'νυκτι' - [šab] meaning night, 'θαλάσσης' - [daryā] 502
meaning sea, or 'πόλεμός' - [jang] meaning war. 503

Some of these intersections are proper nouns; however, contrary to the high overlap of 504
Greek proper nouns that we see in the English translations, most proper nouns do not 505
match in the Persian translations. Part of these differences is caused by the influence of 506

the mediating language on pronunciation and others by the limits and characteristics of the Persian writing system. Still, few names have only one writing, such as Zeus or Apollo (in Persian, [*āpolon*]). Examples of proper names with multiple spellings are:

Greek	English	Nafisi	Kazzazi	Shamsian
Ἀχιλλεύς	Achilles	(āxilus) أخيلوس	(āšil) آشیل	(āxile'us) آخیلئوس
Ἀτρείδης	Son of Atreus	فرزند آتره (farzand-e ātre)	پور آتره (pur-e ātre)	پور آترئوس (pur-e ātre'us)
Χρύσης	Chryses	کریزس (kerizes)	شریسه (šerise)	خروسس (xoruses)
Χρύση	Chryse	کریزه (kerize)	شریسا (šerisa)	خروسه (xoruse)

Figure 12: Variations of proper names in translations of the *Iliad*

5. Conclusions and Future Work

In this paper, we presented a preliminary insight into what could be done with the resources provided by translation alignment and a tool like Ugarit, specifically applied to the study of translations of ancient texts. To sum up, we presented a combination of the following criteria as measures of the interaction between translation and original, and across various translations: 1) number of non-aligned tokens in both languages and the ratio between the two; 2) intersection data, implemented with lemmatization, across competing translations; 3) ratios of translation pairs; 4) POS tags and their intersection, limited to Ancient Greek and English. We used a combination of these criteria to examine trends across our translations. For *Hippolytus*, these observations led to the somewhat surprising conclusion that a 1910 translation, despite a completely different set of stylistic and linguistic choices, was in fact more literal and adherent to the original than the modern and academic ones by all accounts. For the *Iliad*, the application of these criteria supported the isolation of phenomena specific to indirect translations, such as the peculiar rendering of proper names.

The work here presented is part of a larger effort in upscaling the functionalities of Ugarit and its user base, and it is conducted in parallel with the development of Alignment Guidelines for various types of projects, often but not exclusively with Ancient Greek as a source language. Our future work is oriented towards considerably expanding the dataset of parallel corpora at our disposal, to apply this methodology to a larger group of texts.

First of all, we are using alignment guidelines and Gold Standards for the development of a multilingual translation model, which should considerably facilitate the collection of translation pairs (Yousef, Palladino, Shamsian, Ferreira, et al. 2022) and alleviate the burden of the manual work that is required at present. Second, the expansion of the corpus will amplify the tolerance to errors in the establishment of translation pairs and in the POS analysis. The bigger the dataset, the less minor mistakes are going to

affect the overall conclusions, while at the same time reducing the demand for intensive manual supervision.

On the other hand, it is likely that lemmatization and lexicalization will be more important with larger corpora, where there will be less space for the analysis of individual subtleties.

Overall, a larger dataset will help limit the incidence of errors in the analysis. However, in the course of this study we also observed that the manual intervention of a scholar in establishing certain kinds of translation pairs is essential to conduct an analytical study. Automatic models tend to privilege 1-1 TPs, but a researcher may be interested in investigating phrasal translations, or in collecting instances of peculiar adjectivization or expansion, and so on. This implies that there needs to be a certain level of manual supervision and intervention, regardless of the size of the corpus, and a philological approach is essential to the design of a dataset that aims at the investigation of translations of ancient texts. Finally, it needs to be emphasized that some languages, such as Persian as described in this paper, do not have the luxury of accurate NLP models: while an alignment-based analysis of translations is still very important for these languages, there is a serious hindrance to the generalized application of automated methods to produce supporting data.

Since Antiquity, translations have been a medium between cultures, not just between languages. This was well known to modern philologists, who reflected upon the necessity of translating the Classics as a cultural problem of transferring an “alien” literature and its values. Wilhelm von Humboldt acknowledged that translations are essential to non-expert audiences: however, he advised to read the Classics by comparing multiple translations, to make sure that the readers could somehow get a sense of the complexity of the original text (Humboldt 1816).

It's an interesting quote, thus it may be worth providing the verbatim quotation in footnote, or at least the page number

Translations are witnesses of the linguistic and semantic complexity of ancient texts. Their very different approaches to them are reflected in the very little consistency and lack of semantic overlap, even in texts that are somehow editorially stable, such as the *Iliad* and Greek tragedy. Our study reflects how translators' choices and decisions create, at all effects, very different texts and very different impressions of the original: while lack of linguistic overlap is to be expected, it is very surprising to see how there is very little consistency in addressing even the least ambiguous types of words, such as personal or place names. Overall, patterns of inconsistency and instability can be detected across the board: in the semantics, word choices, grammatical constructs, and in the establishment of word correspondences. Certain translations, such as Theodoridis' *Hippolytus*, emerge as peculiar because of the stylistic choices of the translator, but they are not any closer to the original than the rest in many respects. In the case of Persian, the mediation of a different tradition, the French one, affects the structure of the text and its relation to the Ancient Greek original.

Traduttore traditore (translator = traitor): translations may appear equivalent on the surface, but they are really different in the way they render the complexities of an ancient

text. As their semantic overlap is minimal, they reflect the individuality of the translators 578
and their specific circumstances, rather than ‘just’ the individual character of the author. 579
However, translations are necessary, and can even be works of art in their own respect. 580
As computational methods become more and more accessible, tools like UGARIT can 581
support an in-depth approach to the relationship between translations and original 582
that was not possible before. By empowering a user-centered and analytical approach 583
to word correspondence, tools like UGARIT can help experts and non-experts engage 584
more deeply with linguistic and semantic differences, encouraging better exchange 585
between translations and originals (Palladino 2020). Not everybody can easily read 586
the *Iliad* or Euripides in Ancient Greek: translation alignment, however, may facilitate 587
a cross-linguistic approach to a text, as it places at its center not the translation or the 588
original as autonomous entities, but the relationship between them, at the linguistic, 589
grammatical, and semantic level. 590

Conference

6. Data availability 591

<https://github.com/UgaritAlignment/JCLS22-Paper> 592

7. Software availability 593

The alignments have been created using UGARIT translation alignment editor 594

<http://ugarit.ialigner.com> 595

8. Acknowledgements 596

We would like to thank Bethany Morgan for providing the first version of the alignments 597

of *Hippolytus*, for selecting the translations, and for her valuable feedback on this article. 598

9. Author contributions 599

Chiara Palladino: Conceptualization, Formal Analysis, Investigation, Writing-original 600
draft, Validation 601

Farnoosh Shamsian: Writing-original draft, Formal analysis, Investigation, Resources 602

Tariq Yousef: Methodology, Software, Data curation, Visualization, Review & Editing 603

References 604

Almas, Bridget and Marie-Claire Beaulieu (July 2013). "Developing a New Integrated 605
Editing Platform for Source Documents in Classics". In: *Literary and Linguistic Com- 606*
puting 28.4, pp. 493–503. ISSN: 0268-1145. DOI: [10.1093/llc/fqt046](https://doi.org/10.1093/llc/fqt046). eprint: [https 607](https://academic.oup.com/dsh/article-pdf/28/4/493/5646635/fqt046.pdf)
[://academic.oup.com/dsh/article-pdf/28/4/493/5646635/fqt046.pdf](https://academic.oup.com/dsh/article-pdf/28/4/493/5646635/fqt046.pdf). URL: 608
<https://doi.org/10.1093/llc/fqt046>. 609

Baker, Mona, G. Francis, and E. Tognini-Bonelli (1993). "'Corpus Linguistics and Trans- 610
lation Studies: Implications and Applications'". In: DOI: [10.1075/z.64.15bak](https://doi.org/10.1075/z.64.15bak). 611

Barreiro, Anabela, Francisco Raposo, and Tiago Luís (2016). "CLUE-Aligner: An align- 612
ment tool to annotate pairs of paraphrastic and translation units". In: *10th Language 613*
Resources and Evaluation Conference (LREC), pp. 7–13. 614

Berti, Monica, ed. (Aug. 2019). *Digital Classical Philology: Ancient Greek and Latin in the 615*
Digital Revolution. en. Publication Title: Digital Classical Philology. De Gruyter Saur. 616
ISBN: 978-3-11-059957-2. URL: [https://www.degruyter.com/view/title/537705 617](https://www.degruyter.com/view/title/537705)
(visited on 06/16/2020). 618

Bettini, Maurizio (2012). *Vertere: un'antropologia della traduzione nella cultura antica*. Ein- 619
audi. 620

- Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation". In: *Computational Linguistics* 19.2, pp. 263–311. URL: <https://aclanthology.org/J93-2003>.
- Caseli, H. M., V. D. Feltrim, and M. G. V. Nunes (2002). *TagAlign: Uma ferramenta de pré-processamento de textos (NILC-TR-02-09)*. Tech. rep. 169. Instituto de Ciências Matemáticas e de Computação (ICMC-USP). URL: <http://www2.dc.ufscar.br/%0helenacaseli/pdf/2002/NILC-TR-02-09.pdf>.
- Celano, Giuseppe G. A., Gregory Crane, and Saeed Majidi (2016). "Part of Speech Tagging for Ancient Greek". In: *Open Linguistics* 2.1. doi: [doi:10.1515/opli-2016-0020](https://doi.org/10.1515/opli-2016-0020). URL: <https://doi.org/10.1515/opli-2016-0020>.
- Crane, Gregory (2019). "Beyond Translation: Language Hacking and Philology". en. In: *Harvard Data Science Review* 1.2. ISSN: , DOI: [10.1162/99608f92.282ad764](https://doi.org/10.1162/99608f92.282ad764). URL: <http://hdsr.mitpress.mit.edu/pub/owxwohyz/release/4> (visited on 06/16/2020).
- Dagan, I., K. Church, and W. Gale (1999). "Robust Bilingual Word Alignment for Machine Aided Translation". en. In: *Natural Language Processing Using Very Large Corpora*. Ed. by Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky. Text, Speech and Language Technology. Dordrecht: Springer Netherlands, pp. 209–224. ISBN: 978-94-017-2390-9. DOI: [10.1007/978-94-017-2390-9_13](https://doi.org/10.1007/978-94-017-2390-9_13). URL: https://doi.org/10.1007/978-94-017-2390-9_13 (visited on 05/08/2019).
- Dou, Zi-Yi and Graham Neubig (Apr. 2021). "Word Alignment by Fine-tuning Embeddings on Parallel Corpora". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 2112–2128. URL: <https://aclanthology.org/2021.eacl-main.181>.
- Eve, Martin Paul (June 2019). *Close Reading with Computers: Textual Scholarship, Computational Formalism, and David Mitchell's Cloud Atlas*. en. Google-Books-ID: JJwtuWEA-CAAJ. Stanford University Press. ISBN: 978-1-5036-0937-2.
- Foradi, Maryam (2019). "Confronting Complexity of Babel in a Global and Digital Age. What can you produce and what can you learn when aligning a translation to a language that you have not studied?" In: *DH2019: Digital Humanities Conference, Utrecht University, July 9-12. Book of Abstracts*. Utrecht.
- Germann, Ulrich (2008). "Yawat: yet another word alignment tool". In: *Proceedings of the ACL-08: HLT demo session*, pp. 20–23.
- Gibert, John (2022). "Review of: Euripides, Children of Heracles, Hippolytus, Andromache, Hecuba". In: *Bryn Mawr Classical Review* (). BMCR ID: 1996.12.02. ISSN: 1055-7660. URL: <https://bmcr.brynmawr.edu/1996/1996.12.02/> (visited on 04/25/2022).
- Gilmanov, Timur, Olga Scrivner, and Sandra Kübler (2014). "SWIFT Aligner, A Multifunctional Tool for Parallel Corpora: Visualization, Word Alignment, and (Morpho)-Syntactic Cross-Language Transfer." In: *LREC*, pp. 2913–2919.
- Graça, João, Joana Paulo Pardal, Luísa Coheur, and Diamantino Caseiro (2008). "Building a Golden Collection of Parallel Multi-Language Word Alignment". In: *LREC*.

- Grimes, Stephen, Xuansong Li, Ann Bies, S. Kulick, Xiaoyi Ma, and Stephanie Strassel (2010). "Creating Arabic-English Parallel Word-Aligned Treebank Corpora at LDC". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. pp. 665–666. doi: [10.3115/2010-1001](https://doi.org/10.3115/2010-1001). URL: <https://www.aclweb.org/anthology/D10-1001>. 665–666
- Homer (1867). *Homere: Iliade ; traduction nouvelle par Leconte de Lisle*. Alphonse Lemerre. 668
- (1962). *Iliade: Traduction de Paul Mazon*. les Belles lettres (Rennes, impr. Oberthur). 669
- (1965). *Liliade: traduction, introduction et notes par Eugene Lasserre*. Paris: Garnier-Flammarion. 670–671
- Humboldt, Wilhelm von (1816). "Einleitung". In: *Aeschylus Agamemnon metrisch Übersetzt*. Leipzig: Fleischer. 672–673
- Jalili Sabet, Masoud, Philipp Dufter, François Yvon, and Hinrich Schütze (Nov. 2020). "SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 1627–1643. doi: [10.18653/v1/2020.findings-emnlp.147](https://doi.org/10.18653/v1/2020.findings-emnlp.147). URL: <https://aclanthology.org/2020.findings-emnlp.147>. 674–679
- Kay, Martin and Martin Röscheisen (Mar. 1993). "Text-translation Alignment". In: *Comput. Linguist.* 19.1, pp. 121–142. ISSN: 0891-2017. URL: <http://dl.acm.org/citation.cfm?id=972450.972457> (visited on 05/08/2019). 680–682
- Kazzazi, Mir jaleddin (1998). *Iliad*. Markaz. 683
- Laviosa, Sara (May 2008). "Corpus-based translation studies: Where does it come from? Where is it going?" en. In: *Language Matters*. Publisher: University of South Africa. doi: [10.1080/10228190408566201](https://doi.org/10.1080/10228190408566201). URL: <https://www.tandfonline.com/doi/abs/10.1080/10228190408566201> (visited on 08/04/2021). 684–687
- Lefevere, André (1992). *Translation, Rewriting, and the Manipulation of Literary Fame*. London; New York: Routledge. 688–689
- Melamed, I. Dan (May 1998). "Manual Annotation of Translational Equivalence: The Blinker Project". In: *arXiv:cmp-lg/9805005*. arXiv: cmp-lg/9805005. URL: <http://arxiv.org/abs/cmp-lg/9805005> (visited on 08/04/2021). 690–692
- Nafisi, Saeed (1958). *Iliad*. Elmi Farhangi. 693
- Nergaard, Siri (1993). *La teoria della traduzione nella storia*. Milano: Bompiani. 694
- Och, Franz Josef and Hermann Ney (2003). "A Systematic Comparison of Various Statistical Alignment Models". In: *Computational Linguistics* 29.1, pp. 19–51. doi: [10.1162/089120103321337421](https://doi.org/10.1162/089120103321337421). URL: <https://aclanthology.org/J03-1002>. 695–697
- Palladino, Chiara (2020). "Reading Texts in Digital Environments: Applications of Translation Alignment for Classical Language Learning". In: *The Journal of Interactive Technology and Pedagogy* 18. URL: <https://jitp.commons.gc.cuny.edu/reading-texts-in-digital-environments-applications-of-translation-alignment-for-classical-language-learning/>. 700–702
- Palladino, Chiara, Maryam Foradi, and Tariq Yousef (Oct. 2021). "Translation Alignment for Historical Language Learning: a Case Study". In: *Digital Humanities Quarterly* 15.3. ISSN: 1938-4122. 703–705
- Shukhoskvili, Maia (Dec. 2017). "Methodology of translation alignment of Georgian Text of Plato's 'Theaetetus'". In: *International Journal of Language and Linguistics* 4.4, pp. 63–69. 706–708

- Straka, Milan and Jana Straková (Aug. 2017). “Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe”. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, pp. 88–99. DOI: [10.18653/v1/K17-3009](https://doi.org/10.18653/v1/K17-3009). URL: <https://aclanthology.org/K17-3009> (visited on 12/10/2021).
- Véronis, Jean, ed. (2000). *Parallel Text Processing: Alignment and Use of Translation Corpora*. Text, Speech and Language Technology. Dordrecht-Boston-London: Springer Netherlands. ISBN: 978-0-7923-6546-4. URL: <https://www.springer.com/la/book/9780792365464> (visited on 05/08/2019).
- Yousef, Tariq, Chiara Palladino, Farnoosh Shamsian, Anise d’Orange Ferreira, and Michel Ferreira dos Reis (2022). *An automatic model and Gold Standard for translation alignment of Ancient Greek*.
- Yousef, Tariq, Chiara Palladino, Farnoosh Shamsian, and Maryam Foradi (2022a). “Translation Alignment with Ugarit”. In: *Information* 13.2. ISSN: 2078-2489. DOI: [10.3390/info13020065](https://doi.org/10.3390/info13020065). URL: <https://www.mdpi.com/2078-2489/13/2/65>.
- (2022b). *Translation Alignment with Ugarit*. DOI: [10.3390/info13020065](https://doi.org/10.3390/info13020065).
- Yousef, Tariq, Chiara Palladino, David J Wright, and Monica Berti (Apr. 2022). *Automatic Translation Alignment for Ancient Greek and Latin*. DOI: [10.31219/osf.io/8epsy](https://doi.org/10.31219/osf.io/8epsy). URL: osf.io/8epsy.

Evaluation of measures of distinctiveness: Classification of literary texts on the basis of distinctive words

Keli Du  ¹

Julia Dudar  ²

Christof Schöch  ²

1. Trier Center for Digital Humanities, University of Trier, Trier (Germany).

2. Department for Computational Linguistics and Digital Humanities, University of Trier, Trier (Germany).

Keywords:

Distinctiveness, keyness, evaluation,
literary texts

Licenses:

This article is licensed under: 

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

1. Introduction

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

What the above observation points to is that comparison is a fundamental operation in

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

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

In this paper, we focus mainly on subgenres of the novel as our distinguishing category. This is motivated both by the fact that subgenres are an important classificatory principle in literary studies² and by our anecdotal observation that human readers of popular literature are able to determine the subgenre of a novel (whether they are reading a crime fiction, sentimental, or science-fiction novel) based on only a relatively small section from a given novel. The classification task we use in this contribution is meant to mirror this ability and asks the following question: How reliably can a machine learning classifier, based on words identified using a given measure of distinctiveness, identify the subgenre of a novel when provided only with a short segment of that novel? The subgenre labels used in this task are derived from publisher data, especially with respect to book series dedicated to specific subgenres of the novel. We test the identification of distinctive words with a wide range of measures of distinctiveness (including measures that can be described as frequency-based, distribution-based, and dispersion-based) and using a broad range of literary corpora in seven different languages.

Specifically for the task at hand, we further hypothesize that dispersion-based measures of distinctiveness should have an advantage over other measures. The reason for this, we assume, is twofold: first, features (single word forms, in our case) identified to be distinctive by a dispersion-based measure have a higher chance of appearing in shorter, randomly-selected segments taken from an entire novel than features identified using

1. See: <https://github.com/Zeta-and-Company/pydistinto> (DOI: <https://doi.org/10.5281/zenodo.6517683>).

2. For a concise introduction to genre theory, see Hempfer (2014) and, with a focus on computational approaches to genre, Schöch (2020).

other kinds of measures, in particular frequency-based measures; second, dispersion-based measures have a tendency to identify content-related words as distinctive, in contrast to (some) frequency-based measures, which tend to identify high-frequency function words as distinctive (as observed in Schöch, Schlör, et al. (2018)).

Our paper is structured as follows: First, we summarize related work (a) describing different measures of distinctiveness and (b) specifically comparing several measures of distinctiveness to each other (section 2). We go on to describe the different corpora we have used for our study (section 3) as well as the methods used to perform the evaluation task and to analyze the results (section 4). We then discuss the results we have obtained, first in a single-language setting, then in a multi-language setting (section 5). We close our contribution by summarizing our key findings and describing possible future work.

2. Related Work

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

2.1. Measures of distinctiveness

The measures of distinctiveness implemented in our framework have their origins in the disciplines of Information Retrieval, Computational Linguistics, and Computational Literary Studies.

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

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

When it comes to the amount and the variety of measures of distinctiveness, Computa-

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

Table 1: An overview of measures of distinctiveness

tional Linguistics (CL) is the most productive domain. However, almost all measures widely used in CL were originally not invented for text analysis, but were adapted from statistics. As they are usually used in CL for corpus analysis, many of them are implemented in different corpus analysis tools.

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

The Chi-squared and Log-likelihood ratio tests are somewhat more sophisticated statistical distribution tests with underlying hypothesis test.³ These measures are widely used in CL and implemented in some corpus analysis tools, such as WordSmith Tools (Scott 1997), Wmatrix (Rayson 2009), and AntConc (Anthony 2005). One problem with these measures is that p-values tend to be very low across the board when these tests are used

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

for comparing language corpora. The more important problem, however, is that they are designed to compare statistically independent events and handle corpora as a bag of words. These tests use the total number of words in the corpus and do not consider an uneven distribution of words within a corpus (Lijffijt, Nevalainen, et al. 2014).

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

Unlike previous measures, the Wilcoxon rank sum test, also known as Mann-Whitney U-test, does not make any assumption concerning the statistical distribution of words in a corpus; in particular, it does not require the words to follow a normal distribution, as assumed by other tests such as the t-Test. Corpus frequencies are usually not normally distributed, making the Wilcoxon test better suited (Wilcoxon 1945; Mann and Whitney 1947; see also Oakes (1998)). It is based on a comparison of a sum of rank orders of texts in two text collections. The rank orders of texts are defined according to the frequency of a target word, without considering to which of both corpora this text belongs (see Lijffijt, Nevalainen, et al. (2014)). In our implementation, it sums up the frequencies per segment of documents; for this reason, we consider it to be a dispersion-based rather than a frequency-based measure.

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

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

comparison corpus. In our framework, we implemented two variants of Zeta: Burrows' Zeta (Zeta_orig, Burrows 2007) and logarithmic Zeta (Zeta_log, Schöch, Schlör, et al. 2018) to compare their performance.

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

2.2. Comparative evaluation of measures

The evaluation of measures of distinctiveness is a non-trivial task for the simple reason that it is not feasible to ask human annotators to provide a gold-standard annotation. Unlike a given characteristic of tokens or phrases in many annotation tasks, a given word type is distinctive for a given corpus neither in itself, nor by virtue of a limited amount of context around it. Rather, it becomes distinctive for a given corpus based on a consideration of the entire target corpus when contrasted to an entire comparison corpus. Furthermore, whether or not a word can be considered to be distinctive depends on the category that serves to distinguish the target from the comparison corpus. Commonly-used categories include genre or subgenre, authorship or author gender as well as period or geographical origin. For any meaningfully large target and comparison corpus, this is a task that is cognitively unfeasible for humans.

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

We provide some more comments on previous work in this area. Kilgariff (2001) gives a detailed overview of statistical characteristics of some distinctiveness measures, such as log-likelihood ratio test, Wilcoxon rank sum test, t-test, TF-IDF. He suggests the chi-squared test as more suitable measure for comparative analysis, but does not provide significant empirical evidence for his claims. Paquot and Bestgen (2009) compare three measures: log-likelihood ratio test, the t-test and the Wilcoxon rank sum test. They apply these measures to find words that are distinctive of academic prose compared

to fictional prose. The authors stress that the choice of a statistical measure depends on the research purpose. In the case of their analysis, the t-test showed better results because the distribution of the words across texts in the corpus was taken into account. One of the most comprehensive evaluation studies of distinctiveness measures is provided by Lijffijt, Nevalainen, et al. (2014). The authors evaluate a wide range of measures, such as log-likelihood ratio test, chi-squared test, Wilcoxon sum rank test, t-test and others. Their evaluation strategy principally relies on a test of the uniformity of p-values designed to identify measures that are overly sensitive to slight differences in word frequencies or distributions (for details, see their paper).

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

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

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

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

3. Corpora

For our analysis we used nine text collections. The first two corpora consist of contemporary popular novels in French published between 1980 and 1999 (160 novels published in the 1980s and 160 novels published in the 1990s). To enable comparison and classification

of texts, we designed these custom-built corpora in a way that they contain the same number of novels for each of four subgroups: highbrow novels on the one hand, and lowbrow novels of three subgenres (sentimental novels, crime fiction and science fiction) on the other. The texts in these corpora are, for obvious reasons, still protected by copyright. As a consequence, we cannot make these corpora freely available as full texts. We are currently preparing, however, their publication in the form of a so-called derived text format (see Schöch, Döhl, et al. (2020); Organisciak and Downie (2021)) suitable for use with our Python library and devoid of any copyright protection.

Another group of text corpora that we used for our analysis consists of 7 collections of novels in 7 different European languages taken from the European Literary Text Collection (ELTeC) produced in the COST Action Distant Reading for European Literary History (see Burnard, Schöch, and Odebrecht (2021); Schöch, Patras, et al. (2021)).⁵ We reuse the English, French, Czech, German, Hungarian, Portuguese and Romanian corpora. From each of these corpora, we selected a subset of 40 novels: 20 novels from the period from 1840 to 1860 and 20 novels from the period from 1900 to 1920.

corpus	corpus size (million words)	no. of types	document length (mean)	document length (standard deviation)	no. of authors
fra_80s	8.83	119,775	55,225	27,161	120
fra_90s	8.48	111,501	53,010	26,976	124
ELTeC_cze	1.98	163,900	49,642	24,734	33
ELTeC_deu	4.62	158,726	115,531	101,915	30
ELTeC_eng	4.66	53,285	116,477	75,672	35
ELTeC_fra	3.31	65,799	82,802	86,926	37
ELTeC_hun	2.44	258,026	61,055	40,513	36
ELTeC_por	2.33	95,572	58,325	38,787	34
ELTeC_rom	2.41	156,103	60,395	36,493	37

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

4. Methods

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

In order to identify distinctive words, we first define a target corpus and a comparison

5. Texts and metadata for these collections are available on Github: <https://github.com/COST-ELTeC>; DOI: 10.5281/zenodo.3462435. On the COST Action more generally, see also: <https://www.distant-reading.net/>.

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

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

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

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

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

5. Results

5.1. Classification of French popular novel collections (1980s and 1990s)

This section describes the classification of French novel segments into four predefined classes: highbrow, sentimental, crime and scifi. Before running the tests on corpora of different languages, we want to check the variance of results within one language. Only by excluding one confounding variable (language) from the test, we can conclude

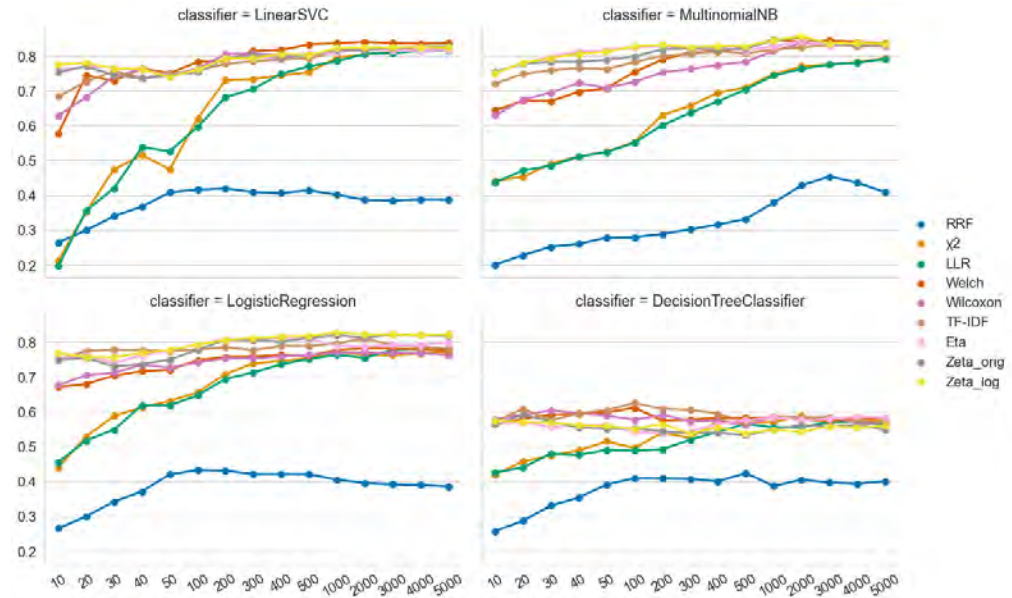


Figure 1: Classification performance of French corpus (1980s) with four classifiers, depending on distinctiveness measure and the setting of N.

that the differences in the performance of measures of ELTeC-corpora are caused by 290
the differences among different languages. That’s why we built two corpora of French 291
novels for our analysis: novels from the 1980s and from the 1990s. 292

First we applied bag-of-words based classification on both parts of the French novel 293
corpus, testing four classifiers: Linear Support Vector Classification, multinomial Naive 294
Bayes, Logistic Regression and Decision Tree Classifier.⁶ 295

Figure 1 shows the classification results of the 1980s-corpus. The Decision Tree Classifier 296
has a clearly lower performance than the other three classifiers. The other three 297
classifiers produce better results with similar trends of F1-scores across different measures. 298
Therefore, in our further experiments we focus on results based on one classifier, namely 299
the Multinomial NB⁷. The classification results of the 1990s-corpus, for this preliminary 300
test, are very similar to the results presented in figure 1 and thus are not shown here. 301

Figure 2 shows the F1-macro score distribution from 10 fold cross-validation for classifi- 302
cation of the French novel segments of the 1980s-dataset. The setting of N varies from 303
10 to 5000. The baseline is visualized as a green line in the plot. It corresponds to the 304
average of the classification results based on N * 8 random words, resampled 1000 times. 305

The classification based on the N most distinctive features leads almost always to better 306
classification results, compared to the baseline. The smaller the number of features, the 307
bigger is the difference between the baseline and performed F1-scores. The baseline 308
approaches the performance of the classifier that uses distinctive words as the number 309

6. LinearSVC, MulinomialNB, LogisticRegression and DecisionTreeClassifier from the Python package
scikit-learn; see: <https://scikit-learn.org/>.

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

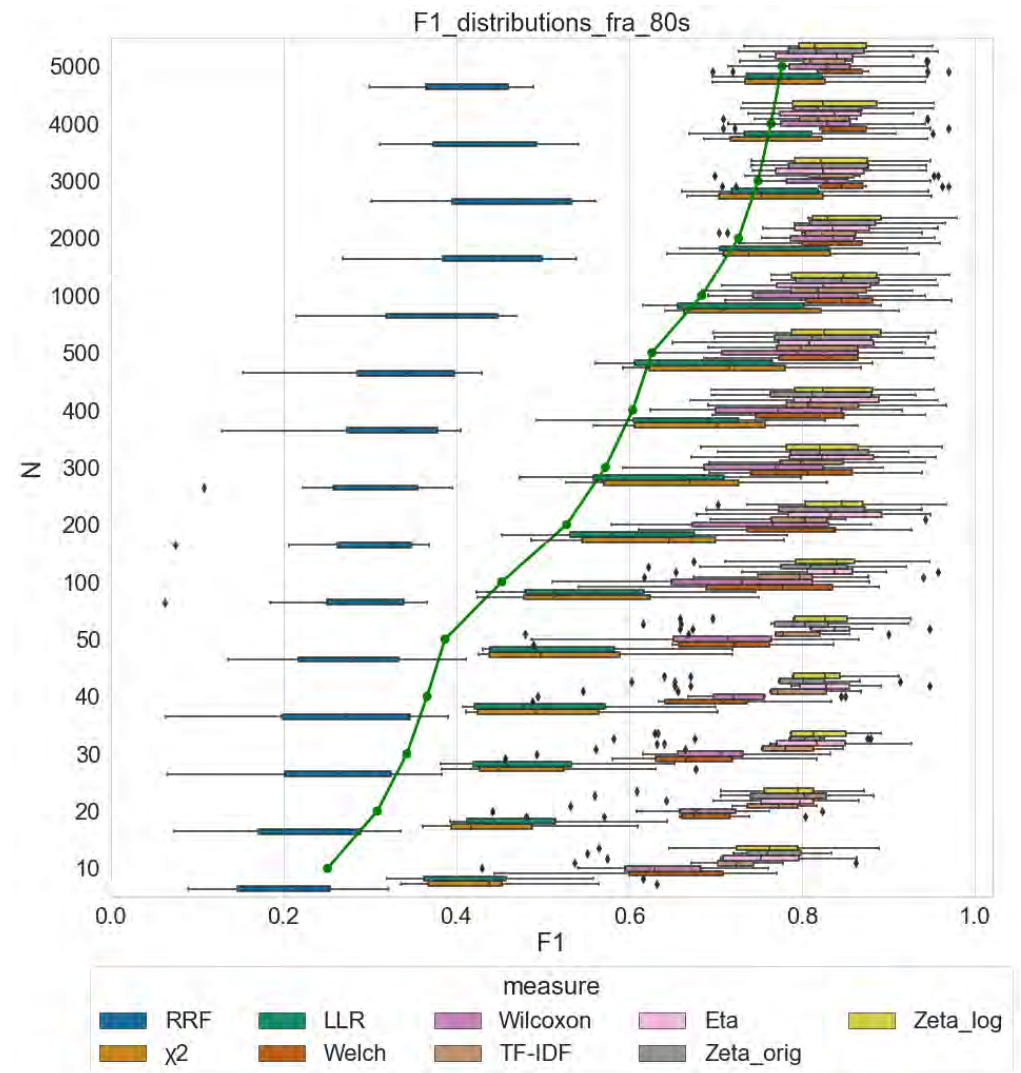


Figure 2: F1-macro score distribution from 10 fold cross-validation obtained by genre classification of French corpus 1980s with Multinomial NB. The green line is a baseline F1-score.

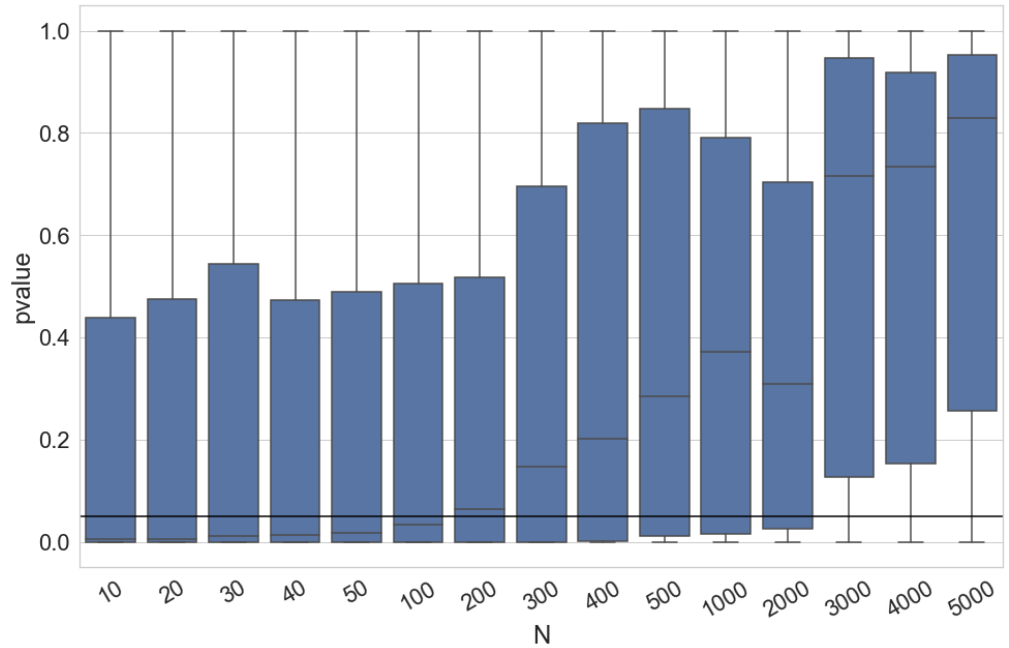


Figure 3: T-test on every pair of the F1-scores distributions of measures. F1 score obtained from classification of the 1980s-corpus. The black line is the significance threshold.

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

Another observation based on figure 2 is that the differences in the variations of F1-score distributions decrease with the increase of N . The measures also show different degrees of variation of results depending on the corpora.⁸ In order to identify which distinctiveness measures produce features that lead to results that are significantly better and more robust, we applied a two-tailed t-test on every pair of the F1-score distributions. The results for the 1980s text collection are shown in figure 3.

In figure 3, each boxplot represents the distribution of 36 p-values (all pairwise combinations of 9 measures) at the setting of the corresponding N . We can observe that with increasing N , the number of p-values smaller than 0.05 (significance threshold) decreases.⁹ This means that the more features are used, the less statistically significant differences exist between classification results. This observation proves our previous

8. Classification of the 1980s-collection leads to lower variations of the F1-scores compared to the classification of the 1990s-collection.

9. When $N = 10$ -100, more than 50% of the p-values are below the threshold of 0.05 and when $N = 300$ or higher, most of the p-values are above the threshold of 0.05.

conclusion, that a high number of features automatically leads to high accuracy and (certainly, according to the p-values, from $N = 3000$) it is not important, in such a scenario, which measure is used.

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

First of all, we can observe in figure 4(a) that for the classification with $N = 10$ features, the F1-scores of RRF, χ^2 and LLR are very low, Wilcoxon and Welch have average performance, while both Zeta variants, Eta and TF-IDF have the highest scores.¹¹

We can also observe, in figure 4(b), that RRF is an outlier and has significantly different F1-scores compared to all other measure. χ^2 and LLR have almost perfect correlation with each other and significantly differ from all other measures as well as from RRF. We can make the same observation concerning the Wilcoxon and Welch measures: they have strong a correlation with each other and significantly different results to other measures with exception of TF-IDF. As for the other measures, we observe a high correlation in F1-scores between TF-IDF, Eta and both Zetas. Combining this information with F1-score distributions at $N = 10$ (figure 4a) lets us affirm that all frequency-based measures (RRF, LLR and χ^2) perform significantly worse compared to the other measures, when we set $N = 10$ for our classification task. Concerning both Zetas, Eta and TF-IDF we can conclude that they have significantly better results compared to other measures. Wilcoxon and Welch have average performance and similar scores, a fact that explains their relatively high correlation.¹²

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

Figure 5 shows that almost all p-values obtained from F1-scores of both Zeta variants, Eta and TF-IDF are greater than the significance threshold of 0.05. The Wilcoxon and Welch have around 25% of p-values below 0.05. This means that the classification results

10. This observation on the 1980s-dataset can also be seen in the results from tests on the 1990s-dataset.

11. RRF median = 0.22, χ^2 = 0.44, LLR = 0.45, Wilcoxon = 0.63, Welch = 0.65, TF-IDF = 0.73, Eta = 0.76, Zeta_orig = 0.77, Zeta_log = 0.77.

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

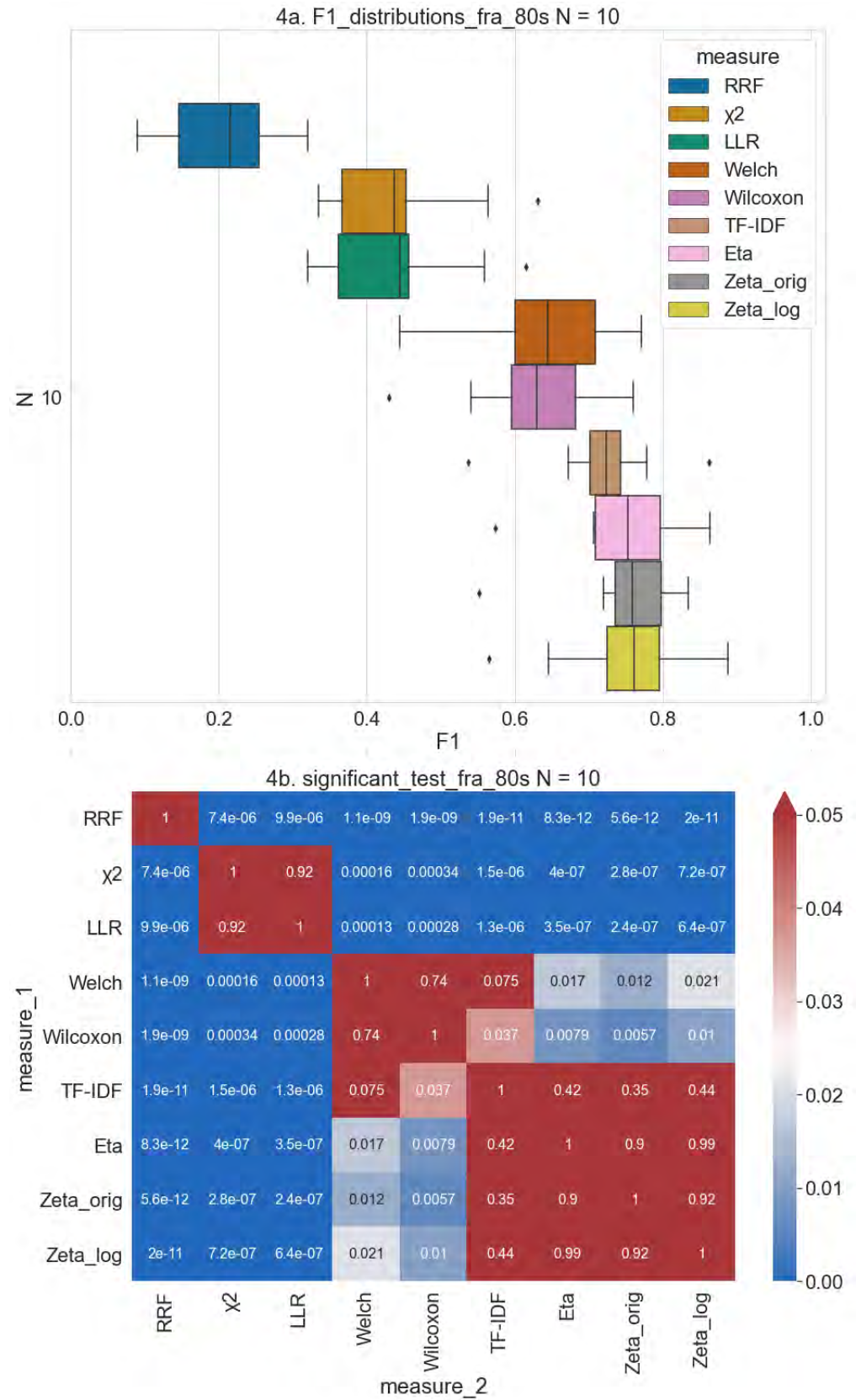


Figure 4: (a) F1-scores distributions for classification with N = 10. (b) p-values obtained from t-test on pairs of the F1-scores distributions of measures. F1 score obtained from classification of the 1980s-corpus with N = 10. Significance threshold is 0.05. Note that all values above 0.05 are shown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

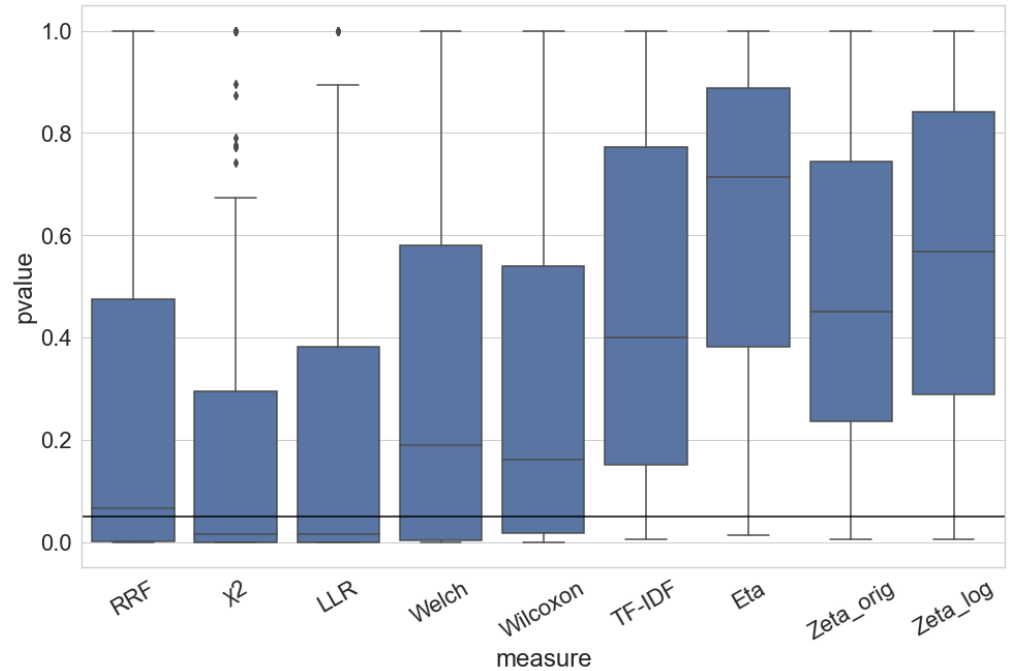


Figure 5: Significance test on F1-scores distributions for each measure. F1-scores obtained from classification of the 1980s-corpus. Black line is significance threshold

based on features extracted by these measures are stable and robust, independently of N. 366
Concerning LLR and χ^2 , there are over 50% of p-values below the significance threshold, 367
RRF has around 50% of p-values below 0.05.¹³ 368

Summarizing the information from the classification of both corpora, we can argue that 369
Zeta_log, Zeta_orig, Eta and TF-IDF have the highest and the most robust performance 370
when using the smallest number of features ($N = 10$).¹⁴ These results mean that 10 371
words identified as distinctive by these measures are sufficient to correctly distinguish 372
over 70% of texts into four groups. 373

It is important to note that this group of the most successful measures have something in 374
common: they are all dispersion-based.¹⁵ (TF-IDF with some restrictions). It appears 375
fair to conclude that in our case, dispersion-based measures can best identify the words 376
that are the most distinctive for a certain genre. The frequency-based measures show 377
a significantly lower and less stable performance. Wilcoxon and Welch show average 378
results.¹⁶ 379

13. The results of the classification of the 1990s-dataset show the same tendency.

14. Zeta_log has the highest mean F1-score (1980s: 0.75, 1990s: 0.72), followed closely by Eta (1980s: 0.75, 1990s: 0.72), and then by Zeta_orig (1980s: 0.75, 1990s: 0.70), TF-IDF (1980s: 0.72, 1990s: 0.71).

15. Dispersion describes the even/uneven spread of words across a corpus or across each particular text in a corpus. We cannot claim, however, that the measures we have used rely exclusively on dispersion; rather, they are also influenced by frequency; see Stefan Th. Gries (2021b).

16. For information about the types of measures, see table 1.

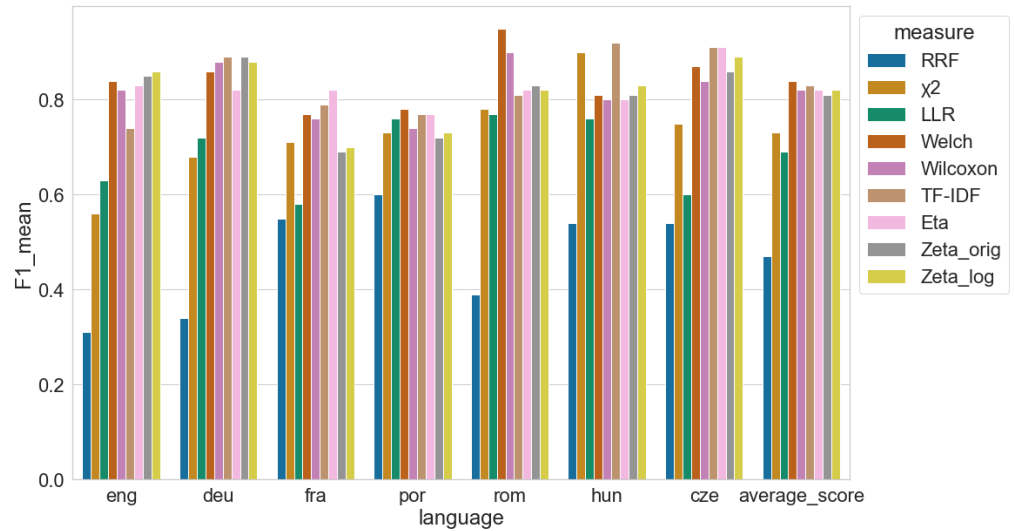


Figure 6: Mean F1-score of classification across 7 ELTeC corpora. (N = 10)

5.2. Experiments on seven ELTeC text collections

380

The above-mentioned conclusion regarding the superior performance of dispersion-based 381
measures when compared to frequency-based measures is based on the specific use-case 382
of our 20th-century French novel corpus. In order to verify whether this claim is also 383
true when corpora in other languages are used, we performed the same tests on several 384
subsets derived from ELTeC (as described above, section 3), namely from the English, 385
French, Czech, German, Hungarian, Portuguese and Romanian collections. 386

The classification task that we use differs from the previous one. We are interested not 387
in classifying the texts by subgenre, but by their period of first publication (1840-1860 388
vs. 1900-1920). The main reason for this is practical: the corpora included in ELTeC do 389
not have consistent metadata regarding the subgenre of the novels included, due to the 390
large variability of definitions and practices in the various literary traditions that are 391
covered by ELTeC. However, all collections cover a very similar temporal scope so that 392
it is possible to use this as a shared criterion to define two groups for comparison. 393

We consider the performance across corpora and measures for N = 10, based on the 394
mean F1-score of the classification task (figure 6). We can observe that among the tests 395
based on seven corpora, five of them could achieve a result of 0.8 or higher. In particular, 396
the dispersion-based and the distribution-based measures can guarantee good or even 397
best results in almost every classification task. The only exception is the classification 398
of the Portuguese corpus. The classification results based on other measures are very 399
similar, except for RRF. Both Zeta variants and Eta are among the best classification 400
results for the English, German, Hungarian and Czech corpora, while Welch and TF-IDF 401
yielded particularly good results when classifying the Romanian corpus. 402

With regard to the frequency-based measures, we can observe that χ^2 has very good 403
results for the Hungarian corpus, but not for the English or German corpora. LLR has 404

relatively high scores for the Portuguese and Hungarian corpora. But in most cases, it is still not as good as dispersion-based measures such as Zeta_log. Compare to all other measures, the 10 most distinctive words defined by the RRF lead to worst results in all classification tasks.

Based on additional data (available in our Github repository: <https://github.com/Zeta-and-Company/JCLS2022/tree/main/Figures>), we consider the difference between F1-score distributions for each measure with varying N. In a similar way to the results from the French novel sets, the differences decrease with increasing N. However, unlike the results from the French novel sets in figure 3, some corpora have more than 75% of the p-values greater than 0.05 when N is greater than 100 (e.g. Czech and German corpora). Some do not have the same results until N is greater than 500 (e.g. English corpus). This indicates that although the results show some variations between the different corpora the overall trend is the same. The larger the value of N, the less important it is which measure is used to select the features (distinctive words) for classification.

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

Summarizing the results described above, we can conclude that dispersion-based and distribution-based measures have been shown again to yield higher performance in identifying distinctive words and to be more stable and robust than other measures. In contrast, the average performance of frequency-based measures (see figure 6) is still considerably lower than the other measures.

6. Conclusion and Future Work

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

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

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

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

The second issue concerns the number and range of measures of distinctiveness implemented in our Python package so far. With 9 different measures, we already provide a substantial number of measures. However, we plan to add several more measures to this list, notably Kullback-Leibler Divergence (a distribution-based measure, see: Kullback and Leibler (1951)), the measure combining dispersion and log-likelihood ratio used by Egbert and Biber (2019), the inter-arrival time measure proposed by Lijffijt, Papapetrou, et al. (2011) as well as a measure yet to be defined that would be based on the pure dispersion measure DP_{nofreq} recently proposed by Stefan Th. Gries (2021b).

Thirdly, it should be considered that almost all previous studies in the area of distinctiveness, our own included, do not allow any conclusions as to whether the words defined by a given measure as statistically distinctive are also perceived by humans as distinctive. Such an empirical evaluation is out of scope for our paper, but would certainly add a different kind of legitimacy to a measure of distinctiveness (for a theoretical take on this, see Schröter et al. (2021)).

Finally, we would of course like to expand our research regarding the elephant in the room, so to speak: not just evaluating statistically which measures perform more or less well in particular settings, but also explaining why they behave in this way. We believe that the distinction between measures based on frequency, distribution and dispersion is

a good starting point for such an investigation, but pushing this further also requires 486
to include measures that really measure only dispersion and not a mix of dispersion 487
and frequency, as recently demonstrated by Stefan Th. Gries (2021b). Measures of 488
distinctiveness have clearly not yielded all their secrets to us yet. 489

Conference

7. Data availability 490

Data can be found here: <https://github.com/Zeta-and-Company/JCLS2022> (DOI: <https://doi.org/10.5281/zenodo.6517748>). 491
492

8. Software availability 493

Software can be found here: <https://github.com/Zeta-and-Company/JCLS2022> 494
(DOI: <https://doi.org/10.5281/zenodo.6517748>). 495

9. Acknowledgements 496

We have conducted our research in the framework of a three-year project called Zeta and 497
Company at Trier University, Germany, funded by the German Research Foundation 498
(DFG) under project identifier 424211690 in the framework of the priority programme 499
Computational Literary Studies (SPP 2207). We thank our anonymous reviewers for 500
their constructive feedback on an earlier version of this article. 501

10. Author contributions 502

Keli Du: Methodology, Investigation, Visualization, Software, Writing - review & editing 503

Julia Dudar: Formal Analysis, Data curation, Writing - original draft, Writing - review 504
& editing 505

Christof Schöch: Conceptualization, Data curation, Funding acquisition and Supervision, 506
Writing - original draft, Writing - review & editing 507

References 508

- Anthony, Laurence (2005). “AntConc: Design and development of a freeware corpus 509
analysis toolkit for the technical writing classroom”. In: Proceedings, International 510
Professional Communication Conference, 2005 (IPCC 2005). Pp. 729–737. DOI: 511
[10.1109/IPCC.2005.1494244](https://doi.org/10.1109/IPCC.2005.1494244). 512
- Burnard, Lou, Christof Schöch, and Carolin Odebrecht (2021). “In search of comity: 513
TEI for distant reading”. fr. In: Journal of the Text Encoding Initiative 14. DOI: 514
[10.4000/jtei.3500](https://doi.org/10.4000/jtei.3500). (Visited on 12/10/2021). 515
- Burrows, John (2002). “Delta: a Measure of Stylistic Difference and a Guide to Likely 516
Authorship”. en. In: Literary and Linguistic Computing 17.3, pp. 267–287. DOI: 517
[10.1093/lc/17.3.267](https://doi.org/10.1093/lc/17.3.267). URL: <http://llc.oxfordjournals.org/content/17/3> 518
[/267.abstract](http://llc.oxfordjournals.org/content/17/3/267.abstract) (visited on 07/26/2011). 519
- (2007). “All the Way Through: Testing for Authorship in Different Frequency Strata”. 520
en. In: Literary and Linguistic Computing 22.1, pp. 27–47. DOI: [10.1093/lc/fqi](https://doi.org/10.1093/lc/fqi) 521
[067](https://doi.org/10.1093/lc/fqi067). (Visited on 12/14/2011). 522

- Craig, Hugh and Arthur F. Kinney, eds. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. en. Cambridge University Press. 523
524
- Damerau, Fred J. (1993). "Generating and evaluating domain-oriented multi-word terms from texts". In: *Information Processing & Management* 29.4, pp. 433–447. 525
526
- Diez, David, Mine Cetinkaya-Rundel, and Christopher D. Barr (2019). *OpenIntro Statistics*. 4th ed. OpenIntro. URL: <https://www.openintro.org/book/os/>. 527
528
- Du, Keli, Julia Dudar, Cora Rok, and Christof Schöch (2021). "Zeta & Eta: An Exploration and Evaluation of Two Dispersion-based Measures of Distinctiveness". In: *Computational Humanities Research 2021 (CEUR Workshop Proceedings)*. Ed. by Maud Ehrmann, Folger Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski, and Joris van Zundert. CEUR. URL: <http://ceur-ws.org/Vol-2989/>. 529
530
531
532
533
534
- Dunning, Ted (1993). "Accurate Methods for the Statistics of Surprise and Coincidence". en. In: *Computational Linguistics* 19.1, pp. 61–74. URL: <http://aclweb.org/anthology/J93-1003>. 535
536
537
- Egbert, Jesse and Doug Biber (2019). "Incorporating text dispersion into keyword analyses". In: *Corpora* 14.1, pp. 77–104. DOI: [10.3366/cor.2019.0162](https://doi.org/10.3366/cor.2019.0162). (Visited on 05/04/2020). 538
539
540
- Gries, Stefan Th (2019). "Analyzing dispersion". In: *Practical handbook of corpus linguistics*. New York, NY: Springer. 541
542
- (2008). "Dispersions and adjusted frequencies in corpora". In: *International Journal of Corpus Linguistics* 13.4, pp. 403–437. DOI: [10.1075/ijcl.13.4.02gri](https://doi.org/10.1075/ijcl.13.4.02gri). 543
544
- (2010). "Useful statistics for corpus linguistics". eng. In: *A mosaic of corpus linguistics: selected approaches*. Ed. by Aquilino Sánchez and Moisés Almela. Frankfurt am Main: Peter Lang, pp. 269–291. 545
546
547
- (2021a). "A new approach to (key) keywords analysis: Using frequency, and now also dispersion". In: *Research in Corpus Linguistics* 9, pp. 1–33. DOI: [10.32714/ricl.09.02.02](https://doi.org/10.32714/ricl.09.02.02). 548
549
550
- (2021b). "What do (most of) our dispersion measures measure (most)? Dispersion?" en. In: *Journal of Second Language Studies*. DOI: [10.1075/jsls.21029.gri](https://doi.org/10.1075/jsls.21029.gri). (Visited on 12/17/2021). 551
552
553
- Hempfer, Klaus W. (Jan. 2014). "Some Aspects of a Theory of Genre". en. In: *Linguistics and Literary Studies / Linguistik und Literaturwissenschaft*. Ed. by Monika Fludernik and Daniel Jacob. Berlin, Boston: De Gruyter. ISBN: 978-3-11-034750-0. DOI: [10.1515/9783110347500.405](https://doi.org/10.1515/9783110347500.405). URL: <https://www.degruyter.com/document/doi/10.1515/9783110347500.405/html> (visited on 05/12/2021). 554
555
556
557
558
- Hoover, David L. (2010). "Teasing out Authorship and Style with t-tests and Zeta". en. In: *Digital Humanities Conference (DH2010)*. URL: <http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-658.html>. 559
560
561
- Kilgariff, Adam (2001). "Comparing Corpora". en. In: *International Journal of Corpus Linguistics* 6.1, pp. 97–133. DOI: [10.1075/ijcl.6.1.05kil](https://doi.org/10.1075/ijcl.6.1.05kil). (Visited on 10/18/2018). 562
563
564
- Klimek, Sonja and Ralph Müller (2015). "Vergleich als Methode? Zur Empirisierung eines philologischen Verfahrens im Zeitalter der Digital Humanities". en. In: *JLT Articles* 565
566

- 9.1. URL: <http://www.jltonline.de/index.php/articles/article/view/758> (visited on 01/05/2021).
- Kullback, Solomon and Richard A. Leibler (1951). "On information and sufficiency". In: The annals of mathematical statistics 22.1, pp. 79–86.
- Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2014). "Significance testing of word frequencies in corpora". en. In: Digital Scholarship in the Humanities 31.2, pp. 374–397. DOI: [10.1093/llc/fqu064](https://doi.org/10.1093/llc/fqu064). (Visited on 05/23/2016).
- Lijffijt, Jefrey, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2011). "Analyzing Word Frequencies in Large Text Corpora Using Inter-arrival Times and Bootstrapping". en. In: Machine Learning and Knowledge Discovery in Databases. Ed. by Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer, pp. 341–357.
- Luhn, Hans Peter (1957). "A statistical approach to mechanized encoding and searching of literary information". In: IBM Journal of research and development 1.4, pp. 309–317.
- Lyne, Anthony A. (1985). "Dispersion". en. In: The Vocabulary of French Business Correspondence: Word Frequencies, Collocations and Problems of Lexicometric Method. Paris: Slatkine, pp. 101–124.
- Mann, H. B. and D. R. Whitney (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". en. In: The Annals of Mathematical Statistics 18.1, pp. 50–60. DOI: [10.1214/aoms/1177730491](https://doi.org/10.1214/aoms/1177730491). (Visited on 12/11/2016).
- Oakes, Michael P. (1998). Statistics for Corpus Linguistics. en. Edinburgh University Press.
- Organisciak, Peter and J. Stephen Downie (2021). "Research access to in-copyright texts in the humanities". In: Information and Knowledge Organisation in Digital Humanities. Routledge, pp. 157–177. DOI: <https://doi.org/10.4324/9781003131816-8>.
- Paquot, Magali and Yves Bestgen (2009). "Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction". In: Corpora: Pragmatics and Discourse. Ed. by Andreas H. Jucker, Daniel Schreier, and Marianne Hundt. Brill and Rodopi. DOI: [10.1163/9789042029101_014](https://doi.org/10.1163/9789042029101_014). (Visited on 09/16/2019).
- Rayson, Paul (2009). "Wmatrix: a web-based corpus processing environment". In: Salton, Gerard and Christopher Buckley (1988). "Term-weighting approaches in automatic text retrieval". en. In: Information Processing & Management 24.5, pp. 513–523. DOI: [10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0). (Visited on 10/13/2018).
- Schöch, Christof (2018). "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie". ger. In: Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven. Ed. by Toni Bernhart, Sandra Richter, Marcus Lepper, Marcus Willand, and Andrea Albrecht. Berlin: de Gruyter, pp. 77–94. DOI: [10.1515/9783110523300-004](https://doi.org/10.1515/9783110523300-004).

- Schöch, Christof (2020). "Computational Genre Analysis". eng. In: Digital Humanities for Literary Studies: Methods, Tools & Practices. Ed. by James O'Sullivan. College Station TX: Texas A&M University Press.
- Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmänn, and Jörg Röpke (2020). "Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen". In: Zeitschrift für digitale Geisteswissenschaften (ZfdG) 5. DOI: http://dx.doi.org/10.17175/2020_006.
- Schöch, Christof, Roxana Patras, Toma Erjavec, and Diana Santos (2021). "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives". en. In: Modern Languages Open 1, pp. 1–19. DOI: [10.3828/mlo.v0i0.364](https://doi.org/10.3828/mlo.v0i0.364). (Visited on 12/17/2021).
- Schöch, Christof, Daniel Schlör, Albin Zehe, Henning Gebhard, Martin Becker, and Andreas Hotho (2018). "Burrows Zeta: Exploring and Evaluating Variants and Parameters". eng. In: Book of Abstracts of the Digital Humanities Conference. Mexico City: ADHO. URL: <https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/>.
- Schröter, Julian, Keli Du, Julia Dudar, Cora Rok, and Christof Schöch (2021). "From Keyness to Distinctiveness: Triangulation and Evaluation in Computational Literary Studies". In: Journal of Literary Theory (JLT).
- Scott, Mike (1997). "PC Analysis of Key Words and Key Key Words". eng. In: System 25.2, pp. 233–245.
- Spärck Jones, Karen (1972). "A statistical interpretation of term specificity and its application in retrieval." In: Journal of Documentation 28, pp. 11–21.
- Tufte, Edward R. (1990). Envisioning information. eng. Cheshire, Conn: Graphics Press.
- Welch, B. L. (1947). "The Generalization of Student's Problem when several different population variances are involved". en. In: Biometrika 34.1-2, pp. 28–35. DOI: [10.1093/biomet/34.1-2.28](https://doi.org/10.1093/biomet/34.1-2.28). (Visited on 10/26/2018).
- Wilcoxon, Frank (1945). "Individual Comparisons by Ranking Methods". In: Biometrics Bulletin 1.6, pp. 80–83. DOI: [10.2307/3001968](https://doi.org/10.2307/3001968). (Visited on 11/01/2018).

Annotation Scheme for Knowledge Transfer

Annotation, Evaluation, and Analysis

Melanie Andresen  ¹

Benjamin Krautter  ²

Janis Pagel  ²

Nils Reiter  ²

1. Institute for Natural Language Processing, University of Stuttgart, Stuttgart.

2. Department of Digital Humanities, University of Cologne, Cologne.

Keywords:

annotation, drama, knowledge,
inter-annotator agreement, network
analysis

Licenses:

This article is licensed under:



Abstract. The distribution of knowledge among characters has been described as an important feature for drama analysis. Many turning points in plays are triggered by a knowledge transfer. However, knowledge transfers in plays have not yet been targeted in a computational way. This paper aims at developing a framework to digitally model processes of knowledge dissemination concerning family and love relations among fictional characters in plays. We approach this as an annotation task and introduce how our composite annotation scheme models knowledge transfers among characters. We present preliminary results and discuss the question of inter-annotator agreement, the calculation of which is not yet standardised for this type of annotation. Finally we showcase an analysis of these dissemination networks on Günderrode's play *Udohla*.

1. Introduction

1

“A play should lead up to and away from a central crisis, and this crisis should consist in a discovery by the leading character which has an indelible effect on his thought and emotion and completely alters his course of action,” (Anderson 1965, p. 116) stated American playwright Maxwell Anderson (1888–1959) in an essay titled *The Essence of Tragedy* (1939). Anderson was, among other things, known for co-authoring the screenplay for the academy award-winning movie *All Quiet on the Western Front* (1930). In his essay, he is in search of a formula for writing a successful play. After producing a number of what he called accidentally successful plays, and some box office failures, Anderson was wondering “whether or not there were general laws of governing dramatic structure which so poor a head for theory as my own might grasp and use,” (Anderson 1965, pp. 114–115) in a bid to reduce “some of the gamble [...] of play-writing.” (Anderson 1965, p. 115) He found his answer in Aristotle's *Poetics*. To be precise, he found it in Aristotle's discussion of recognition scenes, i. e., “a change from ignorance to knowledge,” (Aristotle 1995, p. 65) which Anderson transferred into a poetology of his

own. With regard to Aristotle's remarks, Anderson characterised scenes of recognition as "essential to tragedy." (Anderson 1965, p. 115) He states that a playwright has to "follow the ancient Aristotelian rule: he must build his plot around a scene wherein his hero discovers some mortal frailty or stupidity in himself and faces life armed with a new wisdom." (Anderson 1965, p. 120) In Anderson's view, then, recognition scenes, which lead to a central crisis, play a major role in shaping the course of action and the play's impact on the audience.¹ Although we are studying recognition scenes in plays, they are a common feature not only of tragedy or drama, but of literature as a whole. They are neither limited to high, middle or low brow literature nor to certain genres or literary periods (cf. Cave 1988, pp. 1–9). The revelation of the perpetrator in a crime novel, and how they are found guilty, can be seen as similar to recognition scenes in plays.

An instructive example for recognition scenes, which we will use not only to illustrate the phenomenon, but also to explain our methodological approach, is Karoline von Günderrode's two-act play *Udohla* (1805). The play revolves around effects that, according to Terence Cave, are substantial for discovery: "knowledge and the means of acquiring it, with secrets, disguises, lapses of memory, clues, signs and the like." (Cave 1988, p. 2) Günderrode and her writings were virtually forgotten until Christa Wolf published selected works by her in the late 1970s (cf. Lipinski 2011, p. 113). *Udohla*, which is one of three plays Günderrode authored, is set in a palace and its adjacent garden in Delhi. The play's constellation of characters is characterised by a "familial muddle" (Engelstein 2004, p. 81), i. e., family relations that are at first not transparent – neither for the audience nor for the characters appearing in the play – and later turn out to be different than expected. The play's plot is initially marked by two important moments, both of which concern the reigning Sultan of the Mughal Empire. First, members of the Sultan's staff, namely the vizier Mangu, the Hindu Sino, and the Dervish, argue whether he is going to marry his recently reappeared sister Nerissa. Intrafictionally, a sibling marriage would violate Mongolian Muslim law, but not that of the hierarchically subordinate Hindu population (cf. Günderrode 1990, pp. 204–205).² The sultan himself is seemingly undecisive and questions the motives of God when asking: "Warum o Schicksal, muß ich diese lieben? / Die Einzige die du mir hast versagt." (Günderrode 1990, p. 209) ("Why oh fate, must I love her? / The only one you have denied me").³ Second, the sultan is also told that the death sentence against Bahadar, a Hindu rebel and political traitor, has been carried out, but Bahadar's two children could escape. Both pieces of information have implications for the further course of the plot. Over several steps of knowledge transmissions, it turns out that Nerissa is not the sultan's long-gone sister. Instead, she is the daughter of the previously executed Bahadar. At the same time,

1. Aristotle considers the recognition to be a play's inherent counterpart to aesthetic norms of writing it. I. e., recognition as an inner-dramatic concept mirrors the demanded stringency of a tragic plot from exposition to resolution on a smaller scale (cf. Kablitz 1998, pp. 456–457).

2. As Stefani Engelstein points out with regard to *Udohla*, "[i]t is not unusual to encounter works from the eighteenth or early nineteenth centuries which claim, falsely, that some distant culture sanctions sibling incest". In German Literature incest would oftentimes occur with a "reference to the orient and cultural hierarchies" (Engelstein 2004, p. 280).

3. All translations by the authors.

it becomes clear that Nerissa is the sister of the titular character Udohla. Pretending to be a relative of the Nawab⁴ and his herald, Udohla is trying to outsmart the Sultan in a bid to free his father from captivity, which – as the audience already knows – is certain to fail from the beginning. As we can see, just as in Aristotle’s prime example *Oedipus Rex* the scenes of recognition in Günderröde’s play focus primarily on family relations.

In our article, we will extend this small-scale example on the connection of family relations, the knowledge about them and a central discovery to a bigger corpus of plays. For this purpose, we present a framework for the formal modelling and quantitative analysis of family related knowledge transfers in German plays of the eighteenth and nineteenth century. By means of (manual) annotation we will operationalise⁵ knowledge transfers and thereby intertwine a content-focused approach with already established procedures of quantitative drama analysis concentrating on structural properties of theatre plays.⁶ We use annotation as a method that enriches texts or text segments with certain information, whereby the annotation data takes on different functions (cf. Pagel et al. 2020, pp. 125–141). On the one hand, we employ it to reflect on further developing and refining established quantitative methods of text analysis. In doing so, the annotation data becomes part of the analysis of a play or a corpus of plays and can support the interpretation. On the other hand, the annotations will serve as training or test data for future machine learning procedures.

In a first step of our article, we will set forth our theoretical framework from a literary studies perspective drawing upon Aristotle’s *Poetics*. Following that, we will secondly introduce our annotation scheme in detail. In doing so, we will illustrate how to identify text passages that include a transfer of knowledge concerning family relations and how to label them with our annotation scheme. Regarding these shifts of knowledge, we focus on changes for characters present on stage as well as the audience. Thirdly, we go on to discuss the calculation of inter-annotator agreement for our annotated data. As there is no standardised procedure yet to convincingly measure the agreement within our annotations, this task is not limited to a practical application, but takes theoretical considerations into account as well. Lastly, we will analyse the data we obtained during our annotation process. Hereby, we will focus on two different perspectives. Analysing our corpus of 20 plays, we will examine at what point in drama new knowledge about family and love relations is distributed. We will further distinguish whether the new information addresses the fictional characters in the internal communication system or whether it addresses the audience and discuss the results in light of drama theory. Our second perspective concentrates on a methodological question: Can we use our annotation data to employ new, more content-based ways of literary network analysis? Can this approach help to identify important characters for a play’s action and is it possible to then better integrate quantitative network analysis with qualitative close

4. In the Mughal Empire Nawab originally referred to an envoy of the emperor or a viceroy.

5. For our understanding of operationalisation cf. Pichler and Reiter 2021, pp. 1–29.

6. Research focusing on these structural properties includes analysing character speech formally (cf. Reiter and Willand 2019 or Krautter and Willand 2021, pp. 111–118), examining the distribution of characters within a play or a corpus of plays (cf. Marcus 1973 [1970]; Yarkho 2019 [1935–1938]) and network analysis (cf. Moretti 2011; Trilcke 2013).

readings? We will discuss these question with regard to Günderrode's *Udohla*.

91

2. Theoretical Framework: The Distribution of Knowledge in Drama 92

The interference of internal and external communication systems in drama, i. e. the communication of the fictional characters on the one hand and the perception of this communication by the audience on the other, is considered one of the central "qualities necessary for identifying dramatic communication" (Pfister 1988, p. 49). As Bernhard Asmuth points out in his introduction to drama analysis, a play as a whole is not only a sequence of actions, but also a multi-perspectival processing of knowledge (cf. Asmuth 2016, p. 114).⁷ In the light of events that may have taken place before the actual plot of the drama's main text a play's characters are – potentially – already set apart from each other by a different degree of knowledge. Herein, we employ a broad understanding of knowledge that is not strictly limited to the classical notion of propositional knowledge as "justified true belief" (e. g., Pollock and Cruz 1999, p. 13 or Ichikawa and Steup 2018) which originated from Plato.⁸ As it is not uncommon for literature to deliberately play with knowledge, facts, beliefs, hearsay, and rumors,⁹ we opt for a more "lightweight sense of knowledge" (Ichikawa and Steup 2018). In our case, this includes beliefs that are both justified and depicted to be true, but might later turn out to be false, e. g., through scenes of recognition.

108

A character's level of knowledge can change continuously in the course of the play. At the same time, the relationship between the audience's level of information and that of the individual characters in the play is constantly adjusted. The exposition, for example, reduces the knowledge gap between the audience and the characters that prevails at the beginning of a play (cf. Asmuth 2016, p. 122). The disparities in the "level of [...] awareness" (Pfister 1988, pp. 49–50) can be attributed primarily to two causal differences between the internal and external communication systems: While the audience in its observer role perceives every scene of the play and can thus compare and aggregate partial knowledge of the characters, it sometimes remains unclear what prior knowledge the characters actually have. This also applies to possible time leaps, for instance between two acts of the drama. Furthermore, it might not be clear to what extent the statements of a character correspond to the 'facts' of the fictional world, i. e., whether the statements are credible (cf. Jeßing 2015, pp. 50–51). Depending on the course of the plot then, the audience can either have an information advantage or an information disadvantage over the characters acting on stage at different times. The relative level of being informed between the audience and a character can change from scene to scene. The same applies to the internal communication system of the plays' characters, when comparing the degree of knowledge different characters have in a certain scene. For this phenomenon, Bertrand Evans coined the term "discrepant

127

7. As we have already illustrated by the example of *Udohla* and its portrayal of Hindu culture not condemning sibling incest, intra-fictional knowledge does not have to be valid outside the represented fictional world.

8. Defining knowledge as 'justified true belief' is controversial in itself (cf. Gettier 1963 and Dutant 2015).

9. The plays of Heinrich von Kleist are a prominent example for failed communication between characters creating rumours that are believed to be true (cf. Dubbels 2012).

awareness" (Evans 1960, p. VIII). This 'discrepant awareness' between two characters can thus lead to rather different evaluations of the same action or situation. If we think of Günderrode's *Udohla*, a character's judgement of the supposed marriage between the Sultan and Nerrisa would greatly depend on whether the character knows that the Sultan and Nerrisa are not actually siblings and on the character's religious views, i. e., him being Hindu or Muslim. In this situation, the lack of knowledge or a perceived, but actually incomplete, awareness will influence the judgement in one way or the other.

The gap between the characters' level of knowledge and that of the audience can be seen as an important element of suspense in drama, as it ensures sustained attention and emotional excitement (cf. Anz 2007, p. 464). This applies to both, the suspense felt when one is curious about *what* is going to come up next and the suspense arising in respect to *how* something that is already known to be happening is going to happen.¹⁰ In this respect, the device of dramatic irony is particularly important, as it grounds precisely on this gap of being informed. The audience's knowledge advantage with respect to an upcoming action is, thus, a prerequisite for dramatic irony. In understanding a remark that is innocuous from the perspective of the speaking character, the audience can interpret the utterance as an allusion to the catastrophe that is later actually realised.¹¹ Consequently, elements such as dramatic irony are closely linked to the drama's effect on the audience: Is the play supposed to convey a moral theorem? Is it meant to purify the audience's affects? Should it educate the audience? Or is it simply meant to entertain? In his *Poetics*, Aristotle already defines drama's (cathartic) effect as the central concern of tragedy.¹² He considers reversal (peripeteia) and recognition (anagnorisis) as important building blocks to evoke pity and fear, the desired affects caused by a tragedy.¹³ Recognition is directly related to Evans' concept of 'discrepant awareness', for Aristotle defines recognition as "a change from ignorance to knowledge, leading to friendship or to enmity, and involving matters which bear on prosperity or adversity." (Aristotle 1995, p. 65) Since such scenes of recognition are ideally linked to the reversal, i. e., "a change to the opposite direction of events" (Aristotle 1995, p. 65), they represent central moments of knowledge transmissions that can be decisive for understanding and interpreting a play. To substantiate our case, the examples Aristotle is using to illustrate recognition and reversal "are taken solely from the field of familial *philia*" (Destrée 2020, p. 117).

10. See DiYanni 2000, p. 22: "One of our main sources of pleasure in plot is surprise, whether we are shown something we didn't expect or whether we see *how* something will happen even when we may know *what* will happen. Frequently surprise follows suspense – fulfilling our need to find out what will happen as we wait for a resolution of a play's action."

11. Contrary to what this wording suggests, dramatic irony is not limited to tragedies, but is often found in comedies as well.

12. There are numerous studies that examine Aristotle's mention of catharsis in great detail. Cf. for instance (Schmitt 2008, p. 333–348 and 476–510).

13. There is a great debate about what the two affects mentioned by Aristotle actually express and how to translate them properly (cf. Schadewaldt 1955, pp. 129–171).

3. Annotating Knowledge Transfers

160

The aim of our research is to model knowledge transfers in German plays by means of annotation. While knowledge is a very broad phenomenon, we restrict our annotation to the domain of knowledge about familial character relations. As we employ a broad understanding of knowledge that does not imply that the information is correct, we also include beliefs. In this section, we will present the annotation scheme that we are currently using. We developed the guideline by annotating 16 plays in the course of roughly a year. The full (German) guideline as used by our annotators is available online.¹⁴ The annotation is performed using the tool CorefAnnotator¹⁵ (Reiter 2018).

Our annotation scheme targets text sections in which knowledge transfers take place. More precisely, we annotate a text section if

- a) the knowledge concerning character relations of at least one of the characters or the audience is changed OR
- b) a character's knowledge about the knowledge of another character is changed.

A case of a) would be a text section in which character A learns that B and C are siblings. An example for b) is a section in which B learns that A knows that B and C are siblings. The latter can be understood as knowledge about knowledge or meta-knowledge.

Annotation spans are not fixed to a specific length. Knowledge transfers can happen in one sentence or even a word, but can also be extended over a whole paragraph, especially when knowledge is distributed implicitly. However, our annotators are encouraged to identify a span that is as short as possible. When a relevant text span is identified, it is annotated with a label that uses this pattern:¹⁶

(1) transfer(SOURCE, TARGET, KNOWLEDGE, ATTRIBUTES)

The SOURCE is usually a character that provides a piece of information, but can also be an object or an action that allows for inferences about character relations, for instance when Saladin recognises the handwriting of his brother in Lessing's *Nathan der Weise* (1779). The TARGET is always a character or a group of characters (and/or possibly the audience) whose knowledge is changed. The item KNOWLEDGE is restricted to knowledge about character relations and, more precisely, to the set of relations presented in Table 1. Optionally, ATTRIBUTES can be added that, for instance, mark the information as a lie or as uncertain. The latter is especially frequent as many dramatic texts play with strong allusions to a fact that is ultimately confirmed only at the end.

In *Udohla* by playwright Karoline von Günderrode, the vizier Mangu lets the audience know that Nerissa is the sultan's sister (which turns out to be wrong). This is annotated as follows:

14. <https://doi.org/10.5281/zenodo.5729706>

15. <https://doi.org/10.5281/zenodo.1228105> for stable release versions, <https://github.com/nilsreiter/CorefAnnotator/> for development versions.

16. This notation is inspired by the syntax of the programming language Prolog.

(2) `transfer(mangu, audience, siblings(sultan, nerissa))` 195

Characters are referenced by the identifier they receive in the *Drama Corpora Project* 196
(DraCor, Fischer et al. 2019). Characters that are not speaking in the play do not have 197
such an identifier. Instead, they are given an identifier by our annotators. Frequently, 198
characters are not introduced by name and their identity is (partly) unclear. We annotate 199
such character mentions as a variable in capital letters. In the play *Magie und Schicksal* 200
(1805) ('Magic and Fate') by G nderrode, the character Cassandra mentions a son 201
whom we did not hear about before. At first, we do not have any additional knowledge 202
about this son and therefore annotate him as a variable: 203

(3) `transfer(cassandra, audience, parent_of(cassandra, CHILD[CASSANDRA]))` 204

Later in the play, it is revealed that the character Ligares is in fact the mentioned child 205
of Cassandra. We can now annotate that the variable `CHILD[CASSANDRA]` and Ligares 206
are identical: 207

(4) `transfer(cassandra, audience, identity(CHILD[CASSANDRA], ligares))` 208

Note that it is also possible to fill any of the positions in the annotation label with a list 209
of several characters by enclosing them in square brackets. This is used extensively, for 210
instance, when Nerissa (in *Udohla*) reveals in the final scene that she is the daughter of 211
the sultan's enemy Bahadar (who was just killed by him): 212

(5) `transfer(nerissa, [sultan, udohla, mangu, sino, audience], 213
child_of(nerissa, "Bahadar"))` 214

As mentioned above, we restrict our annotation to the domain of knowledge about 215
character relations, i. e. family and love relations. Table 1 gives an overview of all 216
character relations that are included in the annotation scheme. Formally, we differentiate 217
directed relations like `parent_of(PARENT, CHILD)`, where the position of characters is 218
important because of the asymmetry of the relation, from undirected relations. When 219
annotating undirected, symmetric relations like `siblings(SIBLING-A, SIBLING-B)`, 220
the order of characters is irrelevant. Semantically, the relations form three groups, the 221
biggest of which are family relations and love relations.¹⁷ The last group of identity 222
relations is not about relations between two characters in the strict sense, but includes 223
a) cases where we learn a (first, or additional) name of a character, and b) cases where 224
two characters are revealed to be the same, as in example 4. All relations in the table 225
can be negated by adding a ! at the beginning, e. g., `!siblings(nerissa, sultan)` to 226
express that Nerissa and the Sultan are not siblings. 227

While the annotation guideline covers most of the knowledge transfers happening in the 228
plays in a way that is accessible to our annotators, some challenges remain. This is related 229
to the rather simplistic communication model that underlies the scheme. Although we 230
try to include rules of pragmatic communication in our annotation decision, we formally 231
conceptualise knowledge distribution as transfer: The knowledge of one character is 232

17. The group of love relations is very heterogeneous and subsumes all relationships motivated by love, sexual or material interest, which one might want to differentiate in follow-up studies.

	Directed Relations	Undirected Relations
Family Relations	parent_of(PARENT, CHILD) child_of(CHILD, PARENT) aunt:uncle_of(AUNT:UNCLE, NIECE:NEPHEW) niece:nephew_of(NIECE:NEPHEW, AUNT:UNCLE)	siblings(SIBLING-A, SIBLING-B) cousins(COUSIN-A, COUSIN-B) relatives(RELATIVE-A, RELATIVE-B)
Love Relations	in_love_with(LOVER, TARGET) widow:er_of(WIDOW:ER, DEAD-PARTNER)	lovers(LOVER-A, LOVER-B) couple(PARTNER-A, PARTNER-B) engaged(PARTNER-A, PARTNER-B) spouses(PARTNER-A, PARTNER-B)
Identities	has_name(A, NAME)	identity(A, B)

Table 1: Character relations covered by our annotation scheme. Where applicable, the prefixes grand-, step-, foster-, god- and ex- as well as the suffix -in-law can be added.

transferred to another character. This assumes that the communicated information is understood in the intended way. While this might be true in many cases, there are possible exceptions. There can be misunderstandings, pieces of information can be interpreted in different ways and different prior knowledge or values can influence the understanding. Currently we also assume that the communicating characters are transparent to all characters involved. This is not true for text passages where characters transfer knowledge to a character whose identity is unclear to the speaker. For instance, in Schiller’s *Braut von Messina* (1803), Don Cesar confesses his love to Beatrice without even knowing her name – not to mention that she is also his brother’s lover AND his sister. The annotation shown in (6) captures the view of the audience but does not conform to the perspective of Don Cesar. While the guidelines do provide solutions for such scenes, future versions might increase their generalisation, once more texts with similar constellations have been annotated.

(6) transfer(don_cesar, [audience, beatrice], in_love_with(don_cesar, beatrice))

All plays are annotated by two student annotators independently, then all deviations between the two versions are discussed with one of the authors. Afterwards, each of the annotators produces a revised version. Contrary to many other annotation projects, the aim of this step is not to create one consensus version of the annotation. The annotation task is complex and many text passages can be interpreted in more than one way. In addition, the annotation scheme sometimes allows for different ways of modelling a knowledge transfer. Therefore, the revision focuses on plausibility, consistency and formal correctness of the annotations. However, additional consensus versions were created for analyses that require one single reference version, because the focus is not on annotation variation (as in section 5.2).

Table 2 gives an overview of the current state of our corpus. We mainly selected plays of which we expected knowledge about character relations to be important for the plot based on prior readings and secondary literature. In order to capture as many potentially relevant phenomena as possible, we have chosen not to limit ourselves to a specific genre of plays or a particular literary period. Instead, we opted to annotate a broad mix

No	Author	Text
1	Johann Wolfgang Goethe	<i>Iphigenie auf Tauris</i>
2	Johann Wolfgang Goethe	<i>Die natürliche Tochter</i>
3	Johann Wolfgang Goethe	<i>Stella</i>
4	Franz Grillparzer	<i>Die Ahnfrau</i>
5	Friedrich Hebbel	<i>Maria Magdalene</i>
6	Hugo von Hofmannsthal	<i>Elektra</i>
7	Hugo von Hofmannsthal	<i>Der Rosenkavalier</i>
8	Heinrich von Kleist	<i>Die Familie Schroffenstein</i>
9	Friedrich Maximilian Klinger	<i>Die Zwillinge</i>
10	Jakob Michael Reinhold Lenz	<i>Der Hofmeister</i>
11	Gotthold Ephraim Lessing	<i>Nathan der Weise</i>
12	Gotthold Ephraim Lessing	<i>Emilia Galotti</i>
13	Johann Gottlob Benjamin Pfeil	<i>Lucie Woodvil</i>
14	Friedrich Schiller	<i>Die Braut von Messina</i>
15	Friedrich Schiller	<i>Die Räuber</i>
16	Arthur Schnitzler	<i>Komtesse Mizzi</i>
17	Luise Adelgunde Victorie Gottsched	<i>Das Testament</i>
18	Karoline von Günderrode	<i>Udohla</i>
19	Karoline von Günderrode	<i>Magie und Schicksal</i>
20	Johanna von Weißenthurn	<i>Das Manuscript</i>

Table 2: Name and author of all dramas that are part of the annotated corpus. The first 16 have been annotated in the process of guideline development, the last four have been used for agreement calculation. The latter group will be further expanded in the future.

of plays from the eighteenth and nineteenth century.¹⁸ In the process of developing the annotation guidelines, we annotated 16 dramatic texts. Once the guidelines were consolidated we started tracking the initial versions of our annotators for the calculation of inter-annotator agreement. The annotation of this second round is ongoing and up to this point, four additional dramatic texts were annotated. Based on these four plays, we are developing a suitable way of determining inter-annotator agreement for a complex annotation task as this. The next section will present and discuss our current measure.

4. Calculating Inter-Annotator Agreement

For manual annotation and coding tasks in a wide range of disciplines, measuring inter-annotator agreement (IAA, sometimes also called ‘inter-coder reliability’) is a standard procedure (cf. Artstein and Poesio 2008; Krippendorff 2004), much like the evaluation of automatic predictions based on machine learning. The goal of this measuring is to have a quantitative view on the agreement between annotators, and ultimately to evaluate the quality of the annotation guidelines, the annotation process or the annotations themselves. Unlike an evaluation of automatic predictions, there is no ‘gold standard’, i. e., no annotation set is considered to be true. Instead, IAA ‘only’ measures the agreement. A corner stone of IAA metrics is to take into account expected agreement (often also called ‘chance agreement’), i. e., agreement that is achieved when making random annotation

18. This decision was also influenced by our future plans on automating aspects of the annotation process.

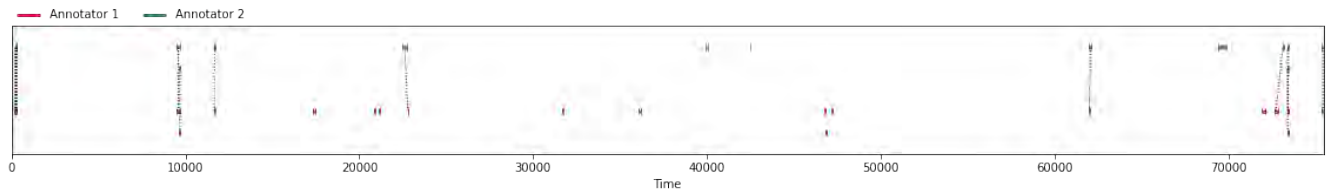


Figure 1: An alignment between the annotations in Günderrodes' *Magie und Schicksal* as established by Gamma

decisions. This is done to compensate for the difficulty of the task itself: If there are more classification categories, the task is considered to be more difficult because there are more decisions to make, and the expected agreement goes down. On structurally simple tasks such as part-of-speech tagging, measuring IAA is well established and understood: Fleiss' Kappa (Fleiss 1971), for instance, can be used to calculate the IAA between n annotators, who assigned one of k categories to each of N items.¹⁹

The annotation task we discuss in this article, however, is more complex: i) Annotation decisions are not made in isolation, but depend on the textual context as much as on previously made decisions. As we are only annotating the transfer of new information to the target, a subsequent mention of the same information by and to the same character is not annotated. As a consequence, each annotation label may only appear once in a text. ii) After having decided that a knowledge transfer takes place (and selecting the exact boundaries), annotators need to make decisions about the source and target of the transfer, the participants of the character relation and its direction, and, finally, potentially about attributes of the annotation (e. g., the transfer being a lie). iii) The annotation is not done on fixed, pre-defined units, but the annotation spans can be defined freely. All three properties make measuring IAA difficult.

The metric Gamma (Mathet, Widlöcher, and Métivier 2015) has been proposed as a versatile, highly adaptable metric for various tasks. It has several properties that make it promising for our use case: i) To calculate expected agreement, it samples a large number of random annotations from the existing annotations. Based on these random annotations, we can compute expected agreement in the same way we calculate observed agreement. This way, expected agreement can be measured empirically instead of theoretically, which makes it less dependent on assumptions and more widely applicable. ii) Equality between annotation categories can be graded: Instead of only recognising that $\text{transfer}(X, Y, \text{parent_of}(P, C))$ is different from $\text{transfer}(X, Z, \text{parent_of}(P, C))$, we can provide a function to express the similarity of the two annotations as a value between zero and one. This allows us to define the similarity of the annotations above to be less than one, but larger than zero. iii) For measuring observed agreement, Gamma first establishes an alignment between the different annotators' annotations. This alignment can also be visualised and inspected, which is a helpful tool in the annotation process. Figure 1 shows an example for the established align-

19. For an overview of annotation metrics that is tailored to readers in computational literary studies, see also Reiter and Konle (2022).

ments. The overall Gamma score is calculated based on pairwise similarity functions 312 between two (or more) annotations that are aligned. Since Gamma is computed over 313 disagreements instead of agreements, we will discuss the calculation of disagreements 314 in the following.²⁰ 315

The final Gamma value is a weighted combination of two aspects of disagreement: 316 Positional disagreement expresses how different the annotations' positions are, and 317 categorical disagreement compares the labels that the annotators have assigned. The 318 exact calculation of positional and categorical disagreement as well as the weighting of 319 these two components can be customised. The two values are not fully independent 320 of each other as the alignment of the annotations already takes the labels into account, 321 i. e., Gamma tries to align annotations with the same label. 322

4.1. Gamma Setup 323

To calculate Gamma, we make use of the "pygamma-agreement"²¹ implementation²¹ with 324 the CBC solver. To adapt Gamma to our purposes, we have defined custom functions 325 for categorical and positional disagreement. 326

For the positional dissimilarity, we consider each annotation that overlaps by at least 327 one character²² as having the same position. Annotations that do not overlap become 328 more dissimilar with increasing distance. Because we are measuring all distances and 329 positions in character offsets and this quickly results in high absolute numbers, the 330 increase in positional disagreement is weighted with 0.001: If the two annotations are 331 10 characters apart, the dissimilarity is only $0.001 \times 10 = 0.01$.²³ 332

For the calculation of categorical disagreement, which is defined for a tuple of anno- 333 tations (u and v , one from each of the two annotators), we look at the six components 334 of the annotated predicate separately: Those are source, target and attribute of the 335 knowledge transfer as well as the literary characters 1 and 2 involved in the relation 336 and the relation name. The disagreement d for each of these components is combined 337 linearly, allowing us to focus on each of them individually by giving them a weight w 338 (Equation 1). 339

20. This distinction is important technically, but conceptually not so much, because we can always convert an agreement score into a disagreement score by subtracting it from 1. The final Gamma score, however, can be interpreted in the same way as other metrics: The higher the score, the better the agreement.

21. <https://github.com/bootphon/pygamma-agreement>

22. In this case, 'character' refers to the graphic symbols of a text, not the literary characters.

23. An important property of Gamma is that no scaling of the dissimilarity values takes place. Both positional and categorical disagreement are expressed on the same scale without any kind of normalisation before combination: If an annotation of Annotator 2 is just next to an annotation of Annotator 1 with the same category, their dissimilarity is just as high as if the annotation had a different category, but the same position. If the annotations are farther apart, their dissimilarity increases proportionally with increasing distance. This in turn makes it important whether the position is counted over characters or tokens: Since we are using character positions to count positional disagreement, an annotation distance of one word might already lead to a large positional dissimilarity – depending on the word length. To cope with this problem we are using custom functions to compute Gamma.

$$\begin{aligned}
d_{\text{cat}}(u, v) = & w_{\text{source}} d_{\text{source}}(u, v) \\
& + w_{\text{target}} d_{\text{target}}(u, v) \\
& + w_{\text{attribute}} d_{\text{attribute}}(u, v) \\
& + w_{\text{character 1}} d_{\text{character 1}}(u, v) \\
& + w_{\text{character 2}} d_{\text{character 2}}(u, v) \\
& + w_{\text{relation name}} d_{\text{relation name}}(u, v)
\end{aligned} \tag{1}$$

The dissimilarity of the individual components is calculated in different ways: The components relation name and attribute are always single values that can be directly compared, returning a value of 0 or 1. For the components containing characters (i. e., source, target, character 1 and character 2), annotators can express lists of characters, and they make use of this frequently (see Example 5). For this reason, we use the Jaccard distance (Jaccard 1912) as a measure of dissimilarity between the two lists (Equation 2). This distance is calculated as the invert of the Jaccard similarity, which measures how many of the elements that appear in at least one of the lists (their union) are present in both list (their intersection), resulting in a value of 1 if the two lists are identical.

$$d_{\text{source}}(u, v) = 1 - \frac{|u_{\text{source}} \cap v_{\text{source}}|}{|u_{\text{source}} \cup v_{\text{source}}|} \tag{2}$$

The Jaccard distance is also employed to measure dissimilarity between character groups for undirected relations. If both annotations specify an undirected relation, we compare the entirety of characters by Annotator 1 with the entirety of characters by Annotator 2.

Once the categorical and positional dissimilarity are calculated, they are weighted against each other in order to receive the final Gamma score. Since we are generally more interested in the categories, we set $\alpha = 1$ and $\beta = 2$, thus categorical disagreement is twice as important as positional disagreement.

4.2. Inter-Annotator-Agreement Results

Table 3 shows Gamma scores for four texts, using different ways of weighting positional and categorical disagreements and of comparing the predicates used in the annotation. For the first column, “Position only”, we set the weight of the categorical agreement to 0, such that the score only depends on the positional agreement and two annotations are considered similar if they occupy the same position, irrespective of their categories. The next six columns evaluate one component at a time, with a weighting of 0.95 of the component of interest, and 0.01 for the other five components.²⁴ The final column, “All”, shows a score for which all components are considered with a uniform weight of

24. The decision to set the weights not to 0 and 1 was made after inspecting some of the alignments that Gamma produced. By specifying a small weight for each component, each component has some influence over the established alignments, and we prevent an alignment that is only based on a single component.

Text	Components of the annotated predicates							
	Position only					Relation		
		Source	Target	Attribute	Char. 1	Char. 2	Name	All
Gottsched: <i>Das Testament</i>	0.403	0.331	0.414	0.400	0.295	0.326	0.243	0.250
Günderrode: <i>Magie und Schicksal</i>	0.525	0.582	0.526	0.521	0.417	0.369	0.507	0.392
Günderrode: <i>Udohla</i>	0.454	0.356	0.246	0.416	0.144	0.199	0.241	0.146
Weissenthurn: <i>Das Manuscript</i>	0.623	0.606	0.476	0.599	0.510	0.488	0.518	0.508

Table 3: IAA scores for Gamma, when various components are taken into account. In column **Position only**, categorical agreement is irrelevant. Column **All** shows scores when all components are uniformly weighted ($\frac{1}{6}$).

$\frac{1}{6} = 0.166$. As discussed above, the scores are calculated on the best possible alignment, which is determined by the Gamma metric itself. This means that every column in Table 3 is (potentially) calculated with a different alignment.

If these scores are evaluated in usual IAA terms,²⁵ they are rather low. Even relatively clear components, such as the source of the transfer (which is often just the character speaking), seem to be more difficult than expected. The variance between texts is also noticeable. Günderrode's *Udohla* seems to be the most difficult one to annotate, while the results for Weissenthurn's *Das Manuscript* are much more promising.

The main reason for the low scores, however, is not a disagreement on individual components of the knowledge transfer, but the fact that many annotations do not have a counterpart at – roughly – the same position in the text (as can also be seen exemplary in Figure 1). This means that many of the annotations are aligned with a dummy annotation which yields maximal categorical dissimilarity. Thus, it seems to be more difficult to decide if an annotation should be made at some position than to decide on the individual annotation's categories.

4.3. Discussion

The calculation of inter-annotator agreement for complex annotation tasks like the one we have presented here is not straightforward. To tackle this issue we decided to use the highly adaptable measure Gamma. Our customised version of Gamma allows for a tentative assessment of the agreement between the two annotations. It permits us to evaluate the difficulty of annotating a play, when compared to other plays. In addition, we get a clearer picture regarding the difficulty of the annotations' different components (like SOURCE vs. TARGET). However, many properties of the annotations are not yet captured in a fully satisfactory way and the highly adaptable nature of Gamma presents us with a large number of choices, not all of which can be motivated theoretically.

The core conceptual question is what to consider as agreement (or disagreement). Two

25. Many publications at this point refer to the table by Landis and Koch, published in the context of diagnostics of multiple sclerosis diagnosis, but even Landis and Koch consider the table "arbitrary" (Landis and Koch 1977, p. 165).

annotators marking the exact same span of text with the exact same label is not very likely and not really necessary. We decided to consider two annotations as an agreement if the annotation spans overlap, because there is usually some key term (like *mother*) that will definitely be annotated while the question of how much syntactic context should be included will be answered differently by different annotators. For some cases, we could go even further and declare two annotations an agreement if they appear in the same scene or act. For love relations that develop gradually finding agreement on which text segment is crucial for knowing that A loves B is especially hard. One annotator might already take the first allusions as justified evidence for an annotation (see example 7 from Weißenturn's *Das Manuskript*, 'The manuscript') while another might wait for a segment that removes the last doubt (example 8). Both decisions can be legitimate and a contrasting analysis of how different readers perceive the development of the relationship could be very fruitful. Our current agreement measure does not account for this scenario.

- (7) EMERIKE etwas verschämt. (*a little shy*)
 Ich kenne einen Andern, den ich gerne glücklich machen möchte.
I know someone else, whom I would like to make happy.
 FLINT.
 Einen – Andern?
 Someone – else?
 EMERIKE. Ja – ich kenne – [...], denn ich möchte Ihnen sagen – Herzlich. daß ich Ihnen recht gut bin.
Yes – I know – [...], because I want to tell you – Sincerely. that I am quite sympathetic to you.
- (8) EMERIKE mit einem Blick auf Flint. (*looking at Flint*)
 Ach nein! er will mich nicht, und ich werde doch keinen Andern lieben.
Oh no! He does not want me and I will still not love anyone else.

With regard to the comparison of annotation labels we also want to incorporate inferences that can be drawn from relations that are logically related or equivalent. For undirected relations, it is obvious, e. g., that *siblings(A, B)* and *siblings(B, A)* are semantically equivalent. As described above, this is already taken into account by our customisation of Gamma. But more complex cases would need to be covered as well. Directed relations oftentimes have a complimentary relation that can be used to express the same fact, like *parent_of(A, B)* and *child_of(B, A)*. Our annotators are asked to base their decision on the textual expression of the relation, but some ambiguities remain. Depending on previous knowledge about familial character relations, other pairs of relations can also be equivalent. In *Die Familie Schroffenstein* by Heinrich von Kleist we encounter such an ambiguity in the list of characters at the beginning of the play:

- (9) Rupert, Graf von Schroffenstein, aus dem Hause Rossitz.
Rupert, count of Schroffenstein, from the house of Rossitz.
 Eustache, seine Gemahlin.
Eustache, his wife.

Ottokar, ihr Sohn. 433
Ottokar their/her son. 434

The pronoun *ihr* can either be plural or singular, feminine, and thus refer to Eustache 435
 and Rupert or only to Eustache. This corresponds to the following two options for the 436
 annotation: 437

(10) `transfer("Dramatis Personae", audience, child_of(ottokar, eustache))` 438
 (11) `transfer("Dramatis Personae", audience, child_of(ottokar, [eustache, 439
 rupert]))` 440

Given that we know that Rupert and Eustache are married, we might want to consider 441
 these annotations a match, even though the surface form is different. To actually com- 442
 pare the readings of the two annotators, we would need to analyse if one reading is 443
 semantically equivalent to the other. We are therefore working on an inference system 444
 that automatically expands the annotated relations to all relations that are logically 445
 inferable. Once this is completed, we can update our notion of agreement and consider 446
 annotations as agreement if they result in the same knowledge base for the characters 447
 involved. This is complicated by the fact that, in example 9, strictly speaking, we cannot 448
 logically infer that Rupert is Ottokar's father. Still, a human reader of this list will most 449
 likely assume this relationship unless presented with contradicting information. 450

For the implementation of Gamma, the choices of weighting need to be further discussed 451
 and refined. Fundamentally, it is necessary to justify how positional agreement and 452
 categorical agreement should be weighted against each other. As our annotation labels 453
 are complex, we additionally have to establish a weighting of the individual compo- 454
 nents. In our current implementation, all six components are considered independently 455
 and have equal weight. This independence assumption raises new questions though: 456
 Currently, the way we compare related characters is determined by the relation name. 457
 If both annotators use *siblings* as a label, we can compare the characters with the 458
 Jaccard index. It is unclear, however, how to proceed if one annotator specified a di- 459
 rected and the other an undirected relation. In addition, the component's independence 460
 can lead to non-intuitive judgements: If one annotator argues for a given text passage 461
 that `parent_of(A, B)`, whereas the other annotator argues that `lovers(A, B)`, this 462
 would be considered a 2/3 match, even though the transmitted information differs 463
 significantly. 464

5. Analysing Annotated Knowledge Transfers 465

The following section is dedicated to analysing our annotated corpus with a focus on 466
 two different aspects of our annotations. In a first investigation we concentrate on the 467
 annotated relations' quantitative properties (5.1): Which relations are annotated most 468
 often by which annotator in our corpus? And when in a play is knowledge distributed 469
 to other characters or to the audience? We discuss the results with regard to established 470
 drama theoretical views. Our second examination makes use of character networks to 471

Annotation 1		Annotation 2	
Relation	Count	Relation	Count
in_love_with	110	in_love_with	109
identity	73	identity	79
child_of	63	child_of	50
parent_of	44	parent_of	44
!in_love_with	39	has_name	44
has_name	31	!in_love_with	32
engaged	30	engaged	30
siblings	21	siblings	25
lovers	17	spouses	16
spouses	16	lovers	11

Table 4: 10 most frequently annotated relations per annotator.

evolve conventional methods of dramatic network analysis, which currently is mostly based on so called configurations (cf. Pfister 1988, pp. 171–176). Doing so, we not only visualise the annotated knowledge transfers as a network, but we also compare different characters in view of centrality measures (5.2). Focusing on Günderrode’s *Udohla*, we will exemplify our network analytical approach on a single play. Using a more content-based form of character networks, we try to chart a path to better integrate quantitative analysis and interpretative reading. As we have argued in the previous section, there can be more than one way of interpreting a text and possibly also more than one way of modelling knowledge transfers in our annotation scheme. While we use both versions of the annotations for our statistical analyses in (5.1), we have created a consensus version for the analysis of Günderrode’s *Udohla*. As the IAA scores already suggest, the two versions of *Udohla* differ rather significantly and would thus result in quite varying networks.

5.1. Quantitative Analysis of the Annotations

In total, our analysed corpus consists of 20 plays (see Table 2 for an overview) annotated by two annotators. It contains 1057 transfer annotations (551 for Annotation 1 and 506 for Annotation 2). On average, there are 26.4 (± 11.5) annotations per play and 1.06 (± 0.56) annotations per 1000 tokens. The standard deviations indicate a substantial variation between the plays. Table 4 shows the ten most frequently annotated relations for each annotator. Overall, the ranking is fairly similar for both annotators and two adjacent relations switch ranks only twice. The relation *in_love_with* is by far the most frequent, with its negation following shortly after. In contrast to most family relations, love relations can change over time. They can be hinted at, be part of rumours or trigger an important conflict for a play’s rising action. Hence, they are talked about more often than other relations. The *identity* relation occupies the second rank. It is most frequently used for characters that are first mentioned without name and therefore annotated by a variable at first that is later unified with their character id. Unsurprisingly, the relations *child_of* and *parent_of* are also frequent and mark the importance of the core family for the plot of our selected plays.

Out of 289 characters in total that appear in the plays, around 50 % are involved in knowledge transfers, with 38 % being the source and around 43 % being the target of knowledge transfers at least once. Looking at female and male characters separately, 58 % of all female characters and 49 % of all male characters are involved in knowledge transfers. Out of the 1057 transfers, in 473 cases (45 %) SOURCE transfers a relation involving themselves and only 56 times (5 %) TARGET learns about a relation concerning themselves. It is evident that characters possess the most knowledge about their own relations and can therefore pass on this knowledge reliably. For the same reason, learning about one's own family or love relations is rather rare, but might point to especially interesting passages of the plot.

Additionally, we also investigate when in a play knowledge transfers happen. Figure 2 shows the number of annotations over the relative position at which they occur and who they are directed at: other characters in the internal communication system, the audience or both. The position in this analysis encompasses the entire text, including the dramatis personae. We bin the number of annotations, so that each bar covers a range of around five percent. Thus, a combined number of 123 annotations were made in the first five percent of each drama with the audience as the target, 65 annotations were made in the next five percent and so on. We can see that the segments right at the beginning and end of a play are the ones with the highest number of annotations. The remaining segments of the plays have a more or less similar distribution of annotations with increases in the middle of the plays and in the final quarter. At the beginning of the plays, the majority of information is transferred to the audience, while this focus shifts to the characters towards the middle and end of the plays.

This observation can be explained conclusively as it supports established drama theory. Beginning and end of a play are central places for the transmission of knowledge both in the internal and the external communication system. "What we understand as the transmission of information at the beginning of a play largely coincides with the classical theoretical concept of the exposition," (Pfister 1988, p. 86) acknowledges Manfred Pfister. He goes on to define the exposition as forwarding of information concerning "events and situations from the past that determine the dramatic present".²⁶ With regard to the audience, the transmission of information in the character's internal communication system or the dramatis personae fulfils at least two functions: On the one hand, it is intended to instigate the audience's attention at the beginning of a play. On the other hand, the audience is provided with the knowledge necessary to understand the subsequent actions (cf. Asmuth 2016, pp. 103–105). As Figure 2 illustrates for our annotated corpus, most of the new knowledge about family and love relations at the beginning of a play is indeed directed at the audience – oftentimes even solely. Similar to the exposition at the beginning, the resolution at the closure of a play is a common section for transmitting unknown information, e. g., through recognition. In such closed endings, deviations in knowledge between characters and the audience, which, e. g., can evoke dramatic irony in the scenes prior to the resolution, are typically dissolved:

26. The exposition, then, is not necessarily limited to a play's introduction. Furthermore, not every information that is transmitted early on serves an expository purpose.

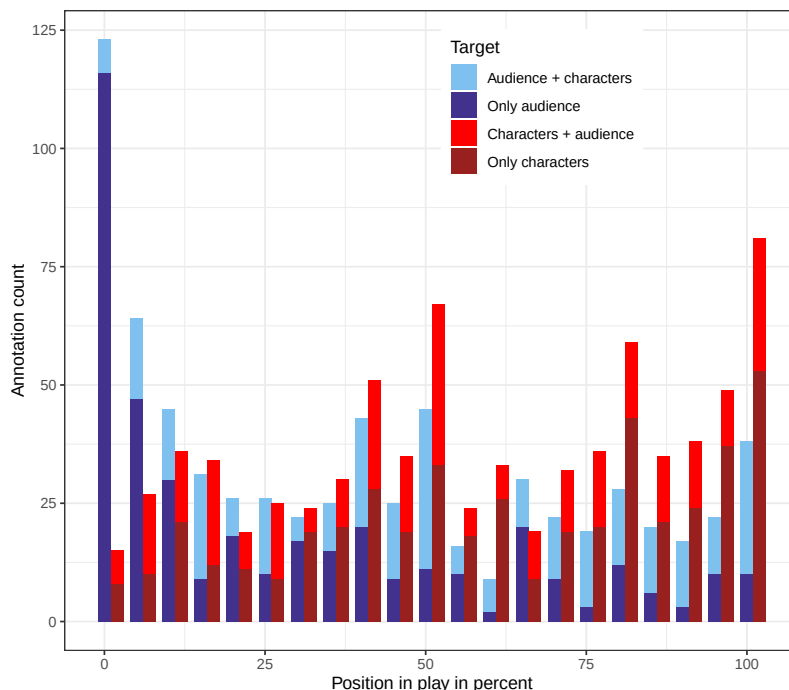
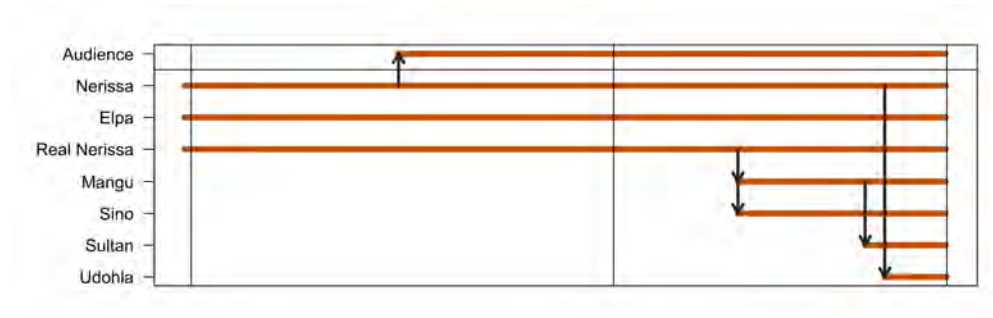


Figure 2: Histogram showing the number of annotations by both annotators at different positions in course of the 20 plays. The annotations are separated by the target of the knowledge transfer: (i) Only the audience is the target, (ii) the audience and one or more characters are the target, or (iii) only characters are the target, but not the audience.

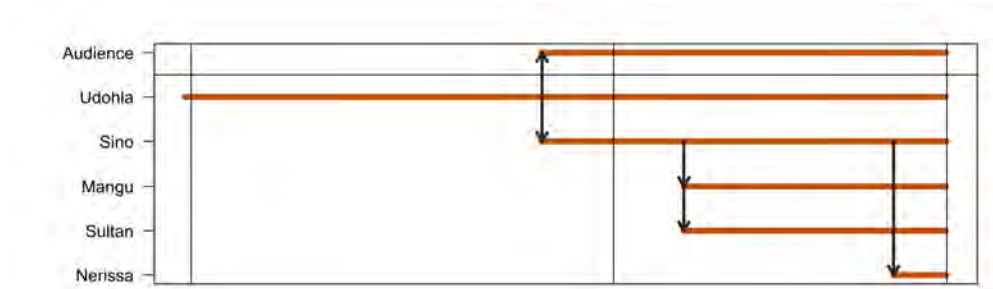
“as a result of either intrigue, self-deception or lack of information”, a character or even 542
a group of characters have gotten into trouble. “This situation then culminates in either 543
a happy or a tragic ending, after additional information has been introduced” (Pfister 544
1988, p. 95). The values in Figure 2 show a shift of direction towards the end of the 545
annotated plays. About halfway through the plays, the number of annotations directed 546
at characters in the internal communication system increases relatively to those directed 547
at the audience. The unknown knowledge then, which is transmitted in the resolution, 548
frequently seems to be addressed at the plays’ characters. The audience, in turn, already 549
possess the information necessary to deduce the probable outcome. Thus, the suspense 550
felt by the audience at the end – at least in our corpus – seems to be in respect to *how* an 551
information they possess influences the characters’ actions. 552

These theoretical considerations can in turn be exemplified by Günderröde’s *Udohla*. 553
By discussing the possible marriage between the Sultan and Nerissa at the beginning 554
of the play, Sinu, Mangu and the Dervish indirectly pass on their knowledge to the 555
audience. In doing so, the audience is also put in the picture that Nerissa and the Sultan 556
are siblings – at least according to the current beliefs of the present characters. As 557
Figure 3a visualises, it is Nerissa herself that indirectly corrects this wrong information 558
for the audience while talking to Elpa. From there on the audience has an information 559
advantage over most of the fictional characters. For the other characters, it takes until the 560
middle of the second act, where Mangu receives a letter of the Sultan’s actual sister, to 561

learn of this fact. Udohla, in the meantime, passes the information that he is Bahadar's 562
son to Sino. Sino then passes the knowledge to Mangu (between the scenes), who in 563
turn tells the sultan (see Figure 3b). The resolution at the end of the play then brings 564
together the knowledge acquired by the various characters in the course of the play. 565
Nerissa reveals that she is the daughter of Bahadar. She is the last character to learn that 566
Udohla, too, is Bahadars child and thus her brother. 567



(a) Distribution of the knowledge `!siblings(sultan, nerissa)`.



(b) Distribution of the knowledge `child_of(udohla, "Bahadar")`.

Figure 3: Knowledge distribution over the two acts of Günderröde's *Udohla*.

5.2. Networks of Knowledge Transfer

568

As a second kind of analysis, we use the annotated knowledge transfers to construct 569
character networks. Networks, which are based on the knowledge about family relations 570
and its dissemination in a play, can help to identify key characters that propel the 571
dramatic plot either by gaining new information or by distributing it. As in these 572
networks each node represents a character (or other sources of information like letters, 573
observations, etc.) and edges between nodes signify that one or more family related 574
knowledge transfer(s) between two nodes have taken place,²⁷ they can be used to 575
complement the information gathered by established configuration based networks. 576
Since there is a SOURCE and a TARGET to each knowledge transfer, the networks are 577
directed. The edges can be weighted with the total number of knowledge transfers 578
that have taken place between two nodes. An example of such a network is shown in 579
Figure 4 for Günderröde's *Udohla*. The nodes are scaled according to their weighted 580

27. To compute the knowledge transfer networks, we only focus on the internal communication system of the dramatic characters. Therefore, we omitted the audience's nodes and the dramatis personæ in the networks.

degree (Barrat et al. 2004), which is a measure that calculates the sum of the weights of all incoming and outgoing edges for each node.

The visualisation in Figure 4 shows Udohla, Sino, a Hindu staff member of the Sultan, and the vizier Mangu, to be the central characters of the network according to their weighted degree. Sino and Mangu being two of the most central characters in the network might seem surprising on a first glance, as for the plot and its resolution there are more important characters, mostly Nerissa, Udohla and the Sultan. How can their central position, i. e., their high weighted degree then be explained? For Sino, there are mainly two reasons: The first reason concerns the intra-fictional progression of the plot. Günderrode conceptualised Sino as Udohla’s only confidant within the Sultan’s palace. Both Sino and Udohla are Hindus and they are linked through a mutual close acquaintance. Naturally, then, Sino is the only character in the play Udohla could trust to share his real identity with, which is important for the play’s final scene as Sino is able to confirm to the Sultan that Udohla is Bahadar’s son. The second reason is that Sino’s role is used to transmit knowledge from the internal communication system to the audience. Herein, Sino becomes the recipient of new information, while in reality the audience is “the intended receiver of the information given.” (Pfister 1988, p. 89) To that effect, Mangu takes on a different role in the network. As he receives the letter of the Sultan’s real sister, he is then able to pass on the information that Nerissa is not the Sultan’s sister to other characters. The audience, however, already knows this fact from an earlier conversation of Nerissa and Elpa.

To further track the development of knowledge in the course of *Udohla*, we bin the play’s text into 10 equal-length segments and create a network based on the consensus version found in each of these segments. On this network, we calculate in- and out-strength. While the strength metric that was used to scale the nodes in Figure 4 uses both incoming and outgoing edges, in-strength only considers incoming, and out-strength only considers outgoing edges for the calculation. Figure 5 shows cumulative curves for the development of both in-strength and out-strength in *Udohla*. Here, *cumulative* means that the networks of each bin are constructed by taking the annotations of the current bin and all previous bins. In this way, we can see which character received and transferred knowledge about family relations at what point in the play. There are some instructive observations that are in need of interpretation: Firstly, Udohla’s high out-strength value is mostly linked to a single scene right in the middle of the play, where he introduces himself as Achmed pretending to be the Nawab’s herald. As he passes this false information to five other characters, it has a big impact on his central position in the network. Secondly, the roles of Sino and Mangu in the knowledge transfer network seem to be roughly comparable. Both receive knowledge about family relations that they in turn pass on to other characters. While Sino can be described as confidant of Udohla, Mangu, being a Muslim, takes on a similar role with regard to the Sultan. All the important information the Sultan receives before the final resolution come from Mangu. Thirdly, the Sultan’s role in view of knowledge distribution is strikingly passive. He is only TARGET of knowledge transfers, never the SOURCE. This underlines a different conceptualisation of the Sultan’s character. Although he does indeed receive some

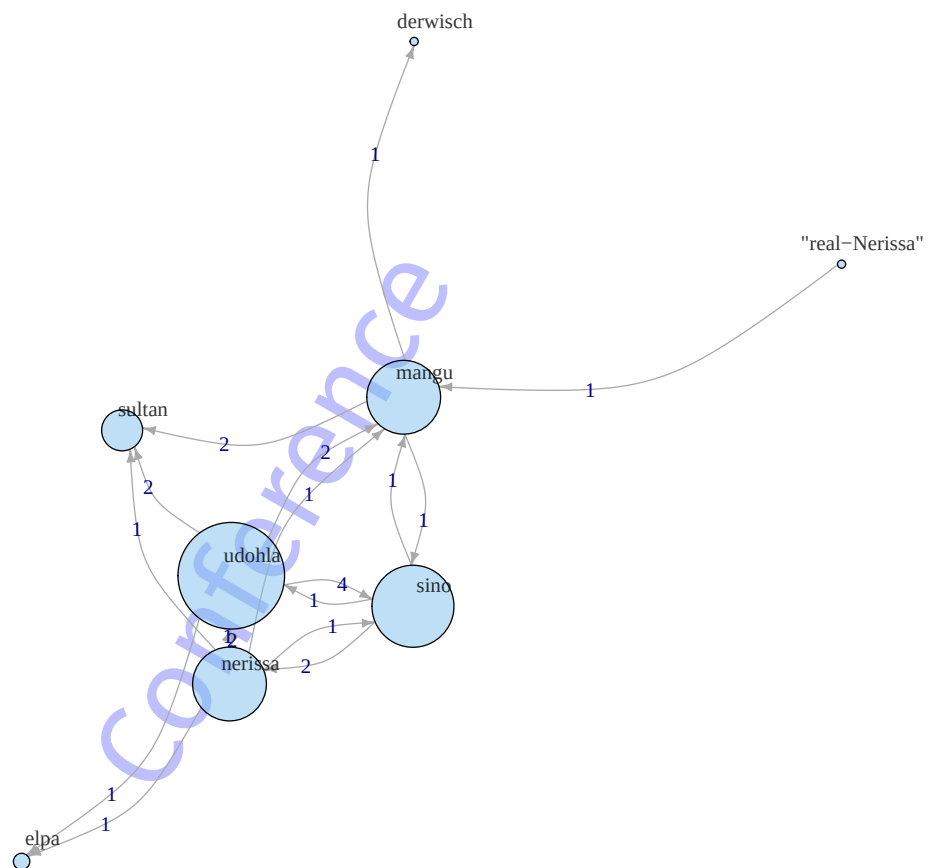


Figure 4: Network of knowledge transfer in Günderrode's *Udohla*. Based on the consensus version.

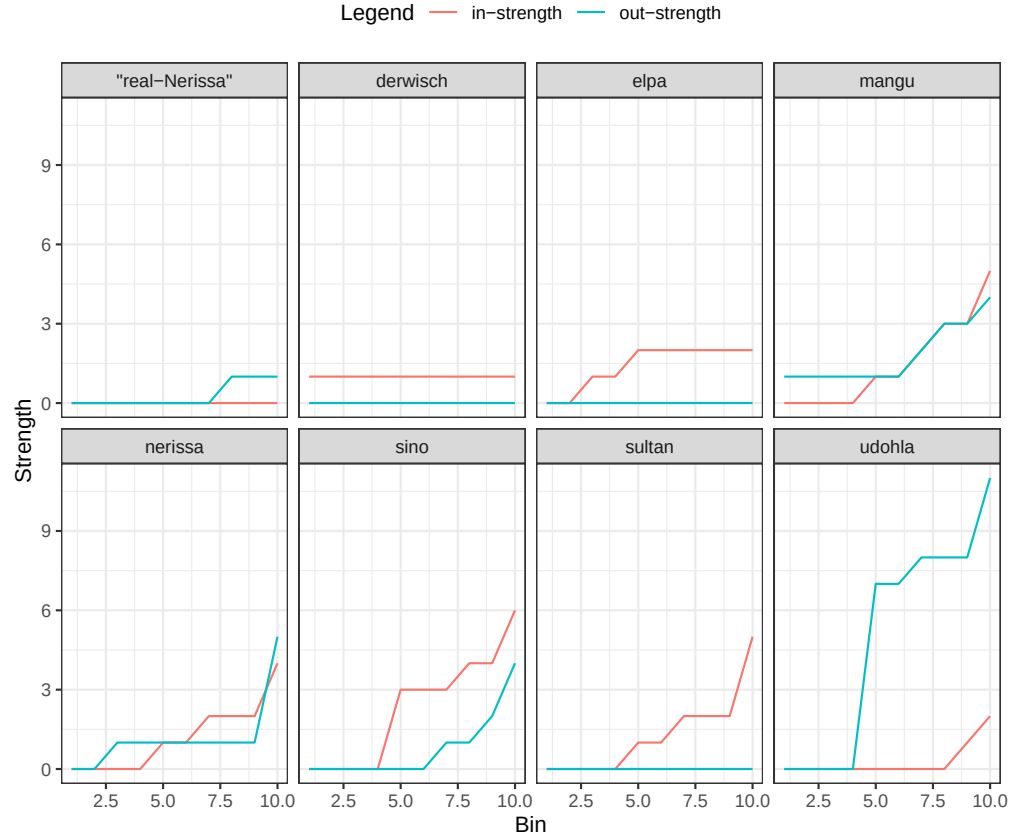


Figure 5: Cumulative in-strength and out-strength in the course of Günderrode’s *Udohla* for all the involved entities.

information in the course of the play, oftentimes he is the last character to be reached by this knowledge. Looking at the play’s resolution this makes sense. Being at the centre of the final recognition scene, the Sultan has to be unaware that Nerissa and Udohla are the children of Bahadar until this point in time. Sino and Mangu, on the other hand, accumulate new knowledge throughout the play and serve as middlemen, bridging the knowledge either to the audience or to the main characters.

Following this, we investigate the so-called betweenness centrality (Freeman 1977) of the network. Betweenness centrality measures how often a node k is part of a shortest path between two other nodes and is formally defined as

$$b(k) = \sum_{i \neq j \neq k}^N \frac{g_{ij}(k)}{g_{ij}} \quad (3)$$

where g_{ij} is the number of shortest paths (or geodesics) between two nodes i and j and $g_{ij}(k)$ is the number of geodesics of these two nodes that passes k (cf. Freeman 1977, p. 37). Figure 6 shows the development of betweenness centrality for *Udohla*.

Since betweenness centrality can be seen as a measure for the flow of communication

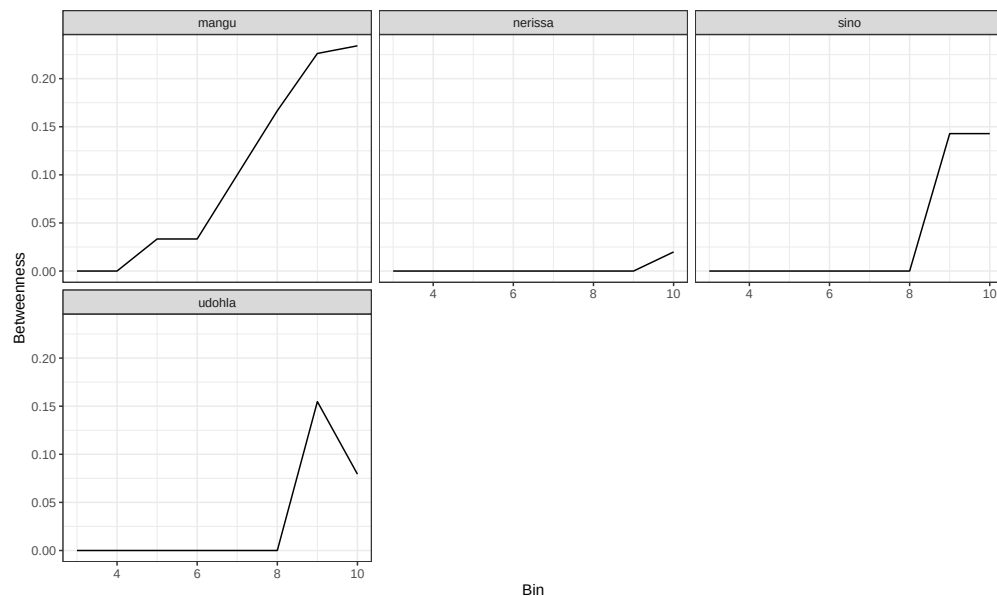


Figure 6: Betweenness centrality in the course of Günderrode’s *Udohla* for all the involved entities. Three characters received a betweenness centrality value of 0 for all positions and were omitted from the graph: derwisch, elpa and sultan.

in a network and how single nodes control the flow of communication, it appears to be especially suited for networks of knowledge transfer. Its “use seems natural in the study of communication networks where the potential for control of communication by individual points may be substantively relevant” (Freeman 1977, p. 40), as Linton Freeman states in his pioneering study. In Figure 6 we can see that Sino and especially Mangu are the characters with the highest betweenness centrality in the play. This further corroborates their role as middlemen in the play. Moreover, the visualisation illustrates that in *Udohla* knowledge transfers responsible for betweenness centrality mostly occur in the second half or even the end of the play. Looking at the structure of a theatre play, this makes sense from a conceptual point of view: As a node has to be both TARGET and SOURCE of at least one knowledge transfer to be part of a shortest path, it is not surprising to find this realised only towards the end of *Udohla*. As shown above, beginning and end of a play are key segments for the transmission of knowledge. In order to have a play’s resolution resulting directly from a recognition scene – as is demonstrated in Günderrode’s *Udohla* – the characters involved must possess a different knowledge base right until that moment.

In summary, the analyses we have shown a fruitful perspective for more extensive investigations on a bigger corpus. They illustrate that our annotation data can provide insights into different structural principles of German plays. Sino’s and Mangu’s central position in the network and their values for in- and out-strength as well as betweenness centrality furthermore show the potential of our methodological approach. As exemplified in *Udohla* we can detect characters that take a key role for the flow of knowledge in the course of the play, without being considered as main characters themselves. Although our networks are based on the transmission of knowledge about family relations, they

depend on co-presence networks. Thus, they can be described as second-order networks. I. e., if in the course of a play two characters are not present on stage together, it is highly unlikely that new information circulates between them. Therefore, we consider a systematic comparison between co-presence networks and knowledge transfer networks as an especially insightful task for future research.

6. Conclusions

In this article, we have presented a composite scheme for the annotation of knowledge transfers about family relations in German plays. As illustrated throughout our article, annotating these knowledge transfers is a complex task, which gives rise to a number of challenges. Our scheme is based on considerations of drama theory on knowledge distribution. As our results are prospectively also intended to be of relevance for research in traditional literary studies, we have refrained from an operationalisation that overly simplifies concepts in light of computation. Instead, we chose an operationalisation that purposefully connects to terms and concepts of drama theory. As a consequence, the scheme is situated at the intersection between annotation and modelling.

At the same time, this project is (to our knowledge) the first that attempts to measure inter-annotator agreement for such a complex annotation task by employing the metric Gamma. We have discerned a number of intricacies that make the application of Gamma tricky and might be relevant for other annotation projects in computational literary studies: While the ability to provide a custom similarity function makes Gamma versatile, this also requires us to make a high number of design decisions that influence the results and decrease comparability with other applications of Gamma. Conceptually, the definition of what we want to consider a positional and/or categorical agreement is not always straightforward because of the (sometimes) vague nature of the target phenomenon, the compositionality of the annotation labels, and dependencies between its components.

As our preliminary analyses have shown, a systematic annotation of knowledge transfers about family relations allows for investigations that go beyond structural features of the play's surface. Herein, we made use of our annotation data to propose an extension to the widely utilised co-presence networks. In specifying the edges as a directed knowledge transmission, networks can be interpreted in light of more tailored research questions as we have hinted at with Günderrode's *Udohla*. The analyses have also revealed clear perspectives for larger corpus studies. This gives rise to future questions concerning literary history: Do patterns of family related knowledge distribution emerge for different dramatic genres? Is it possible to characterise the scenes where changes of knowledge occur in more detail? How many characters are on stage in these scenes? How many of them are actively involved in passing on knowledge? What kind of characters do pass on the knowledge?

Our future work mostly focuses on two aspects. Firstly, we are currently implementing a system to automatically infer all deducible family relations from our annotations. As

the annotations only cover the transmission of new information from one character to 701
another (or to the audience), this inference system is needed to have a full account 702
of what all characters and the audience know at all times during the drama. Having 703
this knowledge base would both benefit the measuring of the IAA – as it would solve 704
certain problems such as using different predicates for the same relation – and the 705
subsequent analysis. Secondly, we are working on automating certain aspects of the 706
annotation process by creating transformer-based machine learning models which learn 707
to predict the positions in a text where knowledge is transferred and the type of family 708
or love relation that is transferred. Applying these models on new data will facilitate 709
the annotation of new texts. Evaluating the performance of the models on existing data 710
can give additional insights into the complexity of the annotation task. 711

Conference

7. Data availability 712

Data can be found here: <https://github.com/quadrama/jcls2022> 713

8. Software availability 714

Software can be found here: <https://github.com/quadrama/jcls2022> 715

9. Acknowledgements 716

The research in this article has been conducted in the Q:TRACK project (<https://quadrama.github.io/index.en>), which is part of the priority programme SPP 2207 *Computational Literary Studies* and funded by the German Research Foundation (DFG). We would like to thank Jonas Hirner and Christian Lantzinger for their annotations.

10. Author contributions 721

Melanie Andresen: Annotation Supervision, Guideline Development, Inter-Annotator Agreement, Writing 722

Benjamin Krautter: Theoretical Framework, Analysis and Interpretation, Writing 724

Janis Pagel: Corpus Statistics, Network Analysis, Writing 725

Nils Reiter: Methodology, Inter-Annotator Agreement, Writing 726

References 727

- Anderson, Maxwell (1965). "The Essence of Tragedy". In: *Aristotle's "Poetics" and English Literature*. Ed. by Elder Olson. Chicago and London, pp. 114–121.
- Anz, Thomas (2007). "[Art.] Spannung". In: *Reallexikon der deutschen Literaturwissenschaft. Neuberarbeitung des Reallexikons der deutschen Literaturgeschichte*. Ed. by Jan-Dirk Müller, Georg Braungart, Harald Fricke, Klaus Grubmüller, Friedrich Vollhardt, and Klaus Weimar. Vol. 3. Berlin and New York: De Gruyter, pp. 464–467.
- Aristotle (1995). "Poetics". In: *Aristotle: Poetics*. Ed. by Stephen Halliwell. Cambridge and London: Harvard University Press, pp. 27–141.
- Artstein, Ron and Massimo Poesio (Dec. 2008). "Inter-Coder Agreement for Computational Linguistics". In: *Computational Linguistics* 34.4, pp. 555–596.
- Asmuth, Bernhard (2016). *Einführung in die Dramenanalyse*. 8., aktualisierte und erweiterte Auflage. Stuttgart: J.B. Metzler.
- Barrat, A., M. Barthélemy, R. Pastor-Satorras, and A. Vespignani (2004). "The architecture of complex weighted networks". In: *Proceedings of the National Academy of Sciences* 101.11, pp. 3747–3752. DOI: [10.1073/pnas.0400087101](https://doi.org/10.1073/pnas.0400087101). URL: <https://www.pnas.org/content/101/11/3747>.

- Cave, Terence (1988). *Recognitions. A Study in Poetics*. Oxford: Clarendon Press. 744
- Destrée, Pierre (2020). "Family Bounds, Political Community, and Tragic Pathos". In: *The Poetics in its Aristotelian Context*. Ed. by Pierre Destrée, Malcolm Heath, and Dana L. Munteanu. London and New York: Routledge, pp. 113–128. 745
- DiYanni, Robert (2000). *Drama. An Introduction*. Boston [MA]: McGraw-Hill. 746
- Dubbels, Elke (2012). "Zur Dynamik von Gerüchten bei Heinrich von Kleist". In: *Zeitschrift für deutsche Philologie* 131.2, pp. 191–210. 747
- Dutant, Julien (2015). "The Legend of the Justified True Belief Analysis". In: *Philosophical Perspectives* 29.1, pp. 95–145. doi: <https://doi.org/10.1111/phpe.12061>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/phpe.12061>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/phpe.12061>. 752
- Engelstein, Stefanie (2004). "Sibling Incest and Cultural Voyeurism in Günderode's "Udohla" and Thomas Mann's "Wälsungenblut"". In: *The German Quarterly* 77.3, pp. 278–299. 753
- Evans, Bertrand (1960). *Shakespeare's Comedies*. Oxford: Clarendon Press. 754
- Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hecht, Christopher Kittel, Carsten Milling, and Peer Trilcke (2019). "Programmable Corpora – Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor". In: *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*, pp. 194–197. doi: [10.5281/zenodo.2596095](https://doi.org/10.5281/zenodo.2596095). 755
- Fleiss, Joseph L. (1971). "Measuring nominal scale agreement among many raters". In: *Psychological Bulletin* 76.5, pp. 420–428. 756
- Freeman, Linton C. (1977). "A Set of Measures of Centrality Based on Betweenness". In: *Sociometry* 40.1, pp. 35–41. doi: [10.2307/3033543](https://doi.org/10.2307/3033543). 757
- Gettier, Edmund L. (1963). "Is Justified True Belief Knowledge?" In: *Analysis* 23.6, pp. 121–123. 758
- Günderode, Karoline von (1990). "Udohla". In: *Karoline von Günderode: Sämtliche Werke und ausgewählte Studien*. Ed. by Walter Morgenthaler. Vol. 1. Frankfurt am Main: Stroemfeld/Roter Stern, pp. 203–231. 759
- Ichikawa, Jonathan Jenkins and Matthias Steup (2018). "The Analysis of Knowledge". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University. 760
- Jaccard, Paul (Feb. 1912). "The Distribution of the Flora in the Alpine Zone". In: *The New Phytologist* 11.2. doi: [10.1111/j.1469-8137.1912.tb05611.x](https://doi.org/10.1111/j.1469-8137.1912.tb05611.x). 761
- Jeßing, Benedikt (2015). *Dramenanalyse: Eine Einführung*. Berlin: Erich Schmidt Verlag. 762
- Kablitz, Andreas (1998). "Wiedererkennung: Zur Funktion der Anagnorisis in der klassischen französischen Tragödie (Corneille: *Cédipe* – Racine: *Iphigénie en Aulide*)". In: *Erkennen und Erinnern in Kunst und Literatur*. Ed. by Dietmar Peil, Michael Schilling, and Peter Strohschneider. Tübingen: Max Niemeyer Verlag, pp. 455–486. 763
- Krautter, Benjamin and Marcus Willand (2021). "Vermessene Figuren: Karl und Franz Moor im quantitativen Vergleich". In: *Schillers Feste der Rhetorik*. Ed. by Peter-André Alt and Stefanie Hundehege. Berlin and Boston [MA]: De Gruyter, pp. 107–138. 764
- Krippendorff, Klaus (2004). *Content Analysis: An Introduction to its Methodology*. 2nd. Los Angeles, California, USA: Sage. 765

- Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1, pp. 159–174. 788
789
- Lipinski, Silke (2011). "Udohla – Plattform für Karoline von Günderrodes philosophische Gedanken". In: *New German Review* 24.1, pp. 113–122. 790
791
- Marcus, Solomon (1973 [1970]). *Mathematische Poetik*. Bukarest and Frankfurt am Main. 792
- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier (2015). "The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment". In: *Computational Linguistics* 41.3, pp. 437–479. 793
794
795
- Moretti, Franco (2011). "Network Theory, Plot Analysis". In: *Pamphlets of the Stanford Literary Lab* 2, pp. 1–12. URL: <https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf>. 796
797
798
- Pagel, Janis, Nils Reiter, Ina Rösiger, and Sarah Schulz (July 2020). "Annotation als flexibel einsetzbare Methode". In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der Creta-Werkstatt*. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. Berlin: De Gruyter, pp. 125–141. DOI: [10.1515/9783110693973-006](https://doi.org/10.1515/9783110693973-006). 799
800
801
802
- Pfister, Manfred (1988). *The Theory and Analysis of Drama*. Trans. by John Halliday. Cambridge and New York: Cambridge University Press. 803
804
- Pichler, Axel and Nils Reiter (2021). "Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse. Eine Annäherung über Norbert Altenhofers hermeneutische Modellinterpretation von Kleists Das Erdbeben in Chili". In: *Journal of Literary Theory* 15.1–2, pp. 1–29. 805
806
807
808
- Pollock, John L. and Joseph Cruz (1999). *Contemporary Theories of Knowledge*. 2nd ed. Lanham et al.: Rowman & Littlefield. 809
810
- Reiter, Nils (2018). "CorefAnnotator – A New Annotation Tool for Entity References". In: *EADH 2018*. EADH 2018. Galway, Ireland. URL: <https://eadh2018.exordo.com/programme/presentation/118>. 811
812
813
- Reiter, Nils and Leonard Konle (2022). "Messverfahren zum Inter-Annotator-Agreement". In: forthcoming. 814
815
- Reiter, Nils and Marcus Willand (2019). "Surveying Shakespeare's Impact on German Drama: Taking a Computational Approach to an Epoch". In: *Anglo-German Dramatic and Poetic Cultures: New Perspectives on Exchange in the Sattelzeit*. Ed. by Michael Wood and Sandro Jung. Bethlehem, PA: Lehigh University Press, pp. 117–143. 816
817
818
819
- Schadewaldt, Wolfgang (1955). "Furcht und Mitleid? Zur Deutung des Aristotelischen Tragödiensatzes". In: *Hermes* 83.2, pp. 129–171. 820
821
- Schmitt, Arbogast (2008). "Kommentar". In: *Aristoteles: Werke*. Ed. by Hellmut Flashar. Vol. 5. Darmstadt, pp. 193–741. 822
823
- Trilcke, Peer (2013). "Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft". In: *Empirie in der Literaturwissenschaft*. Ed. by Philip Ajouri, Katja Mellmann, and Christoph Rauen. Münster: Mentis, pp. 201–247. 824
825
826
- Yarkho, Boris I. (2019 [1935–1938]). "Speech Distribution in Five-Act Tragedies (A Question of Classicism and Romanticism)". In: *Journal of Literary Theory* 13.1, pp. 13–76. 827
828
829

Anna Mareike Weimer  ¹

Florian Barth  ²

Tillmann Dönicke  ²

Luisa Gödeke  ¹

Hanna Varachkina  ³

Anke Holler  ¹

Caroline Sporleder  ²

Benjamin Gittel  ¹

1. Department of German Philology, University of Göttingen, Göttingen.

2. Göttingen Centre for Digital Humanities, University of Göttingen, Göttingen.

3. Göttingen State and University Library, University of Göttingen, Göttingen.

Keywords:

literary theory, narratology,
commentary, operationalisation,
annotation, supervised machine
learning

Licenses:

This article is licensed under:



Abstract. This paper explores how both annotation procedures and automatic detection (i.e. classifiers) can be used to assess the consistency of textual literary concepts. We developed an annotation tagset for the "literary comment"—a frequently used but rarely defined concept—and its subtypes (interpretative comment, attitude comment and meta-narrative/metafictional comment) and trained a multi-output and a binary classifier. The multi-output classifier shows FScores of 28% for attitude comment, 36% for interpretative comment and 48% for meta comment, whereas the binary classifier achieves FScores up to 59%. Crucially, both our annotation and the automatic classification struggle with the same subtypes of comment, although annotation and classification follow completely different procedures. Our findings suggest an inconsistency in the overall literary concept "comment" and most prominently the subtypes "attitude comment" and "interpretative comment". As a best-practice-example, our approach illustrates that the contribution of Digital Humanities to Literary Studies may go beyond the automatic recognition of literary phenomena.

1. Introduction

1

While Computational Literary Studies received much attention in recent years, the potential for collaboration between traditional Literary Studies and the Digital Humanities has not yet been fully explored. Arguments about the benefits of digital methods—often framed as promises for the future of Literary Studies—flourish, including: (1) a systematic application of concepts developed within Literary Studies (in what follows: "literary concepts") in the process of annotation leads to refining their definitions (cf. Gius and Jacke 2015; Gius and Jacke 2017); (2) the use of quantitative methods may

2

3

4

5

6

7

8

lead to “new forms of evidence” for literary phenomena and to a ‘scientification’ of Literary Studies (Jockers 2013, 5–10, here: 8); (3) insofar as automatic recognition of literary phenomena succeeds, a large number of examples can be readily retrieved and submitted to qualitative analysis (cf. Piper et al. 2021); (4) if an automatic recognition of literary phenomena in representative diachronic corpora is successful, it is possible to model developments in literary history (cf. Underwood 2016; Underwood 2019) and justify claims about generic literary entities like ‘the novel’ (cf. Piper 2018, p. xi). Digital Humanities could help to examine which literary concepts are useful for quantitative empirical research, thus potentially reducing the abundance of literary concepts. As has been shown in recent years, however, by no means all attempts to operationalise and automatically detect such literary concepts were successful. Some seem to resist operationalisation and/or automatic detection (cf. Herrmann, Dalen-Oskam, and Schöch 2015; Willand, Gius, and Reiter 2020).

By the “consistency of a concept”, we mean that a) comparatively homogeneous phenomena fall under it and b) that the concept is methodologically guiding in the sense that these phenomena are intersubjectively and automatically recognisable. Inconsistent concepts, on the other hand, describe comparatively heterogeneous phenomena, which cause hardly or not at all surmountable difficulties when trying to recognise them intersubjectively and/or automatically.

We will presume that a consistency study is feasible and demonstrate this using a concrete example: the “literary comment”, which lends itself very well to such an approach. Comments can be used to clarify a narrator’s/character’s attitude, to steer the reader’s attention, interpret or explain plot elements, reflect about the real world, the narration or the literary work (Gittel to appear), or signal an “overt-narrator” (Chatman 1980). As intuitively easy to understand as “comment” may seem at first glance, it nevertheless turns out to be surprisingly imprecise due to sketchy definitions and competing conceptualisations. In order to explore the consistency of a literary concept it is not sufficient to operationalise, annotate and detect it automatically and evaluate whether annotation and automatic detection succeeded or failed (by measuring inter-annotator agreement or a classifier’s performance). Rather, we seek to explain patterns of the annotation and automation experiments using qualitative and quantitative evidence such as measures of the relation between available training data and classifier-performance, the features found to be predictive for automatic classification as well as assessment and contextualisation of the relevant conceptualisations based on textual examples. Specifically, a consistency study carries out inferences to the best explanation where the inconsistency of a literary concept is a possible hypothesis among others that may explain certain outcomes. We utilise empirical observations from annotation and automation to gain insights on the consistency of the theoretical concept itself.

In the next section, we will scrutinize narratological research on the literary comment (section 2). We then operationalise the notion of “comment” and follow Chatman 1980 in distinguishing subtypes (section 3). We report the results of a collaborative annotation effort (section 4) and an automatic classification (section 5). Finally, we

discuss limitations and challenges of consistency studies of literary concepts in general (section 6.1), the results of our consistency study for the three types of literary comment (section 6.2), and conjectures on the overarching concept of "comment" (6.3).

2. Theoretical Background

Although the concept of "comment" is known in Literary Studies, it has not yet been systematically considered through the perspective of a consistency study.¹ Rather, the concept "comment" is often used in its commonplace understanding and would benefit from a more detailed analysis. In narratology, a comment usually is associated with the narrator making remarks on what is narrated, that interrupt the narration (cf. Zeller 1997) or with authorial intrusion (cf. Dawson 2016a). Its function goes beyond the description of action. Comments explain the meaning of a narrative element, make value judgments, and/or refer to the real world (cf. Prince 2003). The following thoughts are essentially based on the influential contributions of Bonheim 1975 and Chatman 1980. While Bonheim draws attention to structural features of comment, Chatman is interested in the multiplicity of phenomena subsumed under the concept of "comment".

Bonheim examines modes of narrative in accordance with their function, distinguishing between dynamic and static modes on the basis of their temporal constitution of discourse. The comment is treated as a static mode, along with the mode "description", and is thus contrasted with the dynamic modes "speech" and "report". Basically, the modes may overlap, but according to Bonheim, however, comment is the most autonomous and thus the purest of the modes and is most often found unblended (cf. Bonheim 1975, p. 332). While Bonheim does not define linguistic indicators for what constitutes a comment, he formulates criteria on the text-structural level: A comment must be embedded in a narrative pause and need not be descriptive (e.g. describing the scenery of the narrative).

In a narrative pause (cf. Lahn and Meister 2013, p. 154 drawing on Genette 1994 [1972]), the narrating time exceeds the narrated time such that readers might get the impression the narrated time stops or slows down extremely, although information is provided. However, the concept of "narrative pause" is not unproblematic, since, for now, there is no objective measure for narrating time (beside the word quantity indicator) and the determination of the narrated time may require complex interpretive decisions in individual cases.

Chatman distinguishes four types of explicit comment not as a mode of narrative, but as a quality of sentences or text passages (cf. Chatman 1980).² In the following, we will

1. Terminologically, both the term "comment" (Bonheim 1975) and "commentary" (Chatman 1980) are applied in literary studies. "Commentary" is a term with multiple meanings, often used colloquially in its narratological sense and with other meanings in historical criticism and journalism. In the following, we use "comment" in the article for the sake of uniformity.

2. Chatman makes an additional distinction between implicit and explicit comment. The former includes statements by unreliable narrators and ironic remarks that must be reconstructed by the reader and interpreted from the context. In the following, we focus on the explicit comments and leave out the implicit communication on account of its complexity.

take a closer look at the four comment types. 85

Generalisation Chatman defines generalising comments as general truths that can apply 86
not only to the fictional but also the real world. He takes his cue from Booth 1983, who 87
speaks of generalisation as the reinforcement of norms. 88

- (1) Sie [Otilie] ward den Männern vorgestellt und gleich mit besonderer Achtung als 89
Gast behandelt. Schönheit ist überall ein gar willkommener Gast. (Goethe 2012 90
[1809])³ 91

From a linguistic perspective, "generalisation" is an umbrella term for phenomena 92
like genericity and (overt) quantification. Thus, several linguistic markers might be 93
associated with "generalisations". We will come back to problems resulting from this in 94
section 3. 95

Interpretation The speaker, mostly the narrator, explains the plot pro- or analeptically 96
and provides additional information to help readers correctly understand what is being 97
told. (2), the end of E.T.A. Hoffmann's *Der Sandmann*, illustrates this usage. We provide 98
context to also clarify the function of the narrative pause. 99

- (2) Als Nathanael mit zerschmettertem Kopf auf dem Steinpflaster lag, war Coppelius 100
im Gewühl verschwunden. - Nach mehreren Jahren will man in einer entfernten 101
Gegend Clara gesehen haben, wie sie mit einem freundlichen Mann, Hand in 102
Hand vor der Türe eines schönen Landhauses saß und vor ihr zwei muntre Knaben 103
spielten. Es wäre daraus zu schließen, dass Clara das ruhige häusliche Glück noch 104
fand, was ihrem heiteren lebenslustigen Sinn zusagte und das ihr der im Innern 105
zerrissene Nathanael niemals hätte gewähren können. (E. T. A. Hoffmann 2012 106
[1816/17]) 107

In the first sentence of this example, plot is conveyed. The dash indicates a time jump. 108
In the following narrative pause we are given an insight into what happened to Clara 109
after the end of the narration: The narrator offers a description of Clara's situation. In a 110
third step (the underlined passage), this description is interpreted by the narrator and 111
therefore a comment. 112

Judgment Evaluative comments formulate the narrator's judgment reflecting values, 113
norms and beliefs. They are intended to confront the reader with ethical aspects included 114
in the story. Chatman distinguishes between interpretation and judgment only on 115
the basis of the (moral) evaluation underlying the judgment, while interpretation is 116
"relatively value-free" (Chatman 1980, p. 237). 117

- (3) Charlotte benutzte des andern Tags auf einem Spaziergang nach derselben Stelle 118
die Gelegenheit, das Gespräch wieder anzuknüpfen, vielleicht in der Überzeu- 119
gung, daß man einen Vorsatz nicht sicherer abstumpfen kann, als wenn man ihn 120
öfters durchspricht. (Goethe 2012 [1809]) 121

3. Translations for all examples are provided in the appendix.

Here, the character's decision to wear down the partner by repeatedly talking through the controversy is commented on by justifying it with a conviction — or at least the narrator strongly assumes this motivation with Charlotte, as is shown by *vielleicht* 'perhaps' which he uses to reflect his own conviction through this comment. This example also contains another type of Chatman's types of comment: generalisation. This is because that the generalisation of what the narrator is convinced of is presented as a universally valid truth.

Comment on the Discourse This type of comment, which we will call meta comment, expresses reflections on the process of writing and/or the existence of the respective work and its fictionality itself.

- (4) Ich verspräche gerne diesem Buche die Liebe der Deutschen. Aber ich fürchte, die einen werden es lesen, wie ein Kompendium, [...] indes die andern gar zu leicht es nehmen, und beede Teile verstehen es nicht. (Hölderlin 2012 [1797])

Meta comment has been extensively studied in other contexts as a category of its own (see for example Fludernik 2003 or Nünning 2005). This includes metanarrative comment and metafictional comment, the latter discussing truth, fictivity and/or fictionality of the respective work.

Given the heterogeneity of phenomena that have been subsumed under the concept of "comment" in narratological research, the question arises, how we may be able to annotate and automatically detect comments in texts. We address this in the next section.

3. Operationalisation

Concepts in Literary Studies including Narratology are often designed from a theoretical point of view and only selectively consider textual examples. Applying them on a larger scale often reveals incompleteness or discrepancies within the theory. Thus, making a literary phenomenon more tangible through annotation requires an iterative process of refinement of the concept utilising complete texts or longer parts of works instead of hand-picked examples (see Gius and Jacke 2017).

The starting point for our operationalisation of the comment are the findings from the previous section: Even if the category "literary comment" seems intuitively coherent and comprehensible, our examination of its conceptualisation revealed that comments are often defined *ex negativo* (see for example Bonheim 1975 or Prince 2003). Interestingly, instead of defining comment, researchers restrict themselves to create open lists of indicators or partial phenomena of comment. Thus, the state of the art seems to suggest that there is no robust concept of "comment", but rather a bunch of related phenomena, that have been subsumed under the overarching concept.

Combining the approaches of Bonheim and Chatman, we assume comment is present if a narrative pause is identifiable (Bonheim) and characteristic features of one of Chatman's comment types are present in it:

comment := narrative pause AND (interpretative passage OR attitude pas- 160
sage OR meta passage) 161

By this procedure we exclude blending of the modes to the extent that we do not include 162
comments if they appear linked to dynamic elements, but can thus achieve a higher 163
comparability of the collected data and lower the amount of interpretation required of 164
the annotators. 165

Let us first look at our approach of detecting a narrative pause. Since these readings are 166
widely unpredictable (section 2), we decided not to pre-determine sentence structures: 167
Our annotation relies completely on intuitive (i.e. form-independent) recognition of 168
narrative pauses.⁴ This procedure enables us to maintain the explorative character of 169
our narrative pause detection.⁵ 170

As described above, Chatman's types of comments are "generalisation", "interpreta- 171
tion", "judgment" (attitude) and "commentary on the discourse" (meta comment). In 172
contrast to his informal usage of the term, we understand "generalisation" as a linguistic 173
phenomenon triggered by e.g. generic terms and quantificational expressions. These 174
can co-occur with any of the subtypes (see e.g. (1) and (2)), but do not constitute 175
a subtype on their own. Since we examine generalisation as a separate category (cf. 176
Gödeke et al. to appear), three typological manifestations of comment emerge, (i) the 177
attitude comment, (ii) the interpretive comment, and (iii) the meta comment. 178

(i) Interpretive Comment The interpretive comment offers an interpretation of events 179
within the diegesis. Sometimes it takes the form of an explanation of events. This type 180
of comment can be recognised by the fact that additional information is provided that 181
re-perspectives, interprets or corrects elements of the plot or events within the diegesis. 182
As shown in (2), Clara's situation at the end of the story is interpreted by the narrator. 183

(ii) Attitude Comment In the attitude comment, an attitude of the speaker (narrator or 184
character) to the diegesis is expressed. By "attitude", we mean the way in which a 185
speaker views something or feels about something. This includes all objects of the 186
narrative, such as characters, the plot, fictional objects and the fictional world (order) as 187
well as self-references. In (3), presented above, the speaker's attitude towards Charlotte's 188
talking through the argument topic becomes clear. Here we have made significant 189
changes to Chatman's broad notion of this subtype of comment, which he calls *judgment* 190
and understands as evaluations being based on norms, values and beliefs of the narrator. 191
He uses this criterion as a demarcation to the comment type "interpretation" which he 192
takes to be "relatively value-free" (Chatman 1980, p. 237). Since the vagueness of this 193
criterion led to difficulties during the annotation we decided to annotate the speaker's 194

4. This procedure includes the understanding that a narrative pause can also occur in direct speech, which we understand as a narrative structure in itself. This allows us to include comments made by characters and not only those made by the narrator or so-called "authorial insertions" (Dawson 2016b).

5. In doing so, our approach differs from, for example, Vauth et al. 2021, who categorise verbal phrases by their eventness from non-event up to change of state.

attitude, as this is more clearly identifiable and the explicit result of the evaluation process. Therefore, we call the subtype "attitude comment" to make the difference clear.

(iii) Meta Comment The meta comment combines two aspects: metafictional and meta-narrative comment. It reveals the narrator's attitude toward the narrative, its process of creation (narrating) or its truth-status. Since its identification relies on direct mentions of the context and circumstances, in which the respective work of literature was created, we consider meta comment easier for the annotators to identify.

Based on the presented typology of comment, we created a tagset and annotation guidelines.⁶ Accordingly, the tagset for comment includes three subtags: INTERPRETATION, EINSTELLUNG (attitude), and META that correspond to (i), (ii), and (iii). The annotators are supposed to assign these subtags to passages, where a passage can comprise one or several clauses. These clauses usually follow one another, but discontinuous annotations are also possible. As for the narrative pause, we do not pre-select any linguistic properties as unique indicators for comment subtags, i.e. the annotation is solely based on a passage's reading and not its form. Comment is a phenomenon that tends to span rather long parts of text. One passage can be labelled with more than one comment subtags. Passages labelled with different subtags can overlap.

4. Corpus and Annotation

Our corpus consists of 19 texts covering the time period from 1616 to 1942. 17 texts serve as training set for the classifiers described in section 5. All six annotators are students with a background in German Philology. In general, the first approximately 200 sentences of each text were annotated by two annotators with the three subtags. Two texts were annotated by all six annotators in order to have a better insight into the feasibility of our approach. We created gold standards for all texts by having 2–3 experts (authors of this paper) collaboratively adjudicate (i.e. review, accept, correct or delete) the initial annotations. Table 1 shows for each text the annotated comment passages and the number of annotated clauses.⁸ Overall, we observe a median of 47 passages and 205 clauses for comments per text.⁹

To evaluate the annotation, we calculate inter-annotator agreement on clause-level with Fleiss' Kappa (κ , Fleiss 1971) and Mathet's Gamma (γ , Mathet, Widlöcher, and Métivier 2015). While κ calculates agreement based on the differences for each clause, γ respects the individual annotated comment passages as units in a continuum, and also partial overlapping passages are compared as units instead of disjointed clauses. We therefore consider that γ better represents the errors made by annotators for a category with rather

6. Our annotation guidelines are available at <https://gitlab.gwdg.de/mona/korpus-public>.

7. We do not show γ for the test texts since the Python package *Pygamma-agreement* ([https://github.com/bbootphon/pygamma-agreement](https://github.com/bootphon/pygamma-agreement)) used for calculation throws a runtime error for 6 annotators.

8. The number of comment passages and clauses for Goethe's *Die Wahlverwandtschaften* is higher since this was our first annotation, where we annotated the complete first four chapters.

9. We use the median rather than the average because the former is robust against outliers (i.e. texts with an extremely high or low number of comments), and thus better resembles the typical number of comments in a text.

Year	Author: Text	#		κ		γ	
		Pa.	Cl.	M.	B.	M.	B.
Training set							
1616	Andreae: <i>Die chymische Hochzeit</i>	47	66	.26	.29	.33	.37
1645	Zesen: <i>Adriatische Rosemund</i>	45	244	.71	.85	.83	.89
1668	Grimmelshausen: <i>Der abenteuerliche Simplicissimus</i>	53	205	.26	.26	.39	.40
1731	Schnabel: <i>Die Insel Felsenburg</i>	73	203	.74	.91	.82	.92
1747	Gellert: <i>Das Leben der schwedischen Gräfin von G.</i>	34	187	.61	.59	.64	.63
1771	LaRoche: <i>Geschichte des Fräuleins von Sternheim</i>	60	282	.33	.33	.42	.42
1797	Hölderlin: <i>Hyperion oder der Eremit in Griechenland</i>	72	313	.41	.76	.67	.86
1802	Novalis: <i>Die Lehrlinge zu Sais</i>	73	400	.61	.71	.76	.85
1809	Goethe: <i>Die Wahlverwandtschaften</i>	138	619	.34	.34	.48	.48
1810	Kleist: <i>Michael Kohlhaas</i>	36	72	.08	.09	.14	.16
1816	Hoffmann: <i>Der Sandmann</i>	37	103	.46	.46	.50	.50
1876	Dahn: <i>Kampf um Rom</i>	43	157	.28	.27	.35	.39
1893	May: <i>Winnetou II</i>	45	79	.55	.68	.64	.71
1898	Fontane: <i>Der Stechlin</i>	54	219	.31	.31	.42	.41
1924	Mann: <i>Der Zauberberg</i>	45	133	.41	.48	.54	.57
1930	Musil: <i>Der Mann ohne Eigenschaften</i>	47	317	.83	.83	.88	.88
1931	Kafka: <i>Der Bau</i>	55	280	.68	.68	.77	.77
		47	205	.46	.52	.56	.60
Test set							
1766	Wieland: <i>Geschichte des Agathon</i>	60	282	.58	.60	–	–
1942	Seghers: <i>Das siebte Kreuz</i>	48	92	.43	.48	–	–

Table 1: For each text, the number of comment passages (Pa.) in the gold standard and the number of clauses (Cl.) overlapping with them, and multi-label (M.) and binary (B.) agreement values in terms of κ and γ . The last row for the training set shows the median counts and the average agreement values.⁷

long passages, and that it measures agreement more adequately. The multi-label values 229 for both scores are based on the agreement between the subtags; binary agreement 230 treats all subtags as a single class (COMMENT). The average binary agreement for κ and γ 231 is moderate (between 0.41 and 0.60; see agreement levels in Landis and Koch 1977). The 232 multi-label agreement is 0.05 lower on average but can still be regarded as moderate. 233

As hypothesised in section 3, META is easier to annotate since it directly addresses 234 either the way content is mediated or the creation of the respective work. This can 235 be observed when calculating agreement scores directly on the individual subtags as 236 shown in Table 2. The subtag META achieves 0.71 (substantial) for κ and 0.81 (perfect) 237 for γ . In contrast, EINSTELLUNG holds moderate values and INTERPRETATION only achieves 238 a fair agreement (> 0.2). As pointed out above, especially the distinction between 239 EINSTELLUNG and INTERPRETATION can be difficult, and a decision for only one of both can 240 cause disagreement between the annotators. This effect can be verified when calculating 241 the binary agreement for EINSTELLUNG+INTERPRETATION, which yields a κ of 0.49 and γ of 242 0.57. 243

Subtag	κ	γ
EINSTELLUNG	.44	.53
INTERPRETATION	.26	.38
META	.71	.81

Table 2: Average agreement for subtags.

5. Automatic Classification

244

To gain insights into the consistency of the category “comment”, we employ diverse linguistically available features that we consider to be potentially relevant based on manual inspection of annotated comment passages. In the following, we describe the feature extraction, the classifiers and their evaluation, and then turn to a comprehensive analysis.

5.1. Feature Extraction

250

We preprocess texts with spaCy,¹⁰ using its default tokeniser, part-of-speech (POS) tagger, lemmatiser and sentenciser for German and adding several custom preprocessing components:

- a dictionary-based normaliser that we trained on the German Text Archive¹¹ to account for spelling variants in older texts
- the Universal Dependency parser, morphological analyser, clausiser and tense–mood–voice–modality tagger from Dönicke 2020
- a direct speech tagger that recognises text between opening and closing quotation marks
- a component that assigns Levin 1995’s categories to verbs and Hundsnurscher and Splett 1982’s categories to adjectives from GermaNet (cf. Hamp and Feldweg 1997)
- the sentiment tagger¹² from Remus, Quasthoff, and Heyer 2010 as well as our own emotion tagger based on the NRC Word-Emotion Associated Lexicon¹³ (Mohammad and Turney 2010; Mohammad and Turney 2013), which assign scores for positive/negative sentiment and Ekman 1992’s basic emotions, respectively, to each token

Inspired by Dönicke 2021’s grammatical feature extraction for discourse segmentation, we extract features clause-wise from the clause, its noun phrases (NPs), the composite verb and free discourse elements (i.e. conjunctions, complementisers, sentential

10. <https://spacy.io/> (version 2.3.2)

11. <https://www.deutschestextarchiv.de/download>

12. <https://github.com/Liebeck/spacy-sentiws>

13. <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>

14. We use word lists to classify overt quantifiers with Dönicke, Gödeke, and Varachkina 2021’s tagset; except for numerical quantifiers, which we identify by POS (NUM) and/or dependency relation (nummod).

Unit	Features
clause	root's dependency relation, root's POS, preceding/inner/succeeding punctuation, first clause of a sentence?, directed distance to superordinate clause, direct speech?
NP	head's dependency relation, head's POS, adpositional?, case, person, number, gender, sentiment, emotion, article's POS, article's lemma, quantifier's POS, quantifier's type ¹⁴ , adjective's POS, adjective's degree, adjective's GermaNet category, adjective's sentiment, adjective's emotion
(composite) verb	main verb's dependency relation, main verb's POS, verb form, tense, aspect, mood, voice, modal verb's lemma, main verb's GermaNet category, sentiment, emotion, quantifier's POS, quantifier's type ¹⁴
free discourse element	dependency relation, POS, at first/middle/last position?

Table 3: Extracted features for different syntactic units.

adverbs). Table 3 shows all features. Grammatical features have been found to work well for the identification of discourse segments—which are also a multi-clause-level phenomenon—in German (cf. Dönicke 2021) and might also include useful features for comment identification. For example, we expect verb categories such as tense or mood to be especially useful since a change in those often marks a narrative pause, as in (2). Here, the narrated time is interrupted, and the present tense in the first sentence changes to the past tense in the second one.

Punctuation is also integrated as feature. In (4), punctuation marks the direct speech, in which a comment is contained. We also integrate semantic categories for verbs and adjectives. Main verbs like *lesen* 'read' and *verstehen* 'understand' belong to Levin's category of cognition, which we assume to be indicative for comments.

Since comment, especially attitude, can be expressed in an emotional manner, we include emotion and sentiment labels as features. (5) shows an excerpt for *EINSTELLUNG* that is highly expressive due to the usage of so-called "thick concepts", such as *offen* 'expansive' and *wundersam* 'miraculous', which "combine evaluation and non-evaluative description" (Väyrynen 2021).

(5) Wer also ihr [der Natur] Gemüth recht kennen will, muß sie in der Gesellschaft der Dichter suchen, dort ist sie offen und ergießt ihr wundersames Herz EINSTELLUNG.
(Novalis 2012 [1802])

5.2. Classifiers

Since a comment passage spans an open number of clauses, we define a classification task on clause level: When vectorising a text $D = (c_1, \dots, c_n)$ with n clauses, we construct feature vectors $\vec{c}_1, \dots, \vec{c}_n$ as described in section 5.1, which we then concatenate to context-sensitive vectors $X_D = (\vec{x}_1, \dots, \vec{x}_n)$ using a window of three clauses: $\vec{x}_i := \vec{c}_{i-1} \circ \vec{c}_i \circ \vec{c}_{i+1}$. Given $\vec{x}_i$, the classifier should predict all tags of passages that contain c_i . In a post-processing step, every maximal sequence of clauses with the same tag is combined into

Setting	Development set	#Clauses
split 1	Grimmelshausen (1668), Schnabel (1731)	408
split 2	Mann (1924), Kafka (1931)	413
split 3	Gellert (1747), Fontane (1898)	405

Table 4: Texts in the development set and number of clauses overlapping with comment passages in each split.

#	Parameter name	Values
Decision tree		
1	maximum depth	5, 10, 15, 20, 25, ∞
2	min samples leaf	1, 2, 5, 10, 15, 20
Logistic regression		
1a	solver	newton-cg (ng), lbfgs (ls), sag (sg), saga (sa)
1b	multi class	multinomial (m), ovr (o)
2	C	.1, .5, 1, 5, 10

Table 5: Values for hyperparameters optimised in the grid search. The parameter number (#) and abbreviations in parenthesis are used in Table 6.

a passage, which is, however, not relevant for the evaluation, see section 5.3. 297

From our training set, we remove two texts as development set. To alleviate the impact 298
of the split, we perform our experiments for three different splits as shown in Table 4. 299
Since the median count of comment clauses per text is 205 (see Table 1), we take two 300
texts with a total number of comment clauses around 410 in each split. Furthermore, 301
split 1 uses two early texts, split 2 uses two late texts, and split 3 uses an earlier and a 302
later text as development set. 303

In each split, we train 1) a multi-output classifier that consists of three independent 304
binary classifiers (one for every subtag), and 2) a binary classifier that only distinguishes 305
comment (any subtags) from non-comment (no subtag). As base classifier we use 306
either 1) a decision tree or 2) a logistic regression, both with balanced class weights. 307
Since the performance of a decision tree strongly depends on its maximum depth and 308
minimum leaf size, we perform grid search on the development set to select the optimal 309
values for these parameters (see Table 5), using the same values for all base classifiers. 310
For the logistic regression, we optimise the solver and multi-class parameter, and the 311
regularisation parameter C.¹⁵ During the grid search, we use (macro-averaged) Fscore 312
(cf. Sokolova and Lapalme 2009) as scoring function, which we also use for evaluation. 313

Additionally, we combine the classifiers from the three splits into one majority classifier. 314
The majority classifier assigns those tags to a clause that are predicted by at least two of 315
the incorporated classifiers. 316

15. We set the maximum iterations of the logistic regression to 500. If not stated otherwise, we use scikit-learn's (<https://scikit-learn.org/stable/>) default parameters for our classifiers.

Setting	Multi								Binary							
	Development					Test			Development					Test		
	#1	#2	P	R	F	P	R	F	#1	#2	P	R	F	P	R	F
Decision tree																
split 1	5	20	.17	.72	.28	.28	.64	.39	25	1	.37	.54	.44	.45	.52	.49
split 2	5	2	.13	.73	.22	.27	.61	.38	∞	10	.35	.69	.46	.43	.54	.48
split 3	10	2	.17	.63	.27	.27	.58	.36	10	10	.38	.70	.49	.48	.63	.55
majority	–	–	–	–	–	.28	.62	.38	–	–	–	–	–	.50	.60	.54
Logistic regression																
split 1	sg/o	.1	.19	.49	.27	.30	.51	.37	ls/m	.1	.40	.57	.47	.53	.64	.58
split 2	sg/o	.1	.14	.57	.22	.29	.53	.37	sa/m	10	.37	.65	.47	.46	.57	.51
split 3	ng/o	.1	.20	.53	.28	.31	.49	.37	ng/o	.1	.44	.64	.52	.55	.65	.60
majority	–	–	–	–	–	.31	.51	.37	–	–	–	–	–	.54	.65	.59

Table 6: Macro-averaged P(recision), R(ecall) and F(score) on the development sets and the test set, for the multi-output and the binary classifier in all settings. Parameter values (#1 and #2, see Table 5) are given as optimised on the development set.

5.3. Evaluation

Table 6 shows Precision, Recall and Fscore for all settings. For the binary classifier, Precision measures how many of the clauses tagged as comment are also annotated as comment in the gold standard; Recall measures how many of the clauses annotated as comment in the gold standard are also tagged as comment. The Fscore is the harmonic mean of Precision and Recall. For the multi-output classifier, Precision, Recall and Fscore are calculated separately for each subtag first and then averaged.

Decision tree and logistic regression show similar results on both the development sets and the test set. The performance of both methods varies across splits, but the majority classifiers alleviate these discrepancies: In all but one setting, the majority classifiers achieve equal or better Fscores than the best of its incorporated classifiers.

Although the Fscores for decision tree and logistic regression are similar, Precision and Recall are not: The decision-tree classifier performs much better in terms of Recall at the cost of a lower Precision, whereas the difference between Precision and Recall is less extreme for the logistic-regression classifier.

Unsurprisingly, both methods achieve higher performance in the binary setting (54% and 59% for the majority classifiers) than in the multi-output setting (38% and 37% for the majority classifiers), where the classifiers have to distinguish subtags of comment.

5.4. Analysis

Somewhat surprisingly, every classifier performs better on the test set than on its development set. Part of an explanation might be that the test set includes more comment clauses than the development sets, see Table 7, and our classifiers are mainly driven by Recall. Table 7 also shows further differences between decision tree and logistic

Setting	EINSTELLUNG		INTERPRETATION		META	
	Development	Test	Development	Test	Development	Test
#Clauses						
split 1	173	201	154	252	172	280
split 2	171	--	216	--	73	--
split 3	258	--	147	--	31	--
Decision tree						
split 1	.24	.28	.25	.38	.34	.50
split 2	.24	.29	.29	.31	.14	.52
split 3	.39	.26	.25	.34	.16	.50
majority	–	.28	–	.35	–	.52
Logistic regression						
split 1	.26	.30	.26	.36	.31	.45
split 2	.24	.28	.29	.35	.13	.46
split 3	.39	.30	.28	.36	.17	.45
majority	–	.28	–	.36	–	.48

Table 7: Number of clauses and Fscores for each subtag on the development sets and the test set, for the multi-output classifier in all settings.

	EINSTELLUNG	INTERPRETATION	META
#Clauses	1887	2154	588
Fscore	.28	.36	.48
Ratio (#/%)	103	60	12

Table 8: Number of clauses in the training set (including the development texts) and Fscore of the logistic-regression majority classifier on the test set for each subtag. The bottom row shows the number of training clauses needed for one percentage point of Fscore.

regression: With a logistic regression, the Fscores on the test set for each subtag are comparatively stable across training/development splits, whereas the decision tree’s Fscores show a greater variance. The majority classifiers achieve performance close to the best individual classifiers for each subtag, resulting in Fscores of 28% for EINSTELLUNG, 35%–36% for INTERPRETATION and 48%–52% for META.

The comparatively high performance for META is outstanding, considering that META is the less frequent comment type in our data. In Table 8, we calculate for each subtag the average number of training clauses that contribute to one percentage point on the test set. We can see that the ratio is significantly lower for META (12) than for EINSTELLUNG (103), with INTERPRETATION inbetween them (60), which illustrates that META is much easier to learn by our classifiers than the other comment types.

Our binary classifier is considerably better than the multi-output classifier. In general, it is not unusual that a classifier performs better for a binary tagset than a more differentiated one. Still, since we observed in the agreement that annotators tend to disagree between EINSTELLUNG and INTERPRETATION while agreeing that a passage is one of both (see section 4), we trained an additional logistic-regression majority classifier

	EINSTELLUNG					INTERPRETATION					META				
#	±	i	Unit	Feature	Value	±	i	Unit	Feature	Value	±	i	Unit	Feature	Value
1	+	0	verb	mood	subj:past	+	0	verb	mood	subj:past	−	0	verb	tense	past
2	+	0	clause	speech	direct	−	−1	NP:nsubj	person	1per	−	0	clause	speech	direct
3	−	−1	clause	punct:inner	:	−	0	NP:obl	pos	PROP	+	1	verb	category	Kommunikation
4	+	0	NP:nsubj	quant:type	NEG	−	−1	NP:obl	pos	PROP	+	1	verb	mood	subj:past
5	−	0	NP:obl	pos	PROP	−	0	NP:nsubj	person	1per	−	1	verb	tense	past
6	+	1	verb	mood	subj:past	+	−1	verb	mood	subj:past	+	0	verb	tense	fut
7	+	0	NP:root	emotion	Trust	+	1	verb	mood	subj:past	−	1	clause	speech	direct
8	−	0	verb	mood	subj:pres	−	1	NP:obj	quant:pos	PRON	+	0	verb	mood	subj:past
9	−	−1	NP:nmod	case	acc	−	1	clause	punct:prec	»	−	−1	clause	speech	direct
10	+	0	NP:advmod	art:pos	DET	−	0	NP:nsubj	person	2per	−	0	NP:nsubj	gender	masc
11	−	1	NP:obj	quant:type	DIV	−	−1	NP:nsubj	person	2per	−	1	clause	punct:succ	!
12	+	0	clause	punct:succ	!	−	0	NP:obj	person	1per	+	−1	verb	tense	fut
13	+	1	NP:root	adj:category	Gefuehl	−	0	verb	mood	imp	+	−1	verb	modal	lassen
14	−	−1	clause	punct:prec	:	+	0	verb	dep	csubj	−	1	NP:nsubj	gender	masc
15	+	1	clause	pos	X	−	1	NP:nsubj	person	1per	+	1	clause	punct:inner	−
16	+	0	NP:root	emotion	Joy	+	0	verb	modal	scheinen	+	1	NP:root	emotion	Trust
17	−	0	verb	tense	past	+	0	NP:nmod	art:lemma	ein	+	1	NP:conj	art:lemma	mein
18	−	0	NP:obj	quant:type	DIV	+	1	verb	modal	scheinen	−	−1	NP:obl	emotion	Fear
19	+	−1	NP:ccomp	numerus	sing	−	0	NP:appos	gender	masc	+	−1	verb	modal	wollen
20	−	1	verb	mood	subj:pres	−	0	clause	punct:prec	«	+	0	verb	modal	wollen

Table 9: Top-20 features for each subtag ranked by absolute value of feature coefficient in logistic regression (split 1). ± is the sign of the coefficient. *i* denotes whether the feature is extracted from the preceding (-1), current (0) or succeeding (1) clause.

that regards EINSTELLUNG and INTERPRETATION as the same tag. (We left out all META 356 passages for this.) This classifier achieves an Fscore of 47% on the test set, which is 19% 357 higher than that for EINSTELLUNG and 11% higher than that for INTERPRETATION. Therefore, 358 we assume that the difficulty of differentiating between EINSTELLUNG and INTERPRETATION 359 applies for both humans and machine-learning methods, whereas a joint category is 360 easier to learn. 361

Table 9 exemplarily shows the most important features for one logistic-regression clas- 362 sifier.¹⁶ Positive features are indicative for a subtag whereas negative features are 363 indicative against a subtag. 364

Tense and mood/modality are learned to be relevant for all subtags. We have seen this 365 in (3), where tense and mood shift from present indicative to past subjunctive to express 366 a comment of type EINSTELLUNG. From the table, we can conclude that all three types 367 often occur in past subjunctive, accompanied by different modal verbs (e.g. *scheinen* 368 ‘seem’, *lassen* ‘let’, *wollen* ‘want’). 369

The comment types also differ in their presence within direct speech. While comments of 370 type EINSTELLUNG frequently occur in direct speech, comments of type META rather occur 371 outside direct speech. An explanation for this might be that utterances of characters 372 in direct speech qualify for EINSTELLUNG, whereas META is mostly produced by the 373 narrator. For INTERPRETATION, the speech feature is not important. Instead, it is learned 374 that comments of type INTERPRETATION do rarely occur after quotation marks (» and 375 «), which makes sense because they indicate a change of the speaker (from narrator to 376 character or vice versa) and an interpretative comment typically follows a statement by 377 the same speaker. 378

16. The most important features show only minor variations between splits.

As anticipated in section 5.1, Example (5), a striking characteristic for EINSTELLUNG is the high importance of features related to emotion (Trust, Joy, and the more general feature *Gefuehl* 'emotion'). For INTERPRETATION, we find that a subject in first person (*I*, *we*) or second person (*you*) is a negative indicator since only in third-person sentences something is told/interpreted about persons, incidents etc. For META comments, past tense is a negative feature. Instead, they often occur in grammatical future tense or with the modal verb *wollen* 'want', which can also express semantic future. This is illustrated in (6), where we can also see the typical use of *Kommunikation* 'communication' verbs, such as *erzählen* 'tell'.

(6) Ein Märchen will ich dir erzählen_{META}, horche wohl. (Novalis 2012 [1802])

In general, our classifiers tend to return many shorter comment passages, with interruptions between them, while we annotate longer passages in the gold standard. This is because we train the classifiers on the clause level, giving only three clauses as input, whereas human annotators can draw connections between clauses that are farther apart. We do not see this as a problem, as long as the relevant passages from a text are returned. (Example 7) compares the gold annotations (a) and the predictions by the logistic-regression majority classifier (b) for an excerpt from Wieland's *Geschichte des Agathon*. For sake of illustration, EINSTELLUNG is **boldfaced**, INTERPRETATION is *italicised* and META is underlined.

(7a) [...] so war es um so viel nötiger ihn auch dieser Probe zu unterwerfen, da Hippias, bekannter maßen, eine historische Person ist, und mit den übrigen Sophisten derselben Zeit sehr vieles zur Verderbnis der Sitten unter den Griechen beigetragen hat. [...]

(7b) [...] so war es um so viel nötiger ihn auch dieser Probe zu unterwerfen, da Hippias, **bekannter maßen**, *eine historische Person ist, und mit den übrigen Sophisten derselben Zeit sehr vieles zur Verderbnis der Sitten unter den Griechen beigetragen hat.* [...]

The excerpt is part of a long META passage in which the narrator reveals the conception of the main character Agathon and his confrontation with the sophist Hippias. The narrator outlines parts of the story, which can be seen as background knowledge that qualifies for an (overlapping) INTERPRETATION passage. Both passages span several (≥ 8) sentences in the gold standard. The classifier detects shorter passages instead. It correctly recognises large parts of the excerpt as META. This is remarkable since the comprehension of the META passage is tied to its long context and our clause-based classifier is able to detect important parts of it. It also identifies large parts as INTERPRETATION, but is missing the beginning and a short interruption. Lastly, it also identifies the EINSTELLUNG in the last part of the excerpt; as well as a short EINSTELLUNG passage which is not in the gold standard. The short passage is a good example for a false positive: It is probably labelled as EINSTELLUNG because it features the evaluative term *bekannter maßen* 'as is well known', but it does not express attitude towards the diegesis and is therefore not an EINSTELLUNG comment in the gold standard.

#	$\pm$	i	Unit	Feature	Value	Subtags (#)
1	+	0	verb	mood	subj:past	EINSTELLUNG (1), INTERPRETATION (1), META (8)
2	-	0	NP:obl	pos	PROPN	INTERPRETATION (3), EINSTELLUNG (5)
3	-	0	NP:nsubj	person	2per	INTERPRETATION (10)
4	-	1	NP:appos	case	nom	
5	-	-1	NP:obl	pos	PROPN	INTERPRETATION (4)
6	+	0	verb	dep	csubj	INTERPRETATION (14)
7	-	0	clause	punct:prec	«	INTERPRETATION (20)
8	+	1	verb	dep	csubj	
9	+	-1	NP:nsubj	emotion	Fear	META (18)*
10	-	0	clause	dep	flat	
11	-	0	verb	tense	past	META (1)
12	+	0	clause	speech	direct	EINSTELLUNG (2), META (2)*
13	+	1	clause	punct:inner	«	
14	+	0	verb	modal	scheinen	INTERPRETATION (16)
15	-	0	verb	mood	imp	INTERPRETATION (13)
16	+	1	verb	mood	subj:past	META (4), EINSTELLUNG (6), INTERPRETATION (7)
17	-	-1	NP:appos	case	nom	
18	-	-1	NP:nsubj	person	2per	INTERPRETATION (11)
19	+	0	NP:obl	quant:pos	PART	
20	+	-1	NP:nmod	art:lemma	dies	

Table 10: Top-20 features for the binary classification ranked by absolute value of feature coefficient in logistic regression (split 1). $\pm$ is the sign of the coefficient. i denotes whether the feature is extracted from the preceding (-1), current (0) or succeeding (1) clause. The last column shows the rank of the features if it appears among the most important features for the individual subtags in Table 9. A star (*) indicates that the feature has the opposite sign in the subtag’s base classifier.

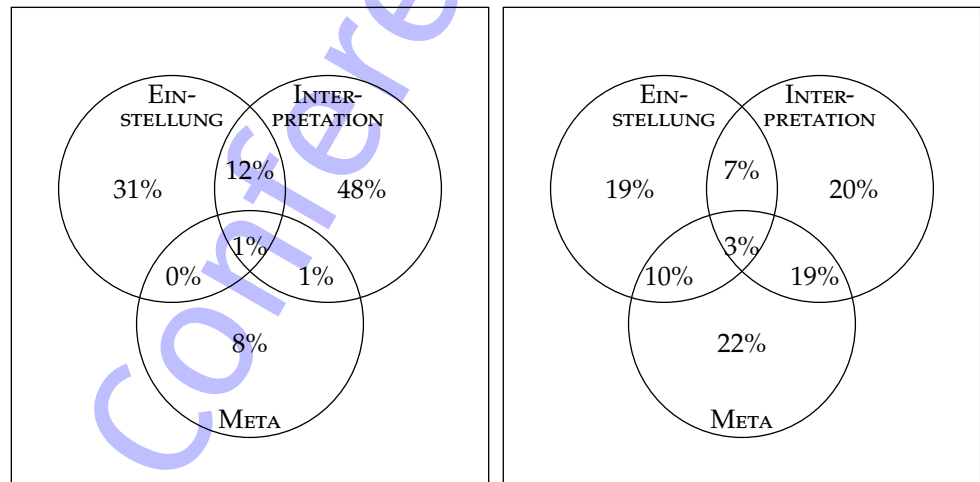


Figure 1: Overlap of comment clauses in the training data (left) and the test data (right).

Table 10 shows the most important features of the binary classifier. It mostly includes 420 important features for INTERPRETATION (see Table 9), which is the most frequent class 421 in the training data. It also includes some important features for EINSTELLUNG, but 422 important features for META are underrepresented, and there are even features with an 423 opposite sign to those for META. This suggests that META passages are not individually 424 learned by the binary classifier. This is not surprising when looking at Figure 1: Only 425 8% of all clauses in the training data are only annotated with META (other META clauses 426 overlap with another comment type). 427

6. Discussion

428

6.1. General Considerations

429

As announced in the introduction, we do not consider the attempt to recognise literary comments as an end in itself. Rather, we want to use this example to illustrate that attempts to operationalise and recognise literary phenomena automatically can shed light on the consistency of the concepts on which they are based.

When speaking of the “consistency of a concept”, we mean that a) comparatively homogeneous phenomena fall under it and b) that the concept is methodologically guiding in the sense that these phenomena are intersubjectively and automatically recognisable. Inconsistent concepts, on the other hand, describe comparatively heterogeneous phenomena, which cause hardly or not at all surmountable difficulties when trying to recognise them intersubjectively and/or automatically. Accordingly, “(in)consistency” is a gradual concept: a concept can be more or less (in)consistent as the phenomena that fall under it have more or less relevant commonalities.

In the following, when we try to judge whether there is (in)consistency of a theoretical concept based on our observations on theory, operationalisation, annotation, and detection of it, we implicitly or explicitly carry out inferences to the best explanation. Generally, inferences to the best explanation have the following structure (see Lipton 2005; Bartelborth 2017, pp. 200–291; here according to Descher 2019, p. 75):

P_1 : X is a fact that requires explanation.

P_2 : The hypothesis H_1 explains X .

P_3 : No competing hypotheses $H_2, H_3, \dots, H_n$ explain X better than H_1 .

C : So H_1 is probably true.

Due to premise P_3 , inferences to the best explanation are not truth-preserving, i.e., a true conclusion does not always follow from true premises. Even if one considers as many relevant alternative hypotheses as practically possible, one may simply miss a hypothesis that explains X better than H_1 . Thus, claims about the consistency or inconsistency of certain concepts based on results of annotation and automation should be understood as hypotheses to be tested by further research.

In the case of comment, two main facts seem to need explaining:

- X_1 : While two subtypes of comment (attitude comment and interpretative comment) can be annotated with little intersubjective agreement and detected with little success automatically, the opposite is true for meta comment.
- X_2 : Automatic detection of comment by the binary tagger works well, although there is (to our knowledge) no robust overarching literary definition of comment and 2 of the 3 comment-subtypes are poorly recognised.

In the following, we first discuss the facts X_1 , then X_2 .

6.2. The (In-)Consistency of each Comment Type

465

One of our most intriguing findings is that annotation and automatic detection struggle 466
with the same types of comment, although the annotation is based on a passage's reading 467
whereas the automatic detection is based on a passage's form. Interpretative comment 468
and attitude comment were annotated with only moderate and fair agreement and 469
their detection also performed poorly with FScores below 40% and 30%. In contrast, 470
meta comment achieves substantial agreement and can be detected well, with FScores 471
close to 50%. Taking into account the relation between available training data for each 472
comment-type and the performance of the multi-output-classifier, showed that meta 473
comment is much easier to learn than the other comment-types (see Table 8; META 474
performing better than INTERPRETATION/EINSTELLUNG by a factor of 5 and 8.5). Do these 475
surprising results suggest an inconsistency of the concepts "interpretive comment" and 476
"attitude comment" as we operationalised them based on several theories of literary 477
comment? 478

With respect to annotation, we suspect that mainly semantic properties (the occurrence 479
of terms such as "narrative," "truth," or "invented") of the meta comment passages are 480
responsible for their good agreement. The vague terms "attitude" and "interpretation", 481
on the other hand, made the development of precise annotation guidelines difficult. 482
What was contentious in the discussion of concrete annotations was not only at what 483
point something is an attitude or interpretation, but also where the difference between 484
the two lies. For clarification, let us recall example (3) from Fontane's *Der Stechlin*: 485

- (8) "Wir glauben doch alle mehr oder weniger an eine Auferstehung" (das heißt, er 486
persönlich glaubte eigentlich nicht daran), "und wenn ich dann oben ankomme 487
mit einer rechts und einer links, so ist das doch immer eine genierliche Sache." 488
(Fontane 2012 [1898]) 489

In this direct discourse passage the attitude of the main character Dubslav von Stechlin 490
to a second marriage becomes clear. As an argument he uses the assertion that everyone 491
more or less believes in the resurrection and the bad reputation of appearing with two 492
wives. In the brackets between the direct speech, however, the narrator formulates a 493
second attitude of Dubslav: he does not believe in the resurrection. Since this statement 494
is not made by Dubslav but the narrator, we annotate it as INTERPRETATION. Chatman 495
attempts to distinguish between these two types on the basis of the judgment/evaluation- 496
criterion. However, he leaves open at what point a statement is evaluative enough to 497
be considered an evaluative "judgment" rather than an interpretation. This turned out 498
to be problematic. For example, is the use of a term like "eagerly" sufficient to show 499
that the speaker has a positive or negative attitude towards someone? We found that 500
annotators, based on their reading impressions, answer such questions differently. 501

For automatic recognition, it is, among other things, the difference between interpretative 502
comment and attitude comment that causes problems. If we train a classifier that treats 503
EINSTELLUNG and INTERPRETATION as one tag (binary classification without META), we 504
obtain FScores that almost approximate the binary score (EINSTELLUNG + INTERPRETATION 505

+META). A problematic indicator of the attitude in both annotation and automatic recognition are "thick concepts" such as *eagerly* or *miraculous*, which "combine evaluation and non-evaluative description" (Väyrynen 2021).

If we exclude obvious alternative hypotheses such as unqualified annotators, inadequate machine-learning models, or errors in the statistical analysis of our classifiers,¹⁷ our findings suggest that (in contrast to meta comment) interpretive comment and attitude comment, as we have operationalised them, are *not* consistent concepts. The phenomena that fall under these two concepts are evidently too heterogeneous to be reliably recognised by humans and computers. Our findings on the automation side also suggest that there may be a consistent concept that encompasses all phenomena that fall under "attitude comment" or "interpretive comment". Defining this concept conclusively, without resorting (exclusively) to the vague terms "attitude" and "interpretation," would be a future task for Literary Studies.

6.3. The Inconsistency of the Generic Concept "Comment"

Although literary theory does not provide a consistent definition of the overarching concept of comment, our binary classifier (differentiating between comment and non-comment) achieves good results (FScores close to 60%). On the one hand, this is not very surprising because the binary classifier has a) more training data per category than the multi-label classifier and b) binary categorisation is less demanding. On the other hand, the classifier seems to accomplish the very thing that literary theory cannot provide (yet): a possibility to identify comment as a general phenomenon. What does this mean for a narratological concept of "comment"? We have already noted in section 3 that comment as a literary phenomenon is sometimes defined *ex negativo*. Therefore, many researchers refrain from defining comment and take an additive approach: Thus, "comment" is understood as a bunch of related phenomena (phenomenon1 OR phenomenon2 OR ...) whose commonalities are rarely discussed.

Our own approach takes a related route, by identifying three comment types that share narrative pause as common feature or prerequisite. Our proposal, having the following logical form: necessary feature AND (feature1 OR feature2 OR feature3), takes the form of what Fricke calls a "flexible definition" (Fricke 1981). However, we have seen that there is reason to believe that two of the criteria that our operationalisation of "comment" uses ("interpretative passage", "attitude passage") are themselves not consistent concepts (see section 6.2). Thus, the question arises whether the generic concept "comment" is a meaningful consistent literary category at all.

It is important to see that automatic detectability is no reliable indicator that there is an underlying consistent concept. Not everything computers can automatically recognise is based on a consistent concept. Suppose we define the concept "apple" as "being an

17. We exclude these alternative hypotheses as improbable on the basis that (i) our annotators have a sound background in German Philology and have considerable experience with annotating works of literature, (ii) employ comprehensive machine learning models, extracting a wide variety of features which range from structural to sentiment features and (iii) employ a well-tested machine learning suite.

apple or a table". This would be a very inconsistent concept because the phenomena that fall under it have little in common except that they are material objects. Nevertheless, one could undoubtedly build a supervised model that recognises "tapples"; it would most probably use the features of apples on the one hand and the features of tables on the other. Please note, that this only *prima facie* contradicts what has been said on inconsistent concepts above. The difficulty with automatic recognition would be that the model would be highly susceptible to bias due to unbalanced training data: If the majority of the training instances are tables, apples will probably not be detected at all, because they share no relevant commonalities with tables.¹⁸

So how does our binary classifier work? Our comparison of the most prominent features between the binary classifier (comment vs. non-comment) and the multi-label classifier (EINSTELLUNG, INTERPRETATION, META) yields an interesting result. 13 of the 20 most prominent features are features that also play a role for the recognition of the comment types (see Table 10). More importantly, only two of these 13 features are among the 20 most prominent features of all three comment types (subjunctive in the current or succeeding clause) and 9 features are indicative of one comment type only (according to the multi-label classifier). If the classifier had learned a general concept of comment, one would expect two kinds of features to dominate: features that are indicative of all three comment types and/or completely new features that played no role for the multi-label classifier. Therefore, our analysis suggests that the binary classifier, at least partly, uses feature combinations that are indicative of certain *types* of comment to recognise comment. The fact that 10 out of 20 most prominent features of the binary classifier are important features for interpretive comment (being the most common type in our training set, see Table 8), dovetails nicely with our expectation that a model that reflects a concept which is to a certain degree inconsistent is highly susceptible to bias due to unbalanced training data. Taken together our results can be regarded as evidence for comment being a rather inconsistent literary concept. The best explanation for the classifier not learning a general concept of comment is that the concept subsumes relatively heterogeneous phenomena, that share not enough relevant commonalities.

We have already underlined that our conclusions in the discussion section are ultimately hypotheses for which we have found some evidence, if we concede certain assumptions. There is one more background assumption that is relevant for our conclusion in this section of the discussion. Like many researchers in the Digital Humanities, we assume that literary phenomena manifest themselves at multiple levels (cf. Underwood 2019, p. 42), meaning that if there were a consistent narratological concept of "comment", it would be reflected in linguistically available features. This assumption, rarely made explicit, may be more justified for an essentially textual phenomenon as comment than for phenomena that include relational properties. Let us suppose this background assumption is justified, so that our results show that "comment" is a rather inconsistent concept. This would mean fundamentally re-examining the category of "comment" and asking whether the important phenomena worthy of investigation that it describes

18. As every analogy, our analogy has its limits. In particular, comment types can overlap, because of their textual extent, but apples and tables as material objects do not.

cannot be grouped differently and/or partially subsumed under other concepts such as "authorial intrusion" (Dawson 2016b), "digression" (Esselborn 1997–2003), "factual discourse"/"serious speech acts in fictional works" (Konrad 2017; Klauk 2015), "reflective passage" (Gittel to appear), or *Sentenz* ('aphorism', Reuvekamp 1997–2003). At least this procedure seems appropriate to us, assuming that literary concepts should be also suitable for quantitative research nowadays.

7. Conclusion

Andrew Piper noted, that we "do not have a clear picture of how emerging quantitative methods speak to the questions that matter within the discipline of Literary Studies." (Piper 2018, p. 10) The present paper addressed this issue by investigating the extent to which inferences about the consistency or inconsistency of textual literary concepts can be drawn from attempts at annotation and automation. Concretely, we operationalised the literary concept of "comment" and phenomena associated with it: attitude passages, interpretative passages and meta passages. We annotated a corpus and trained classifiers for the automatic recognition of comment and its subphenomena. We were able to show that the concepts of the subphenomena vary in consistency. While meta comments are readily identifiable, clear overlaps emerge between interpretative and attitude comment. We also discussed the extent to which comment in and of itself can be understood as a consistent concept or as a catch-all for rather heterogeneous phenomena and found evidence in favor of the second assumption. We thus illustrated one way in which digital methods can contribute to humanities research in general and to a better understanding of "comment" as a literary concept in particular. We not only examined an important literary phenomenon more closely and made it identifiable, we also addressed the question of why concepts such as the "literary comment" are sometimes difficult to operationalise, investigating how far the success or failure of operationalisation and automation can help exploring their consistency.

8. Appendix: Translations of Examples

- (1') She [Otilie] was introduced to the gentlemen, and was at once treated with especial courtesy as a visitor. Beauty is a welcome guest everywhere. (J. W. v. Goethe 19–?)
- (2') When Nathanael lay on the stone pavement with a shattered head, Coppelius had disappeared in the crush and confusion. Several years afterwards it was reported that, outside the door of a pretty country house in a remote district, Clara had been seen sitting hand in hand with a pleasant gentleman, while two bright boys were playing at her feet. From this it may be concluded that she eventually found that quiet domestic happiness which her cheerful, blithesome character required, and which Nathanael, with his tempest-tossed soul, could never have been able to give her. (E. Hoffmann 1885)

- (3') The next day, as they were walking to the same spot, Charlotte took the opportunity of bringing back the conversation to the subject, perhaps because she knew that there is no surer way of rooting out any plan or purpose than by often talking it over. (J. W. v. Goethe 19-?)
- (4') I'd happily promise this book the love of the Germans. But I fear some will read it like a compendium and be overly concerned with the fabula docet, whilst others will take it too lightly, and neither party will understand it. (Hölderlin 2019)
- (5') Whosoever wills to be well acquainted with her [the Nature's] Soul must seek her company with the Poet, for to him she is expansive and pours out her miraculous heart EINSTELLUNG. (Novalis 1903)
- (6') I will tell thee a tale META. Listen! (Novalis 1903)
- (7') [...] so it was all the more necessary to subject him also to this test, since Hippias, as is well known, is a historical person, and, with the other sophists of the same time, contributed very much to the corruption of morals among the Greeks. [...]
- (8') Happy days awaited him there, the happiest of his life. But they were of brief duration; the very next year his wife died. Taking another was not for him, in part because of a sense of order and in part for aesthetic considerations. "After all," he maintained, "we all believe more or less in a resurrection (which is to say he personally really did not), and if I put in an appearance up there with one woman on my right and another on my left, well, that's always sort of an embarrassing business." (T. Fontane 1995)

9. Data availability 644

Data can be found here: <https://doi.org/10.5281/zenodo.6467062>. 645

10. Software availability 646

Software can be found here: <https://doi.org/10.5281/zenodo.6466328>. 647

11. Author contributions 648

Anna Mareike Weimer: Data curation, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing 649

Florian Barth: Project administration, Data curation, Formal analysis, Resources, Software, Writing – original draft, Writing – review & editing 651

Tillmann Döncke: Data curation, Formal analysis, Resources, Software, Validation, Writing – original draft, Writing – review & editing 653

Luisa Gödeke: Data curation, Methodology, Writing – original draft, Writing – review & editing 655

Hanna Varachkina: Data curation, Investigation, Resources, Writing – original draft, Writing – review & editing 657

Anke Holler: Funding acquisition, Supervision, Writing – review & editing 659

Caroline Sporleder: Funding acquisition, Methodology, Supervision, Writing – review & editing 660

Benjamin Gittel: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing 662

References 664

- Bartelborth, Thomas (2017). *Die erkenntnistheoretischen Grundlagen induktiven Schließens. Induktion, Falsifikation, Signifikanztests, kausales Schließen, Abduktion, HD-Bestätigung, Bayesianismus*. Routledge (International library of philosophy). URL: <https://nbn-resolving.org/urn:nbn:de:bsz:15-qucosa-220168>. 665
- Bonheim, Helmut (1975). "Theory of narrative modes". In: *Semiotica* 14.4, pp. 329–344. 669
- Booth, Wayne C. (1983). *The Rhetoric of Fiction*. Chicago: The University of Chicago Press. 670
- Chatman, Seymour Benjamin (1980). *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell paperbacks. Cornell University Press. 671
- Dawson, Paul (2016a). "From Digressions to Intrusions: Authorial Commentary in the Novel". In: *Studies in the Novel* 2, pp. 145–167. 673

- Dawson, Paul (2016b). "From Digressions to Intrusions: Authorial Commentary in the Novel". In: *Studies in the Novel* 48.2, pp. 145–167. 675
676
- Descher Stefan und Petraschka, Thomas (2019). *Argumentieren in der Literaturwissenschaft. Eine Einführung*. Ditzingen: Reclam. 677
678
- Dönicke, Tillmann (Oct. 2020). "Clause-Level Tense, Mood, Voice and Modality Tagging for German". In: *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*. Düsseldorf, Germany: Association for Computational Linguistics, pp. 1–17. DOI: [10.18653/v1/2020.tlt-1.1](https://doi.org/10.18653/v1/2020.tlt-1.1). URL: <https://aclanthology.org/2020.tlt-1.1>. 679
680
681
682
683
- (Nov. 2021). "Delexicalised Multilingual Discourse Segmentation for DISRPT 2021 and Tense, Mood, Voice and Modality Tagging for 11 Languages". In: *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 33–45. URL: <https://aclanthology.org/2021.disrpt-1.4>. 684
685
686
687
688
- Dönicke, Tillmann, Luisa Gödeke, and Hanna Varachkina (2021). "Annotating Quantified Phenomena in Complex Sentence Structures Using the Example of Generalising Statements in Literary Texts". In: *17th Joint ACL-ISO Workshop on Interoperable Semantic*, p. 20. 689
690
691
692
- Ekman, Paul (1992). "An argument for basic emotions". In: *Cognition & emotion* 6.3-4, pp. 169–200. 693
694
- Esselborn, Hartmut (1997–2003). "Digression". In: *Reallexikon der deutschen Literaturwissenschaft*. Ed. by Harald Fricke, Klaus Grubmüller, Jan-Dirk Müller, and Klaus Weimar. Vol. 1. Berlin [u.a.]: De Gruyter, pp. 363–364. ISBN: 9783110156645. 695
696
697
- Fleiss, Joseph L (1971). "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5, p. 378. 698
699
- Fludernik, Monika (2003). "Metanarrative and Metafictional Commentary: From Metadiscursivity to Metanarration and Metafiction". In: *Poetica* 35, pp. 1–39. 700
701
- Fontane, Theodor (1995). *The Stechlin*. translated by William L. Zwiebel. Germ Series. Camden House. ISBN: 9781571130242. URL: <https://books.google.de/books?id=7yh0ZJjNgxoC>. 702
703
704
- (2012 [1898]). "Der Stechlin". In: *TextGrid Repository*. Digitale Bibliothek. URL: <https://hdl.handle.net/11858/00-1734-0000-0002-AECD-2>. 705
706
- Fricke, Harald (1981). *Norm und Abweichung. Eine Philosophie der Literatur*. München: Beck. 707
708
- Genette, Gérard (1994 [1972]). *Diskurs der Erzählung [Discours du récit]*. München, pp. 9–191. 709
710
- Gittel, Benjamin (to appear). "Reflexive Passagen in fiktionaler Literatur. Überlegungen zu ihrer Identifikation und Funktion am Beispiel von Wielands "Geschichte des Agathon" und "Goethes Wahlverwandtschaften"". In: *Euphorion* 116, pp. 1–16. 711
712
713
- Gius, Evelyn and Janina Jacke (2015). "Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse". In: *Zeitschrift für digitale Geisteswissenschaften* 1. 714
715
- (2017). "The Hermeneutic Profit of Annotation. On Preventing and Fostering Disagreement in Literary Analysis". In: *International Journal of Humanities and Arts Computing* 11, pp. 233–254. 716
717
718

- Gödeke, Luisa, Florian Barth, Tillmann Döncke, Anna Mareike Weimer, Hanna Varachkina, Benjamin Gittel, Anke Holler, and Caroline Sporleder (to appear). "Generalisierungen als literarisches Phänomen. Charakterisierung, Annotation und automatische Erkennung". In: *Zeitschrift für digitale Geisteswissenschaften*. 719-722
- Goethe, Johann Wolfgang von (2012 [1809]). "Die Wahlverwandtschaften". In: *TextGrid Repository*. Digitale Bibliothek. URL: <https://hdl.handle.net/11858/00-1734-0000-0006-6A93-D>. 723-725
- (19-?). *Elective affinities : a novel*. URL: <https://archive.org/details/electiveaffinities00goetuoft/page/68/mode/2up>. 726-727
- Hamp, Birgit and Helmut Feldweg (1997). "Germanet-a lexical-semantic net for german". In: *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. 728-730
- Herrmann, Julia Berenike, Katrin Van Dalen-Oskam, and Christoph Schöch (2015). "Revisiting Style, a Key Concept in Literary Studies". In: *Journal of Literary Theory* 9, pp. 25-52. 731-733
- Hoffmann, E. T. A. (2012 [1816/17]). "Der Sandmann". In: *TextGrid Repository*. Digitale Bibliothek. URL: <https://hdl.handle.net/11858/00-1734-0000-0003-6A92-6>. 734-735
- (1885). *The Sand-Man*, translated by J.Y. Bealby. New York: Charles Scribner's Sons. 736
- Hölderlin, Friedrich (2019). *Hyperion, or the Hermit in Greece*. Ed. by Howard Gaskill. Open Book Publishers. ISBN: 978-1-78374-655-2. URL: <https://www.openbookpublishers.com/product/941>. 737-739
- (2012 [1797]). "Hyperion oder der Eremit in Griechenland". In: *TextGrid Repository*. Digitale Bibliothek. URL: <https://hdl.handle.net/11858/00-1734-0000-0003-7CC8-A>. 740-742
- Hundsnurscher, Franz and Jochen Splett (1982). *Semantik der Adjektive des Deutschen. Analyse der semantischen Relationen*. 743-744
- Jockers, Matthew Lee (2013). *Macroanalysis. Digital Methods and Literary History*. Urbana, Chicago, and Springfield: University of Illinois Press. 745-746
- Klauk, Tobias (2015). "Serious Speech Acts in Fictional Works". In: *Author and Narrator: Transdisciplinary Contributions to a Narratological Debate*. Ed. by Dorothee Birke and Tilmann Köppe. De Gruyter, pp. 187-212. URL: <https://doi.org/10.1515/9783110348552.187>. 747-750
- Konrad, Eva-Maria (2017). "Signpost of Factuality: On Genuine Assertions in Fictional Literature". In: *Art and Belief*. Ed. by Ema Sullivan-Bissett, Helen Bradley, and Paul Noordhof. London: Oxford University Press, pp. 42-62. ISBN: 9780198805403. 751-753
- Lahn, Silke and Jan Christoph Meister (2013). *Einführung in die Erzähltextanalyse*. Stuttgart und Weimar: Metzler. 754-755
- Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, pp. 159-174. 756-757
- Levin, Beth (1995). "English verb classes and alternations". In: *A preliminary Investigation* 1. 758-759
- Lipton, Peter (2005). *Inference to the Best Explanation*. London: Routledge (International library of philosophy). 760-761

- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier (2015). "The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment". In: *Computational Linguistics* 41.3, pp. 437–479.
- Mohammad, Saif and Peter Turney (June 2010). "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon". In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles, CA: Association for Computational Linguistics, pp. 26–34. URL: <https://aclanthology.org/W10-0204>.
- (2013). "Crowdsourcing a word–emotion association lexicon". In: *Computational intelligence* 29.3, pp. 436–465.
- Novalis (1903). *The Disciples at Sais and Other Fragments*, translated by F.V.M.T. and U.C.B., introduction by Una Birch. London: Methen & Co. URL: <https://archive.org/details/disciplesatsais00nova/>.
- (2012 [1802]). "Die Lehrlinge zu Sais". In: *TextGrid Repository*. Digitale Bibliothek. URL: <https://hdl.handle.net/11858/00-1734-0000-0004-6129-B>.
- Nünning, Ansgar (2005). "On Metanarrative: Towards a Definition, a Typology and an Outline of the Functions of Metanarrative Commentary". In: *The Dynamics of Narrative Form*. Ed. by John Pier. Berlin: De Gruyter, pp. 11–58.
- Piper, Andrew (2018). *Enumerations: Data and Literary Study*. Chicago and London: University of Chicago Press. ISBN: 9780226568898. URL: https://books.google.de/books?id=0_FLDwAAQBAJ.
- Piper, Andrew, Sunyam Bagga, Laura Monteiro, Andrew Yang, Marie Labrosse, and Yu Lu Liu (2021). "Detecting Narrativity Across Long Time Scales". In: *Proceedings http://ceur-ws.org ISSN 1613, p. 0073*.
- Prince, Gerald (2003). "Commentary". In: *A Dictionary of Narratology*. Loncoln and London: University of Nebraska Press, p. 1980.
- Remus, Robert, Uwe Quasthoff, and Gerhard Heyer (2010). "SentiWS – a Publicly Available German-language Resource for Sentiment Analysis". In: *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, pp. 1168–1171.
- Reuvekamp, Silvia (1997–2003). "Sentenz". In: *Reallexikon der deutschen Literaturwissenschaft*. Ed. by Harald Fricke, Klaus Grubmüller, Jan-Dirk Müller, and Klaus Weimar. Vol. 3. Berlin [u.a.]: De Gruyter, pp. 425–427. ISBN: 9783110156645.
- Sokolova, Marina and Guy Lapalme (2009). "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4, pp. 427–437. ISSN: 0306-4573. DOI: <https://doi.org/10.1016/j.ipm.2009.03.002>. URL: <https://www.sciencedirect.com/science/article/pii/S0306457309000259>.
- Underwood, Ted (2016). "The Life Cycles of Genres". In: *Journal of Cultural Analytics*.
- (2019). *Distant Horizons. Digital Evidence and Literary Change*. Chicago and London: The University of Chicago Press.
- Vauth, M., H. Hatzel, Gius E., and C. Biemann (2021). "Automated Event Annotation in Literary Texts". In: *Accepted for: Computational Humanities Research (CHR)*.

- Väyrynen, Pekka (2021). "Thick Ethical Concepts". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University. 804
805
806
- Willand, Marcus, Evelyn Gius, and Nils Reiter (2020). "SANTA: Idee und Durchführung". In: *Reflektierte Algorithmische Textanalyse*. De Gruyter, pp. 391–422. 807
808
- Zeller, Rosmarie (1997). "Erzählerkommentar". In: *Reallexikon der deutschen Literaturwissenschaft*. Band I. Berlin/New York: DeGruyter, pp. 505–506. 809
810

Conference

Validating Topic Modeling as a method of analyzing sujet and theme

Julian Schröter  ¹

Keli Du  ²

1. Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, University of Würzburg, Würzburg.

2. Trier Center for Digital Humanities, University of Trier, Trier.

Keywords:

sujet, theme, validation, topic modeling, content

Licenses:

This article is licensed under:



Abstract. In Computational Literary Studies (CLS), several procedures for thematic analysis have been adapted from NLP and Computer Science. Among these procedures, topic modeling is the most prominent and popular technique. We maintain, however, that this procedure is used only in the context of exploration up to date, but not in the context of justification. When we seek to prove assumptions concerning the correlation between genres, methods of computational text analysis have to be set up in research environments of justification, i.e. in environments of hypothesis testing. We provide a holistic model of validation and conceptual disambiguation of the notion of aboutness as sujet, fabula, and theme, and discuss essential methodological requirements for hypothesis-based analysis. As we maintain that validation has to be performed for individual tasks respectively, we shall perform empirical validation of topic modeling based on a new corpus of German novellas and comprehensive annotations and draw hypothetical generalizations on the applicability of topic modeling for analyzing aboutness in the domain of narrative fiction.

1. Introduction

1

Determining what literary texts are about is an essential part of interpreting literary texts and is also fundamental to investigating literary history. In Jockers 2013, which has been one of the most controversially received monographs in the last decade in computational literary studies (CLS), Jockers starts with a comprehensive and pretheoretical notion of theme, which is subsequently explored using topic modeling. Topic modeling is currently the most prominent tool for investigating aspects of aboutness in CLS. As it is based on unsupervised machine learning, topic modeling does not depend on our assumptions with regard to themes in texts. Hence, topic modeling has become a popular tool for exploring corpora. In several contexts, this tool has also been used in classification tasks for testing concrete hypotheses on genres or other text categories (e.g. Schöch 2017). The central claim of our paper is that topic modeling is still lacking justification to be used for hypothesis-driven research on specific aboutness claims in the domain of literary studies. Although this criticism on topic modeling is not new (e.g. Da 2019, Shadrova

2021), it has not yet been taken as a reason to overcome the desideratum. The task of this paper is to elaborate on this thesis and to prepare the methodological framework for solving this desideratum.

This desideratum affects the specific kind of interpretation that is at work when a concrete topic, which consists of a list of weighted words, is interpreted as, for example, a topic of "female fashion" in Jockers and Mimno 2013, or of "love as a challenge and a reward" in Schöch 2017. The use of topic modeling relies - at least implicitly - on the following three axioms in order to interpret lists of weighted words as genuine representations of aboutness:

1) A pre-theoretical notion has to be introduced to denote what topic modeling is expected to reveal in terms of humanities research. Our initial observation that Jockers 2013 starts from a general notion of 'theme' can be reversed. Theme is commonly considered to be the qualitative correlate of computationally generated topics. This holds also for Blei 2012, Jockers and Mimno 2013, Weitin and Herget 2017, and Schöch 2017. Hence, we take the linkage between the notion of theme and topic modeling to be the current state in CLS.

2) A specific theory of the structure of topics has to be developed. The formalized concept of topic in topic modeling can be outlined as follows: the core of topic modeling, Latent Dirichlet Allocation (LDA), comes from computational linguistics. It is a generative model and describes a fictional process in which a document is generated. It is based on the assumption that a text is a mixture of different topics with different probabilities, where each topic represents a probability distribution over a fixed set of words. A word can belong to one or several topics with certain probabilities. To generate a document, a probability distribution over topics is chosen randomly. Then, a topic is randomly chosen from all the topics and a word is randomly assigned to it. Thus, a single word of the document is determined. This process is then repeated until the document is finally generated (Blei 2012). LDA topic modeling in practice can then be understood as the inverse of the above described generative process. Given a text collection, the unseen topic-word distribution and the topic-document distribution are to be inferred by topic modeling (Blei and Lafferty 2009). There is often a semantic relationship between words that occur together in texts. These words are more likely to be grouped into one topic through topic modeling. Therefore, topics, which have in effect the form of lists of weighted words, are supposed to be interpretable as themes and to reflect the hidden content structure of the text collection.

3) A general theory is needed that justifies that themes are properly represented as lists of weighted words (topics), whose distribution in the text is similar. The best candidate of such a general theory seems to be distributional semantics, which holds that meaning consists of distributions of words (Harris 1954, Firth 1957, Evert 2005).

Based on these three steps, topic modeling is expected to return representations of 'theme' in a genuine sense of aboutness. However, our central claim that topic modeling lacks justification so far entails that topic modeling does not represent the genuine sense

of aboutness in literary studies. In other words, the predicate "interpret a topic as a topic of..." is commonly used only in a loose sense, which means that the reader is reminded of a specific aboutness claim when reading a topic and expresses a subjective impression. If we want to use topic modeling for hypothesis-driven research on specific aboutness claims, the predicate should be used in a stricter sense that treats topics as an exact representation of specific aboutness claims. Section two of our paper elaborates and justifies our central claim. If we assume, for the moment, that our claim is correct, then topic modeling is, at best, an approximation to aboutness under certain conditions. It is an approximation to aboutness if it can be substantiated with a more refined validation strategy. In general, the call for more validation is characteristic of CLS (Swafford 2015, Piper 2015, and Hammond 2017). Such call for validation points to a methodological gap that arises when methods from domains such as statistics or computer linguistics are transferred to CLS. This gap can be described as the ignorance of the equivalence of two procedures. For aboutness, it is the ignorance of whether topic modeling detects themes in a way that is equivalent to the human practice of determining the respective themes based on reading. This ignorance of equivalence has two dimensions: firstly, the internal dimension of the operative structure of the procedure itself, and, secondly, the external dimension of the results (Hammond 2017). For the former, the claim of ignorance means that there is no evidence that a quantitative procedure performs the same operative steps as human minds do. With regard to the second dimension, there is the problem that we do not know whether the results are equivalent because the results have different forms. In other words, the ignorance consists of the problem that the output of both procedures are incommensurable. To bridge the methodological gap we shall propose a ternary model of operationalization and validation, which is visualized in Figure 1. This model is more comprehensive and, as we shall demonstrate, more powerful than the established binary conceptions of validation. In this way, our contribution fundamentally differs from the general criticism of topic modeling as it has been put forward in recent criticism (Shadrova 2021) that rejects topic modeling based on the claim that the concept of topic in topic modeling would have to be identical to the concept of topic in the domain where the procedure shall be used (in our case, the concept of aboutness in the domain of literary studies), and on the finding that there is no such conceptual identity, i.e. no co-extensionality (identity of references) and co-intensionality (identity of definitions) between both (i.e. the concepts of topic and aboutness). We maintain that Shadrova's requirement is far too strong. It is true but also obvious that the notion of topic in topic modeling is not co-extensional and co-intensional to the concept of aboutness in literary studies. We rather seek to develop a strategy of applying topic modeling in a hypothesis-based design that allows to investigate aboutness independently of the notion of topic. We maintain that the following model facilitates such kind of hypothesis-based analysis.

The figure shows three units (qualitative concept, annotated texts, and quantitative procedure) with three binary relations between each two of these units. These three relations, one between the quantitative procedure and the intension (i.e. the definition) of the qualitative concept, another between the qualitative concept and the annotated

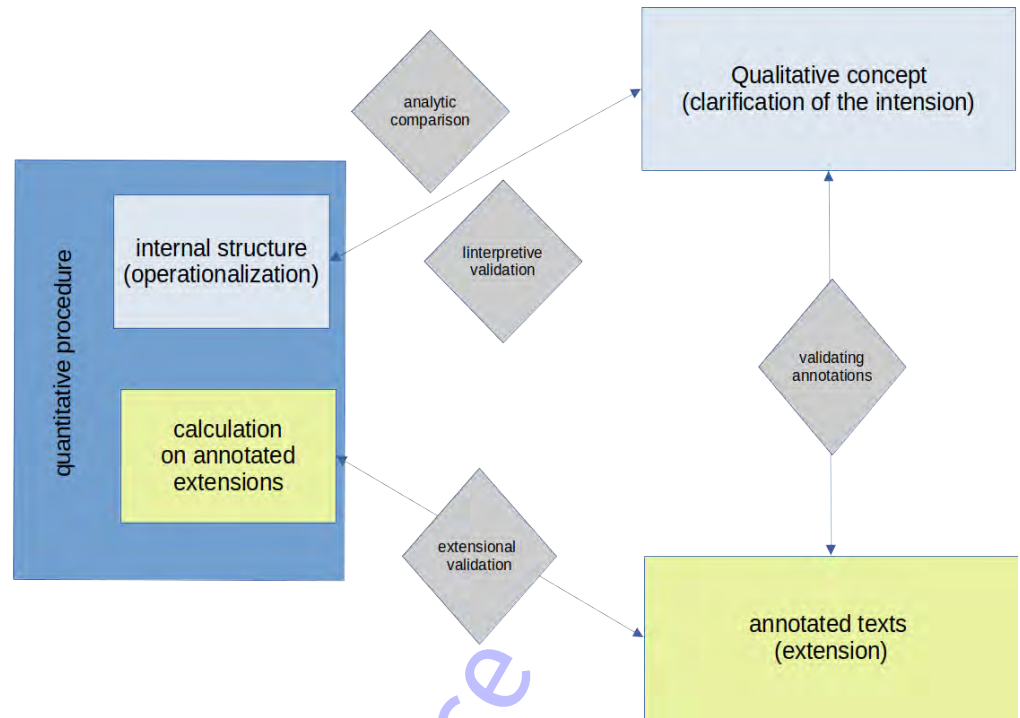


Figure 1: Ternary model of validation

texts, and the final relation between the results of the quantitative procedure and the 99
extension (i.e. the scope of objects the concept refers to) of the annotated texts, mark 100
the locations where different kinds of validation are required. So far, discussion on 101
validation usually has limited itself to one of these three relations respectively.¹ We 102
maintain that a full understanding of the impact of topic modeling as a technique of 103
analyzing aboutness in the context of hypothesis-driven research (and not only in that 104
of exploring corpora) necessitates that all three relations be modeled and validated. In 105
the following sections, we shall demonstrate the general methodological requirements for 106
ternary validation by discussing the three relations successively. Our methodological 107
discussion will be empirically supported and illustrated by a new and large corpus of so 108
far unknown 19th-century German novellas.² 109

2. Disambiguation and internal validation 110

We first discuss the relation between the intension of the qualitative concept and the 111
quantitative procedure, which is, according to the first axiom, the relation between 112
aboutness and the internal structure of topic modeling. The theoretical reason why we 113
consider this relationship to be problematic is that we question the third axiom of the 114
adequacy of distributional semantics. In other words, the following disambiguation shall 115
demonstrate that topic modeling does not exactly reproduce aboutness in the way the 116

1. Swafford 2015 focuses on the relation between the intension of the procedure and the qualitative concept, Piper 2015 on that between concept and annotations.
2. For a description of the corpus see the data repository.

concept of aboutness is used in literary studies. We do not contest that distributional semantics can be an appropriate and satisfactory theory within specific domains of linguistics, in particular for scenarios focusing on word similarity and synonymy or concerning usefulness in the context of information retrieval. From the perspective of literary studies, however, the distributional idea of semantics does not suffice to define the notion of aboutness because it can take different forms. We, therefore, have to think in scenarios of aboutness-claims. For this purpose, literary theory provides helpful terminological distinctions.

2.1. Conceptual clarification: aboutness as sujet, fabula, and theme

The list of notions that are often used synonymously to indicate aboutness could be extended with 'subject', 'subject matter', or, in more specific contexts, 'issue', or 'problem'. Concerning the general grammatical structure, aboutness occurs as about-p-assertions such as 'this novel is about love'. Two terminological distinctions from literary theory are relevant in the first instance, that between subject and theme (Lamarque 2009), and that between sujet/syuzhet and fabula in the tradition of Russian Formalism (introduced by Tomaševskij [1931] 1985), which has been translated to the distinction between story and plot in narratology. We take the latter distinction as a specification of Lamarque's notion of subject so that we can focus on three terms: fabula, sujet, and theme. Tomaševskij defines fabula as the temporal and causal sequence of events. In large parts, this notion corresponds to that of Lamarque's idea of subject: "To say what a work is about at subject level is in effect to retell the story or, in the case of non-narrative works, to redescribe the occasion or emotion presented." (Lamarque 2009, 150) In short, fabula is the plot-based aspect of aboutness. In contrast to fabula, both, Lamarque and Tomaševskij, define theme as the rather abstract unity of a literary work. This unity is, in most cases, not obvious but a result of interpretation. Sujet, which is a widely but heterogeneously used term in literary studies, is defined by Tomaševskij as the way the fabula is presented on the level of discourse including not only digressions, analepses and prolepses, as it has been emphasized in narratology, but also the setting (the place and situation of the fabula), the time, and the way characters are described (and, for example, dressed) and so on. In his illustrative analyses, Tomaševskij uses sujet to denote those aspects of the setting and surrounding that are not part of the fabula itself. In Aristotelian terms, sujet can in practice be used as the sum of the accidentia of the fabula.

We discuss the operationalizability of theme, fabula, and sujet based on the following illustrative extraction of several claims and interpretative hypotheses from different discursive contexts on one of the most canonical novellas of the period of Realism in 19th-century German literature, Keller's *Romeo und Julia auf dem Dorfe* (1855/75):

- (a. love-1) The novella "treats the theme of love and death" (Saul 2003, 138).
- (b. love-2) The novella is about the tragic conflict between ideal, absolute, and unconditional love in contrast to social constraints (Kaiser 1971, 30).

(c. love-3) The novella is about the problematic concept of love itself that has been internalized by the protagonists (Holub 1985, 476).	157 158
(d. love-4) The novella is about structural incest in terms of Freud's psychoanalytic theory (Holub 1985, 481).	159 160
(e. sujet) The novella is an instance of the set of texts that are located in a rural surrounding (Stocker 2007, 72), it takes place "in an isolated rural 'Dorfgeschichte' location" (Saul 2003, 133).	161 162 163
(f. social-1) The novella is about a devastating destiny caused by a violation of ownership (Menninghaus 1982 according to Walter Benjamin)	164 165
(g. social-2) Based on the symbolic meaning of the character of the black fiddler, the message of the novella is that "in all members of the community [...] is an inner Gypsy, in all those secure in their unreflected homely identity lies hidden the exotic other" (Saul 2003, 139).	166 167 168 169
(h. structure) The aesthetic value of the novella results from reflexivity on semiotic processes and intertextuality, which is a step from realism to aestheticism. (Stocker 2007, 69-75, Saul 2003).	170 171 172
The first claim (a) is an aggregation of fabula that is extended in the subsequent claims on the theme of love to a more complex structural thematic claim. All but (e) and (h) are aboutness claims. The latter does not point to the theme but the semiotic structure of the text. The contrast between the love claims (b to c), the psychoanalytic thesis (d), and the claims on social issues (f and g) shows that thematic claims are often controversial, sometimes absurd, and, in all cases, the result of intensive interpretive work. The claim on sujet (e) is a description of the text regarding general literary forms. As there is a tendency in literary studies towards giving interpretations of theme a higher prestige than analyzing sujet, ³ we shall address the possible objection that claims on sujet are not aboutness claims in a proper sense. It seems to be clear that Keller's novella is a tragic love story but not a village story. This objection implies that aboutness relates to theme or fabula, but not to sujet. It is, however, also true that the novella is about love in a rural setting. Hence, sujet can be part of aboutness claims. Such claims have the logical structure 'x is about p in setting s'. As p refers to fabula or theme in claims of that type, theme, fabula, and subject can be nested. Our illustrative example at the end of this paper demonstrates that sujet can be significant to literary history, too.	173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189

2.2. Comparing procedure and conceptual intension 190

The first relation that has to be validated requires the operationalization of a technical procedure that promises to approximate the conceptually clarified notion of aboutness.	191 192
---	------------

3. Lamarque 2009, who highlights the relevance of eternal and universal themes for assessing literary value, is representative of the tendency in literary studies to regard thematic interpretation as the more prestigious task compared to analyzing sujet and fabula.

We distinguish three steps of operationalization, (1) that of selecting a promising quantitative technique or method, (2) that of adjusting factors that could impact the output of the selected procedure, which includes not only the parameters of the algorithm but also operations such as preprocessing textual data, and, if topic modeling is used, (3) that of selecting promising candidate topics. The subsequent fourth step is commonly labeled as 'internal validation' (Hammond 2017). It would be equally correct to label this type as 'intensional validation' because the internal structure of a quantitative procedure is compared to the intension of a qualitative concept from literary studies.

We start discussing internal validation concerning sujet:⁴ Several sujets such as surrounding, furnishing, or dressing, that are denoted by a limited set of descriptive terms or named entities, can be expected to be expressed satisfactorily by lists of weighted words. Romanesque environment, which is relevant to German novellas, can be expected to be approximated by words including named entities of cities or regions.⁵ Another relevant sujet, that of a 'rural surrounding', can be expected to be expressed by nouns that denote typical buildings or the specific social structure in villages, or nouns and verbs that express or refer to typical activities such as agriculture. Prior to validation, the degree of strength between a specific word and sujet should be taken into account in terms of a theory of meaning. Of course, the occurrence of words is neither sufficient nor necessary for any sujet in a strict sense because lists of weighted words are not the proper representation of sujet but rather an approximation. Named entities, however, which are proper names in contrast to general terms,⁶ are almost inevitable for an author if a story shall be located in a certain setting. It is hardly possible to tell a story that takes place in Paris without referring to the name 'Paris' or to entities that clearly refer to places, buildings, well-known events, or prominent historical persons in Paris⁷. This strong relationship between named entities and sujet, which can be expected for a Romanesque setting does not hold for the sujet of a rural surrounding because it has to be approximated by general terms rather than named entities. Therefore, a heuristic distinction between sujets that shall be approximated mostly by singular terms and sujets that shall be approximated by general terms is useful for estimations prior to validation. Such estimation will also instruct the process of operationalization and of preprocessing because it requires that named entities are not removed from the corpus. According to (Tomaevskij [1931] 1985, 220), local or dynamic sujet, which is present only in particular scenes of a story, can be distinguished from global or static sujet that is prevalent over the whole text. The former requires that the texts be split up into segments. Prior to validation, we can assume that topic modeling performs best for

4. Existing research occasionally interpreted concrete topics as indications of sujet (Schöch 2017), but did not yet provide a theoretical account of the relationship between topic modeling and sujet.

5. 'Romanesque environment' means that it is fictional that the story is located either in France, Italy, or Spain.

6. This distinction can be traced back to Frege 1892.

7. We can, of course, think of counterexamples. The story of Flaubert's *Madame Bovary*, for example, is not located in Paris but the main character Emma often thinks of Paris and longs for living there. The novel has 75 hits for Yonville, the village where the action takes place, 74 hits for Rouen, the town that serves Emma as a replacement for her desire for Paris, and 34 hits for Paris. Of course, it would be mistaken to infer that the story is situated for 20% in Paris and, respectively 40% in Rouen and Yonville, as the absolute word counts suggest. It is nevertheless true that the novel is, in part, about a female protagonist's thoughts about Paris.

stereotypical and homogenous global sujets that are approximated by named entities. 228
The more local or dynamic and the more abstract and heterogeneous a sujet, the smaller 229
the chances of success and the higher the efforts for parameter adjustment and for text 230
manipulation in the process of preprocessing. 231

The third step is that of selecting the *prima facie* best topics after generating a topic 232
model. This step is necessary because of two restrictions: Firstly, the previous paragraph 233
demonstrated that only several sujets can be expected to be approximated by topic 234
modeling. Secondly, not all topics are good candidates for approximating specific sujets⁸. 235
Fortunately, topic modeling is capable of returning several promising village topics for 236
our 19th-century novella corpus. The most promising candidate (topic no. 64, see 237
code repository) starts with the nouns 'Dorf' (village), 'Haus' (home), 'Mann' (man), 238
'Knecht' (servant), 'Leute' (people), 'Feld' (field), 'Wald' (forest), 'Wagen' (carriage), 239
'Pferd' (horse), 'Bauer' (peasant), 'Stall' (barn), 'Arbeit' (labor). These words may 240
create the impression of a good approximation to the sujet of a rural surrounding. For 241
the sujet of a Romanesque surrounding, however, we were not able to identify any 242
promising candidate topic. The occurrence of the names of cities, regions, or other 243
entities that refer to French, Italian, or Spanish surroundings is not distributed with 244
sufficient frequency and density in the text. In place of topic modeling, we developed 245
another method of generating lists of semantically related words by manually drawing 246
up a list of expected words such as 'France', 'Italian', or 'Naples', and determining the 50 247
nearest vectors to each of the words in the initial list, based on a SpaCy language model. 248
Then, we summed up all nearest vectors for all words and selected the 30 most frequent 249
words, which yield the final embedding-based list. Then, for all texts in the corpus, we 250
calculated the relative share of this list by counting all lemmatized words of the text 251
that are in the respective list and dividing by the sum of all word tokens in the text.⁹ 252
Although both the village topic as well as the embedding-based list can be expected to 253
be competing for approximations to specific sujets, no reliable insight is gained unless 254
the techniques are validated also with regard to the remaining two relations. 255

With fabula, things are more complicated than with sujet. As fabula is defined as 256
the causal progression of events, it implies a change in situation. In the case of love 257
stories, events of falling in love are followed by a threat to the love relationship, and, 258
finally, either by the elimination of the threat or of the failure of love. As for sujet, the 259
proper representation of fabula is not a list of words, but rather a summary. Recently, 260
more advanced methods of automatic summarization have been developed. "Automatic 261
summarization seeks to present given information in a more compact form, determining 262
the key messages of the text and eliminating unnecessary details and filler sentences." 263
(Alexandr et al. 2021) The earlier approaches are mostly focused on extracting key 264
sentences or passages as the summary of a document (Neto, Freitas, and Kaestner 2002, 265
Ribeiro et al. 2013). Such approaches have improved thanks to the recent development 266

8. This latter limitation, that a considerable number of topics in each topic model does not approximate semantic content but rather condenses rhetorical, stylistic expressions or verbs of communications, etc., is well reflected in all studies on topic modeling and expressed by the distinction between interpretable and non-interpretable topics.

9. Code and the resulting lists are documented and explained in the code repository.

of deep-learning-based pre-trained language models. By identifying the key concepts and entities in the source document, automatic summarization combines the word-embedding-based representation of the input document and other linguistic features such as part-of-speech and named-entity tags (Nallapati et al. 2016). For its automatic evaluation, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) has been suggested in Lin 2004. The idea is to count the overlapping textual units between the generated summary and a set of gold reference summaries. For the human evaluation of automatic summarization, Kryciski et al. 2019 suggested that the summaries should be evaluated from four perspectives: Coherence, Consistency, Fluency, and Relevance. Based on this instruction, automatically generated summaries are rated by human annotators on a Likert scale.

Such methods of summarization can be expected to be better approximations to fabula than topics. As this paper focuses on the scope of topic modeling, we can ask, nonetheless, whether several plot structures have semantic consistency over the whole text irrespective of situative changes during the progression of events. Although topics and other types of word lists are not proper representations of fabula, there can be pragmatic reasons for using word lists as rough approximations to static kinds of fabula such as crime, love, Western, or seafaring stories¹⁰. As this holds only for several plot structures, the rationale for this consistency has to be reflected in terms of semantic theory: In many love stories, the aspect of love can be expected to be present globally over the whole text. For love stories, it is not the setting but rather the mode of communication and its characteristic forms of address that justify prior assumptions of semantic unity on the level of word lists. For stories about seafaring, western (Jannidis, Konle, and Leinen 2019, 169), and several other highly stereotypical plots, in turn, it is not the plot structure itself that is represented in the topic, but rather the global sujet of a surrounding that is strongly connected to the plot. According to the terminological disambiguation we introduced in this section, it would be more appropriate to say that there are several text types such as Western or Seafaring stories that are characterized by a specific plot structure as well as by a specific global sujet. In such cases, topic modeling does not identify fabula but rather sujet, which are, however, connected to fabula in the case of specific genres.

For theme, things are even more complicated than with fabula and sujet. Our illustration of interpretative claims on Keller's novella shows that several abstract concepts can serve as an abbreviation either for a typical plot structure or for thematic theses, where two operations can be observed: In our example, the core concept of love is integrated into the structural claim that there is a conflict between love and another abstract entity. Moreover, claims (b) and (c) indicate that one of several different general ideas of love is actualized in the text: in (b) that of radical and absolute romantic love, in (c), in contrast, that of not sufficiently radical love. The scope of both claims can only be understood properly if competing concepts of love are held present in the horizon of expectation. We refer to one of the most advanced theories of semantic change,

10. The latter is an example in (Jockers 2013, 125). In our topic model, there is a highly conspicuous seafaring-topic (no. 98), too.

Luhmann's *Liebe als Passion* (Luhmann 1982), which distinguishes (1) idealized love in medieval culture (*fin amour*), which is based on ratio based idolization mediating the difference between animalistic sexuality and sublime love, (2) paradoxical passionate love based on the idea of kurtosis and excess (*amour passion*), (3) love as friendship, (4) romantic and radically individualized love that is not concentrated on the character of the beloved, but on self-referential love itself, (5) the trivialization and ideology of reproduction where love as passion and romantic love appear as a problem that is transformed towards comradeship so that love becomes a matter of matrimonial viability mediating between the individual and social restraints.

Schöch 2017 in his study on the correlation between topic and genre identifies three different love topics that correlate with different dramatic sub-genres. When he notes that "each of the 'love' topics actually represents quite a different perspective on the theme of love", he interprets these candidate topics as representations of different abstract ideas of love, for example, "love as challenge and reward". Based on our terminological disambiguation, we can see more clearly that this exploratory strategy of interpreting topics starting from the resulting word lists is ambiguous. It may be the case that different love topics indeed approximate different abstract ideas of love. It is, however, also possible, according to the correlation between topic and genre that Schöch verifies, that all love topics refer to the same abstract idea of love but rather indicate different courses of fabula: One topic may include words that refer to a tragic ending whereas another refers to a happy ending. It is likewise possible that different love topics refer to different *sujets* such as different surroundings (for example, love in a rural versus urban milieu). Different topic word lists that are semantically related to love do not provide any information as to whether that topic approximates different concepts of love or different *sujets* or fabula aspects. For the general semantic relation between word lists and aboutness, we assume the following relation: the stronger the process of abstraction from fabula and *sujet* to theme and the more complex the propositional structure of thematic claims, the smaller the chances of success that thematic claims can be represented in lists of weighted words. Hence, we should not expect topic modeling to reveal thematic claims.

2.3. Interpretive or intensional validation

If topic modeling shall be applied in the context of testing hypotheses regarding the presence of specific *sujets* or concepts at the core of themes, a further step of interpretive validation after operationalization is common practice (e.g. Rhody 2014, Navarro-Colorado 2018). We illustrate this strategy by adapting it to the case of rural surroundings in correlation with different candidate topics. According to current evaluation strategies (Newman et al. 2010, Mimno et al. 2011, Aletras and Stevenson 2013), topics can be manually evaluated through a questionnaire. Table 1 shows in an illustrative manner the first lines of such a questionnaire for three candidate topics of a rural surrounding. A common scenario for the application of interpretive validation in Digital Humanities is that people acquire a rough knowledge of several texts of an object area with a rough idea of typical, *sujets*, fabulae, and themes (in our example the knowledge of novellas and the

idea that some novellas are about love, some are situated in a rural surrounding, etc.). 350
 Each row in the questionnaire contains the 20 most frequent words of the respective 351
 topic. One task is to identify all words that do not belong to the respective sujet, fabula, 352
 or theme. The other task is to decide whether the respective topic words approximate 353
 the annotator's qualitative notion of the respective sujet, fabula, or theme. 354

id	topic words	words that do not belong to village topic	interpretable as village topic
28	alt tag alte hoch gut kapitel bamme rot beginnen seidentopf groß hohen-vietz herz nehmen bild tür jung dorf schritt hohen-vietzer	alt, tag, alte, gut, kapitel, ...	No
64	dorf haus mann hof knecht leute schloß kommen rufen feld sagen wald förster wagen stehen pferd bauer sehen stall arbeit	haus, schloß, rufen, sagen, stehen, sehen	Yes
38	dorf mühle hand ameile fränz schauen haus welt marann furchenbauer mund sagen gehen bauer frau vater hof stube munde bruder	ameile, fränz, schauen, ...	Yes

Table 1: Questionnaire of manual evaluation of topics

Two coefficients can be calculated from this type of questionnaire: Firstly, a ranking of 355
 words that are most often expected for village topics across all evaluated topics and all 356
 annotators, secondly, the average number of the minimum of words that must belong 357
 to a topic of a specific sujet, fabula, or theme can be determined. In this way, an 358
 empirical link between topics and qualitative concepts on the level of intension can be 359
 achieved. We have to concede here that such validation is much more complicated for 360
 more complicated sujets or concepts of love. For different ideas of love, expected words 361
 have to be articulated in advance. For fin amour in Luhmann's terms, descriptions of 362
 perfection, and expressions of admiration have to be expected in combinations with 363
 articulations of being in love. For amour passion, descriptions and expressions of passion 364
 as well as of feigning love are to be expected, for romantic love the singularity of the 365
 love itself, and for love as companionship nouns that express or denote friendship and 366
 descriptions of the reality of matrimonial and family live. 367

Irrespective of the practical difficulties for more complex sujets and themes, there are, 368
 however, to our mind, critical shortcomings of this strategy if it shall be transferred to the 369
 domain of literary studies: The presented type of evaluation has been developed within 370
 and for computational linguistics according to its proper needs: "For our purposes, the 371
 usefulness of a topic can be thought of as whether one could imagine using the topic in a 372
 search interface to retrieve documents about a particular subject" (Newman et al. 2010). 373
 This particular strategy has then been adapted to the specific domain of information 374
 retrieval and relies on a rather restricted idea of the usefulness of topics. In the domain of 375
 information retrieval, this strategy may be appropriate. In the realm of literary studies, 376
 however, readers are more likely to adjust their expectations concerning aboutness to 377

the presented lists of weighted topic words in a way that departs from the way they would estimate the presence or absence of specific sujets or themes if they were not confronted with topic word lists. Although the presented type of interpretive validation seems to be promising, it does not guarantee that the validated topics are actually about the respective sujet or theme, which is identified by close reading without looking at the results of quantitative procedures. Therefore, external validation is necessary.

3. Validating annotations

The relation between the intension of a qualitative concept (such as amour passion) and the practice of identifying and annotating the presence of that concept in literary texts has to be clarified in an intermediate step. This clarification is not part of the quantitative procedures and of operationalization itself. In many scenarios, however, CLS cannot dispense with this dimension of validation (Schröter et al. 2021) and there is the possibility of validating this relation. There is, however, further need for a more systematic assessment of the methodologically controversial aspect of this type of validation. It is not entirely consensual how aboutness is represented in terms of reader response. Readers' judgments with regard to the aboutness of literary works are, as Piper 2015 points out, subjective in general and often arbitrary or idiosyncratic. In such cases, there is, in statistical terms, high variance and low agreement between readers, which cannot be ignored as normal noise. As all people have different positions in the world,¹¹ Piper 2015 rightly stresses the a priori subjective character of readers' judgments. If, however, judgments were completely arbitrary, reader response would be the expression of totally private feelings but not a response to texts as existing objects. From a pragmatic point of view, there are always fields of more consensual descriptions and there are domains of wider spread and lower inter-annotator agreement. Therefore, two further aspects have to be introduced. Firstly, the distinction between the psychological and the hermeneutic side of reader response. Secondly, the scaling from the intensional subjectivity of single annotations to extensional intersubjectivity.

For the first aspect, the dimensions of epistemic genesis and epistemic validity have to be distinguished. Concerning validity, aboutness is relevant either as a mental representation in concrete readers or as an objective property of a text as an entity. With regard to epistemic genesis, in contrast, aboutness is measured based either on empirical reader-response analysis or expert judgement or technical procedures. This dual distinction of validity and genesis is represented in table 2, which records proponents and opponents of the possible positions.

Both objectivism and perspectivism are legitimate frames for different research interests. However, objectivist interests necessitate reasonable and regulated annotations, whereas perspectival interest makes sense only based on perspectival data. Perspectival data

11. This is what Davidson 2001, 39, calls the rational and unproblematic form of relativism in contrast to conceptual and epistemological relativism.

12. We do not distinguish between the currently dominating nominalist version and the outdated perspective based on a realism of universals, Stegmüller 1969, XXI.

genesis validity	empirical reader-response study	hermeneutic reasoning	technical procedure
insight into the object itself ¹² (objectivism)	Mellmann and Willand 2013 as proponents; rejected as 'psychologism' by Frege 2021 and Husserl [1900] 2009.	Lamarque 2009	Carnap 1950 (cf. Schröter et al. 2021)
insight into a perspective on objects (perspectivism)	Piper 2015	relativist or constructionist professional reading, Barthes 1971	Underwood 2019

Table 2: Modeling the difference between epistemic genesis and validity

can, for example, be extracted from contemporary reception documents such as reviews, 415
articles, diaries, or letters for historical cultures, or from annotations, interviews, or 416
surveys for present cultures. 417

Concerning the second aspect, that of transforming subjective and intensional reader 418
response to extensional and intersubjective judgments, things are different for the 419
relationship between objectivism and perspectivism. For both, it will be essential 420
to calculate the spread of inter-annotator agreement in order to assess the degree of 421
intersubjective consensus versus subjective arbitrariness. Under an objectivist interest, 422
the spread of inter-annotator agreement is a strong benchmark of validity of annotations. 423
Low agreement between annotators is problematic because it shows that the intension 424
of the concept that shall be annotated has either not been sufficiently clarified prior 425
to the task of annotating or that it is not clear in itself. Hence, a high spread should 426
lead to revising the intension and the rules of annotation. If inter-agreement cannot be 427
achieved, external validation will not be possible. 428

For perspectival modeling, in contrast, a low agreement between historical agents 429
indicates that the concept was not well defined in contemporary culture. In the specific 430
design of perspectival modeling (Underwood 2019), validating the historical perspective 431
concerning intensions is not necessary. It is, in general, not necessary if the meaning of 432
the historical perspectives does not need to be articulated in analytical terms of literary 433
studies. Validation is necessary, in contrast, if a historical practice or a quantitative 434
procedure or both shall be expressed in terms of literary studies. This is the case for 435
interpreting topic modeling as an approximation to sujet, fabula, and theme. 436

For operationalizing sujet, fabula, and theme as properties of texts and not as historical 437
perspectives based on topic modeling, an objectivist design is necessary. For sujet 438
and fabula, a higher inter-annotator agreement can be expected than for theme, which 439
highly depends on abstraction and imports of external theories (such as psychoanalytical 440
theory in the thematic claim d or historical materialism in claim f of section 2.1). For 441
abstract ideas such as different concepts of love within structurally complex thematic 442
claims, a sufficiently high agreement between annotators will require extensive training 443

on foundational theories. For the validation study that we present in the final section, the sujets of a rural surrounding and Romanesque environment as well as the idea of romantic love were disambiguated, in case of the latter concept according to Luhmann (see section 2.2), and transferred into rules for annotating about 100 novellas.¹³

4. External or extensional validation

The final and most important relation that has to be validated is that between the extension of the qualitative concept and the extension of the quantitative procedure. Hence, we shall refer to this type, which is sometimes called external validation in linguistics (Gries 2008, 427), as extensional validation. Based on annotations (or, in case of perspectival modeling, on reader-response analysis of reception evidence) as described in the preceding section, the extension of texts with a specific sujet, fabula, or theme in qualitative terms has to be provided and compared to the results of the quantitative procedure. There is an important restriction to this type of validation. As Shadrova 2021 points out, the results of this type of validation cannot be generalized. This is certainly true with regard to the inductive structure of empirical inference in general. In our case, the results for extensional validation of the quantitative procedure for operationalizing a specific sujet, for example Romanesque setting, cannot be generalized for the relationship between topic modeling and all sujets. Shadrova, however, over-emphasizes this restriction. We maintain that it is possible to articulate systematic hypotheses on generalizability based on specific empirical validations. Such hypotheses have to be proved in subsequent case studies. Hence, we shall present an example for an extensional validation and discuss possible generalizations in the conclusion of this paper. Our case study is based on our novella corpus and. Its results are recorded in table 3. The disambiguated qualitative concept of the respective sujet or theme is recorded in the first column, its translation into samples based on annotations, according to the process of transforming intensions to extensions as elaborated in section 3, is recorded in the second column.

A methodological issue arises as we have to relate a categorical variable (presence or absence of a sujet, fabula, or theme) with a metric value of quantitative procedures. Accordingly, there are two options: The weaker and easier option is to calculate the share of the respective word list for the contrary groups based on annotation. According to the distribution (mean and standard deviation) of the dominance of words of the list in both contrary samples, a T-Test (here Welch's t-test for samples with different variance) is calculated. Its t-statistic and p-value are recorded in the third and fourth column for scaled data. This first option is applicable in contexts of weak comparative hypotheses. The stronger the difference for the share between the contrary samples, the higher the probability that a high value for individual texts indicates that a text has the respective sujet, theme, or fabula recorded in the first column. The second option is more demanding and it is required in contexts, where the quantitative results, which are in their very structure metric, can be interpreted categorically in a way that a threshold

13. The results are stored in the data folder in the code repository.

qualitative sujet, fabula, or theme	Annotated sam- ples (size)	quantitative approxima- tion	t-test, t- statistic	t-test, p- value	classifica- tion (LR), accuracy score
rural sur- rounding	'located in a vil- lage' (46) ver- sus urban milieu (56)	topic no. 64	1.899	0.061	0.511
		topic no. 38	1.233	0.222	0.404
		topic no. 28	-0.556	0.580	0.399
		list of words, based on em- bedding	2.962	0.004	0.616
Romanesque setting	'located either in Spain, France, or Italy' (25) ver- sus 'located else- where' (78)	list of words based on em- bedding	5.542	5.448e-7	0.786
romantic love	a story featuring romantic love (82) versus stories that are not love stories (36)	topic no. 36	-0.587	0.559	0.401
		topic no. 47	2.951	0.004	0.627
		topic no. 34	3.211	0.002	0.628
		list of words based on em- bedding	3.871	2.107e-4	0.636

Table 3: Extensional evaluation of rural surrounding, Romanesque setting, and romantic love

facilitates classifying texts as having a specific sujet, fabula, or theme. For our examples, 484 we performed a classification task with the metric value of the topic share or the word 485 list share as the independent predictor variable and the qualitative sujet, fabula, or 486 theme as the dependent predicted variable based on a logistic regression algorithm, with 487 cross-validation and a custom-made bootstrapping method with 10,000 iterations of 488 resampling, training, and calculating the accuracy scores for predications on a validation 489 set. For each sample of contrary subsamples of the same size, with the larger subsample 490 reduced to the size of the smaller subsample randomly, 80% of the documents were used 491 for training and the remaining 20% for validation. The final column records the accuracy 492 scores for predictions on the validation set. For comparison, we conducted a simple 493 bag-of-words based classification to set a baseline. The classification for annotated sets 494 of rural surrounding, Romanesque environment, romantic love are 0.401, 0.400 and 0.540 495 for the 5000 most frequent and tf-idf normalized words as features, respectively.¹⁴ 496

For the statistical significance of the hypothesis that both samples are from different 497 populations (which means that texts with a specific sujet, fabula, or theme are different 498 from texts without that sujet, fabula, or theme) as well as for the results of the 499 classification task, we see that the candidate topics selected from our topic model 500 performs better than the baseline of classifying annotated samples based on a document 501 term matrix of the 5000 most frequent tf-idf normalized word types but worse than our 502 generated word lists based on word embedding. In a future study, we shall address the 503 methodological ground for such embedding-based lists. With regard to our theoretical 504

14. All details of the significance test and the classification are documented in the code repository.

discussion in section 2, we can understand why the Romanesque setting based on a list generated by word embedding has the best performance and why no candidate topic word for this sujet could be generated. Words that indicate Romanesque surroundings do not appear with sufficient frequency and equal dispersion in the texts concerning topic modeling. If such words (for example, named entities of cities and regions) appear in a text, however, these words are highly specific to and indicative of a Romanesque setting. Also for romantic love, the embedding-based word lists outperform topic modeling. For rural surroundings, the best candidate topic has the same performance as the embedding-based word list. If two sufficiently large annotated validation samples were available, a more refined strategy would be advisable. The first sample could be used as a test set in a grid search for optimizing parameters such as the total amount of topics, length of chunks, and hyperparameters of the algorithm itself. According to the results of the grid search, candidate topics with the best performance in the discussed classification task can be identified. With the second set as a validation sample, the optimized topic model could be validated as discussed in this section.

Against this proposed strategy of extensional validation, one could object that the aboutness of texts does not have to correlate with high dominance of specific topics. With regard to sujet, this objection can be appropriate because it can be necessary for local sujets to calculate the share of topic dominance not for whole documents but only for specific segments. In general, however, this objection amounts - intentionally or unintentionally - to the claim that topic modeling would be completely irrelevant concerning aboutness. If this objection holds true, the dominance of specific topics for singular documents would not have any meaning. It was the aim of this paper, however, to provide the ground for strategies that allow proving whether there is such a meaning of the dominance of topics with regard to the question of what texts are about.

5. Conclusion

This paper has a methodological impact as well as an empirical result: With regard to the first, we claim that it is common practice in CLS to distinguish between thematically interpretable and uninterpretable topics. This dichotomy of interpretability versus non-interpretability has two weaknesses: Firstly, it is imprecise because our disambiguation demonstrated that 'theme' (from Jockers 2013) often means 'fabula' or 'sujet' and that both notions refer to different types of textual properties. The second weakness is that it has not yet been validated whether topics really approximate specific sujets, fabulae, or concepts within thematic claims. In this paper, we maintain that validation is not, as methodological discussion in CLS suggests so far, either internal or external. It is rather located on a relation between (a) the intension and (b) the extension of a qualitative concept and (c) a quantitative procedure. On each relation of this triangle, conceptual clarification, explication, and operationalization are important methodological units and are interlinked with different tasks of validation. Hence, we do not claim that everything is validation or that validation is everything, but, rather, that validation pops up at all three relations of a holistic research design. Disambiguating different forms of aboutness

is necessary and limiting oneself to specific aspects (such as certain *sujets*) is useful 546
because quantitative procedures are expected to behave unequally to different *sujets*, 547
fabulae, and themes so that different forms of aboutness need different operationalization 548
and individual validation. 549

Although singular validation results cannot be generalized in a simple way and without 550
further empirical proof, our illustrative example in the fourth chapter can serve as a 551
starting point for generalizations that have to be proved in forthcoming studies. Based 552
on rational reflection and the results of our case study, we expect *sujet* to be better 553
operationalizable with topic modeling than fabula, and fabula to be operationalized in 554
specific cases such as seafaring or Western as *sujet*. Such cases may be well operationalized 555
because of their homogeneous setting, which is linked with fabula according to genre rules. 556
In such cases, it is rather *sujet* than fabula that is represented by word lists. With regard 557
to theme, only the isolated abstract concepts that have a basis on the level of *sujet* or 558
fabula in a text (such as love) can be expected to be operationalized with word lists. We 559
suggest that the practice of operationalization should be regarded as a recursive process 560
that repeatedly compares the intension of the qualitative concept with the internal 561
structure of the quantitative procedure and adjusts the parameters of that procedure 562
based on such comparison. Therefore, one of our potential future works is to test LDA 563
with different parameter settings and also to test more advanced quantitative methods 564
such as Deep Neural Networks-based topic models (Zhao et al. 2021) or state-of-the-art 565
language models, to find out whether more complex aboutness-claims in literary corpora 566
could be operationalized. 567

In technical terms, topic modeling reduces the dimensions of a document-term matrix of 568
a corpus. Internal validation with reference to the intension of the qualitative concept 569
is the most common and often an appropriate form of validation in computational 570
linguistics. However, as we discussed in the paper, only some of the topics can be used 571
as the representation of a small part of the distribution of aboutness in literary corpora. 572
In CLS, internal validation can be useful but it is not sufficient because it does not 573
guarantee that topics are capable of identifying texts that have the respective *sujet*, 574
fabula, or theme from the perspective of hermeneutics. Our results for the extensional 575
validation support this suspicion. 576

The empirical result is that, based on extensional validation, topic modeling did not 577
perform with statistical significance in all cases. However, the calculated t-statistic has a 578
positive value for all but one candidate topics, which implies that topics mostly indicted 579
the expected tendency. From this empirical result, we can draw several hypothetical 580
generalizations. We assume that topic modeling is not able to identify aboutness for all 581
sorts of *sujet*, fabula, and theme in a strict sense. A two-step validation strategy based on 582
two different annotated validation samples and a grid search for optimizing parameters 583
could, however, yield better results for topic modeling in future research. As the 584
discussion of conceptual intension and interpretive validation in section 2 demonstrated, 585
it is hardly possible to generate promising topics as approximations to *sujets* such as 586
Romanesque setting. For other *sujets* that have promising approximations as topics, 587

the method performs more poorly than the method generating lists based on word 588
embedding. As simple word lists with equally weighted words are less complex than 589
topics with differently weighted words, this result may be astonishing. Based on analytic 590
reasoning and for the sujet of a Romanesque surrounding, however, this result comes as 591
no surprise. Whereas existing studies examined the applicability of topic modeling in 592
different domains (e.g. Navarro-Colorado 2018), we applied it to the domain of narrative 593
fiction and come to the preliminary conclusion, that in the realm of analyzing aboutness 594
topic modeling may be most appropriate to operationalize fabula related sujets such 595
as Western or Seafaring because of the homogeneity of setting-references and the high 596
frequency of these references. Non-fabula based sujets such as location in a specific 597
cultural environment may be operationalizable with dictionary or word-embedding based 598
word lists. These results do not reduce the applicability of topic modeling for domains 599
different from aboutness, for example for analyzing historical (Lee 2019) or philosophical 600
(Nichols et al. 2018) discourses. Therefore, we do not share Shadrova's general scepticism 601
against the non-generalizability of topic modeling. If the statistical characteristics of 602
each quantitative procedure are taken into account and related to the terminological 603
definitions of philological notions of fabula, theme, and sujet, there is new epistemic 604
ground for articulating hypothetical generalizations of the particular empirical results of 605
validation studies. If these hypothetical generalizations can be proved in further studies, 606
stronger empirical evidence for the appropriateness of specific quantitative procedures 607
for analyzing general types of aboutness can be gained. 608

6. Data availability 609

Data can be found here: <https://github.com/julianschroeter/19CproseCorpus> 610

7. Software availability 611

Software can be found here: https://github.com/julianschroeter/evaluating_topic-modeling_for_sujet_and_theme 612
613

8. Acknowledgements 614

The construction of the corpus and the task of annotation was performed by Theresa 615
Valta, Johannes Leitgeb, and Julian Schröter and has been funded by the Forschungsfonds 616
der Philosophischen Fakultät der Universität Würzburg from 2018 to 2020 in the course 617
of a larger project on the history of German novellas ([https://www.germanistik.uni-wuerzburg.de/en/lehrstuehle/computerphilologie/mitarbeiter/schroeter](https://www.germanistik.uni-wuerzburg.de/en/lehrstuehle/computerphilologie/mitarbeiter/schroeter/forschung-publikationen/habilitationsprojekt/) 618
[/forschung-publikationen/habilitationsprojekt/](https://www.germanistik.uni-wuerzburg.de/en/lehrstuehle/computerphilologie/mitarbeiter/schroeter/forschung-publikationen/habilitationsprojekt/)). 619
620

9. Author contributions 621

Julian Schröter: Conceptualization, Formal Analysis, Writing - original draft, Data 622
curation, Validation, Software 623

Keli Du: Methodology, Writing - review & editing, Formal Analysis, Validation, Software 624

References 625

- Aletras, Nikolaos and Mark Stevenson (Mar. 2013). “Evaluating Topic Coherence Using 626
Distributinal Semantics”. In: Proceedings of the 10th International Conference on 627
Computational Semantics (IWCS 2013) Long Papers. Potsdam, Germany: Association 628
for Computational Linguistics, pp. 13–22. URL: <https://www.aclweb.org/anthology/W13-0102> (visited on 05/05/2020). 629
630
- Alexandr, Nikolich et al. (2021). “Fine-Tuning GPT-3 for Russian Text Summarization”. 631
In: Proceedings of the Computational Methods in Systems and Software. Springer, 632
pp. 748–757. 633
- Barthes, Roland (1971). S/Z. Paris: Seuil. 634
- Blei, David M. (Apr. 2012). “Probabilistic Topic Models”. en. In: Communications of the 635
ACM 55.4, p. 77. ISSN: 00010782. DOI: [10.1145/2133806.2133826](https://doi.org/10.1145/2133806.2133826). URL: <http://dl.acm.org/citation.cfm?doid=2133806.2133826> (visited on 05/17/2019). 636
637
- Blei, David M. and John D. Lafferty (2009). “Topic Models”. In: Text Mining. Chapman 638
and Hall/CRC, pp. 101–124. 639
- Carnap, Rudolf (1950). Logical Foundations of Probability. Chicago: University of 640
Chicago press. 641

- Da, Nan Z. (Mar. 2019). "The Computational Case against Computational Literary Studies". In: *Critical Inquiry* 45.3. Publisher: The University of Chicago Press, pp. 601–639. ISSN: 0093-1896. DOI: [10.1086/702594](https://doi.org/10.1086/702594). URL: <https://www.journals.uchicago.edu/doi/full/10.1086/702594> (visited on 07/19/2021).
- Davidson, Donald (2001). "The Myth of the Subjective". In: *Subjective, Intersubjective, Objective*. Oxford: Clarendon Press, pp. 39–52.
- Evert, Stefan (2005). *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. Stuttgart: Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.
- Firth, John R. (1957). *Papers in Linguistics*. 1934/1951. London: Longmans.
- Frege, Gottlob (1892). "Über Sinn und Bedeutung". In: *Zeitschrift für Philosophie und philosophische Kritik* 100, pp. 25–50.
- (2021). *Die Grundlagen der Arithmetik*. Breslau: Wilhelm Köbner.
- Gries, Stefan Th. (2008). "Dispersions and Adjusted Frequencies in Corpora". In: *International Journal of Corpus Linguistics* 13.4, pp. 403–437. DOI: [10.1075/ijcl.13.4.02gri](https://doi.org/10.1075/ijcl.13.4.02gri).
- Hammond, Adam (2017). "The Double Bind of Validation: Distant Reading and the Digital Humanities' Trough of Disillusionment". In: *Literature Compass* 14.8, pp. 1–13.
- Harris, Zellig S. (1954). "Distributional Structure". In: *Word* 10.2-3, pp. 146–162.
- Holub, Robert C. (1985). "Realism, Repetition, Repression: The Nature of Desire in *Romeo und Julia auf dem Dorfe*". In: *MLN* 100.3, pp. 461–497.
- Husserl, Edmund ([1900] 2009). *Logische Untersuchungen, Erster Band: Prolegomena zur reinen Logik (1900/1913)*. Hamburg: Meiner.
- Jannidis, Fotis, Leonard Konle, and Peter Leinen (2019). "Makroanalytische Untersuchung von Heftromanen." In: *DHd Konferenzabstracts*, pp. 167–173.
- Jockers, Matthew L. (Apr. 2013). *Macroanalysis: Digital Methods and Literary History*. Englisch. Urbana: University of Illinois Press. ISBN: 978-0-252-07907-8.
- Jockers, Matthew L. and David Mimmo (2013). "Significant Themes in 19th-century Literature". In: *Poetics* 41.6, pp. 750–769.
- Kaiser, Gerhard (1971). "Sündenfall, Paradies und himmlisches Jerusalem in Kellers *Romeo und Julia auf dem Dorfe*". In: *Euphorion* 65.1971, pp. 21–48.
- Kryciski, Wojciech et al. (2019). "Neural Text Summarization: A Critical Evaluation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 540–551.
- Lamarque, Peter (2009). *The Philosophy of Literature*. Malden (USA), Oxford (UK), Carlton (Australia): John Wiley & Sons.
- Lee, Changsoo (Apr. 2019). "How are immigrant workers represented in Korean news reporting? A text mining approach to critical discourse analysis". In: *Digital Scholarship in the Humanities* 34.1, pp. 82–99. ISSN: 2055-7671, 2055-768X. DOI: [10.1093/dsch/fqy017](https://doi.org/10.1093/dsch/fqy017). URL: <https://academic.oup.com/dsh/article/34/1/82/5039835> (visited on 05/17/2019).
- Lin, Chin-Yew (2004). "Rouge: A Package for Automatic Evaluation of Summaries". In: *Text summarization branches out*, pp. 74–81.

- Luhmann, Niklas (1982). *Liebe als Passion: zur Codierung von Intimität*. Frankfurt am Main: Suhrkamp. 686
687
- Mellmann, Katja and Marcus Willand (2013). “Historische Rezeptionsanalyse. Zur Empirisierung von Textbedeutungen”. In: *Empirie in der Literaturwissenschaft*. mentis, pp. 263–281. 688
689
690
- Menninghaus, Winfried (1982). “Romeo und Julia auf dem Dorfe. Eine Interpretation im Anschluss an Walter Benjamin”. In: *Artistische Schrift. Studien zur Kompositionskunst Gottfried Kellers*. Frankfurt am Main: Suhrkamp. 691
692
693
- Mimno, David et al. (2011). “Optimizing Semantic Coherence in Topic Models”. In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 262–272. 694
695
696
- Nallapati, Ramesh et al. (2016). “Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond”. In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290. 697
698
699
- Navarro-Colorado, Borja (2018). “On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry”. In: *Frontiers in Digital Humanities* 5. ISSN: 2297-2668. DOI: [10.3389/fdigh.2018.00015](https://doi.org/10.3389/fdigh.2018.00015). URL: <https://www.frontiersin.org/article/10.3389/fdigh.2018.00015>. 700
701
702
703
- Neto, Joel Larocca, Alex A. Freitas, and Celso AA Kaestner (2002). “Automatic Text Summarization Using a Machine Learning Approach”. In: *Brazilian symposium on artificial intelligence*. Springer, pp. 205–215. 704
705
706
- Newman, David et al. (2010). “Automatic Evaluation of Topic Coherence”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 100–108. 707
708
709
710
- Nichols, Ryan et al. (2018). “Modeling the Contested Relationship between Analects, Mencius, and Xunzi: Preliminary Evidence from a Machine-Learning Approach”. In: *The Journal of Asian Studies* 77.1, pp. 19–57. 711
712
713
- Piper, Andrew (2015). *Validation and Subjective Computing* [Blog post]. URL: <http://txtlab.org/?p=470>. 714
715
- Rhody, Lisa (2014). “The Story of Stopwords: Topic Modeling an Ekphrastic Tradition.” In: *DH*. 716
717
- Ribeiro, Ricardo et al. (2013). “Self Reinforcement for Important Passage Retrieval”. In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 845–848. 718
719
720
- Saul, Nicholas (2003). “Keller, Romeo und Julia auf dem Dorfe”. In: *Landmarks in German Short Prose*. Ed. by Peter Hutchinson. Bern: Peter Lang, pp. 125–140. 721
722
- Schöch, Christof (2017). “Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama.” In: *DHQ: Digital Humanities Quarterly* 11.2. 723
724
- Schröter, Julian et al. (2021). “From Keyness to Distinctiveness Triangulation and Evaluation in Computational Literary Studies”. In: *Journal of Literary Theory* 15.1-2, pp. 81–108. 725
726
727
- Shadrova, Anna (Oct. 2021). “Topic models do not model topics: epistemological remarks and steps towards best practices”. In: *Journal of Data Mining & Digital Humanities* 728
729

2021. DOI: [10.46298/jdmdh.7595](https://doi.org/10.46298/jdmdh.7595). URL: <https://jdmdh.episciences.org/8608>. 730
8. 731
- Stegmüller, Wolfgang (1969). *Wissenschaftliche Erklärung und Begründung*. Berlin, Heidelberg, New York: Springer-Verlag. 732
733
- Stocker, Peter (2007). “Romeo und Julia auf dem Dorfe. Novellistische Erzählkunst des Poetischen Realismus”. In: Gottfried Keller. *Romane und Erzählungen*. Ed. by Walter Morgenthaler. Stuttgart: Reclam, pp. 57–77. 734
735
736
- Swafford, Annie (2015). Why Syuzhet Doesnt Work and How we Know. [Blog post]. URL: <https://annieswafford.wordpress.com/2015/03/30/why-syuzhet-does-nt-work-and-how-we-know/>. 737
738
739
- Tomaevskij, Boris ([1931] 1985). *Theorie der Literatur. Poetik*. Wiesbaden: Seemann. 740
- Underwood, Ted (2019). *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press. 741
742
- Weitin, Thomas and Katharina Herget (2017). “Falkentopics. Über einige Probleme beim Topic Modeling literarischer Texte”. In: *Zeitschrift für Literaturwissenschaft und Linguistik* 47.1, pp. 29–48. 743
744
745
- Zhao, He et al. (Feb. 2021). “Topic Modelling Meets Deep Neural Networks: A Survey”. In: arXiv:2103.00498 [cs]. arXiv: 2103.00498. URL: <http://arxiv.org/abs/2103.00498> (visited on 04/22/2022). 746
747
748

Keywords:

events, narrativity, plot, annotation,
narrative theory

Licenses:

This article is licensed under:



Abstract. In this paper, we introduce a new computational narratology approach to modeling plot. It is based on events, or, more precisely, on the narrativity of event representation at the level of discourse, or the how of narration. For presenting the approach, we first discuss the notion of event in narrative theory and its relation to narrativity and plot. We then show how events and narrativity are operationalized in accordance to these assumptions as discourse phenomena. In the last section, we optimize the parametrization of our narrativity graphs by relating them to summaries and thus relating the how to the what of narration in order to account for a comprehensive notion of plot.

1. Introduction

1

In narrative theory events are conceived of as the constituents of narratives, i.e. the source ingredient from which narratives are built. Events are therefore considered the smallest units of narrations. Accordingly, models for the so called ‘narrative constitution’ explain the genesis of a narrative based on events. These models describe how events are turned into the text of a narration with a series of (idealized) processes such as permutation and linearization. In this contribution, we discuss the possibility to represent plot on the base of events. Our computational narratology approach to event annotation has already been automated (cf. Vauth, Hatzel, et al. (2021)) as well as adapted by (Chihaia 2021) for the analysis of the representation of the Mexican State of Sinaloa in newspaper reports. Here, we elaborate on the theoretical background of our operationalization and optimize our parametrization for future applications for text analysis. We consider this to be a strongly discourse-based, and thus easier to implement, alternative to the recent and important outline of narrative theory and NLP by Piper, So, and Bamman (2021).

At the center of our efforts is the operationalization of the event concept in narrative theory. We aim at implementing it for large scale text analysis by building a step by step procedure from the determination of events in narrative texts to their subsequent application for the analysis of narrativity and plot. The presented work involves two separate, but connected steps: First, we outline the concept of events, and the possibility of modeling plot based on events against the background of narratological assumptions

and then operationalize events and narrativity. This results in the convertibility of the annotations of these narrative micro phenomena to narrativity graphs that encompass whole texts. Second, we present procedures for optimizing this approach, again by relying on narratological assumptions, especially about narrativity, tellability and plot. We consider narrativity as a property of events and event chains. Tellability, which is a narratological concept used to assess the degree of narrativity of a text passage, we quantify via text summaries as reception testimonies and thus transform narrativity structures into plot representations. In doing so, we model plot as defined by i) the degree of narrativity of events and their representation as graphs over the text course and ii) as a the most tellable event sequences of a narration.

Our focus on the representation of eventfulness in the events and thus on the discourse level of narrations differs from many current approaches tackling events, narrativity or plot in one of these two regards: While many approaches model plot or narrativity by approximation via other variables (such as sentiment as in Jockers (2015) or function and “cognitive” words in Boyd, Blackburn, and Pennebaker (2020)) we address narrativity as a feature of representation, and thus address it directly, and build our operationalization of plot on top of that. Secondly, we do not rely on readers’ inference in a first place or for evaluation purposes, but instead start with textual properties and only use reader based information in a subsequent step for optimizing the approach.

2. Modeling Plot by Narrativity of Events

2.1. Events as Basic Units of Narrative

In narratology, an event is typically described as “a change of state”. Moreover, events are considered the “the smallest indivisible unit of plot construction” (cf. Lotman 1977, p. 232) and “one of the constitutive features of narrativity” (Hühn 2013, p. 1). The way events are organized into narratives is commonly described in models of narrative constitution relating the fictional world (i.e., the *what* of narration) to its representation in the text (i.e., the *how* of narration). These two levels of narratives have been introduced in the 1920s by Tomaševskij and other formalists. Since then, they have been addressed in a variety of – partly even contradictory – terms: among the most prominent ones are *fabula/sjužet* (Tomasevskij 1971), *histoire/recit* (Genette 1980), and *story/text* (Rimmon-Kenan 1983).¹ Whereas these terms refer to models of narrative constitution with two levels, some of the models of narrative constitution define even further differentiate the *histoire* (what) or the *discours* (how). For example, Stierle (1973) and Bal (1985) both introduce three levels and Schmid (2008) proposes even four levels of narrative constitution. Regardless of the number of levels assumed and their specific conception, all models of narrative constitution are – at least implicitly – based on events. Therefore, events can be seen as a core element in narrative.

Nevertheless, up to date only very few approaches in computational literary studies

1. For a more comprehensive overview of the variety of terms and differences in scope cf. Schmid (2008, p. 241), Martínez and Scheffel (2016, p. 26) and Lahn and Meister (2013, p. 215).

have addressed the narratological understanding of events in an adequate manner. This is probably due to their granularity and ubiquity as well as the conceptual challenges connected to events. Since events in narratology are seen as a sort of atoms of narrative, they are difficult to tackle both pragmatically and conceptually. Pragmatically, event analysis is a question of resources: Analyzing events means identifying and classifying a presumably large number of segments in possibly many narratives in order to be able to make qualified statements about events. While the concrete labor connected to such a manual analysis would be less of an issue for automated approaches, there the conceptual fuzziness is a so far unsolved problem. Therefore, also approaches for automated event recognition are still little developed in computational literary studies (and probably also beyond). An exception are Sims, Park, and Bamman (2019) and the implementation in Bamman (2021). In their overview Sims, Park, and Bamman (2019, p. 3624) point out that event detection in literary texts so far focuses on characters and their relations or on the modeling of plot through sentiment. In natural language processing of non-literary texts, by contrast, there is long tradition of analyzing events based on an understanding of events that is not or only partly related to a narratological understanding. These NLP approaches are focused on extracting events according to semantic categories (e.g., the automatic content extraction task, cf. Doddington et al. 2004, Walker et al. 2006) or identifying possibly relevant events from texts, whereas in our view a narratology-based approach should include aspects beyond that. Especially the belonging of events to both *histoire* and *discours* results in the conceptual challenge that should be tackled for a wider engagement with events in computational narratology.

It is exactly the relation to *histoire* and *discours* that led the Hamburg Narratology Group to distinguish events with regard to their features and functions in *event I* and *event II*. As Hühn (2013, par. 1) elaborates, *event I* is any change of state and thus a general type of event without further requirements, whereas *event II* is an event that needs to satisfy certain additional conditions. While the presence of an *event I* can be determined by its — explicit or implicit — representation in a text, an *event II* has additional features that need to be determined with “an interpretive, context-dependent decision”. These features differ in detail but they typically are related to qualities such as relevance, unexpectedness or other kinds of unusualness of the event in question (cf. Table 1 for features of events).

		state(s)	process in time	change of state	physical	mental	anthropomorphic agent	intentional	unexpected
		<i>event I features</i>			<i>additional features for event II</i>				
Prince (2010)	Stative Event	x							
	Active Event	x	x	(x)					
Ryan (1986)	Change of physical state	x	x	x	x		(x)	(x)	
	Mental act	x	x	x		x	x		
Ryan (1991)	Happening	x	x	x					
	Action	x	x	x			x	x	
Martínez and Scheffel (2016)	Happening	x	x	x					
	Action	x	x	x			x	x	
Lahn and Meister (2013)	Happening	x	x	x					
	Event	x	x	x					x
Schmid (2008)	Change of state	x	x	x					
	Event	x	x	x			x	x	(x)

Table 1: Features of events in narrative theory

The differentiation between *event I* and *II* also is connected to two distinct definitions of narrativity: “The two types of event correspond to broad and narrow definitions of narrativity, respectively: narration as the relation of changes of any kind, and narration as the representation of changes with certain qualities” (cf. Hühn 2013, par. 1). The latter goes back to Aristotle’s characterization of plot of tragedies by a decisive turning point and is also present in Goethe’s conception of the novella as based on an unheard-of occurrence (“unerhörte Begebenheit”). The description of *event II* narrativity as “the representation of changes” also highlights the representational character of events.

2.2. From the Representation of Events to Narrativity – and Plot

As we have pointed out, in narrative theory, events are not only considered constituents of narratives, but they are also connected to their narrativity. Additionally, they are connected to representationality. Our event-based approach to modeling plot builds on these aspects, i.e., the constituency of events, their relation to narrativity and their representational character in texts. For this, we focus on the discourse level, or the *how* of narrations and not, as most approaches do, on the story level, or *what* of narrations. From a narrative theory perspective, with this we tackle an important aspect of plot that is typically overlooked, even though it is implicitly and explicitly addressed in the definition of plot:

The term “plot” designates the ways in which the events and characters’

actions in a story are arranged and how this arrangement in turn facilitates
identification of their motivations and consequences. These causal and
temporal patterns can be foregrounded by the narrative discourse itself or
inferred by readers. Plot therefore lies between the events of a narrative on
the level of story and their presentation on the level of discourse. (Kukkonen
2014, par. 1)

The foregrounding of the narrative discourse and the representation of events on the
discourse level described by Kukkonen (2014) is what we try to tackle with our ap-
proach. Besides the goal to complement the approaches that focus on story aspects
and thus provide a theoretically more comprehensive understanding of plot, it is also a
decision driven by pragmatic reasons. While there is a variety of more or less structured
conceptions of how narratives are build out of defined (story-related) elements, none of
these provides an understanding of narrative construction that can be operationalized
for possibly general purposes. Not only is Propp's *Morphology of the Folktale* focused on
the very specific area of (Russian) folktales. Also supposedly general approaches like
Greimas' actantial model, Bremond's narrative roles or Pavel's move grammar are based
on rather schematic assumptions about narrations. If implemented more generally, a
considerable amount of interpretatory work is necessary in order to detect features qual-
ifying as general structural elements of narratives, just as for Levi-Strauss' structuralist
theory of mythology and the identification of "kinship". Even though there might be
a way to make these – in the widest sense: structuralist – approaches applicable, their
operationalization is certainly not a straightforward task. For being able to apply any
of these concepts of narrative construction in an automated and comprehensive event-
based approach to plot, first a clearer idea on how events are combined into narratives
would have to be developed on their basis.

Therefore we consider it not yet feasible to generally address plot by building on story
world related features. Instead, we focus on the easier to grasp representational aspect
of events in narratives. Models of narrative constitution like the one by Schmid (2008)
describe the way events are turned into their representation in the narrative texts. As
Schmid points out, within the narrative constitution model it is these very texts that
are the only accessible level and from which the underlying levels belonging to the
histoire of narrative need to be inferred. Since the analysis of textual phenomena is
easier to implement than the analysis of underlying semantics, or even story world
knowledge, we consider it reasonable to focus on the textual representation of events.
Even more so, because the analysis of events is a starting point for further analysis. Thus
the event analysis needs to be as solid as possible in order to reduce perpetuating (or
even multiplying) of errors in the subsequent steps. Therefore, there is a theoretical
reason for designing our approach based on the textual representation of events. This
takes into account narrative theory and its focus on textual representation of narratives
and its interference with the narrated world or plot.

The core phenomenon here is narrativity, which is, as we will show, our approach to
the modeling of plot. Narrativity is, again, a narrative theory term that is employed in

a variety of senses. All notions can be described as concerning the “narrativeness” of narrative(s) (Abbott 2014) and they can be grouped by their usages into two understandings: Narrativity is either understood as a kind of essence of narratives or as a quality narratives have in comparison to other narratives (or within them). This means that narrativity can be understood as a general phenomenon of narrative (as distinct from argumentation, description, etc.), or as something that particular narratives display and that can be determined by comparing them to other narratives. Therefore, the narrative theory discussion about narrativity can be put, as Abbott (2014, par. 5) says, “under four headings: (a) as inherent or extensional; (b) as scalar or intensional; (c) as variable according to narrative type; (d) as a mode among modes”. In other words, while (a) is concerned with the question of narrativity *as such*, the other three are more interested in discerning specific *characteristics* of narrativity within or between texts.

From an operationalization perspective, the latter are the more interesting notions since they can be used for classifying or clustering narrative texts. When implementing a scalar understanding of narrativity (b), one would typically be interested in the degree of narrativity of texts, whereas the classification of texts according to narrativity features (c) may be helpful for identifying subgroups of texts like genres, and an operationalization of narrativity as a mode (d) could look at the share of narrative passages in a text (or more texts). As we will discuss in the next section in more detail, our approach is based on the concept of narrativity as scalar property. Moreover, it is designed as a heuristic both for approaches based on an understanding of narrative as (c), a variable, and (d), a mode.

While none of these narrativity conceptions addresses plot in a first place, there is a connection between events and narrativity described in narrative theory that opens up the possibility of modeling plot based on events. As already discussed, the two event types introduced by Hühn (2013, par. 5) are connected to specific understandings of narrativity. *Events I* clearly relate to Abbott’s concept of narrativity as (a) inherent property of narrative texts since the mere fact that a text consists of *events I* makes this text a narrative texts. The connection between event types and narrativity concepts (b)–(d), on the other hand, can certainly be connected to the analysis of *events II*. It seems reasonable to infer from the quality and quantity of *events II* to narrative properties of texts and thus use *events II* for the operationalization of narrativity. But also *events I* can be used for this, if operationalized in an adequate manner. This is important for our approach, because the building on *events I* enables us to focus on representational aspects and to ignore story related aspects (as well as extratextual information) that would be needed for *event II* analysis.

The prerequisite for building an understanding of narrativity as property “integral to a particular type of narrative” (Hühn 2013, par. 5) without direct reference to *events II* is the possibility to identify only certain *events I* as relevant. Here, the concept of tellability provides a possibility to operationalize the way events are “ foregrounded by the narrative discourse itself” (Kukkonen 2014, par. 1) and thus to relate *events I* to plot. Tellability, just like plot, is not only connected to story, but also to discourse:

“Tellability [refers] to features that make a story worth telling, its ‘noteworthiness.’ [...] 195
 The breaching of a canonical development tends to transform a mere incident into a 196
 tellable event, but the tellability of a story can also rely on purely contextual parameters 197
 (e.g., the newsworthiness of an event). [...] Tellability may also be dependent on 198
 discourse features, i.e., on the way in which a sequence of incidents is rendered in a 199
 narrative”. (Baroni 2012, par. 1) This possibility of defining tellability with regard to the 200
 very representation of a narrative enables us to focus on *event I*. This is an alternative to 201
 the concept of narrativity developed by Piper, So, and Bamman (2021, p. 3) (“Someone 202
 tells someone somewhere that someone did something(s) [to someone] somewhere 203
 at some time for some reason”). While we consider their definition of narrativity 204
 helpful for furthering computational approaches, it entails the development of a series 205
 of approaches (to characters, time, place, action, representation mode, etc.) that need to 206
 be combined into one approach before being applicable as narrativity analysis. On the 207
 contrary, our approach is more straightforward to apply since it is directly based on the 208
 representation of events and their narrativity. On the long run, both approaches should 209
 be combined. 210

We will now show how we put into practice our approach to modeling plot based on 211
 events and their narrativity. 212

3. Operationalizing Events and Narrativity 213

3.1. Narratological Operationalization of Events 214

Our approach to the annotation of events considers events as “any change of state 215
 explicitly or implicitly represented in a text” and is therefore based on *event I* which is 216
 “the general type of event that has no special requirements” (Hühn 2013, par. 1). In our 217
 operationalization we further differentiate between event types in order to provide for 218
 narrativity analysis and we classify the events according to their representation.² 219

The differentiation of event types is based on the first three event criteria listed in Table 1, 220
 namely being a state, a process in time and a change of state. Being a state as well 221
 as being a process in time are typically considered prerequisites for changes of state. 222
 Since Prince also introduces the notion of a stative event (which is neither a process 223
 nor a change of state), we consider it sensible to use all three criteria and base three 224
 different event types on them: states, processes in time and changes of state. With this 225
 more fine-grained solution we can incorporate more theoretical positions in our event 226
 operationalization such as the one by Prince (2010) or the consideration of processes 227
 of speaking, thinking and movement which are often not considered event candidates. 228
 Moreover, we also provide a possibility to distinguish different levels of narrativity 229
 according to the three event types. Changes of states have the highest level of narrativity, 230
 processes in time have lower and states lowest narrativity. We additionally introduce 231
 non-events as category for enabling the comprehensive annotation of texts.³ 232

2. Cf. Vauth and Gius (2021) for a detailed annotation guideline.

3. We also use additional properties derived from the criteria in Table 1 and additionally determine whether

The annotation is guided by the explicit representation of these event types in finite verbs, i.e., the question whether the verb points to a state, a process or a change of state or none of these in the fictional world. The annotation units are defined as minimal sentences including all words which can be assigned to a finite verb. Thus, there are no overlapping annotations. The determination of verbal phrases as annotation units and the finite verb as central also entails that the change of state needs to be expressed in a single verbal phrase.

The four event categories are determined as follows (examples are taken from Kafka's *Metamorphosis*):

1. 'Changes of state' are defined as physical or mental state changes of animate or inanimate entities as for example "Gregor Samsa one morning from uneasy dreams awoke".
2. 'Process events' cover actions and happenings not resulting in a change of state (e.g., processes of moving, talking, thinking, and feeling) as for example "found he himself in his bed into a monstrous insect-like creature transformed".
3. 'Stative events' refer to physical and mental states of animate or inanimate entities as for example "His room lay quietly between the four well-known walls".
4. 'Non-events' have no reference to facts in the story world and typically comprise questions or generic statements or counterfactual passages as for example "She would have closed the door to the apartment".

With this operationalization, we implement a discourse based approach to events that includes the narrativity of events. From a narrative theory perspective, our approach connects events to plot by basing the event identification on narrativity. Moreover, we focus on the discourse level and most importantly, we annotate neither linguistically nor do we make assumptions about facts in the narrated world beyond the facts represented in the event in question. It is rather the representation of the story and thus the representation of eventfulness in discourse that is being tackled by this approach. This becomes obvious when looking at one of the examples above: While "found he himself in his bed into a monstrous insect-like creature transformed" relates certainly to the most impactful change of state of the whole *Metamorphosis* (i.e., the metamorphosis of Gregor Samsa into an insect), it is here only represented as a process of perception (of a change of state). This example illustrates our focus on event representation and an important advantage of this approach: We avoid the relatively strong interpretations necessary when primary relating to the story world 'behind' its representation in the narrative. With our event type annotations, we do not want to decide whether Gregor's physical transformation is a fact in the narrated world, but stick to its representation as perception (and thus take seriously that Kafka integrates a decisive ambiguity into the

events are irreversible, intentional, unpredictable, persistent, mental or iterative (cf. Vauth and Gius (2021) for the comprehensive description of the annotation categories and tagging routines). This is not discussed here since it is directed towards *event II* detection and integration of (more) story world knowledge in possible further steps and thus beyond the scope of this contribution.

beginning of his novella). 270

Additionally to this discourse orientation, our approach to annotate the whole text 271
implements the narrative theory understanding of events as basic elements of narratives. 272
Therefore, our approach is suitable for testing the assumptions of narrative theory with 273
regard to their applicability. From a quantification perspective, our conception of events 274
as scalar with regard to their narrativity, together with the complete annotation of texts, 275
enables us to further compute our event annotations. 276

3.2. Representing Plot as Narrativity Graphs 277

The annotations of event types are used to model the narrativity of a text as timelines 278
and by that its plot. To do this, we use a scaling of the narrativity of our event types 279
and a smoothing procedure. Scaling and smoothing are also used to optimize the plot 280
modeling in [section 4](#). For this reason, we present both operationalization steps briefly. 281

3.2.1. Narrativity Values 282

As already discussed above, we implement a scalar notion of narrativity. This is real- 283
ized by assigning each of the four event types a narrativity value. In doing so, every 284
annotation and by that every text span gets a narrativity value. Beyond implementing 285
the underlying theoretical assumptions about the narrativity of the event types, this 286
also allows to compute the event annotations. From a statistical perspective, categories 287
should only be transposed into numbers if that can be done in a meaningful way. In our 288
case, we have an obvious ranking of categories. The rank starts with no narrativity for 289
non events and extends to highest narrativity for changes of state. 290

Since the determination of an absolute value of the event categories is a bit less obvious, 291
we used predefined narrativity values for a first exploration: 292

- **Changes of states:** narrativity value 7 293
- **Process events:** narrativity value 5 294
- **Stative events:** narrativity value 2 295
- **Non events:** narrativity value 0 296

These values represent not only our intuition about the relevance of the event types 297
for a text's narrativity, but are also oriented to the discussion about description and 298
narration as text modes with different narrativity (Herman 2005). Nevertheless, it is an 299
open question if these values are appropriate. 300

3.2.2. Smoothing 301

The concepts of narrativity in literary studies do not describe micro phenomena on word 302
or sentence level, but rather larger text passages in the size of a couple of paragraphs 303
and beyond. Due to that, we use a cosine weighted smoothing approach to model 304

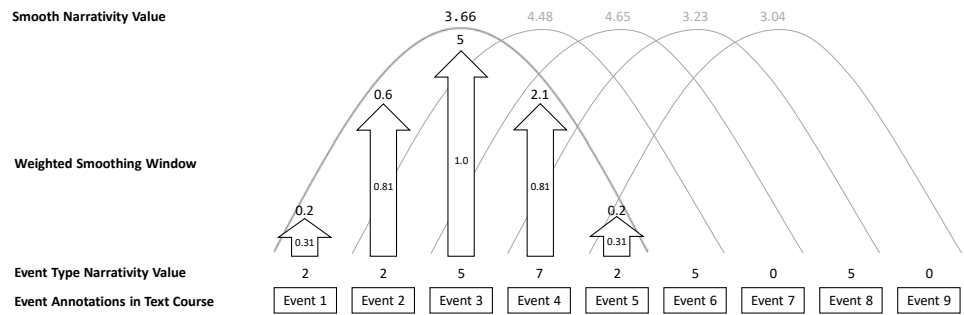


Figure 1: Illustration for the cosine weighted smoothing with a smoothing window of 5 events.

the narrativity of longer text passages. With this, the smoothing process generates narrativity values that can be used to draw an interpretable timeline graph, representing the narrativity in the text's course or, to use the terminology of narratology, in narration time. Figure 1 shows how we compute a smooth narrativity value for each event. Due to the cosine weighting, for the computation of the smoothed narrativity values the unsmoothed narrativity values (Event Type Narrativity Values) of the events in the outer parts of the smoothing window are included to a lesser extent. In doing so, we assume that context influences the narrativity of a text's passage, but that this influence diminishes the further away the contextual events are.

As for the scaling, the size of the smoothing window is an operationalization decision that can be used for optimization. For our exploration we set the value of the window to 100, assuming that passages of 100 events (i.e., verb phrases) are a reasonable size with regard to narrativity.

3.2.3. Evaluation by Exploration

In *anonymized_2*, we evaluated our narrativity timeline graphs by simple exploration of the graph's peaks. As discussed above, the values of the event types were set to 0, 2, 5, and 7, and the smoothing window covers 100 events. Figure 2 shows that the highest peaks of the timeline representing the narrativity in Franz Kafka's *Metamorphosis* are located in text passages where we find actions that are somehow related to the *event II* concept. At least, it would be reasonable to say that these passages are central for the development of the plot. Therefore, this explorative evaluation indicates that narrativity graphs have the potential to model a narrative's plot as timeline and can detect eventful text parts.

4. Optimizing Narrativity for Plot Representation

Our main idea for optimizing the narrativity graphs in their capability to model plot is a quantitative comparison to the text passages that are mentioned in summaries of the

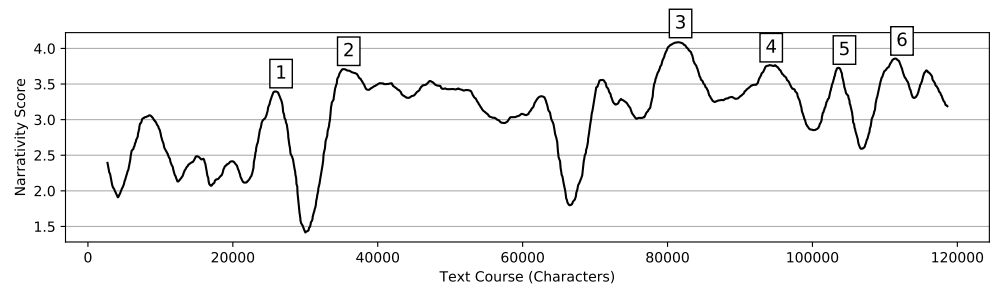


Figure 2: Peaks in Franz Kafka’s *The Metamorphosis* as an Evaluation by Exploration (anonymized_2). The annotated Peaks are:

1. After the metamorphosis, Gregor exposes himself for the first time to his family and colleague.
2. Gregor leaves his room, his mother loses consciousness, the colleague flees and his father forces him back into his room.
3. Gregor’s father throws apples at him. Gregor gets seriously wounded. Escalation of the father-son conflict.
4. Three tenants move into the family’s flat.
5. Gregor shows himself to the tenants, who then flee.
6. Gregor dies.

annotated texts. With this, we can improve our graphs as plot representations which represent the degree of narrativity of events over the text course and highlight the most tellable event sequences of a narration.

This approach is based on the assumption in narrative theory that narrativity and tellability are strongly related. At the same time, we use this procedure to test whether our approach of modeling narrativity on the basis of event representation is suitable also from a story-related notion of narrativity and can thus be considered a comprehensive approach. This is because the summaries should primarily refer to the level of *histoire* and not consider the mode of representation.

For optimization, we used four manually annotated German texts: *Das Erdbeben in Chili* by Heinrich von Kleist, *Die Judenchilde* by Annette von Droste-Hülshoff, *Krambambuli* by Marie von Ebner-Eschenbach and the already mentioned *Die Verwandlung* by Franz Kafka.

4.1. Setting

4.1.1. Resources to Optimize Narrativity Graphs

Our first approach was to base our optimization on summaries by expert readers. For this purpose, we collected summaries written by literary scholars and published in the *Kindler Literatur Lexikon* (Arnold 2020), the most popular encyclopaedia for German literature. However, when reviewing these summaries, it became apparent that a high proportion of them consisted of interpretative passages and comments on the text, and the very summary of the texts was only a small part that also varied in its realization. Because of that, these expert summaries were inappropriate for our purpose to model

plot. 353

As a second attempt, we collected summaries of our manual annotated stories from 354
Wikipedia. These summaries turned out to be more focused on the text's plot. Still, we 355
 noticed that some of these summaries seemed to place an arbitrary emphasis on the 356
 summarized parts of the stories. Therefore, it is not reasonable to assume that these 357
 summaries only focus on the most tellable happenings. 358

To compensate for the randomness of a single text summary, we finally had students 359
 write summaries for four of our manually annotated texts. The work assignment was: 360

- Read the selected primary text. 361
- Write as simple a summary as possible. 362
- Summarize the main events of the text from your point of view. 363
- Do not use any aids but the narrative itself. 364
- Do not write more than 20 sentences. 365

By that we received between 9 and 11 independent summaries for each of our four 366
 texts. Based on these we could now evaluate which passages of the stories have been 367
 mentioned frequently, also assuming that those passages that have been mentioned by 368
 more readers display a higher degree of tellability. 369

4.1.2. Annotation of Summaries 370

To measure the frequency and by that the tellability of text passages, we annotated the 371
 text spans referred by a summary for each sentence of the summary. For this, only 372
 the sentences that refer to clearly identifiable happenings were taken into account. For 373
 example, a sentence like "Kafka's Metamorphosis tells the story of Gregor's expulsion 374
 from the civilized world" is a too general summary, while the reference of a sentence like 375
 "Gregor is wounded by his father" can be located in the narrative without any problems. 376

Figure 3 shows the annotations of summaries for the four summarized texts. The 377
 shortest of these texts, *Krambambuli*, has a length of about 25,000 characters, while the 378
 longest, *Die Verwandlung* ("The Metamorphosis"), has a length about 120,000 characters. 379
 This has an impact on the summaries and their usage for our optimization task. As 380
 Figure 3 shows, many summaries of the two shorter texts, *Erdbeben* and *Krambambuli*, 381
 refer to relatively large parts of the narratives. In the shorter texts, however, the multiple 382
 mention of an event is not a strong indication that these are particularly tell-worthy 383
 passages. It is simply caused by the fact that large parts of the texts are mentioned by all 384
 summaries. For the optimization of the narrativity graphs, however, it is important that 385
 the summaries are as selective as possible, because in the next step we will determine 386
 how many summaries refer to the same events. 387

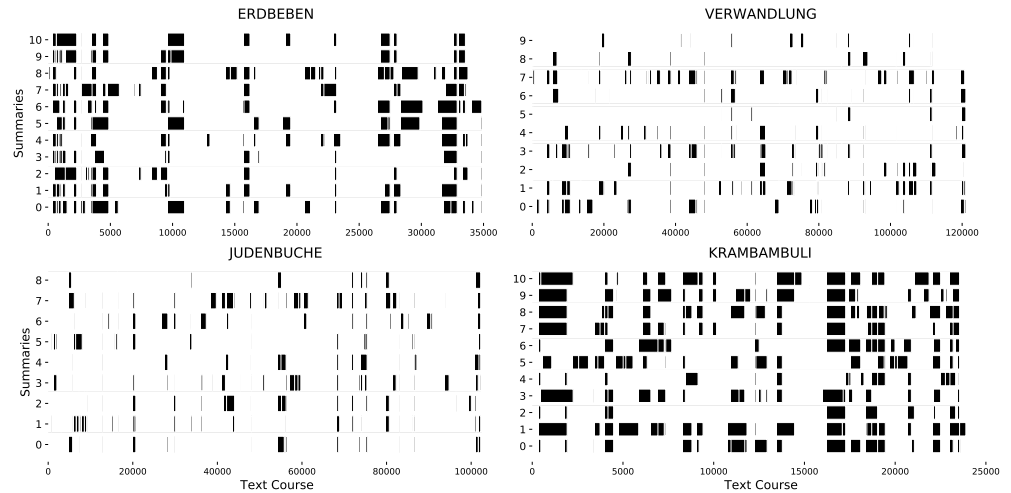


Figure 3: Summary Annotations in Text Course

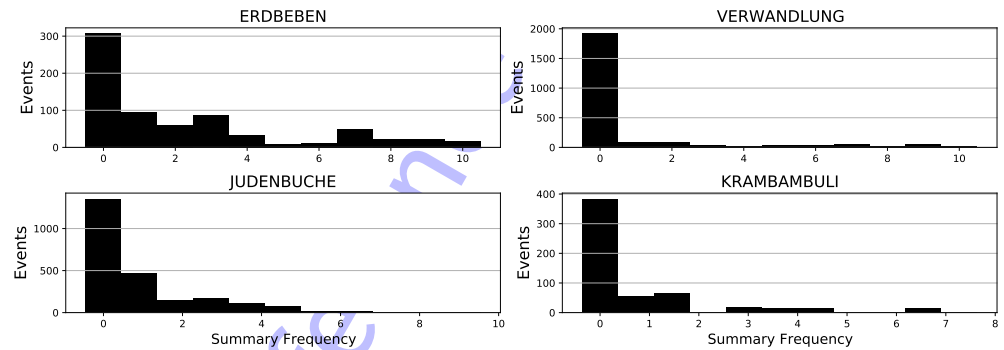


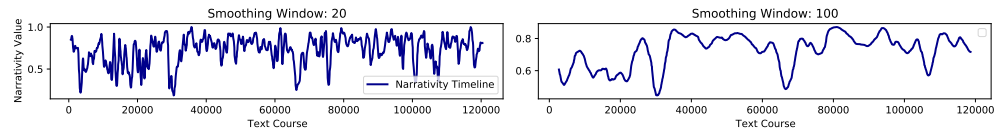
Figure 4: Summary Frequency per Event. With this Figure for each event annotation the number of summary mentions is counted.

4.1.3. Optimization Method

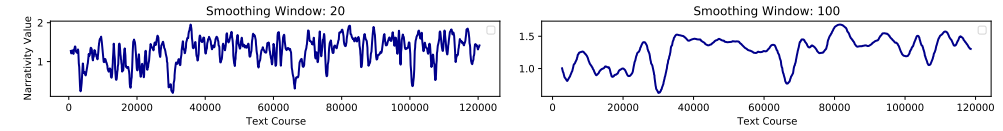
388

Our optimization approach is based on the comparison of the event-based narrativity 389 graphs that we presented at the end of the last section with the tellability scores of 390 individual text passages quantified based on the summaries. For this purpose, we 391 determined for each event annotation (subsection 3.1), in addition to the smoothed 392 narrativity value, in how many summaries the event is mentioned. This resulted in 393 a tellability value for events defined by the number of summaries in which the event 394 in question is mentioned. With this, narrativity and tellability values are available for 395 the entire text, and their correlation can be tested. For optimization, we adjusted the 396 narrativity graphs or the generation of the narrativity values in such a way that the 397 correlation between narrativity and tellability is as high as possible. However, given the 398 assumption that tellability is connected to high narrativity it is particularly important 399 that the passages with high tellability values are also assigned a high narrativity value. 400

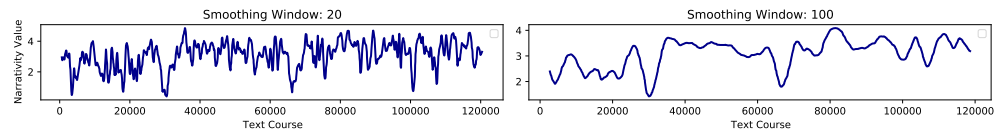
For finding a setting that possibly raises the correlation of narrativity and tellability, 401 we adjusted the narrativity scaling on the one hand and the size of the smoothing 402



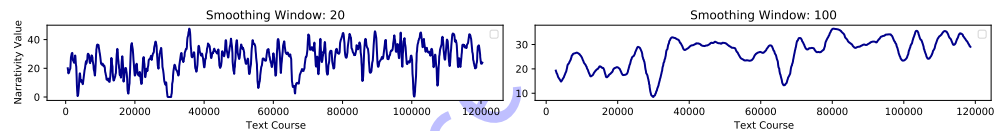
(a) non_event = 0; stative_event = 1, process_event = 1, change_of_state = 1



(b) non_event = 0; stative_event = 1, process_event = 2, change_of_state = 3



(c) non_event = 0; stative_event = 2, process_event = 5, change_of_state = 7



(d) non_event = 0; stative_event = 0, process_event = 50, change_of_state = 50

Figure 5: Narrativity Timelines for Kafka's *Metamorphosis* as an example for the impact of smoothing and event type scaling

windows on the other hand. Figure 5 shows with two different smoothing windows and 403
four different event type scales that the structure of the narrativity graphs is especially 404
affected by the settings for the size of the smoothing window. For example, a comparison 405
between the timelines in Figure 5a, 5b, 5c, and 5d, shows considerable differences from 406
character 40,000 to 60,000 between the graphs on the left (smoothing window 20) and 407
those on the right (smoothing window 100). In certain passages, also the change of 408
narrativity scales influences the graph. 409

Figure 6 indicates a relation between the frequency of event mentions in different sum- 410
maries and its narrativity. The depicted heat maps show the relation of narrativity values 411
and summary frequency for each constellation of smoothing window and narrativity 412
value shown in Figure 5. Due to the changing event type scales, the narrativity values 413
differ from Figure 6a, to 6b, 6c and 6d. The summary frequency is in all subplots of 414
course the same. For all eight heat maps a tendency that frequently mentioned events 415
have a relatively high narrativity value is visible. At least, the events which get men- 416
tioned in more than five summaries have a narrativity score higher than the overall 417
average. 418

At the same time, the share of frequently mentioned events is comparatively small. This 419
is relevant for our optimization method. Because the proportion of frequently mentioned 420
events is so small, they have little effect on the correlation measurements we use for 421
optimization. Instead, the predominant number of not frequently mentioned events 422

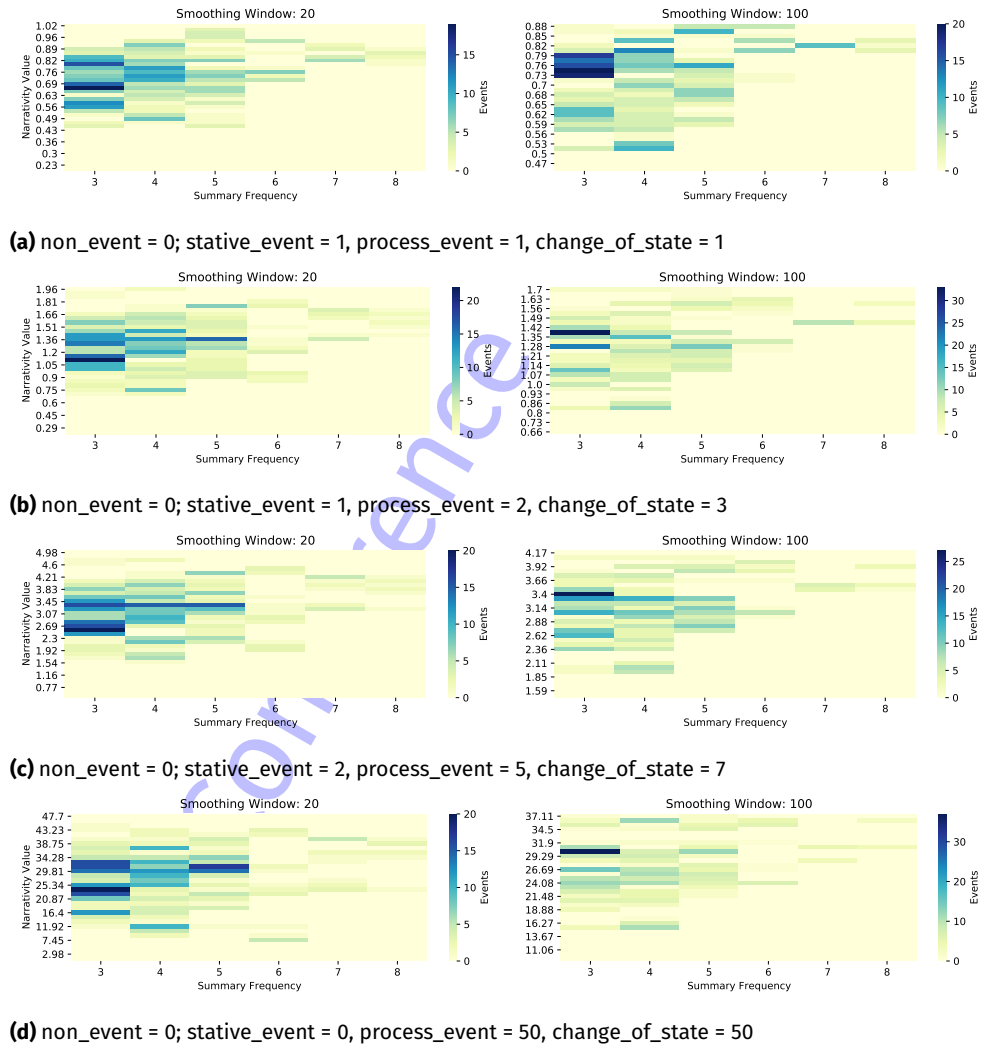


Figure 6: Narrativity of tellable events in Kafka's *Metamorphosis* as an example for the impact of smoothing and event type scaling

heavily affects the correlation. For this reason, no particularly high correlation between tellability and narrativity can be expected, especially, if we take all event annotations of a text into account. However, since this is an optimization process and not an evaluation process, it is rather the improvement of the correlation than the overall value that is of interest.

4.2. Optimizing Smoothing

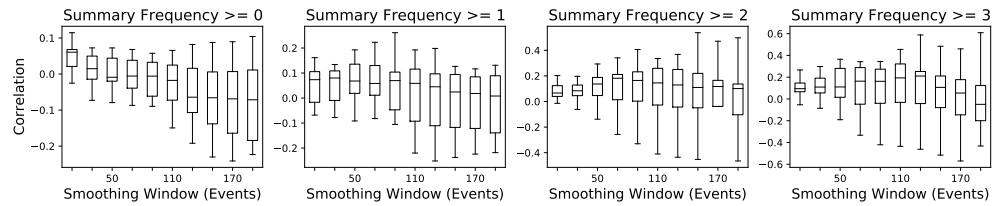


Figure 7: The Impact of Smoothing for the Correlation of Narrativity and Summary Frequency per Event.

For the optimization of the smoothing windows, we used smoothing windows from 10 to 190 events for each of the four texts in combination with different narrativity scalings for the four event types. With respect to the scalings, it was determined that non events always had a narrative value of 0, while process events and changes of state were assigned a narrative value of at least 1 and at most 20. This results in the number of 1,750 permutations. The combination of those with the different smoothing windows (in steps of 20) leads to a total number of 105,000 constellations per text.

With regard to optimizing smoothing windows, the influence of the window size on the correlation of the narrativity and the tellability of an event is shown in Figure 7. The correlation values of the four texts for each smoothing window were combined into a single box plot. If all events of the texts are considered (Summary Frequency ≥ 0), the correlation lies roughly between -0.2 and 0.1 depending on the smoothing window. Again, these low correlation values are due to the fact that the Summary Frequency is 0 for the majority of the events (see Figure 4), whereas the narrative values of these events still vary and thus influence the correlation.

For this reason, we have used different filter settings for the box plot visualization. If only the events mentioned in at least one summary (Summary Frequency ≥ 1) are taken into account, the maximum correlation value rises above 0.2 with a smoothing window of 90 events. Correlation values increase even more if only the events mentioned in at least 2 or 3 summaries are included. In the first case, the maximum correlation is close to 0.5 and in the second case even 0.6. This confirms our assumption that single summaries would not have been sufficient resources for the optimization (see 4.1.1).

In all four subplots, we can observe that both the maximum correlation and the median of the correlation values of a smoothing window at a certain point decrease with an increasing size of the smoothing windows. Considering the median values of all four subplots, a first conclusion of this optimization procedure is therefore that smoothing

windows of a size larger than 100 events are not useful.

As for the scattering of the correlation values, which we depict with individual box plots, it is important to note that it is not primarily caused by the fact that we summarize the constellations for four texts in one box plot. Instead, it is mainly caused the varied narrativity scaling.

4.3. Optimizing Event Type Values

For each of the 1,750 event value combinations and each smoothing window, we determined the average correlation of narrativity and tellability for the four texts. The five highest correlation values for these configurations are listed in Table 2. Here, we perform the same filtering as in Figure 7 and determine the highest correlation values considering the minimum summary frequency. This again results in increasing correlation values according to the filtering, i.e., more frequently mentioned events have a higher correlation.

More interesting for our purposes, however, are the scaling trends in the four sections of Table 2 corresponding to the four tellability values (i.e., the Summary Frequencies $\geq 0, 1, 2, 3$). Here, the best constellations show differences with regard to narrativity value scaling that seem to be related to the tellability values. For constellations where all events are considered (Summary Frequency ≥ 0), there is no difference in scaling between stative events and process events. A similar situation applies to events mentioned in at least three summaries (see the fourth section of the table). In these two sections, stative events and process events have (almost) the same narrative values. The opposite is true for the second section (Summary Frequency ≥ 1) and, with some reductions, also for the third section (Summary Frequency ≥ 2), where process events have 1.66 times the narrativity value of stative events.

	non_event	stative_event	process_event	change_of_state	smoothing window	correlation mean
Summary Frequency ≥ 0	0	7	7	14	10	0.0597
	0	7	7	13	10	0.0597
	0	6	6	10	10	0.0597
	0	11	11	18	10	0.0597
	0	10	10	16	10	0.0597
Summary Frequency ≥ 1	0	0	12	12	50	0.1124
	0	0	18	18	50	0.1124
	0	0	19	19	50	0.1124
	0	0	10	10	50	0.1124
	0	0	11	11	50	0.1124
Summary Frequency ≥ 2	0	11	18	18	70	0.1356
	0	9	15	15	70	0.1356
	0	12	20	20	70	0.1356
	0	10	17	17	70	0.1356
	0	3	5	5	70	0.1356
Summary Frequency ≥ 3	0	19	20	20	50	0.1623
	0	18	19	19	50	0.1623
	0	17	18	18	50	0.1623
	0	11	12	12	50	0.1622
	0	7	7	7	50	0.1622

Table 2: Optimizing Event Type Scaling. Maximum average correlation for the four manually annotated texts and the tested event type scalings.

That the maximum average correlation values in Table 2 are lower than the maximum correlation values in the box plots of Figure 7 is due to the fact that in the latter different

configurations of event type scaling result in the highest correlation values for the individual texts. There, the correlation values for every text have been taken into account, whereas the values in Table 2 are average values comprising all four texts. which results in high correlation values for the four texts. However, none of the correlation values for the four tellability values in Table 2 is significantly below the median values of the box plot evaluation in Figure 7. Also, this cross-text optimization can be regarded more adequate since our approach to plot modeling is intended for the analysis of larger automatically annotated corpora.

However, it is debatable whether one should rather follow the scaling in the first and fourth sections or the scaling in the third and fourth sections. We consider the latter more conclusive. The first section takes all events into account and thus, as we have explained above, the resulting correlation values are not very meaningful. On contrast, the correlation values in the fourth section have a limited significance because they are based on a comparatively small number of events. Here we refer again to the histograms in Figure 4 that show the number of events with regard of their mentions in summaries.

5. Conclusion

For our optimization goal, the considerations and measurements we have presented when discussing Figure 7 and Table 2 yield two main results:

- The smoothing windows should include between 50 and 100 events. This had been indicated by the box plot evaluations and the best configurations in Table 2 have confirmed that.
- For the scaling, a clear weighting gradient of non events and stative events on the one hand and process events and changes_of_state on the other hand is important.

We have presented an approach to plot that is based on the representation of events and narrativity. With this, we add an — up to now little explored — aspect to the computational analysis of events, narrativity and plot, namely their discourse-oriented operationalization.

This focus on the representation of events allows us to leave aside story-related issues to a great extent. Thus, we avoid problems typically arising when analyzing plot with regard to story where reader related information is needed or, alternatively, a rather complex analysis is not possible yet and needs to be approximated by instrumental variables. Instead, we have shown how the establishment of narrativity graphs can build on our event concepts including scalar narrativity and how this can be related to the modeling of plot. As a second point, the parametrization of the narrativity graphs has been optimized with regard to the tellability of events assessed in readers' summaries of narratives.

The outcome of this work is a heuristic firmly rooted in narrative theory with which we now can analyze narratives. With regard to the narrativity notions in Abbott (2014)

discussed above, we have operationalized a scalar notion of narrativity and can use this 519
now for the analysis of narrativity as a variable and a mode of and within narrative 520
texts. Even more so, since our approach has proven to be automatable to a satisfactory 521
extent.⁴ 522

In addition to the development of an approach for analyzing the narrativity of texts, 523
this contribution shows how theoretical concepts and their computational implemen- 524
tation can be closely connected. With regard to the concept of events and narrativity, 525
the operationalization of events as scalar based on their narrativity together with our 526
optimization efforts have show the plausibility of the underlying assumptions from 527
narrative theory. In our view, this connection between theory and implementation is an 528
aspect of computational literary studies that should be emphasized more. 529

Conference

4. We have reached a F1 Score of 0.71 for the event classification on unseen texts (cf. **anonymized_2** which results in a correlation of narrativity graphs typically reaching between 0.8 and 0.9).

6. Data availability 530

Data can be found here: https://anonymous.4open.science/r/event_based_plot_model-14F4/ 531
532

7. Software availability 533

Software can be found here: https://anonymous.4open.science/r/event_based_plot_model-14F4/ 534
535

8. Acknowledgements 536

We would like to thank our student annotators who have annotated tens of thousands 537
of events, Gina-Maria Sachse, Michael Weiland and Angela Nöll, as well as our students 538
in the course “Literarische Konflikte” during the winter term 2021/2021 at Technical 539
University of Darmstadt for the writing of the summaries of the analyzed texts and 540
Gina-Maria Sachse for the annotation of the summaries. 541

This work has been funded by the German research funding organization DFG (grant 542
number GI 1105/3-1). 543

9. Author contributions 544

Evelyn Gius: Conceptualization, Formal Analysis, Writing – original draft, Writing – 545
review & editing 546

Michael Vauth: Conceptualization, Formal Analysis, Writing – original draft, Writing – 547
review & editing 548

References 549

- Abbott, H. Porter (2014). “Narrativity”. In: *the living handbook of narratology*. Ed. by 550
Peter Hühn, John Pier, Wolf Schmid, and Jörg Schönert. Universität Hamburg. URL: 551
<http://www.lhn.uni-hamburg.de/article/narrativity>. 552
- Arnold, Heinz Ludwig, ed. (2020). *Kindlers Literatur Lexikon (KLL)*. de. Stuttgart: J.B. 553
Metzler. ISBN: 978-3-476-05728-0. DOI: [10.1007/978-3-476-05728-0](https://doi.org/10.1007/978-3-476-05728-0). URL: <https://link.springer.com/10.1007/978-3-476-05728-0> (visited on 12/15/2021). 554
555
- Bal, Mieke (1985). *Narratology. Introduction to the Theory of Narrative*. Trans. by Christine 556
van Boheemen. Toronto, Buffalo, London: Toronto Univ. Press. 557
- Bamman, David (2021). *BookNLP. A natural language processing pipeline for books*. URL: 558
<https://github.com/booknlp/booknlp> (visited on 11/10/2021). 559
- Baroni, Raphaël (Mar. 2012). “Tellability”. In: *the living handbook of narratology*. Ed. 560
by Peter Hühn, John Pier, Wolf Schmid, and Jörg Schönert. Hamburg: Hamburg 561
University Press. URL: <http://www.lhn.uni-hamburg.de/article/tellability>. 562

- Boyd, Ryan L., Kate G. Blackburn, and James W. Pennebaker (Aug. 2020). "The narrative arc: Revealing core narrative structures through text analysis". In: *Science Advances* 6.32. Publisher: American Association for the Advancement of Science, eaba2196. DOI: [10.1126/sciadv.aba2196](https://doi.org/10.1126/sciadv.aba2196). URL: <https://www.science.org/doi/10.1126/sciadv.aba2196> (visited on 12/21/2021).
- Chihaiia, Matei (Dec. 2021). "Sinaloa in der ZEIT. Computergestützte Analyse von Ereignishaftigkeit und Erzählwürdigkeit in einem Korpus journalistischer Erzählungen". de. In: *DIEGESIS* 10.1. Number: 1. ISSN: 2195-2116. URL: <https://www.diegesis.uni-wuppertal.de/index.php/diegesis/article/view/425> (visited on 12/21/2021).
- Doddington, George R., Alexis Mitchell, Mark A. Przybicki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel (2004). "The automatic content extraction (ACE) program-tasks, data, and evaluation". In: *Proceedings of LREC*. Vol. 2. 1. Lisbon, Portugal, pp. 837–840.
- Genette, Gérard (1980). *Narrative discourse: an essay in method*. Ithaca, NY, USA: Cornell University Press.
- Herman, David (2005). "Events and Event-Types". In: *The Routledge Encyclopedia of Narrative Theory*. London: Routledge, pp. 151–152.
- Hühn, Peter (2013). "Event and Eventfulness". In: *the living handbook of narratology*. Ed. by Peter Hühn, John Pier, Wolf Schmid, and Jörg Schönert. Universität Hamburg. URL: <http://www.lhn.uni-hamburg.de/article/event-and-eventfulness>.
- Jockers, Matthew (2015). *Revealing Sentiment and Plot Arcs with the Syuzhet Package*. URL: <http://www.matthewjockers.net/2015/02/02/syuzhet/> (visited on 07/16/2019).
- Kukkonen, Karin (Mar. 2014). "Plot". In: *the living handbook of narratology*. Ed. by Peter Hühn, John Pier, Wolf Schmid, and Jörg Schönert. Hamburg: Hamburg University Press. URL: <http://www.lhn.uni-hamburg.de/article/plot>.
- Lahn, Silke and Jan Christoph Meister (2013). *Einführung in die Erzähltextanalyse*. Stuttgart; Weimar: Metzler.
- Lotman, Jurij (1977). *The Structure of the Artistic Text*. English. Michigan Slavic Contributions 7. Ann Arbor: Dept. of Slavic Languages and Literature, University of Michigan.
- Martínez, Matías and Michael Scheffel (2016). *Einführung in die Erzähltheorie*. 10th ed. C.H. Beck Studium. München: C.H.Beck. ISBN: 3-406-69969-3.
- Piper, Andrew, Richard Jean So, and David Bamman (2021). "Narrative Theory for Computational Narrative Understanding". en. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 14.
- Prince, Gerald (2010). *A grammar of stories: An introduction*. Vol. 13. De proprietatibus litterarum Series minor. The Hague: Mouton. ISBN: 978-3-11-081590-0.
- Rimmon-Kenan, Shlomith (1983). *Narrative Fiction. Contemporary Poetics*. London: Routledge.
- Ryan, Marie-Laure (1986). "Embedded Narratives and Tellability". In: *Style* 20.3, pp. 319–340.
- (1991). *Possible worlds, artificial intelligence, and narrative theory*. Bloomington: Indiana University Press.

- Schmid, Wolf (2008). *Elemente der Narratologie*. 2., verb. Aufl. Berlin [u.a.]: De Gruyter. 607
- Sims, Matthew, Jong Ho Park, and David Bamman (July 2019). “Literary Event Detec- 608
tion”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational 609
Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3623–3634. 610
DOI: [10.18653/v1/P19-1353](https://doi.org/10.18653/v1/P19-1353). 611
- Stierle, Karlheinz (1973). “Geschehen, Geschichte, Text der Geschichte”. In: *Geschichte – 612
Ereignis und Erzählung*. Ed. by Reinhart Koselleck and Wolf-Dieter Stempel. Poetik 613
und Hermeneutik 5. München: Fink, pp. 530–534. 614
- Tomasevskij, Boris (1971). *Teorija literatury. Poetika*. Rarity reprints. Letchworth: Bradda 615
Books. 616
- Vauth, Michael and Evelyn Gius (July 2021). *Richtlinien für die Annotation narratologischer 617
Ereigniskonzepte*. URL: <https://doi.org/10.5281/zenodo.5078174> (visited on 618
07/08/2021). 619
- Vauth, Michael, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann (Nov. 2021). “Auto- 620
mated Event Annotation in Literary Texts”. In: *CHR 2021: Computational Humanities 621
Research Conference*. Amsterdam, The Netherlands, pp. 333–345. URL: [http://ceur-w 622
s.org/Vol-2989/short_paper18.pdf](http://ceur-ws.org/Vol-2989/short_paper18.pdf). 623
- Walker, Christopher, Stephanie Strassel, Julie Medero, and Kazuaki Maeda (2006). “ACE 624
2005 Multilingual Training Corpus”. In: *Linguistic Data Consortium, Philadelphia 57, 625
p. 45.* 626

Analyzing the Positive Sentiment Towards the Term “Queer” in Virginia Woolf through a Computational Approach and Close Reading

Heejoung Shin  ¹

Ted Underwood  ¹

1. Information Science and English, University of Illinois, Urbana-Champaign.

Keywords:

Virginia Woolf, queer, modernism,
sentiment analysis, word
embedding model, Word2Vec

Licenses:

This article is licensed under:



Abstract. This article validates the thesis that Virginia Woolf’s usage of the term “queer” is positive, and that the author is more progressive with her idea of things conceived as “queer” in the era characterized as literary Modernism and in English fiction as a whole from 1850s to 1990s. Using Word2Vec, a word embedding model, I locate the top 100 words semantically closest to “queer” in Woolf’s works and in the works of other modernist authors, James Joyce, F. Scott Fitzgerald, D. H. Lawrence, Gertrude Stein, and Katherine Mansfield. I then measure the net positivity of each author’s list and compare Woolf’s with the individual authors’, and then with words closest to “queer” in English fiction from 1850 to 2000. In demonstrating the usefulness of applying word embedding models in literary criticism, a field that has traditionally primarily relied on interpretation, this article aims to serve as a case study of how a computational approach can benefit close reading.

1. Introduction

1

The word “Queer” appears more than 200 times in Virginia Woolf’s published novels, short stories, and essays. This number may be statistically insignificant, but is nonetheless important for literary critics who aim to identify forms of repetition that do not constitute a cultural reproduction of rigid identity categories. This article thus explores how “queer” is deployed in Woolf’s oeuvre against the backdrop of the history of English fiction, using Word2Vec, a powerful word embedding model (WEM) recently developed in the field of computational linguistics. This article particularly aims to look at whether Woolf’s usage of the term “queer” is typical of her era characterized as literary Modernism and whether she is progressive in her treatment of queerness throughout the history of English fiction. To accomplish this goal, I compare the top 100 words semantically closest to “queer” in Woolf’s works and in the works of other modernist authors, James Joyce, F. Scott Fitzgerald, D. H. Lawrence, Gertrude Stein, and Katherine Mansfield. Then, I measure the net positivity of each author’s list of these words and compare them with that of Woolf’s. I also analyze the associations around the term in English fiction as a whole from 1850 to 2000 to identify a larger pattern. As “queer” has

a rich semantic history, having been used to indicate existing and emerging identity categories associated with what is out of sync with normativity proper, attending to the sentiment towards the term in literature reveals how normativity is operative in a discursive field and how it is destabilized by its own operation. In demonstrating the usefulness of natural language processing (NLP) using word embeddings in literary criticism, a field that has historically relied on interpretation, this article aims to serve as a case study of how a computational approach can benefit more nuanced literary analysis, beyond identifying topics that appear frequently.

I chose literary Modernism in launching this investigation, as it is a site where the term "queer" is deployed across the broadest spectrum in literary history. Originating in 16th-century England to refer to something strange, odd, eccentric, or illegitimate, "queer" began to suggest sexual practices that fell outside of the normative form of sexuality and gender in the 19th century (Barker and Scheele 2016, 24-27). By the late 19th and early 20th centuries, along with its sister terms, "fairy," "trade," and "gay," it had become a distinct identity category and a codeword within the gay male subculture in London, although its conventional usage as a term to denote "out of the ordinary" was still predominant among the British public (Houlbrook 2005, 162-163). During this period, "queer" had also gained a pejorative connotation for homosexuality and bisexuality (Houlbrook 2005, 179). The earliest known record of the usage of the term as such is from a letter written in 1894 by the Marquess of Queensberry to accuse Oscar Wilde of having an affair with his son, Alfred Douglas: "Snob queers like Rosebery" "corrupted my sons" (Barker and Scheele 2016, 27). It is also worth noting that early 20th-century Britain is when queer expressions of any sort do not necessarily correlate to a homosexual desire. Homosexuality and lesbianism themselves were more "permitted forms of sexuality" back then, although the latter was much less visible than the former (Houlbrook 2005 10).

In the British penal system, engaging in homosexual behaviors or importuning for homosex in public places were largely treated within a broader category of moral indecency, along with its twin problem of female prostitution.... It was only in the two decades after the Second World War that the forms of understanding that we often assume to be timeless – the organization of male [and female] sexual practices and identities around the binary opposition between homo and heterosexual.... solidified (Houlbrook 2005, 10).

Modernist authors wrote at this interesting moment where the term had not yet fully come into a rigid binary configuration of gender and there was still an overlapping assemblage of its usage. In their published works and in their often-suppressed manuscripts, letters, and diaries, "queer" is deployed in a variety of contexts, to denote homosocial/homoerotic desire, their own desired authority and authorship, and more broadly, whatever is at odds with normativity proper in terms of ethnicity, gender, nationality, etc. Yet, each writer's stance and sentiment towards what they call "queer" may radically differ. Gertrude Stein, for instance, constructs what she disavows in her characters'

nationality and class around the notion of queerness in *The Autobiography of Alice B. Toklas* and in *Making of Americans*. In T. S. Eliot's suppressed poems, "queer" is almost always deployed in a self-deprecatingly comic and crude tone of voice to imagine stronger authority in association with racial otherness and homosexual desire, to complement what the poet views as a weakness in his own authority.¹ In Woolf, "queer" is usually described positively and is often associated with peculiar modes of existence, resistance, or self-expression shaped by one's moment-to-moment experience with the tyranny of the norm.² Demonstrating that this interpretation can be quantified will not only answer the question of whether Woolf is ahead of the curve against the backdrop of English literature and how our sense of what we consider to be queer has evolved across history, but also paves a new ground to frame research questions around racial and gender binaries, topics tremendously important in the Humanities field

A detailed discussion of the data, models, and methods used in this research follows, along with my interpretation of the modernist authors in question. Through this research, I validate the thesis that "queer" is more positive for Woolf than for her contemporaries explored in this article, and that Woolf's use of the term was ahead of her time, and possibly still is ahead of current usage of the term. Potentially, a meaningful discovery made throughout the research is that Joyce's works demonstrate the next most positive use of "queer" among this peer group. Indeed, the t-test performed for the positivity of "queer" for Woolf and Joyce cannot formally confirm that Woolf's use of "queer" is always positive than Joyce's, although the mean of positivity based on the ten samplings drawn from each author's corpus is higher for Woolf than for Joyce. It is also quite noteworthy that the other two women authors' usage – Stein's and Mansfield's – exhibit the most negative. This suggests that computational approaches to literature can facilitate a more nuanced understanding and "interpretation" of gendered notions and literary Modernism in the current literary climate, where the tendency to parcel authors into a generalized narrative category of male/female/queer anxiety or hysteria to impose a homogeneous identity, purely based on interpretation, is predominant.³

1. T. S. Eliot had written homoerotically-charged bawdy poems and sexual ribaldry (where he himself is imagined as feminized) and circulated them within his coterie which was exclusively comprised of his close male friends, Ezra Pound, Wyndham Lewis, and Conrad Aiken throughout his life, as a way to keep himself inspired. To see a detailed interpretation of how "queer" is figured among Eliot's coterie, see Introduction and Chapter One of my dissertation titled *Granite and Rainbow: Queer Authority and Authorship in T. S. Eliot, W. B. Yeats, and Virginia Woolf*. To see how "queer" is coupled with homosexual desire and Caribbean blacks, see T. S. Eliot's suppressed "Columbo and Bolo Verses," recently published in their entirety after the death of Eliot's wife, Valery Eliot. Interestingly, "queer" is nowhere to be found in Eliot's major poems that brought him fame. For more information about this, see all volumes of *Letters of T. S. Eliot* published by Yale University Press.

2. Mrs. Dalloway is representative of the positive construction of "queer" in Woolf. In the novel, "queer" often emerges in Clarissa's consciousness to describe the rainbow aspect of life. It is also employed to depict the novel characters' modes of life that falls outside of the conventional norm. In her first notes to the novel, Woolf writes, "Mrs. D. seeing the truth. SS [Septimus Warren Smith] seeing the insane truth." (Woolf and Wussow 1996, 450). Here, Woolf highlights the fact that truth is only seen by those who are categorized as queer.

3. In the last two decades, this has been a trend among literary critics who write in the intersection between queer theory and literature, across periods and genres.

2. Data

87

Sentiments, however positive or negative, are relative and exist on a spectrum. For this reason, the corpora of James Joyce, D. H. Lawrence, F. Scott Fitzgerald, Gertrude Stein, and Katherine Mansfield each are included as a comparison group to validate my thesis that Woolf's use of "queer" is more positive. Although different in nationalities, these authors all wrote in Europe. There is also a fair amount of usage of "queer" in these authors' works. Authors like T. S. Eliot, whose usage of the term is only visible in private letters are excluded, although Eliot's works are rich with queer tensions and thus merit investigation from the perspective of queer theory. As the semantic meaning of "queer" had radically evolved in the first half of the 20th century, I also limited my selection to authors who wrote at roughly the same time as Woolf between the 1910s and the 1940s. Joyce, Fitzgerald, and Lawrence meet this condition. Stein and Mansfield are selected to verify that modernist authors' sentiments towards what was considered "queer" may not necessarily correlate with their gender, although their creative activities spanned slightly differently from Woolf's and with Stein, "queer" is visible mostly in *The Making of Americans*.

The very last point in the previous paragraph is particularly relevant to my choice of Joyce as part of comparison data. As sentiments around "queer" can also vary among male authors, I hoped to select male authors that are representative of the broader spectrum of the sentiment towards "queer." Joyce is an ideal candidate to accomplish this goal. As Joyce scholars and biographers suggest, in real life, Joyce's stance towards homosexuality remained fairly neutral; while Joyce was not above deriving entertainment from his homosexual friends, he was neither sympathetic nor unsympathetic to homosexuality (Norris 1994, 357). Nonetheless, what is intriguing about Joyce's works is that the centrality of feminized (often racialized and satirized) male characters and masculinized female counterparts amid an intense desire for homosocial and homoerotic affiliation emerges as one of the most visible themes. Joyce was also rebellious against the norm of his time and place. He condemns three Irish norms – family, Irish nationalism, and the Catholic Church – as stifling and detrimental to the development as an artist. I was not entirely certain about Fitzgerald's and Lawrence's sentiments, although plenty of existing research demonstrates that Lawrence writes more in a heteronormative convention while Fitzgerald views what he calls queer as an essential human condition:

Begin with an individual, and before you know it you find that you have created a type; begin with a type, and you find that you have created-nothing. That is because we are all queer fish, queerer behind our faces and voices than we want anyone to know or than we know ourselves. When I hear a man proclaiming himself an "average, honest, open fellow," I feel pretty sure that he has some definite and perhaps terrible abnormality which he has agreed to conceal-and his protestation of being average and honest and open is his way of reminding himself of his misprision (Fitzgerald 1989, 317).

All modernist authors' texts utilized in this project are drawn from Project Gutenberg

Australia.⁴ The data on Woolf contains most of her published novels, short stories, 128 and essays. Like the data on Woolf, data on Fitzgerald, Lawrence, and Mansfield each 129 consists of the corresponding author's major novels, short stories, plays, and essays. 130 For Joyce, I use three novels, *Dubliners*, *A Portrait of the Young Artist as a Young Man*, 131 and *Ulysses*, available on Project Gutenberg Australia. Similarly, for Stein, I use *The* 132 *Autobiography of Alice B. Toklas*, *The Making of Americans*, *Three Lives*, *Geography* 133 and *Plays*, available on the same site. As expected, there are differences in the size 134 of each author's corpus. Woolf's corpus accounts for 1,760,779 words in total, Joyce, 135 417,765, Fitzgerald, 615,126, Lawrence, 2,371,834, Stein, 699,562, Mansfield, 239,166. 136

The entire English fiction dataset from 1850 to 2000 (Google N-Grams eng-fiction-all) I 137 use for this study is from the dataset developed as part of the study titled "HistWords: 138 Word Embeddings for Historical Text" (Hamilton, Leskovec, and Jurafsky 2016). I use 139 HistWords' pre-trained word embeddings to extract the top 100 words closest to "queer" 140 for each decade, to compare them with Woolf's list. 141

3. Models and Methods 142

3.1. Associations Around Queer in Woolf and in Joyce, Fitzgerald, Lawrence, Stein, 143 and Mansfield 144

One way to measure Woolf's and others' sentiments towards the term "queer" is to 145 compile a list of the top 100 words semantically close to "queer" in the texts of each 146 author and compare their net positivity. Word embedding models (WEM) are optimized 147 for this task. Unlike topic models that map a text as a network of words based on co- 148 occurrences, word embedding models map a text as relationships between words so 149 that they "enable searching for spatial relations embedded in words," a framework, I 150 would argue, essential to close reading highlighting the particular, effected by close 151 attention to the relationship between words (Schmidt 2015). 152

To develop and train word embeddings specific to each author, each author's oeuvre 153 was combined into a separate single text file. While it is widely known to be effective to 154 adapt word embeddings trained on large collections of texts for predictive purposes, it is 155 worth highlighting again that it is each author's individual sentiment to a certain word 156 that emerges within the works of his or her creation that is being analyzed, and that 157 in literature as a peculiar genre, plethora of figurative words and styles are employed 158 and destruction of normative usage of language, experimented.⁵ If, for example, an 159 author consistently uses "queer," "miracle," and "loving" to describe, say, "pebbles," 160 these four words are closer in meaning and are thus placed closer within the space of 161 the particular author's corpus.⁶ For other authors, however, "pebbles" may not likely 162 be queer at all; they may likely be ordinary objects. This implies that, as Laura Burdick, 163

4. For a complete list of literary works used to create corpora data, see Appendix I: Complete List of Literary Texts Used in this Study from Project Gutenberg Australia.

5. With profusion of styles and the quantity of allusions, modernist authors' works are, in general, experimental and difficult to interpret, with Joyce's *Ulysses* being one of the most appropriate examples.

6. Here, I use this rather strange example to remind the reader that words are essentially signs.

Jonathan K. Kummerfeld, and Rada Mihalcea also aptly point out, word embeddings 164
change if different authors' texts or different collections of texts are used as input, as 165
words have different connotations when employed to discuss different topics (Burdick, 166
Kummerfeld, and Mihalcea 2018). This further suggests that no matter how precise or 167
sophisticated they are, using word embeddings trained from a large number of texts 168
that have nothing to do with each author might be risky. For this reason, I took the 169
path of developing and training word embeddings specific to each author, although this 170
choice inevitably raises a question about the relatively small size of individual authors' 171
corpus and methodology. 172

In developing and training word embeddings for each author's corpus, I chose Word2vec, 173
using Gensim, a Python library, which contains many variants of word embeddings 174
(Řehůřek n.d.). Specifically, Gensim's Word2vec is well maintained and takes the single 175
text file containing each author's corpus as input. My use of Gensim's Word2vec was 176
primarily to transform the authors' corpora into semantic spatial vectors, so I could 177
extract "queer"'s semantic vector and its top 100 closest words. 178

As I was working with limited amounts of texts, there may also be a dispute about the 179
choice of Word2vec, which generally requires more text input. To offset this concern, I 180
followed the best practice recommended by Ben Schmidt for those working with rel- 181
atively smaller corpora on Word2vec: "Run many iterations. A hundred, maybe. If 182
your model trains in less than a minute, it's probably no good" (Schmidt 2017). Experi- 183
menting with the size of vector dimensionality was also useful in getting meaningful 184
embeddings.⁷ Additionally, as it was uncertain how much the Word2vec training runs 185
across sentence boundaries, sentence triplets were used instead of single sentences to 186
minimize information loss. I ran 100 iterations in developing and training the models for 187
all authors except Mansfield, for whom I ran 200 iterations given the small corpus. The 188
training time for all models vary due to the corpus size. The longest training time was 8 189
minutes 19 seconds for Lawrence. The shortest training time was 1 minute 23 seconds 190
for Mansfield. Stop words were not removed from the compiled text files because in the 191
case of Word2Vec models, they can provide contextual information. The model can also 192
indirectly learn the sentence representation while feeding the context as the output or 193
input (Paul 2019). 194

By the time the models for each author had been created and trained, I was able to 195
extract the top 100 "synonyms" of "queer" from the corpus of each author. Yet, before 196
measuring the net positivity of each word, it was necessary to ensure that the words 197
identified as synonyms of "queer" were not dependent on one or two instances. I thus 198
ran ten models on different subsamples (sentence triplets) of the authors' corpora. As 199
each word is given its own vector (position) in the space of the corpus specific to a certain 200
author, I measured the distance from "queer" to positive words, and to negative words, 201
and ultimately, the difference between the distances. Since we were measuring distance 202

7. According to one entry from stackoverflow, in general, smaller vector dimensionality works better for smaller corpora. For smaller corpus, vector-dimensionality should be no more than the square-root of the count of unique words. To read more of this, visit <https://stackoverflow.com/questions/66267818/minimum-number-of-words-in-the-vocabulary-for-word2vec-models>

rather than similarity, the positivity of "queer" (net positivity) could be assessed as
such:

$$\text{Positivity of "queer" (net positivity)} = \text{negative distance (distance from negative)} - \text{positive distance (distance from positive)}$$

In terms of the positive and negative words, I used a list created by Bing Liu in 2005, which contain roughly 5,000 positive and negative words respectively (Liu, Hu, and Cheng 2005). I performed a t-test for the positivity of "queer" for Woolf and individual authors respectively to formally confirm the stability of the pattern I observed.

3.2. Associations around Queer in Woolf and in English Fiction from the 1850s to the 1990s

To situate Woolf's use of "queer" in the broader context of English fiction beyond literary Modernism and see whether Woolf was progressive with her ideas of queerness, I measured how different Woolf's associations are from other authors' associations across the collective history of English fiction from the 1850s to the 1990s, using "English Fiction (1800s-1990s) (from Google N-Grams eng-fiction-all)," one of the pre-trained word embeddings developed by William L. Hamilton, Jure Leskovec, and Dan Jurafsky for their project titled HistWords.⁸ As the vector of "queer" itself is missing in HistWords' dataset between the 1800s and the 1840s, this period was excluded.⁹ I took the path of extracting the top 100 words closest to "queer" from each decade from the 1850s to the 1990s and measured the sentiment of those the same way I did for my selected authors. In other words, for each decade, I measured the distance from "queer" to positive words, and to negative words, and calculated the difference between the distances, using Liu's lists. The positivity of "queer" (net positivity) was similarly assessed as (negative distance - positive distance.)

4. Results

4.1. Stability Test and P values from T-Tests

As can be seen from the visualization (Figure 1) on the next page, the stability tests for each author all returned positive results. Notably indeed, for Woolf, all 10 runs returned positive numbers.

8. To borrow Hamilton's description, the goal of the HistWords project is to facilitate quantitative research in diachronic linguistics, history, and the digital humanities. They release pre-trained historical word embeddings spanning from 1800 to 2000 for multiple languages - English, French, German, and Chinese. Embeddings constructed from many different corpora and using different embedding approaches are also included. To read more about this project or access their tools and datasets, visit their site titled HistWords: Word Embeddings for Historical Text on <https://nlp.stanford.edu/projects/histwords/>

9. This is what I see as a limitation of current pre-trained word embeddings. Both topic models and word embedding models tend to suppress low-frequency data, the very data that the close readers may want to explore. The fact that the token "queer" is entirely missing in the dataset of the first half of the 19th century reveals how the norm has operated in a discursive field to oppress those not considered to be the norm. Apparently, the term "queer" was in existence and in use in the early 19th-century.

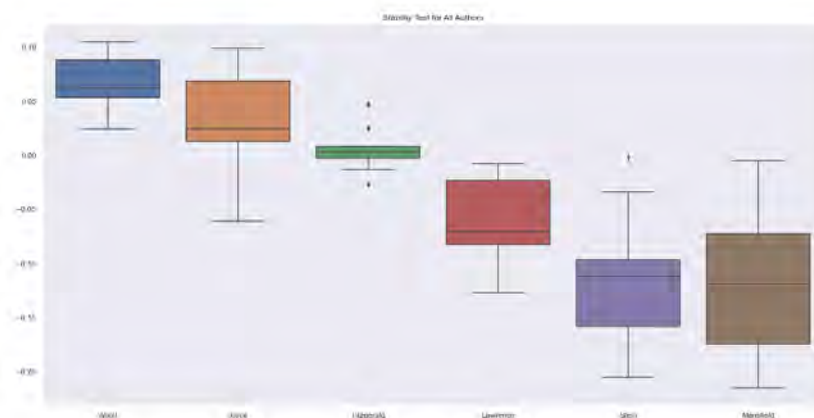


Figure 1: Box plot showing net positivity of the term "queer" for Woolf, Joyce, Fitzgerald, Lawrence, Stein, and Mansfield based on ten tests

It is not a big difference, but there's usually a lean toward the positive (90-100 percent), even when we run the model multiple times and compare all runs. This output shows that for Woolf, "queer" is always more positive than negative. For Joyce, the test outcome is consistently positive although it varies in degree. For Fitzgerald, it is mostly positive, although it is less positive than for Woolf. For Lawrence, Stein, and Mansfield, it is consistently negative and, like Joyce's data, there is a large variance.

The p values each from the t-tests for the net positivity in Woolf and other authors are as follows: Woolf and Joyce: 0.0882141903600226, Woolf and Fitzgerald: 0.0007540572823165007, Woolf and Lawrence: 1.2082467883608112e-07, Woolf and Stein: 5.961016119615059e-07, and Woolf and Mansfield: 9.993124513616901e-07. Except for the case of Woolf and Joyce, the p values are much smaller than 0.05. This shows us that in the cases of Woolf and Fitzgerald, Woolf and Lawrence, Woolf and Stein, and Woolf and Mansfield, the difference of means between these samples would not be likely to occur by chance if these samples were drawn from populations that actually had the same mean value. In short, we can claim with statistical confidence that "queer" is more positive in Woolf than it is in Fitzgerald, Lawrence, Stein, and Mansfield. However, we cannot claim with assurance that Woolf's usage of queer is always more positive than Joyce's, although the p value indicates some statistical significance, at 0.0882141903600226.

4.2. Associations around Queer from the 1850s to the 1990s from Histwords' Word Embeddings

The visualization (Figure 2 on the next page) reveals some interesting patterns about the associations around "queer" in the history of English fiction.

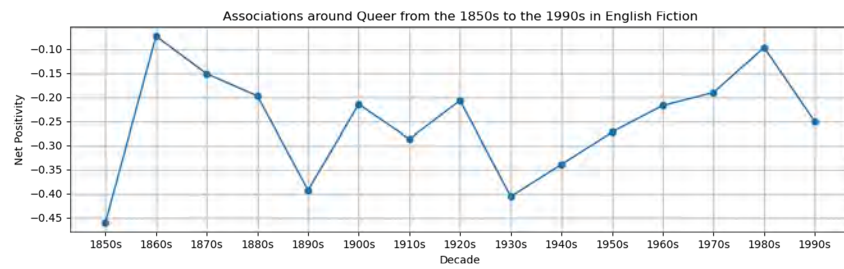


Figure 2: Visualization showing associations around the term "queer" from the 1850s to the 1990s in English fiction, which had always been negative

First, historically, the term "queer" consistently had negative connotations, indicated 254 by the negative net positivity numbers. Interestingly, there was a big shift towards 255 the positive in the 1860s. After that, until the 1890s, it consistently moved further 256 negative. We observe a consistent movement towards positive from the 1930s to the 257 1980s, although the general sentiment towards the term was still negative. Intriguingly, 258 however, there was a move back towards negative in the 1990s. Viewed together with 259 both Google Books Ngram Viewer's and bookworm: HathiTrust's data (Figures 3 and 260 4) in regard to the frequency of "queer" across English fiction between 1930s and 1990s 261 below, this movement merits investigation. 262

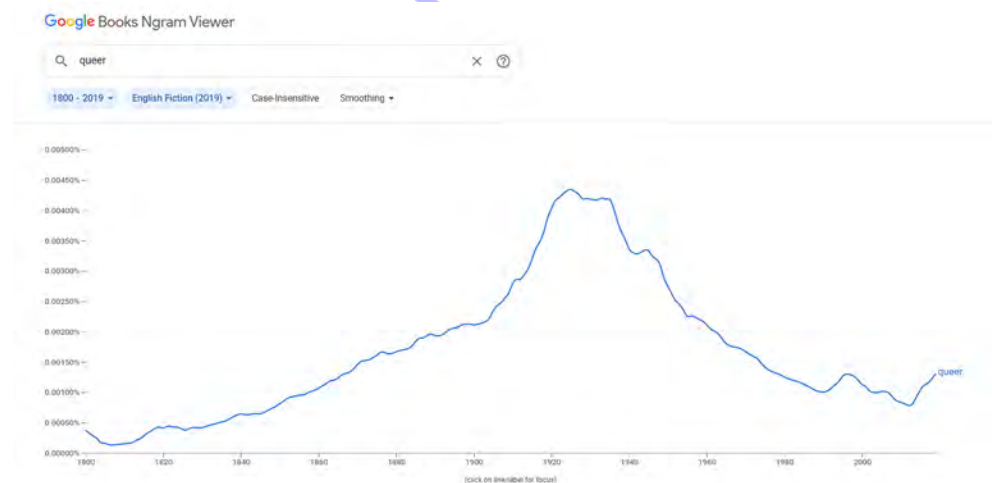


Figure 3: Google Books Ngram Viewer in regards to the frequency of "queer" across English fiction from 1800 to 2000 and beyond

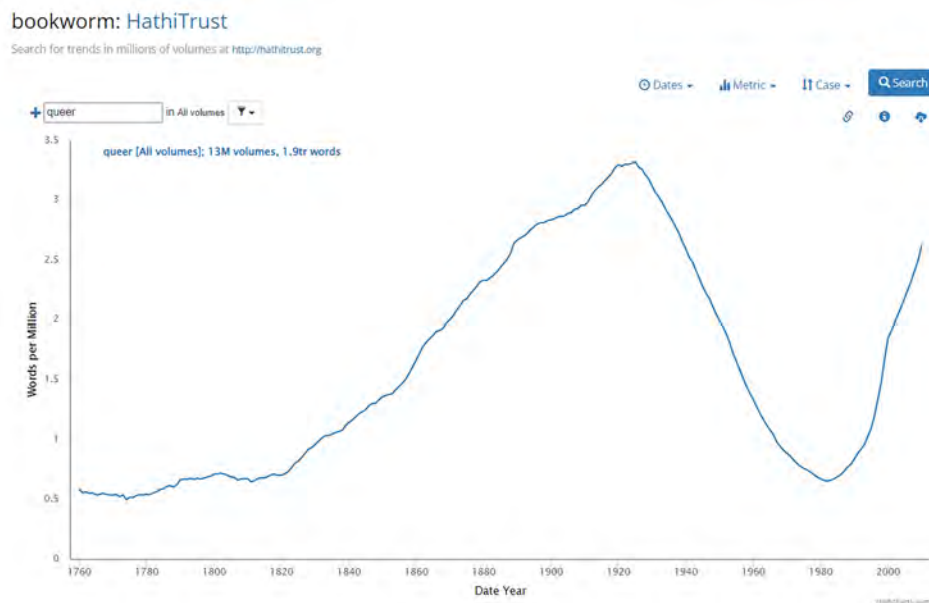


Figure 4: bookworm: HathiTrust's data in regards to the frequency of "queer" across English fiction from 1760 to 2000 and beyond

Going back to our discussion of Figure 2, "queer" became less and less frequently 263
represented in English fiction from the 1930s until its frequency increased back again 264
in the 1990s. That is to say, during this period, the frequency of "queer" and the net 265
positivity of "queer" moved in opposite directions. Without data on the 2000s and the 266
2010s, it is difficult to determine whether the move further negative in the 1990s was 267
part of a larger trend. It might be due to a conservative backlash against the LGBT rights 268
movements¹⁰ that became increasingly visible following the Stonewall riots of 1969, 269
which requires a separate investigation (Boag 2021). One claim I can still confidently 270
make, though, is that Woolf was more positive about the things that were viewed as 271
"out of the ordinary," and that her use of the term was progressive compared to its use 272
in English literary history from the 1850s to 1990s. 273

5. Discussion

Below is the list of the top 100 words closest to "queer" and their corresponding vectors 275
for Woolf from one model. Strikingly, the words identified as closest to "queer" are 276
not simply adjectives but include nouns and proper nouns. For example, Maisie and 277
Walsh are characters from Mrs. Dalloway, and Richard indicates Richard Dalloway who 278
appears both in Voyage Out and Mrs. Dalloway. The relative proportion of positive, 279
neutral, and negative words varies by model. 280

[('awfully', 0.385651171207428), 281

10. Several studies were conducted on the conservative backlash against the LGBTQ movements in the late 1890s and the 1990s, among which Peter Boag's "Gay and Lesbian Rights Movement" is one of the most representative. This phenomenon was universal across the globe.

('sized', 0.37115126848220825),	282
('young', 0.36801040172576904),	283
('suspected', 0.3652426600456238),	284
('absorption', 0.35453736782073975),	285
('posing', 0.3511541485786438),	286
('nice', 0.3499513566493988),	287
('maisie', 0.3435806930065155),	288
('oblivion', 0.3387223780155182),	289
('horrors', 0.32626646757125854),	290
('speeches', 0.323294073343277),	291
('evanescent', 0.31067633628845215),	292
('reputed', 0.3062557578086853),	293
('just', 0.3011806011199951),	294
('blotted', 0.30039259791374207),	295
('buzzing', 0.29972130060195923),	296
('dreaded', 0.2958635687828064),	297
('basins', 0.29578498005867004),	298
('perennial', 0.2948506474494934),	299
('assuring', 0.29381296038627625),	300
('booming', 0.2924637198448181),	301
('bent', 0.2912786900997162),	302
('hailed', 0.29016220569610596),	303
('tender', 0.28999805450439453),	304
('twice', 0.2898043096065521),	305
('lampshere', 0.28842049837112427),	306
('walsh', 0.2876507043838501),	307
('heavens', 0.28620970249176025),	308
('kissing', 0.2838011384010315),	309
('caen', 0.28266850113868713),	310
('pockets', 0.2819019556045532),	311
('painters', 0.2804553210735321),	312
('cocking', 0.2803754508495331),	313
('masculine', 0.2802823781967163),	314
('stogdon', 0.2793353199958801),	315
('exploded', 0.27923551201820374),	316
('comparison', 0.27765434980392456),	317
('deleterious', 0.27695566415786743),	318
('slang', 0.2765229046344757),	319
('squirrels', 0.27599194645881653),	320
('this', 0.2759091258049011),	321
('plans', 0.2755843997001648),	322
('significant', 0.2742382287979126),	323
('asquith', 0.2739396393299103),	324
('persian', 0.27252721786499023),	325

('negligently', 0.27113574743270874),	326
('tirade', 0.2710320055484772),	327
('armenians', 0.27095192670822144),	328
('invalids', 0.27037501335144043),	329
('omitting', 0.2695590555667877),	330
('proof', 0.2687683403491974),	331
('immovable', 0.2673490345478058),	332
('game', 0.2669960856437683),	333
('richard', 0.2666792869567871),	334
('convict', 0.26635780930519104),	335
('porous', 0.2660123109817505),	336
('fountains', 0.2658593952655792),	337
('that', 0.2658226490020752),	338
('affinity', 0.26559382677078247),	339
('sucked', 0.26312384009361267),	340
('cleanliness', 0.2629077136516571),	341
('contamination', 0.26289427280426025),	342
('about', 0.26252618432044983),	343
('happened', 0.26180094480514526),	344
('equitable', 0.2607996463775635),	345
('toy', 0.26030367612838745),	346
('vacancy', 0.26024338603019714),	347
('innocence', 0.2593679130077362),	348
('seeming', 0.25934281945228577),	349
('hovering', 0.2585065960884094),	350
('smiles', 0.255267858505249),	351
('hives', 0.25512927770614624),	352
('suits', 0.25401997566223145),	353
('roused', 0.25368526577949524),	354
('transferred', 0.25327783823013306),	355
('falsehood', 0.25286927819252014),	356
('accomplishment', 0.25260496139526367),	357
('hideous', 0.2523527145385742),	358
('anyhow', 0.25209107995033264),	359
('dog', 0.2512334883213043),	360
('different', 0.25067323446273804),	361
('albanians', 0.2502787411212921),	362
('craftsman', 0.24852287769317627),	363
('escaped', 0.24813099205493927),	364
('cheerless', 0.24807970225811005),	365
('ascertained', 0.24756474792957306),	366
('solicitous', 0.24712622165679932),	367
('judd', 0.24654719233512878),	368
('crabs', 0.24588340520858765),	369

(‘elms’, 0.24558645486831665), 370
 (‘mingling’, 0.24504920840263367), 371
 (‘dangled’, 0.244972825050354), 372
 (‘incompatible’, 0.2448458969593048), 373
 (‘ceremonial’, 0.2445395588874817), 374
 (‘withheld’, 0.2444373220205307), 375
 (‘groom’, 0.24392169713974), 376
 (‘hitching’, 0.24377071857452393), 377
 (‘diction’, 0.2436273694038391), 378
 (‘mentioned’, 0.24321354925632477), 379
 (‘tidy’, 0.24286895990371704)] 380
 381

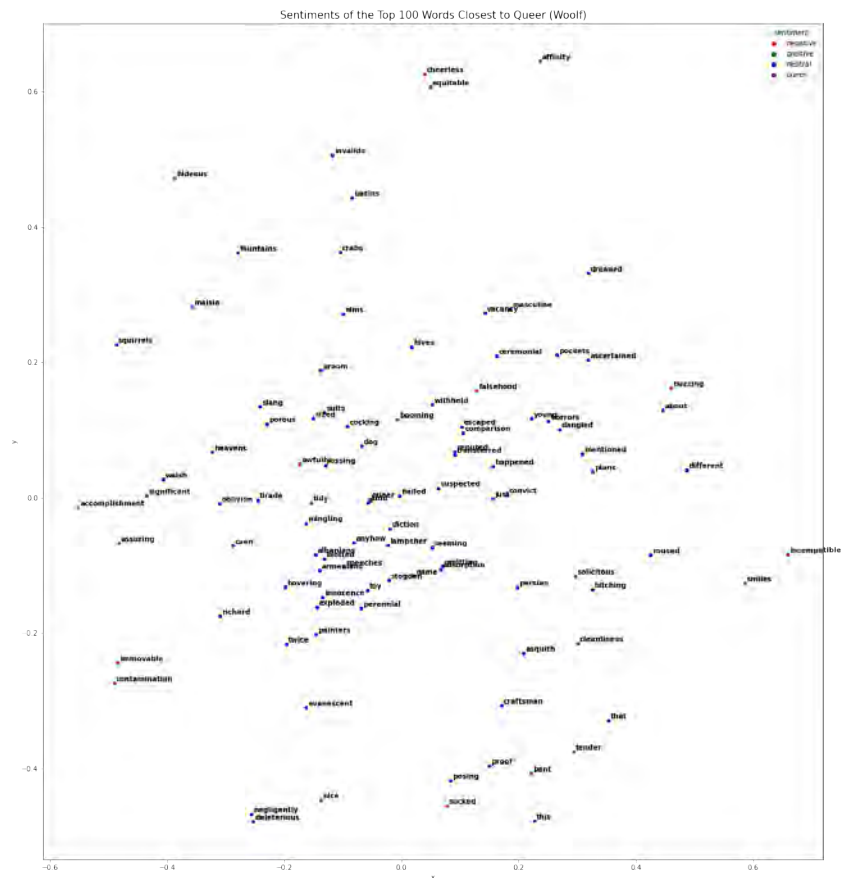


Figure 5: Plot of sentiments of the top 100 words closest to queer in Woolf's text

In the visualization above, words carrying a positive sense are plotted in green, words 382

with negative connotation, in red, and words that are neutral, that is, not present in Liu's 383
positive or negative words lists, in blue. It is worth noting that on Liu's lists of positive 384
and negative words, "queer" is categorized as negative. As that is unlikely to be the 385
case for Woolf, it is marked as a separate category on the graph in purple. The X and Y 386
axes are used to represent semantic vectors specific for each word. Thus, words plotted 387
closer to "queer" on the graph indicate their closer proximity to "queer" in meaning in 388
Woolf. Principal Component Analysis (PCA) was used to reduce the dimensions for 389
the plot. 390

We can see that, while most words are categorized as neutral, there are slightly more 391
positive words than negative ones: 12 vs. 10. This appears to be a small difference. Yet, 392
it is important to remember that what we measured earlier is the net positivity of the 393
words closest to "queer." This means that in the ten samplings drawn from Woolf's 394
corpus, positive words always outnumber negative words among the top 100 words 395
identified as closest to "queer," regardless of the proportion of neutral words. Another 396
potentially important discovery we can make is that, as I mentioned earlier, the model 397
identifies a significant number of proper nouns and nouns as words close to "queer." 398
Proper nouns and nouns are extremely important in literary analysis, as they are the 399
locus in which our interpretation of literature is anchored, whether it is about themes, 400
tropes, characters, or sentence structures. 401

Undeniably, for Joyce as well, "queer" is consistently used positively. For Fitzgerald, 402
6 models return positive outcomes. For interested readers, the plots of Joyce's and 403
Fitzgerald's top 100 words closest to "queer" are provided in Figures 6 and 7. Similar to 404
Woolf's list, we see nouns and pronouns present in Joyce's and Fitzgerald's lists. One 405
can notice, however, that the proportion of positive and negative words decreases in 406
both Joyce and Fitzgerald, compared to Woolf. How 100 individual terms are deployed 407
around "queer" in the texts of Joyce and Fitzgerald used in this research, along with 408
their idea of (hetero)normativity proper, will require a separate in-depth exploration. 409
A point that should be noted here is that in *Ulysses*, "queer" is often deployed within 410
the male protagonist Leopold Bloom's stream of consciousness as a reference to the 411
intricacies of life, which resist a facile, binary categorization. Above all, in the case of 412
Joyce, that all ten models return positive outcomes testifies Norris' depiction of Joyce 413
as unbiased with the matter of homosexuality to a certain degree. Norris argues that, 414
not being one of his own personal predilections, homosexuality is an aspect of human 415
behavior to which Joyce did not devote a great deal of attention (Norris 1994, 357). 416
Indeed, Joyce views homosexuality as a product of the social system, rather than as a 417
personal trait that should be abhorred. In his essay "Oscar Wilde: The Poet of Salome," 418
written approximately around the same time as *A Portrait of the Artist as a Young* 419
Man, Joyce describes Wilde's homosexuality as the "logical and inevitable product" 420
of sexual "secrecy and restrictions" and "unhappy mania" endemic to British public 421
schools (Valente 2004, 215). Similarly, Colleen Lamos views the matricidal fantasies 422
that often emerge throughout *Ulysses* as the author's defensive gestures that attest to 423
the violent consequences of the modern disavowal of same-sex desire (Lamos 1998, 15). 424

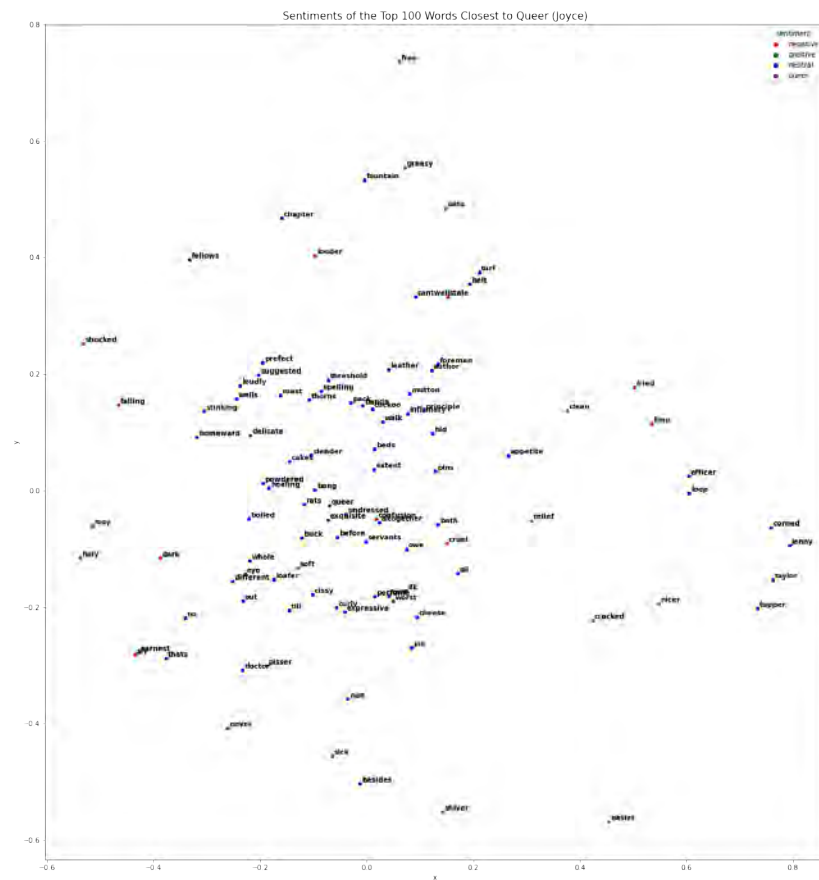


Figure 6: Plot of sentiments of the top 100 words closest to queer in Joyce's text

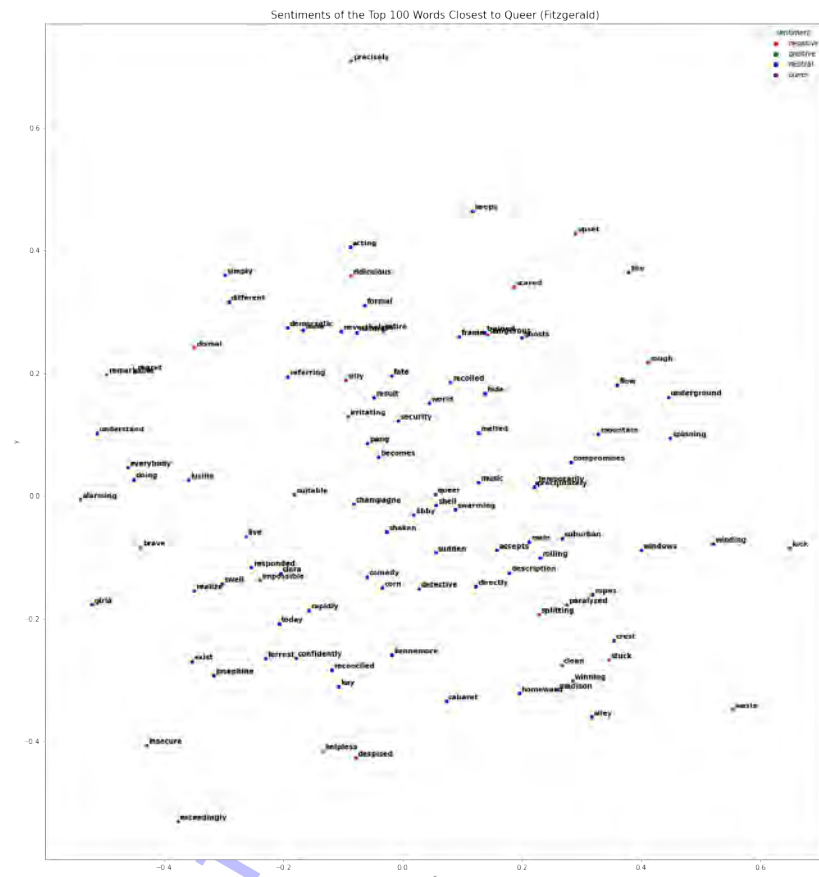


Figure 7: Plot of sentiments of the top 100 words closest to queer in Fitzgerald's text

On the other hand, the Lawrence plot from one model, seen in Figure 8 on the next page, 425 shows us that compared to Woolf and Joyce, there are a significantly greater number of 426 negative associations around "queer." 427

Intriguingly, while running multiple iterations of the model, I could see "savage" and 428 "barbaric" several times as one of the top 100 terms closest to "queer" for Lawrence. This 429 is a meaningful discovery given that Lawrence is notorious for having written in the 430 heteronormative convention and for associating whatever is at odds with conventional 431 femininity with the primitive. Indeed, in Lawrence's narrative strategy, what Gayle 432 Rubin terms as "traffic in women" strongly operates. In other words, in Lawrence, the 433 feminine trope is deployed only to strengthen the bond between males or celebrate 434 conventional ideas of masculinity and femininity, with *Women in Love*, *Sons and Lovers*, 435 and "The Fox" being only a handful of examples (Rubin 1975, 180). In *Women in Love*, 436 for example, the sisters, Ursula and Gudrun – particularly, Ursula's certainty, idealism, 437

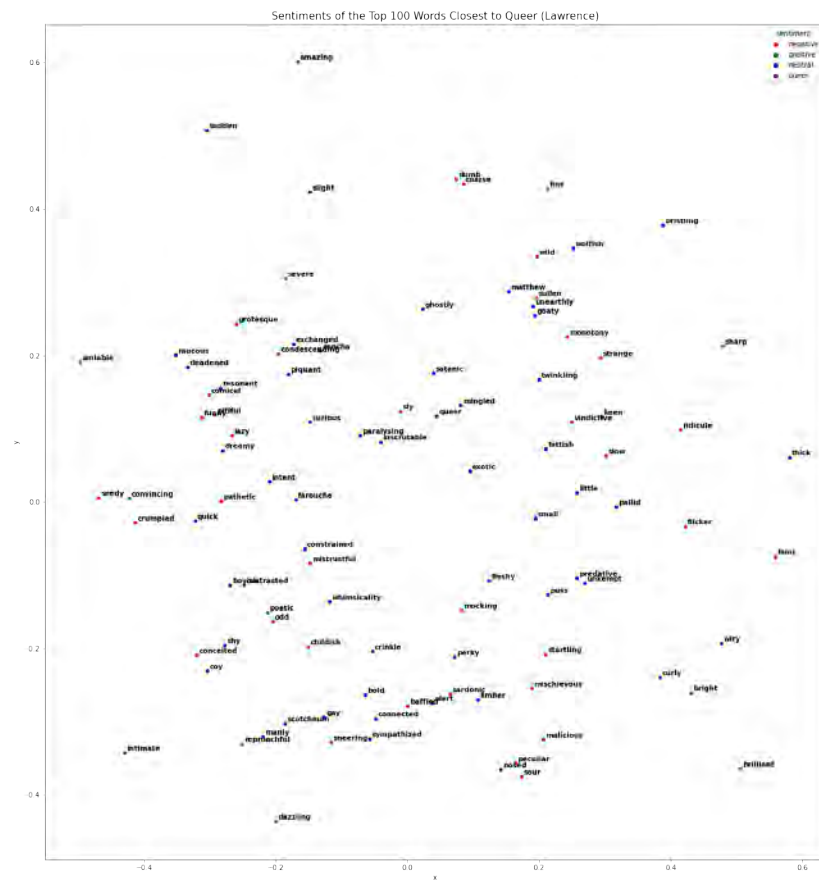


Figure 8: Plot of sentiments of the top 100 words closest to queer in Lawrence's text

and womanliness and Gudrun's sexuality and rebellious personality – are deployed 438
to ultimately strengthen the bond between Birkin and Gerald. The novel ends with 439
Birkin's mourning over the loss of Gerald who freezes to death after his violent fight 440
with Gudrun. 441

What is so intriguing in this narrative strategy is the construction of Gudrun's unruly 442
nature, along with Gerald's cruelty and death drive, as savage and destructive. After 443
all, in the novel, Gudrun is depicted as an artist known for her primitive, savage art. 444
Indeed, Marianna Torgovnick is correct in pointing out that in Lawrence, there are two 445
versions of the primitive (Torgovnick 1991, 159). The first is a feminine version: the 446
primitive as "dangerous," "irrational," "something to be feared," and "the idealized 447
noble savage" (Torgovnick 1991, 159). The second is a masculine version: the primitive 448
as "regeneration" (Torgovnick 1991, 159). The emergence of "savage" and "barbaric" 449
as words closest to "queer" in Lawrence, along with Lawrence's negative sentiment 450
towards "queer," thus demonstrates that "queer," for Lawrence, is associated with the 451
negative version of the primitive – the feminine –, which is shaped by his frustration 452
with disappearing Western values – the conventional idea of masculinity and femininity 453
where the former is associated with regeneration, the latter, reproduction– with the 454
arrival of the modern (Torgovnick 1991, 153). 455

For interested readers, Stein's and Mansfield's plots are also provided below. How and 456
why each corpus exhibits this pattern, other than what I mentioned earlier about Stein's 457
tendency to align class and nationality with "queer," requires a separate investigation. 458
Nonetheless, the outcome that the net positivities of these women authors' corpus are 459
the lowest suggest that female authors do not necessarily have a positive sentiment 460
towards what is considered "out of the ordinary," that the author's gender does not 461
necessarily correlate with their sentiment towards "queer." 462



Conf

f sentiments of the

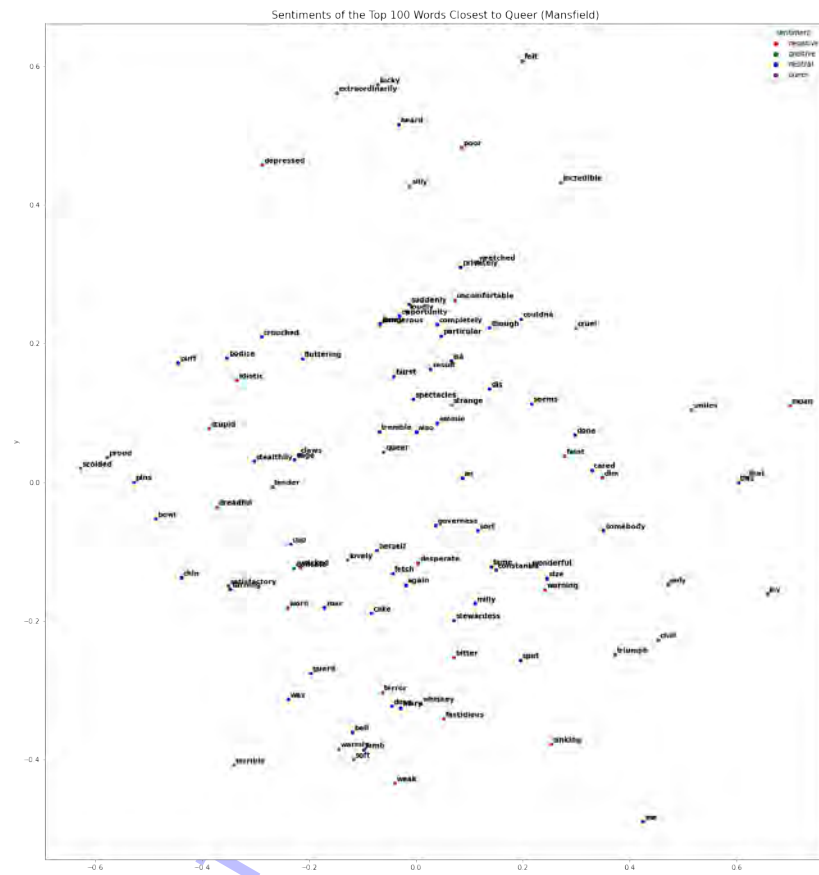


Figure 10: Plot of sentiments of the top 100 words closest to queer in Mansfield's text

6. What "Queer" Represents in Woolf

463

Here, I take the approach of a literary critic, to validate my outcome with close reading, 464
to argue that Woolf, as a renowned feminist writer and queer author, had a keen sense 465
of how the norm manifests itself as various forms of power to oppress those who do not 466
conform to it. Unlike the male authors who were spoiled for choice, Woolf grappled 467
with the absence of a strong female tradition and keenly sensed herself in conflict with 468
the masculinist, heteronormative climate of the British Empire and as permanently 469
in exile. A series of medical treatments she had received due to her recurrent mental 470
and physical illness, albeit a disaster in her personal life, offered her a powerful tool to 471
interrogate the tyranny of the norm as a form of social repression (Lee 1997, 186). 472

As a form of resistance, Woolf deploys "queer" to create desires, personalities, and 473

relationships – bodily, aesthetic, and epiphanic – that exist outside of the paradigmatic markers dictated by normativity. In her diary entry on December 21, 1925, Woolf employs "queer" to mean both bodily consummation and esthetic fulfillment after spending her first night with Vita Sackville-West at Long Barn:

There is her maturity & full breastedness... there is some voluptuousness about her. But then she.....so lavishes on me the maternal protection which, for some reason, is what I have always most wished from everyone.... I shall be hung about with trailing clouds of glory from Long Barn wh. always disorientates me & makes me more than usually nervous: then I am—altogether so queer in some ways. One emotion succeeds another (Woolf 2018, 11654).

In Mrs. Dalloway, "queer" is employed in a sympathetic and lovable note to describe the truth behind her characters who are viewed as failures by the social norm. Earlier, we saw "Maisie" plotted as one of the top 100 words closest to "queer" in Woolf's corpus, along with "invalids." Maisie is a low-class woman from Edinburgh, who appears very briefly at the beginning of Mrs. Dalloway. What is specifically remarkable is the tangible link that Woolf establishes between the term "queer" and those who were parceled into the category of "queer" in the oppressive British interwar regimes, through Maisie Johnson's stream of consciousness in her first encounter with London.

They seemed queer, Maisie Johnson thought. Everything seemed very queer. In London for the first time, come to take up a post at her uncle's in Leadenhall Street, and now walking through Regent's Park in the morning, this couple on the chairs gave her quite a turn; the young woman seeming foreign, the man looking queer..... For she was only nineteen and had got her way at last, to come to London; and now how queer it was, this couple she had asked the way of, and the girl started and jerked her hand, and the man—he seemed awfully odd; quarrelling, perhaps; parting forever, perhaps; something was up, she knew; and now all these people (for she returned to the Broad Walk), the stone basins, the prim flowers, the old men and women, invalids most of them in Bath chairs—all seemed, after Edinburgh, so queer (Woolf 1981, 26).

Remarkably, in this short passage, "queer" is employed five times in total. Here, Maisie Johnson calls Septimus Warren Smith, a veteran of World War I and his Italian wife Rezia each queer and then all the people she comes across in Regent's Park: "the old men and women, invalids most of them in Bath chairs." For Maisie Johnson, "queer" is a term that binds all these people who appear out of time and out of place – invalids sitting in Bath chairs, the foreign (Rezia), and the awfully odd and mad (Septimus), who suffers from shellshock. Remarkably, as the story unfolds, readers also notice that a link between "queer" and a same-sex desire is tellingly made in Septimus when his close relationship with his wartime officer Evans is repeatedly highlighted. Ultimately, Septimus commits suicide in defiance of Dr. Holmes and Sir. William Bradshaw's desire to "straighten" his "shell shock," his madness. Here, in his triumphant choice of death

over treatment, we see "being queer" is also equated with a willing choice and a vehicle for resistance.

Maisie, Septimus, and Rezia are not the only characters associated with "queer" on a sympathetic note. In numerous instances throughout Mrs. Dalloway, the characters' impregnable queerness – Clarissa's bisexuality, Richard's anxiety over his masculinity, the adventurous queer child within Peter Walsh and Elizabeth, and Miss Kilman's misandry and obsession with food – is directly described as "queer" or finds its way out as spatial metaphors in its askew relation and stubborn resistance to normativity. Earlier, we saw "Richard," a politician who is also Clarissa's husband, and Peter "Walsh," Clarissa's friend, identified as terms close to "queer" in Woolf's corpus. Strikingly, in Woolf's manuscript of Mrs. Dalloway, Richard emerges as a queer trope out of place: "Richard had all the marks of that queer breed" (Woolf and Wussow 1996, 75). Indeed, it is repeatedly implied throughout the novel that politics does not suit Richard's simple character and love for nature. Throughout the novel, Richard's nostalgia for Norfolk's sky and movements of grass and breeze is constantly placed in opposition to his awkwardness in London. When Richards unreluctantly visits a jewelry shop with Hugh Whitebread on Conduit street on their way back from Lady Bruton's luncheon in Mayfair, for instance, he feels old and "torpid," unable to "think or move."

With Peter, Woolf goes further. Like Maisie Johnson, Peter Walsh sees through other people's queerness. Elizabeth's bisexuality is remarkably hinted at by Peter's observation: "She's a queer-looking girl, [Peter] thought, suddenly remembering Elizabeth as she came into the room and stood by her mother" (Woolf 1981, 56). Woolf also constructs Peter as a rebellious queer child who takes pleasure in cruising through the city and refuses to conform to normative developmental stages, by stubbornly holding onto his "youth." Notably, as we already saw, the term "young" is identified as one of the closest terms to "queer" in Woolf's model.

& Peter Walsh, thought Peter, I haven't felt so young for years, thought Peter; & yet he was no child could have had yet it was not youth, young, this feeling of irresponsible adventure; rather it was <not> a child's feeling: but a man's; & it & not a normal man's but.... a queer man's.... who after being wound himself about with ties & responsib[ilities] duties, burdens, & privileges, suddenly perceives their vanity <&> his freedom, as a child.... but only for a moment. <second> (Woolf and Wussow 1996, 15).

Another instance where the queer child in Peter is tellingly evoked is when Clarissa meets Peter after 30 years, she thinks "Exactly the same....; the same queer look; the same check suit; a little out of the straight his face is, a little thinner, dryer, perhaps, but he looks awfully well, and just the same" (Woolf 1981, 40).

We can locate another important theme that runs through Woolf's works when looking closely at a subset of words that models on Woolf each identify as a word close to "queer": "painter," "dressmaker," "craftsman," and "archaeologist." Indeed, in A Room of One's Own, To the Lighthouse, Three Guineas, The Years, and "Craftsmanship,"

Woolf deploys "queer" to imagine women author tropes – novelist, painter, archeologist, and dressmaker – who work to uncover the truth beyond the established "archives and repositories of knowledge" by reading between the lines of "patriarchal discourse" (Kaufman 2018, 333).

Elsewhere, the term "queer" is evoked to represent a beautiful harmony made out of incompatible things in life: "the voices of birds and the sound of wheels chime and chatter in a queer harmony, grow louder and louder and the sleeper feels himself drawing to the shores of life" (Woolf 1981, 69). In *Orlando*, "queer" is used to imply the spontaneous, private, and fictional side of all sort of things with respect to their factual, public, normative sides:

Nature, who has played so many queer tricks upon us, making us so unequally of clay and diamonds, of rainbow and granite, and stuffed them into a case, often of the most incongruous, for the poet has a butcher's face and the butcher a poet's; nature, who delights in muddle and mystery, so that even now (the first of November 1927) we know not why we go upstairs (Woolf 1928, 58).

It is notable to note that "diamonds" and "rainbow" are words identified as close to "queer" in certain model iterations on Woolf. "Diamonds" is also a recurring trope in *To the Lighthouse*, which signifies security and privacy out of sync with publicity. So is "rainbow" in *Orlando* and "New Biography," which is directly placed in sharp opposition to cold facts, public school, and diplomacy. For Woolf, "queer" is almost always placed in fierce confrontation with normativity.

7. Conclusion

As the above analyses and discussion demonstrate, I was able to statistically prove my thesis that "queer" is more positive than negative for Woolf, and that Woolf's idea of "queerness" was progressive, a thesis that would otherwise rely solely on interpretation. The top 100 words closest to "queer" that the model on Woolf extracts turned out to be also extremely useful, when used to aid close reading of the author's works. I hope my paper helps identify a space where data science and the Humanities can be brought together to enrich Digital Humanities.

8. Appendix I: Complete List of Literary Texts Used in this Study from Project Gutenberg Australia

Fitzgerald, F. Scott. "The Adjuster." 1926.
 ——. "The Complete Pat Hobby Stories." 1940-41.
 ——. *Collected Stories*.
 ——. *The Great Gatsby*. 1944.
 ——. "The Guest in Room Nineteen." 1937.

———. "Hot and Cold Blood." 1926.	594
———. "Presumption." 1926.	595
———. "The Pusher-in-the-Face." 1925.	596
———. "Shaggy's Morning." 1935.	597
———. "A Snobbish Story." 1930.	598
———. "Strange Sanctuary." 1939.	599
———. Tender is the Night. 1933.	600
———. "Three Acts of Music." 1936.	601
———. "Too Cute for Words." 1936.	602
Joyce, James. Dubliners. 1914.	603
———. A Portrait of the Artist as a Young Man. 1916.	604
———. Ulysses. 1922.	605
Lawrence. D. H. Aaron's Rod. 1922.	606
———. Amores: Poems. 1916	607
———. Birds, Beasts and Flowers. 1923.	608
———. Bay: A Book of Poems. 1919.	609
———. The Captain's Doll. 1923.	610
———. Collected Short Stories.	611
———. A Collier's Friday Night. 1934.	612
———. The Daughter-in-law. 1912.	613
———. David. 1926.	614
———. England My England. 1922.	615
———. Etruscan Places. 1932	616
———. Fantasia of the Unconscious. 1922	617
———. The Fight for Barbara. 1912.	618
———. The Fox. 1923.	619
———. Kangaroo. 1923.	620
———. Lady Chatterley's Lover. 1928.	621
———. Look! We Have Come Through! 1917.	622
———. The Lost Girl. 1920.	623
———. The Ladybird. 1923.	624
———. The Man Who Died. 1929.	625
———. The Married Man. 1926.	626
———. The Merry-go-round. 1912.	627
———. Mornings in Mexico. 1927.	628
———. New Poems. 1918.	629
———. The Plumed Serpent. 1926.	630
———. The Prussian Officer and Other Stories. 1914.	631
———. The Rainbow. 1926.	632
———. St Mawr. 1925.	633
———. Sea and Sardinia. 1921.	634
———. Sons and Lovers. 1913.	635
———. Tortoises. 1921.	636
———. Touch and Go. 1920.	637

———. The Trespasser. 1912.	638
———. Twilight in Italy. 1916	639
———. The Virgin and the Gypsy. 1930.	640
———. The White Peacock. 1911.	641
———. The Widowing of Mrs. Holroyd. 1914.	642
———. The Woman Who Rode Away and Other Stories. 1928.	643
———. Women in Love. 1920.	644
Mansfield, Katherine. Bliss and Other Stories, 1920.	645
———. The Doves' Nest, and Other Stories, 1923.	646
———. The Garden Party and Other Stories, 1922.	647
———. In a German Pension, 1911	648
———. Something Childish and Other Stories (1924)	649
Stein, Gertrude. The Autobiography of Alice B. Toklas, 1933.	650
———. The Making of Americans, 1925.	651
———. Geography and Plays, 1922.	652
———. Three Lives 1909.	653
Woolf, Virginia. Between the Acts. 1941.	654
———. Collected Essays.	655
———. Collected Short Stories.	656
———. The Common Reader. 1925.	657
———. The Common Reader Second Series. 1935.	658
———. The Death of the Moth and Other Essays.	659
———. Flush: A Biography. 1933.	660
———. The Haunted House and Other Short Stories.	661
———. Jacob's Room. 1922.	662
———. The Moment and Other Essays. 1947.	663
———. Monday or Tuesday. 1921.	664
———. Mrs. Dalloway. 1925.	665
———. Night and Day. 1919.	666
———. Mrs. Dalloway. 1925.	667
———. Orlando: A Biography. 1928.	668
———. A Room of One's Own. 1929.	669
———. To the Lighthouse. 1927.	670
———. Three Guineas. 1938.	671
———. The Voyage Out. 1915.	672
———. Walter Sickert: A Conversation. 1934.	673
———. The Waves. 1931.	674
———. The Years. 1937.	675

9. Data availability 676

Data can be found here: https://github.com/heejoungs/woolf_queer 677

10. Software availability 678

Software can be found here: https://github.com/heejoungs/woolf_queer 679

11. Acknowledgements 680

In conducting this research, I received tremendous support from Ted Underwood, 681
Professor of Information Science and English at the University of Illinois at Urbana- 682
Champaign, who volunteered to serve as a technical consultant. This research has 683
been shaped by our discussion of close reading and computational approaches, and 684
his generous guidance on testing the statistical significance of my thesis, as well as his 685
help with code. I am especially thankful for situating my idea on Woolf and literary 686
Modernism to a larger literary context. 687

12. Author contributions 688

Heejoung Shin: Conceptualization, Writing – original draft 689

Ted Underwood: Technical Consultant 690

References 691

- Barker, Meg-John and Julia Scheele (2016). *Queer: A Graphic History*. London: Icon. 692
- Boag, Peter (2021). “Gay and Lesbian Rights Movement”. In: *Oregon Encyclopedia*. doi: 693
[https://www.oregonencyclopedia.org/articles/gay_lesbian_rights_move](https://www.oregonencyclopedia.org/articles/gay_lesbian_rights_movement/#conservative-backlash-1980s-1990s) 694
[ment/#conservative-backlash-1980s-1990s](https://www.oregonencyclopedia.org/articles/gay_lesbian_rights_movement/#conservative-backlash-1980s-1990s). 695
- Burdick, Laura, Jonathan K. Kummerfeld, and Rada Mihalcea (2018). “Factors Influenc- 696
ing the Surprising Instability of Word Embeddings”. In: *NAACL-HLT*. doi: [https:](https://ai.engin.umich.edu/2018/07/23/word-embeddings-and-how-they-vary/) 697
[//ai.engin.umich.edu/2018/07/23/word-embeddings-and-how-they-vary/](https://ai.engin.umich.edu/2018/07/23/word-embeddings-and-how-they-vary/). 698
- Fitzgerald, F. Scott (1989). *The Short Stories of F. Scott. Fitzgerald*. New York: Scribner. 699
- Hamilton, William L., Jure Leskovec, and Dan Jurafsky (2016). “HistWords: Word 700
Embeddings for Historical Text”. In: *The Stanford NLP Group*. doi: [https://nlp.sta](https://nlp.stanford.edu/projects/histwords/) 701
[nford.edu/projects/histwords/](https://nlp.stanford.edu/projects/histwords/). 702
- Houlbrook, Matt (2005). *Queer London: Perils and Pleasures in the Sexual Metropolis, 1918-* 703
1957. Chicago: University of Chicago Press. 704
- Kaufman, Mark David (2018). “True Lies: Virginia Woolf, Espionage, and Feminist 705
Agency”. In: *Twentieth-Century Literature* 64.3, pp. 317–346. 706
- Lamos, Colleen (1998). *Deviant Modernism: Sexual and Textural Errancy in T. S. Eliot, James* 707
Joyce, and Marcel Proust. Cambridge: Cambridge University Press. 708

- Lee, Hermione (1997). *Virginia Woolf*. New York NY: Alfred A. Knopf. 709
- Liu, Bing, Minqing Hu, and Junsheng Cheng (2005). "Opinion Observer: Analyzing and Comparing Opinions on the Web". In: *Proceedings of the 14th International World Wide Web conference (WWW-2005)*. doi: <https://www.cs.uic.edu/~liub/publications/www05-p536.pdf>. 710 711 712 713
- Norris, David (1994). "The "Unhappy Mania" and Mr. Bloom's Cigar: Homosexuality in the Works of James Joyce"". In: *James Joyce Quarterly* 31.3, pp. 357–373. doi: <https://www.jstor.org/stable/25473572>. 714 715 716
- Paul, Arindam (2019). "Do I Have to Remove Stop Words in in Order to Train Word Vectors Using Word Embedding?" In: *Quara*. doi: <https://www.quora.com/Do-I-have-to-remove-stop-words-in-order-to-train-word-vectors-using-word-embedding>. 717 718 719 720
- Řehůřek, Radim (n.d.). In: *Gensim: Topic Modeling for Humans* (). doi: <https://radimrehurek.com/gensim/index.html#>. 721 722
- Rubin, Gayle (1975). *The Traffic in Women: Notes Toward a Political Economy of Sex*. Toward an Anthropology of Women. Monthly Review. ISBN: 9780853453994. 723 724
- Schmidt, Ben (2015). "Vector Space Models for the Digital Humanities". In: *Ben's Bookworm Blog*. doi: <http://bookworm.benschmidt.org/posts/2015-10-25-Word-Embeddings.html>. 725 726 727
- (2017). "Word2Vec Workshop". In: *benschmidt.org*. doi: http://benschmidt.org/slides/Word2Vec_Workshop.html. 728 729
- Torgovnick, Marianna (1991). *Gone Primitive: Savage Intellectuals, Modern Lives*. Chicago IL: University of Chicago Press. 730 731
- Valente, Joseph (2004). *Joyce and Sexuality*. The Cambridge Companion to James Joyce. Cambridge University Press. ISBN: 9780521545532. 732 733
- Woolf, Virginia (1928). *Orlando: A Biography*. Orlando: Harcourt. 734
- (1981). *Mrs. Dalloway*. Orlando: Harcourt. 735
- (2018). *Virginia Woolf: The Complete Works*. MyBooks Classics. 736
- Woolf, Virginia and Helen Wussow (1996). *Virginia Woolf "The Hours": the British Museum Manuscript of Mrs. Dalloway*. New York: Pace University Press. 737 738

On the multi-dimensional differences between German
poetry of realism and modernism

Anton Ehrmanntraut  ¹

Thora Hagen  ¹

Fotis Jannidis  ¹

Leonard Konle  ¹

Merten Kröncke  ²

Simone Winko  ²

1. Institut für Deutsche Philologie, Justus-Maximilians-Universität Würzburg, Würzburg.

2. Seminar für Deutsche Philologie, Georg-August-Universität Göttingen, Göttingen.

Keywords:

short text, similarity, poetry,
modernism, realism

Licenses:

This article is licensed under:



Abstract. This study contributes to the ongoing discussion on how to operationalize text similarity for the purposes of computational literary studies by defining, justifying theoretically and employing a multi-dimensional text model. Additionally, we evaluate a set of strategies to implement this model for very short texts like poetry using a range of methods from weighted sparse vectors up to very recent neural sentence embeddings based on annotations of emotions, genre and similarity. And finally, we show the relevance of using such a complex text model by applying the best method to a research question about the development of early modernism in German poetry. While we can confirm some important hypotheses from literary studies, we are also able to differentiate or relativize others. In particular, our findings suggest that the change from realism to modernism was, contrary to what many researchers assume, an evolutionary transition rather than a revolutionary „rupture“.

1. Introduction

1

This paper pursues two equally important goals: First, to find a suitable state-of-the-art method to model and analyze text similarity for poetry, and second, to contribute to the field of literary studies by studying the transition from realist to modernist poetry using the concept of similarity. The perception of similarity between texts is the basis for the construction of many literary terms like genre, author, or, as in our case, period. Grouping texts according to these terms usually presupposes that these texts have something in common and that these groups can be distinguished via these commonalities from other texts. Though the concept of similarity is ubiquitous in the practice of literary studies it has seldom been analyzed explicitly. One conspicuous exception are scholars in Comparative Studies who reflected on this term as part of their

2

3

4

5

6

7

8

9

10

11

discipline defining practice (e.g. Corbineau-Hoffmann 2013). Similar attempts to model text similarity beyond the aspect of content have also been undertaken in computational linguistics (e.g. Bär, Zesch, and Gurevych 2011). So one of the major contributions of this paper is our attempt to bring these discussion threads together. But while it is possible to discuss these dimensions on a very abstract level, it is not possible to evaluate them on the same level. When we talk about structural aspects of a text, we look at very different elements depending on the genre we look at: speaker, stage directions, dramatis personae, etc. for drama, or stanza, verse, rhyme, etc. for poetry. Therefore, in order to discuss the phenomenon not only theoretically, but also to be able to apply it practically – and that means, above all, to include an evaluation method - it is more productive to limit the task to one genre - in our case, poetry.

The second goal of our research is to provide a broad foundation for a literary history of the beginnings of modernism. In the last years, we assembled a corpus of German poetry consisting of poems from realist and modernist anthologies. We are analyzing this corpus under the perspective of whether we can contribute to the discussion about the transition from realism to (early) modernism. We are using these period terms, as is the custom nowadays in literary studies, as useful constructions. That means: On the one hand it is understood that real breaks and disruptions are very rare and that history can be better understood as an evolutionary, gradual process with many small changes at each step. On the other hand, we assume that this process is not happening at the same speed all the time and that many of the changes in one time segment show some commonalities. Specifically, we will use the concept of similarity to describe the changes between the texts from the different corpora.

We structure our paper as follows: In a theoretical section, we first develop a four-dimensional model of textual similarity for poetry (chapter 2). We then describe our corpora; mainly the digitized anthologies of the poetry of realism and early modernism mentioned above (chapter 3). A selection of these poems was previously manually annotated with a hierarchical system of emotion labels. Within the context of our work, a subset of this selection was then additionally annotated using the dimensions of similarity described in the theoretical section. The following section discusses how each of these four dimensions can be measured in poetry (chapter 4). Poetry presents specific computational challenges even for semantics, a relatively traditional dimension of similarity. The main issue poses the shortness of the texts. Semantic similarity, which is usually modeled by using weighted terms to locate a document in vector space, does not work reliably on short texts. Additionally, working with poetry entails having to adapt to its specific language. This includes a high percentage of figurative speech, which makes the analysis of semantic similarity especially difficult, and also a high percentage of archaic words and expressions. For each of the four dimensions of similarity, we discuss and evaluate different methods to measure them using our poetry corpus in a first step. Among our methods used are traditional sparse document vectors, short dense feature vectors, and dense document embeddings, created either by computing them from token vectors or by using the recently proposed approach for sentence embeddings. In a second step, we adapt the best-performing models to each of the four dimensions (chapter 4.4). In the last section, we employ our final models from

this two-step approach to assess the degrees of similarity and difference between realist and modernist poetry (chapter 5). In particular, we take up three specific research questions from literary studies and discuss our results with respect to the predominant hypotheses within the field. These questions are:

1. How does naturalist poetry relate to realism and modernism?
2. How homogeneous are realist and modernist poems?
3. How revolutionary is early modernism?

In summary, this study contributes to the ongoing discussion on how to operationalize text for computational literary studies by defining, theoretically justifying, and employing a multi-dimensional model of similarity. Additionally, we evaluate a set of strategies to implement this model for poetry using a range of methods from weighted sparse vectors up to the recent neural sentence embeddings based on extensive annotations of emotions, genre, and similarity. And finally, we show the relevance of using such a complex text-based model by employing the best method to provide new input for the continued research on the development of early modernism in German poetry.

2. Theoretical considerations

As far as we can see, in literary studies, text similarity has been discussed mainly by Comparative Studies, where the concept of ‘comparison’ has been closely linked to ‘similarity’ (e.g. Zelle 2005). There seems to be a consensus that comparison is only possible on the basis of similarity in some specific aspects. Though principally many different aspects have been and can be used to compare literature, some have been established as especially useful for the study of literature. Corbineau-Hoffmann (2013), for example, groups them under three headings:

- [I.]Content (1. theme, 2. motifs, 3. settings, 4. characters, 5. concepts) Text-organization (1. narrative/description, 2. poetry/prose, 3. style levels, 4. instances of speech, 5. discourse) History (1. influences, 2. epochs, 3. other arts, 4. sciences, 5. genre).

While the first two groups are aspects of a text, the last group refers to typical contexts, often established again by analyzing groups of texts. To avoid the recursive loop hidden here, we focus on the two first aspects, ‘content’ and ‘text-organization’. It is important to note that these are open lists. There are other interesting aspects, but the ones mentioned are often used when people compare literature. The terms grouped under ‘content’ can be seen as parts of text semantics in general. A text has a theme, or there are specific motifs in a text, but usually, the meaning of text is more than each of these, it encompasses all of them. The terms grouped under ‘text-organization’ on the other hand cover quite heterogeneous aspects – even if you substitute the more common ‘form’ for it. In our experience, especially the term ‘style’ is hard to subsume under the same dimension as other text-organizational aspects.

Semiotics and linguistics support this position as they also distinguish between form and style (Nöth 2008; Sandig 2006), and the three aspects – content, structure, and style – are also distinguished in one of the very few attempts in computational linguistics to model text similarity (Bär, Zesch, and Gurevych 2015). We propose to add one dimension which can only be subsumed with difficulties under one of the three headings and which is usually highly important, especially for literature and especially for poetry, which has been defined as the prototypical medium to express subjective feelings: emotion.¹ Content, Form, Style, and Emotion are the four dimensions of similarity which we will use to describe the relations between texts. From the perspective of this study, it is more useful to explicate the dimensions via operationalizations and examples rather than ‘exact’ definitions. To this end, the annotation guidelines (see section 3) list specific components that make up the dimensions. Content consists of components such as theme, character, or setting; form is operationalized primarily through stanza structure, meter, and rhyme; style, in contrast, refers to components such as register or metaphor; and for emotion, we consider, among other things, the extent to which emotions are represented and their polarity. In further studies, these components could be analyzed individually and be integrated into an even more complex model of text similarity. The heterogeneity of the four dimensions will have a direct influence on the inter-annotator agreement and the performance of any machine learning model trained to detect these aspects automatically. From a theoretical perspective, it is unclear how the dimensions relate to each other, or in the language of statistics, how much they correlate. Winko (2003), for example, assigns the aspect ‘linguistic shaping of emotions’ via the aspect ‘presentation of emotions’ to what is called ‘style’ in our model, while she assigns it to content via the aspect ‘thematization of emotions’. From this perspective, a relatively high correlation of emotion with content and style is to be expected.

3. Corpus and Annotation

The corpus is a collection of anthologies of contemporary poetry from the two epochs ‘realism’ and ‘modernism’.² The collections contain poems that the anthologists, i.e. contemporary experts in poetry, consider to be particularly typical, outstanding, or representative among other aspects. From the large amount of poetry anthologies in both epochs, the corpus was compiled³ according to the following **criteria**: The collections contain **contemporary** poetry, have **no thematic restrictions**, and are all aimed at a **general audience** rather than a particular target group. The criteria minimize the risk that thematic constraints or specific addressee orientation could influence the poem selection as systematic factors. The corpus contains texts by both canonical and non-canonical authors. We call authors ‘canonical’ if they are frequently mentioned in recent literary histories. For early modernism, this applies to Stefan George, Hugo von

1. Why emotion is a dimension of its own for the analysis of text is discussed in Winko (2003).

2. Since this epoch is characterized by a multitude of literary trends, the more neutral label ‘turn of the century around 1900’ is preferred in literary studies. We choose the term ‘modernism’ because the anthologies we include claim to present modern poetry. In the following, ‘modernism’ always means ‘early modernism’, i.e. literature before expressionism.

3. For our corpus selection we used Günter Häntzschel’s comprehensive bibliography (c.f. Häntzschel 1991).

Hofmannsthal, Arno Holz, Else Lasker-Schüler, and Rainer Maria Rilke. 131

(1) sub-corpus ‘Realism’: The first sub-corpus consists of 7 anthologies with German poems from the realist epoch: Prutz 1859; Polko 1860; Kneschke 1865; Willatzen 1875; Bern 1877; Moltke 1882; Avenarius 1882. The poems included in the anthologies cover the period under study, 1850 to 1880. Some of the anthologies, but especially Elise Polko’s widely distributed collection, also contain some poems written before the period of study; these have been excluded. This sub-corpus consists of 3039 poems by a total of 484 different authors. 132 133 134 135 136 137 138

(2) sub-corpus ‘Modernism’: Of the 941 anthologies of German-language poetry published in first edition between 1885 and 1912 (cf. Häntzschel 1991, pp. 587-589), twelve anthologies meet the selection criteria: Arent 1885; Bierbaum 1893; Bierbaum 1894; Tille 1896; Gemmel 1898; Jacobowski 1899; Renner 1899; Benzmann 1904; Bethge 1905; Bonsels et al. 1905; Federmann 1908; P. Friedrich 1911; Huch 1911. They all claim to contain ‘modern poetry’. This sub-corpus consists of 2882 poems by a total of 361 authors. 139 140 141 142 143 144 145

We annotated 1278 poems from both sub-corpora for emotion and thematic genre. Thematic genres such as love poetry or nature poetry provide information about the content of the poems.⁴ The annotated emotions are not the readers’ emotions, but rather the emotions expressed in the text itself. The annotators used a list of 40 discrete emotions which we categorized into 6 groups, inspired by the emotion hierarchy in Shaver et al. 1987: love, joy, agitation/surprise, anger, sadness, and fear. First, emotions and genres were annotated independently by two annotators, then they merged annotations manually into a consensus annotation. Their agreement before creating the consensus annotation, measured with γ (Mathet, Widlöcher, and Métivier 2015), was 0.6445 for individual emotions, 0.7491 for the emotion groups, and 0.69 Krippendorff’s alpha (Krippendorff 2011) for the thematic genres. 146 147 148 149 150 151 152 153 154 155 156

Additionally, we annotated the similarity of the poems.⁵ The task was not to annotate absolute similarities (“These two poems are not at all/a little/very similar”), but relative similarities (“Poem A is more similar to poem B than poem C”), which is much easier. For each triple of poems, the annotators had to judge for each similarity dimension (content, form, style, and emotion) and for a comprehensive ‘overall’ category whether the focus poem was more similar to the one on the left, to the one on the right, or equally (dis)similar to both. The annotation guidelines specify for each dimension which components should be taken into consideration, e.g. stanza structure, rhyme, meter, and text length in case of the formal dimension, and which of these aspects are typically most important. Nevertheless, the annotators ultimately had to weigh the components on a case-by-case basis, which required considerable literary expertise. We annotated 470 triples, consisting of a total of 866 poems. One constraint for the 157 158 159 160 161 162 163 164 165 166 167 168

4. As the annotation is still ongoing to cover more poems, the entire corpus and a detailed report on the annotation guidelines for emotions and genre will be published at a later date.

5. The annotation guidelines can be found here: https://github.com/cophi-wue/jcls2022-poem-similarity/blob/main/annotation_guidelines_text_similarity.pdf.

selection of the triples was that the poem length had to be quite short due to technical prerequisites.⁶ In addition, we selected triples for which we expected a strong similarity of the middle text with either the left or right text based on formal features such as text length or previous annotations of thematic genres and emotions. This second constraint ensured that the annotators could deal with reasonably clear cases. There were 400 triples covering both constraints available in our annotated poems. 70 triples were additionally annotated without similarity expectations. Each triple was annotated by at least two people.

The agreement, measured with Krippendorff's alpha, was 0.53 for content, 0.68 for form, 0.44 for style, 0.32 for emotion, and 0.48 for overall. Possible reasons for the differences in agreement are that the dimensions with lower agreement are more dependent on interpretation or that the weighting of the components is more ambiguous in their case. An experiment showed, however, that three annotators who created a consensus annotation after annotating 60 triples were able to increase their agreement when annotating another 30 triples from 0.49 to 0.63 on content, from 0.48 to 0.69 on emotion, from 0.32 to 0.41 on style, and from 0.45 to 0.68 on overall (only the agreement on form deteriorated from 0.77 to 0.71, but still remained high). Since the creation of consensus annotations seems to significantly improve the annotation quality, we plan to create consensus annotations for all triples in the future. Until then, for the triples without consensus annotations, we will only use annotations that the majority of annotators agree with. In the evaluation of the following section, we also omit all annotations that the majority of annotators found that 'the middle text is equally (dis)similar to both the left and the right text'.⁷ That leaves us with 346 usable annotations for content, 388 for form, 331 for style, 359 for emotion, and 381 for overall, with every annotation stating that the middle text is more similar to either the left text or the right text.

Some of the similarity dimensions correlate strongly with each other, according to the annotations. The 'overall' dimension correlates most strongly, especially with content and style. This is understandable since the annotation of the 'overall' dimension is usually based on annotations of the other similarity dimensions. Another relevant correlation exists between content and style. The most independent dimension is form, whose correlation with the other dimensions is the weakest.

4. Dimensions of the Similarity of Poems

Measuring the similarity of poems along the dimensions discussed above poses two challenges: first, the shortness of the texts makes it difficult to apply well-established approaches with high reliability. Research in natural language processing has proposed a set of methods for the measurement of short text similarity (Prakoso, Abdi, and Amrit

6. The length of poems is bound to a maximum of 124 sentence-piece tokens used as input for paraphrase-xlm-r-multilingual-v1

7. More precisely, for each triple and similarity dimension, we calculate the mode of the annotation results. We use 'The middle text is more similar to the left text' (from now on: 'left') as the final annotation if the mode is 'left', but also if it is 'left' and at the same time 'The middle text is equally (dis)similar to both the left and the right text'. The same is true in reverse for annotations on the right. All other annotations are discarded.

Content	1	0.24	0.68	0.42	0.82
Form	0.24	1	0.22	0.43	0.48
Style	0.68	0.22	1	0.4	0.76
Emotion	0.42	0.43	0.4	1	0.66
Overall	0.82	0.48	0.76	0.66	1
	Content	Form	Style	Emotion	Overall

Figure 1: Pearson correlation in annotated dimensions (majority vote).

2021) usually complementing the texts with other sources which compensate for the lack of information in the text themselves. But research on text similarity, in general, focuses on the ‘content’ aspect. So the second challenge lies in finding methods that can be used to model the other dimensions.

Overviews of the research on short text similarity classify the methods in four groups: string-based, corpus-based, knowledge-based, and hybrid-based (Prakoso, Abdi, and Amrit 2021; Gomaa and Fahmy 2013): 1) String-based methods use only the word or character tokens to create a representation of the text. We use *tfidf* and *mfw*. 2) Knowledge-based methods use an external knowledge base like WordNet. Our two models *features-formal* and *features-emotional* can be seen as variants of this approach. 3) Corpus-based methods use an external corpus to create information-rich representations, nowadays usually word embeddings: *FastText*, *Glove*, *GBert*. Additionally, we experiment with document embeddings using different sentence embedding methods: *XLM-R*, *mpnet*, *MiniLM*, *cross-en-de-roberta*. The drawback of this approach is that we are limited to an input of 127 tokens, but it is reported to be the best representation for short texts. 4) Hybrid approaches, combining some of the strategies outlined above.⁸

The small number of poems we had annotated under the perspective of similarity made it inadvisable to use the typical finetuning approach. Instead, we opted for broader testing of how different text representations are able to mirror the results from our annotations, select the best representations, and then tweak the vector spaces with similarity learning based on our dimension annotations. So in sections 4.1 to 4.3, we introduce the different models we were able to use and their evaluation based on our similarity annotations. In section 4.4, we apply similarity learning to the best performing models.

8. Bär, Zesch, and Gurevych 2015 distinguish between compositional measures, which usually “compute pairwise word similarity between all words, and aggregate the resulting scores to an overall similarity score” (Bär, Zesch, and Gurevych 2015, p. 5), and non-compositional measures, which project the texts into a shared space like the vector space model (Salton and McGill 1983). We concentrate here on the latter.

4.1. Models

229

We evaluate the following embeddings, which can be roughly categorized into more 230
simple baselines on the one hand, and some more complex embeddings derived from 231
sophisticated deep-learning language models on the other. The baseline embeddings 232
are defined as follows: 233

TFIDF- $\{1000,10000,20000\}$: Poems are represented by a vector, where the dimensions 234
correspond to the 1000 etc. most frequent terms in our corpus. Each individual vector 235
component is the relative term frequency of that term in the poem, weighted by the 236
inverse document frequency. 237

MFW- $\{100,200,500,1000\}$: Defined like the embedding above, but the term frequencies 238
are z-standardized for each term over all poems. 239

Features-Formal: Poems are represented by a vector of the following four formal fea- 240
tures: stanza count, verse count, word count, average stanza length in verses - each 241
z-standardized over all poems. 242

Features-Emotional: Each poem is embedded with a vector of its verse-level relative 243
frequency of shaver emotions (see section 3). These emotions derive either from anno- 244
tations or, if no annotations are available, predicted by a machine learning model.⁹ 245
The following deep-learning embeddings are derived from pre-trained static type-based 246
embeddings: 247

$\{\text{FastText}, \text{GloVe}\}$ - $\{\text{mean}, \text{median}, \text{meannorm}, \text{sif}\}$: For each term in the poem (minus 248
stopwords), we obtain the embedding vector for that term with FastText (trained on 249
the German OSCAR corpus with $d = 1536$ as proposed by Ehrmanntraut et al. 2021, 250
resp. a GloVe model with $d = 300$ provided by Deepset.¹⁰) Finally, on that set of vectors, 251
we compute the arithmetic mean (resp. median, resp. meannorm (Ehrmanntraut et al. 252
2021), resp. arithmetic mean weighted by smooth inverse frequency (Arora, Liang, and 253
Ma 2017)) to obtain a single vector for a particular poem. 254

Similarly, the following embeddings are derived from the output of BERT, which gener- 255
ates vectors for each token, but also takes into consideration the textual context of the 256
entire input sequence. 257

GBERT-lastlayer- $\{\text{mean}, \text{median}, \text{meannorm}\}$: For a particular poem, we plug in the 258
tokenized poem into GBERT_{Base} (Chan, Schweter, and Möller 2020), the currently best 259
performing German BERT model. BERT then computes a contextualized output vector 260
(i.e., the output of the last layer) for each token. We now aggregate all vectors by taking 261
the arithmetic mean (resp. median, resp. meannorm), just like above. This results in a 262
vector with 768 dimensions. 263

GBERT-alllayers- $\{\text{mean}, \text{median}, \text{meannorm}\}$: Defined just like above, except that we 264
not only consider the final output vector, but the outputs of all layers. That is, for each 265
token, we concatenate the input embedding with the 12 Transformer outputs to derive a 266
vector with $d = 13 \times 768$. Then, like above, we aggregate this sequence of token vectors 267
into a single vector for a particular poem. 268

In contrast, the following embeddings result from pre-trained language models follow- 269

9. The model achieves a performance of 0.73 (f1 score).

10. <https://www.deepset.ai/german-word-embeddings>

ing a Sentence-BERT-architecture, as proposed by Reimers and Gurevych 2019; Reimers and Gurevych 2020.

paraphrase-XLM-R: This model is a multilingual XLM-RoBERTa model that is fine-tuned to imitate Sentence-BERT-paraphrases, as described by Reimers and Gurevych 2020. We let paraphrase-XLM-R interpret our poems as sentences, which outputs a vector representation for each poem with $d = 768$. Note that in the case of paraphrase-XLM-R and all following Sentence-BERT models, fine-tuning was only performed for input sequences no longer than 126 SentencePiece tokens. Therefore, we also restrict our evaluation of these models to poems that are no longer than 126 SentencePiece tokens. (These are 29% of all poems in our corpus.)

paraphrase-mpnet, paraphrase-MiniLM, cross-en-de-roberta: Similarly, these are pre-trained Sentence-BERT models trained on a wide variety of sentence pair datasets and parallel multilingual data. Specifically, we use publicly available variants paraphrase-multilingual-mpnet-base-v2, paraphrase-multilingual-MiniLM-L12-v2 provided by Reimers and Gurevych 2019, and cross-en-de-roberta-sentence-transformer provided by T-Systems online. Again, we use these on poems of length ≤ 126 SentencePiece tokens to obtain vector representations with $d = 768$.

4.2. Evaluation Setup

Evaluating the embeddings as described above requires us to formulate a task that probes each embedding space for its ability to represent certain dimensions of (dis)similarity of poems via their distances in that particular embedding space, taking into consideration and comparing against the human annotations. As the embeddings define no particular distance function, we evaluate every embedding with each of the following three distance functions: Euclidean (L2), Manhattan (L1), and Cosine Distance. We opted to replicate our prompts for the annotators by formulating a binary classification problem on a particular dimension of similarity, and checking if the model can replicate the majority vote. Note that these votes either take the value ‘annotated left’ or ‘annotated right’. Assume some embedding space and some distance function d fixed. For some annotated triple, let *left*, *anchor* and *right* denote the corresponding vectors in that embedding space. Now, we make the following prediction:

- 1. Predict ‘annotated left’ if $d(\text{anchor}, \text{left}) < d(\text{anchor}, \text{right})$, i.e., *left* is closer to *anchor* than *right*.
- Otherwise, predict ‘annotated right’.

To compare the true majority annotations with the predicted ones, we use the *balanced accuracy* (arithmetic mean over the recall of both classes, cf. Grandini, Bagli, and Visani 2020) as our metric. Note that the random ‘no skill’ classifier has a balanced accuracy score of 0.5.

We remark that variations on the above operationalization are possible as well, particularly if we do not omit cases where the majority of annotators chose ‘The middle text is equally (dis)similar to both the left and the right text’. However, while experimenting we observed that when including this third class ‘same’ in the operationalization, the

balanced accuracy significantly drops. (We made the different balanced accuracies comparable by rescaling to the range $1/(1 - \text{\#classes})$ to 1, so that performance at random scoring is always at 0.) We suspect that this difference in performance is caused by the complexity of the triples that were labeled with ‘annotated same’: Human annotators agree on the features which make them classify a text as ‘more similar to the focus text’. But the label ‘same’ is given when neither of both comparison texts shows obvious similarities to the focus text, but that does not imply that the comparison texts have any features in common; they can be different to the focus text in very diverse ways. In particular, we experimented with the following two variations of the original operationalization:

(a) Probe whether the embedding can predict ‘annotated equally (dis)similar’ vs ‘annotated left or right’ by evaluating $|d(\text{anchor}, \text{left}) - d(\text{anchor}, \text{right})| < \epsilon$ against some optimal decision boundary ϵ .

(b) In a 3-class classification setup, probe whether the embedding space admits a classification using an optimal symmetric decision boundary ϵ . That is, predict ‘annotated left’ when $d(\text{anchor}, \text{right}) - d(\text{anchor}, \text{left}) > \epsilon$ (left is closer to anchor than right by at least ϵ). Symmetric, when $d(\text{anchor}, \text{left}) - d(\text{anchor}, \text{right}) > \epsilon$, predict ‘annotated right’. And otherwise, when $|d(\text{anchor}, \text{left}) - d(\text{anchor}, \text{right})| \leq \epsilon$, predict ‘annotated equally (dis)similar’.

As outlined above, variant (a) is solved with lower balanced accuracy than the original operationalization throughout all embeddings and variant (b) with even lower accuracy.

4.3. Results

The results show for all dimensions except ‘form’ a clear increase with the complexity of text representation: Word Embeddings are better than sparse representations - with dynamic embeddings based on BERT showing a better performance than static embeddings - and sentence embeddings are better than word embeddings. The best sentence embedding is showing an acceptable performance, especially if the cosine is used. As almost all the strategies of text representation, which we applied here, have been developed with the main focus on the semantic aspect, it is not too surprising that the best model is the best in all dimensions. The one big exception is form. Using only a very small set of features is enough to match the annotations. Discussions with the annotators revealed that they usually based their decision on a very small set of observations. The best model is *paraphrase-mpnet*. To evaluate all German sentence embedding models,¹¹ which are available at this moment, we use the rd. 9.000 sentences of the Sick dataset (Marelli et al. 2014) which we had translated into German with DeepL. Our results show *paraphrase-XLM-R* (correlation with human annotations: 0.82) slightly ahead of *paraphrase-mpnet* (0.8165), which is why we include these two models and the best model based on static word embeddings (*FastText-mean*) in the next step.

11. The multilingual models in Huggingface’ sentence transformers; see <https://huggingface.co/sentence-transformers>.

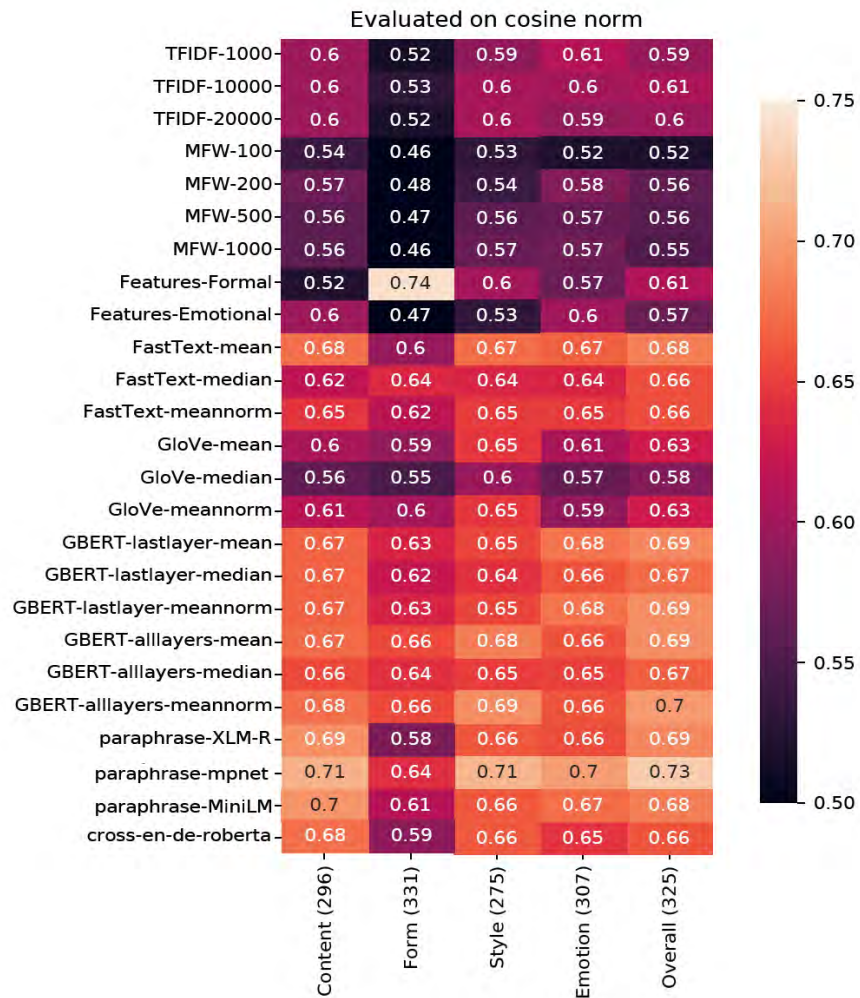


Figure 2: Balanced Accuracy Score for each model and dimension. Numbers on the x-axis indicate class support. For information on results with other distance metrics see the appendix.

4.4. Similarity Learning

349

To adapt the text representations to the specific textual dimensions (content, form, style, 350 and emotion), we additionally apply similarity learning. The goal of this step is to 351 learn a transformation of the vectors presented in the previous chapter that allows for 352 better reproduction of the annotation. We use a siamese neural network (Bromley et al. 353 1993) for this purpose, which we modeled following the maaten network structure from 354 (Szubert et al. 2019). The base model consists of three dense layers (500, 500 and 2000 355 neurons) each followed by a normalization activation function (see Klambauer et al. 356 2017) and dropout. The input for the network consists of our annotated poem triples. 357 Regardless of the original size of its vector representation, each poem is transformed 358 into a space with 128 dimensions. The loss, and hence the optimization objective of the 359 network, is to maximize the difference between the focus text and the negative example 360 while also minimizing the difference between the focus text and the positive example, 361 i.e. the text which has been annotated as being more similar to the focus text. In short: 362

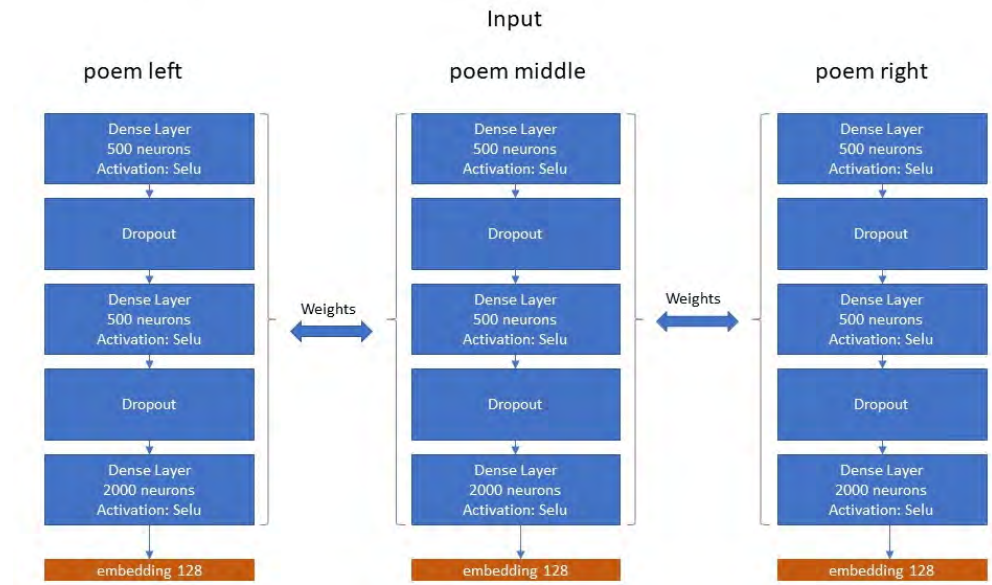


Figure 3: Architecture of the Siamese Neural Network used for Similarity Learning

in Euclidean distances ($\text{dist}(\text{anchor}, \text{negative}) - \text{dist}(\text{anchor}, \text{positive})$).¹² Learning rate decrease is bound to a *reduce on plateau* mechanism, which leads to strong performance gains compared to more common choices like constant or time-based decrease rates. The network’s performance is measured via the amount of correctly identified positive examples (accuracy).

Model	Content	Form	Style	Emotion	Overall
paraphrase-XLM-R	.69→.81	.58→.76	.66→.79	.66→.76	.69→.79
paraphrase-mpnet	.71→.75	.64→.68	.71→.71	.70→.74	.73→.74
FastText-mean	.66→.77	.59→.67	.65→.72	.66→.74	.66→.72
Formal-Features	-	.81→.81	-	-	-

Table 1: Similarity Learning results (Accuracy in 10-fold cross-validation). Format: best performance before similarity learning (see Fig. 2) → performance afterwards.

4.5. Discussion

With our two-step approach, we are able to achieve good results for a complex task. It is probably open to discussion whether the restriction to 127 input tokens is acceptable compared to the small gain in performance. Future work will either improve on the input size or find a reliable way to compute representations for longer texts. Using one representation for three of the four aspects in the first step made us ask whether the representations after the second step are actually different. The correlations of distances (Fig. 4) show a high correlation between content and the category ‘overall’, but only moderate positive correlations between content and style, content and emotion, or style

12. Triplet margin loss

and emotion. In other words, the vector space was attuned to the specific dimension by similarity learning. The close relationship between ‘content’ and ‘overall’ was already noticed by the annotators.

Content	1	0.6	0.62	0.58	0.79
Form	0.6	1	0.66	0.47	0.64
Style	0.62	0.66	1	0.55	0.71
Emotion	0.58	0.47	0.55	1	0.64
Overall	0.79	0.64	0.71	0.64	1
	Content	Form	Style	Emotion	Overall

Figure 4: Pearson correlation of distances in vector space after similarity learning.

It is unclear to us, why the different embeddings show significantly different improvements in the second step (mean values): 0.126 for *paraphrase-XLM-R*, 0.026 for *paraphrase-mpnet*, and 0.08 for *FastText-mean*; on what factors does this capability for improvement depend? Which training data and training regime for the sentence embeddings enables the text representation to be adaptable to the text dimensions beyond content? The results from Figure 2 show that the best results are obtained using language models with transformer architecture and that they increase even more if those have previously been fine-tuned for sentence similarity. With the additional adaptation by similarity learning, we now perform a third tuning step of representations created this way. A next step would be instead of using the frozen output vectors of those networks, to include the network in the learning process and model the similarity learning as a fine-tuning step. Likewise, we should add another layer of pertaining before similarity learning and perform a domain adaptation (Gururangan et al. 2020) to our corpus.

5. (Dis)similarity between the poetry of realism and the poetry of modernism

5.1. Hypotheses from Literary Studies

In the following, we continue a discussion in German literary studies about the relationship of poems of realism to those of early modernism and the special position of naturalistic poetry in this development. We hope to contribute to this discussion by a mix of explorative methods and hypothesis testing. To enable the latter, we will condense positions in the debate into three hypotheses related to this transformation.

Hypothesis 1: The poetry of naturalism, as represented in the anthology “Moderne Dichter-Charaktere”, is predominantly traditional rather than modernist. The question of where exactly naturalism can be located between realism and modernism has been debated many times. In this context, the anthology “Moderne Dichter-Charaktere”, which is part of our corpus, is considered central to naturalist poetry. The anthology’s introductions emphatically assert the novelty and revolutionary character of the texts (especially Conradi 1885: I-III). Research, on the other hand, is mainly of the opinion that these statements are exaggerated and that the poetry of the anthology is, on the whole, traditional (Vietta 1992: 294; Fähnders 1998: 36 f.; Sprengel 1998, 1998: 621; Austermühl 2000: 350 f.; Lamping 2000: 145 f.; Andreotti 2014: 17). However, some scholars, even if they consider the anthology as a whole to be traditional, argue that it was at least innovative in terms of *content* since new themes such as ‘big cities’ or ‘social issues’ were addressed (e.g. Fähnders 1998: 36 f.).

Hypothesis 2: Modernist poetry is heterogeneous, that is, more heterogeneous than realist poetry. While the poetry of realism, or at least the mass-produced poetry of this period, is considered by researchers to be relatively homogeneous (e.g Stockinger 2010: 88), modernist poetry is highly diverse, according to many scholars, given the simultaneity of a wide variety of literary movements (Anz 2007: 330 f.; Becker and Kiesel 2007: 30; Fähnders 1998: IX, 4). But the hypothesis of modernist heterogeneity has its limitations. For example, some researchers support the view that modernism is homogeneous at least insofar as it responds to the same social-cultural problems (Vietta 1992: 30 f; Fähnders 1998: 9 f; Becker and Kiesel 2007: 30; for further statements on the homogeneity of modernist poetry see H. Friedrich 1992: 140-2; Lamping 2008: 13). One researcher, therefore, argues that the period around 1900 was characterized by a “homogeneity of the heterogeneous” (Fähnders 1998: 11). Despite these limitations, most scholars would probably agree that modernist poetry is at least more heterogeneous than the poetry of realism.

Hypothesis 3: There is a fundamental “rupture” between modernist poetry and earlier, more traditional poetry. This view was already held by contemporary authors, critics, and anthologists, who spoke of a ‘revolution’ in poetry (as an example from the corpus anthologies see Bethge 1905: 13f.; cf. on contemporary statements H. Friedrich 1992: 141; Anz 2007: 333; Lamping 2012; Wieland 2019: 17). Many researchers also emphasize major differences between modernism and previous literary periods, often using the metaphor of “rupture” (H. Friedrich 1992: 20; Kiesel 2004: 141 f.; Frick 2007: 97 f.; Goltschnigg 2007: 169; Lamping 2012; Andreotti 2014: 5; without this metaphor: Klinger 2002: 160; Lamping 2000: 140; Lamping 2008: 11,13). But the “rupture”-thesis is also partly qualified. For example, it is emphasized that modernism still refers to traditions (even though it uses them in new ways) (Kiesel 2004: 142 f.; Frick 2007: 98 f.; Goltschnigg 2007: 169). Others argue that many relevant authors were located somewhere *between* realism and modernism or that they combined traditional as well as new elements, which implies a smoother transition between periods (see for C. F. Meyer Selbmann 1999: 149, 152; for Fontane (!) Selbmann 2007: 201; for Baudelaire,

Rilke, Hofmannsthal, and Kafka Lamping 2012). Still, others relativize the novelty of modernism in general (Hiebel 2005: 27; Anz 2007: 333). Thus, hypothesis 3 is partly controversial in research.

It is possible to combine the aforementioned hypotheses in a visual model. The purpose of this model is threefold: it visually summarizes the research hypotheses, it relates the hypotheses to one another, and it demonstrates that all hypotheses about similarity and dissimilarity combined offer a fairly comprehensive interpretation of the transformation from realism to modernism, again underscoring the relevance of similarity as a category of analysis. In the model, each point represents a poem. The greater the

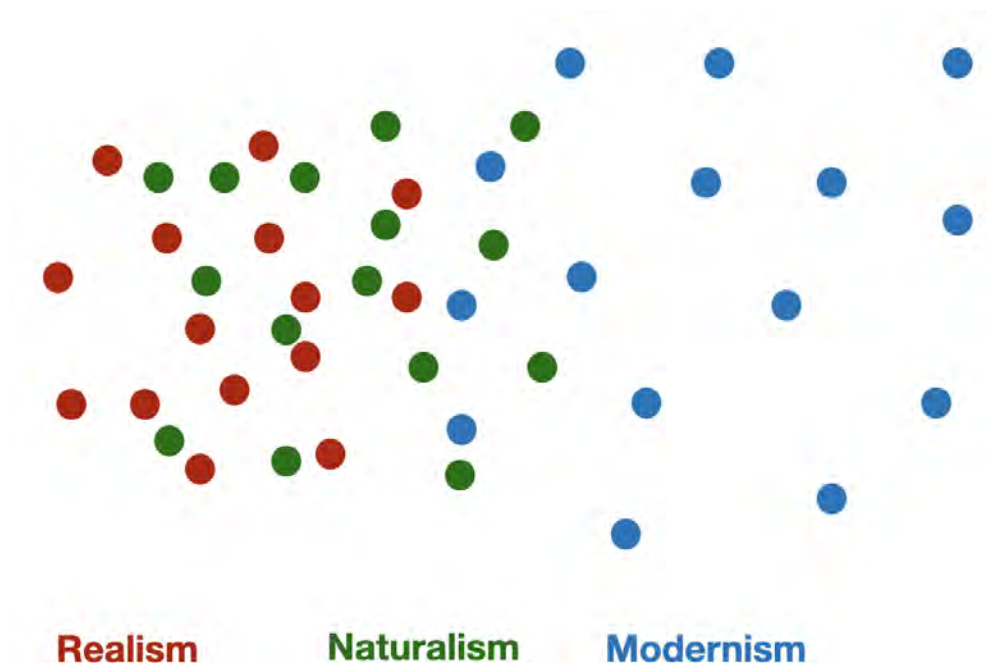


Figure 5: Model of the distances between poems of realism, naturalism and modernism according to research.

distances between the points, the more dissimilar the texts. The distances are not based on calculations but on a hermeneutic understanding of the research and are meant as rough approximations of general ideas. One can see that the distances within realism are smaller than the distances within modernism. It is also visible that there is a strong division between realism and modernism and that the naturalist poems tend to gravitate more towards realism than modernism.¹³

Admittedly, this model is not explicitly advocated in research. Only rarely does a single scholar state all the hypotheses that the model synthesizes. Like any model, it represents only a section of reality and neglects other aspects, such as the differentiation of individual dimensions of similarity, or synchronic and diachronic period-internal differentiations of, for example, individual authors, groups of authors, or literary movements. Some aspects of the model are, as explained, controversial in research, but it is

13. Distances within naturalism should not be given any further significance; no research hypotheses were considered in this regard.

all the more interesting to examine whether our results fit the model and the underlying hypotheses.

5.2. Results

For a first exploration of the (dis)similarities between realism and modernism, we project the poems into a two-dimensional space (Fig. 6). Some similarities with the model derived from research (Fig. 5) become apparent. In particular, a distinction between realism and modernism is evident, even though the separation is far from perfect since there are numerous overlaps between the two periods. Furthermore, it is consistent with the research model that the naturalist poems tend to stay within the realist spectrum and hardly enter 'decidedly modernist' areas. However, it is necessary to test the hypotheses from literary studies more precisely than just by explorative means.

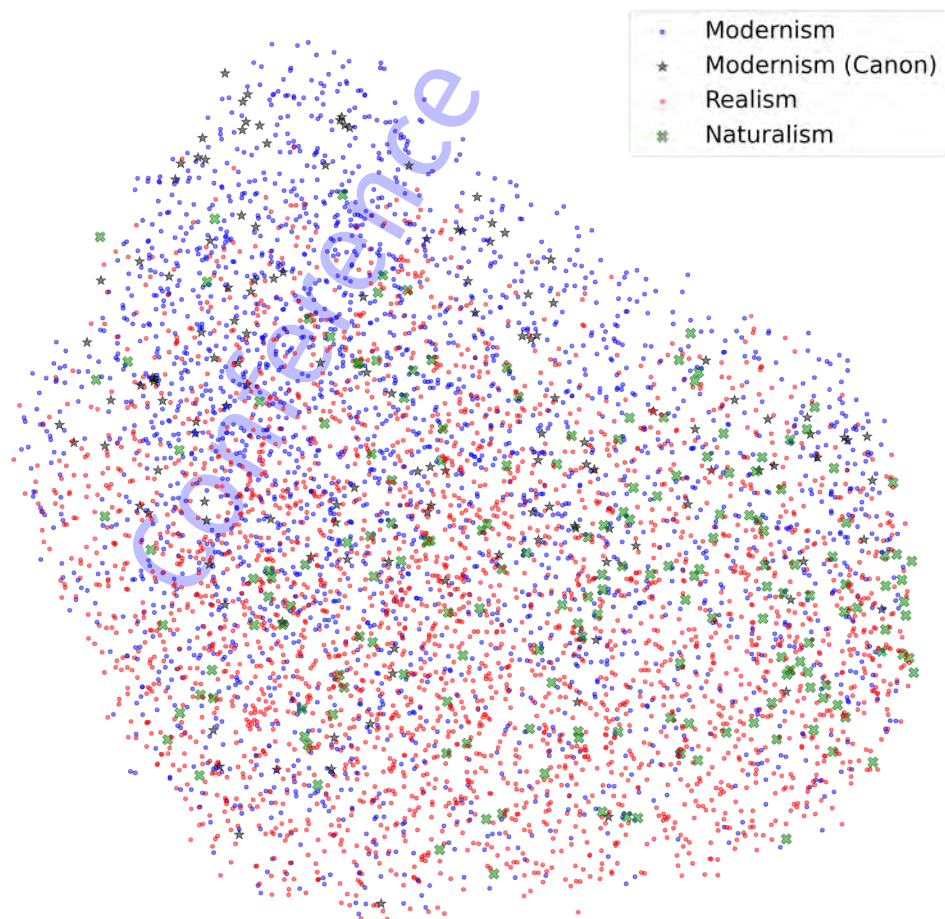


Figure 6: Poems embedded with both vanilla *GBERT-alllayers-meannorm* (see. Fig. 2) and *FastText-meannorm* transformed to reflect the aspect 'content' (see table 1) projected in 2-dimensional space using UMAP (McInnes et al. 2018).

Hypothesis 1: The poetry of naturalism, as represented in the anthology “Moderne Dichter-Charaktere”, is predominantly traditional rather than modernist To test the first hypothesis, we examined the similarity of the programmatically naturalistic anthology “Moderne Dichter-Charaktere” to the realism and modernism corpora. In addition, we measured the distances between the poems within the latter two corpora to be able to assess the comparative analyses more accurately. The boxplots show that overall and for each

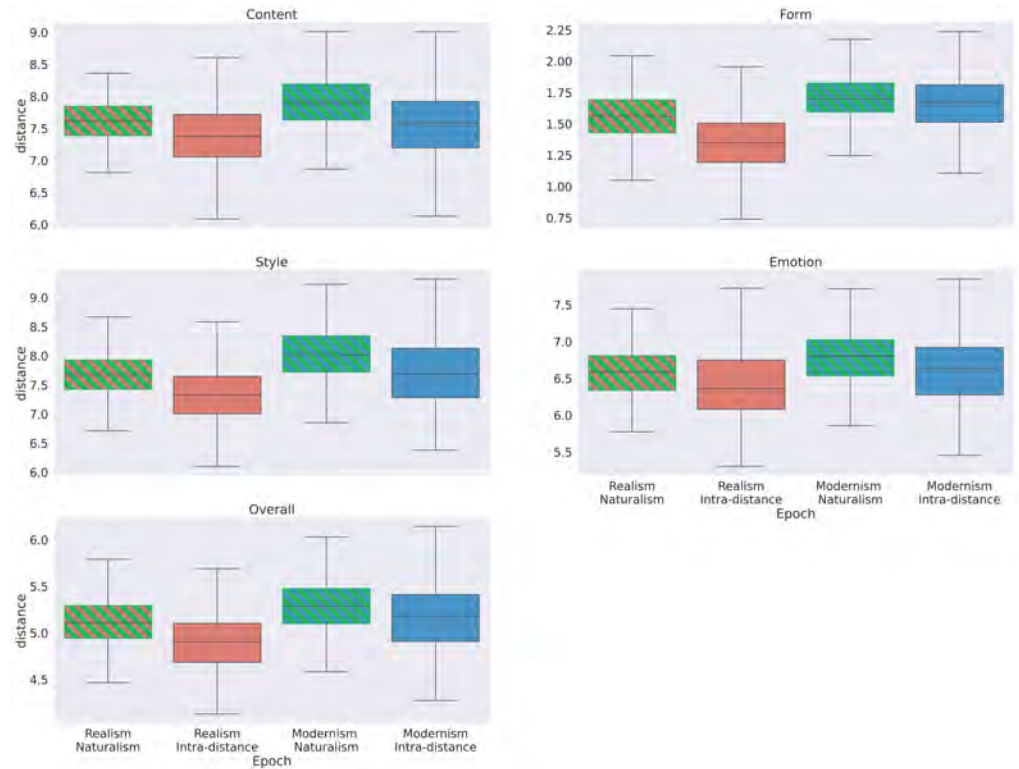


Figure 7: Distances between poems from Realism/Naturalism and Modernism/Naturalism and poems within Realism and Modernism. Distances in ‘content’, ‘style’, ‘emotion’ and ‘overall’ are measured in the space of *paraphrase-XLM-R* embeddings transformed via similarity learning (see section 4.4). Distances in ‘form’ are measured in the Feature-Form embedding space (see section 4.1). Each boxplot represents pairwise euclidean distances of 2000 samples with a size of 20 poems.

individual dimension ‘content’, ‘form’, ‘style’, and ‘emotion’ the distances between naturalism and realism are smaller than the distances¹⁴ between naturalism and modernism. At the same time, the distance between the naturalism and realism corpus is larger than the distance between the poems within the realism corpus. Surprisingly, in the dimension ‘content’ no higher proximity to the modernism corpus is seen. A stronger similarity between naturalist and modernist poems would have been expected based on the literary-historical theses we have mentioned above. As expected, the analyses support the thesis that the naturalism corpus is more similar to the realism corpus

14. We tested for significance and all differences are highly significant. To make sure this is not solely an effect of the large sample size we randomly selected 100 texts, but the differences stay significant. New guidelines usually recommend complementing p-values with effect size. In our case this is not easy to apply, because the measure is not grounded in an intuitively comprehensible unit.

than to the modernist corpus. However, a more detailed look shows that differences can be found in the individual dimensions. This could indicate that the naturalistic poems probably do not use the same means as realistic poems. What exactly these differences are should be investigated in a further study. However, equating naturalist with realist poetry falls short in any case since the internal distance in the realism corpus is smaller than that in the comparison between naturalism and realism. It should be emphasized that we have studied the effect only for the anthology “Moderne Dichter-Charaktere” and only using its short poems, as stated above. Further study would have to take into account that the modernism corpus also contains some naturalistic poems.

Hypothesis 2: Modernist poetry is heterogeneous, that is, more heterogeneous than realist poetry. From now on, when we compare realism with modernism, we no longer include the naturalist poems in our calculations and visualizations, since we have seen that naturalism is located somewhere between realism and modernism. However, we now distinguish in modernist texts canonical and non-canonical authors in order to point out some peculiarities of the canonical poems.¹⁵

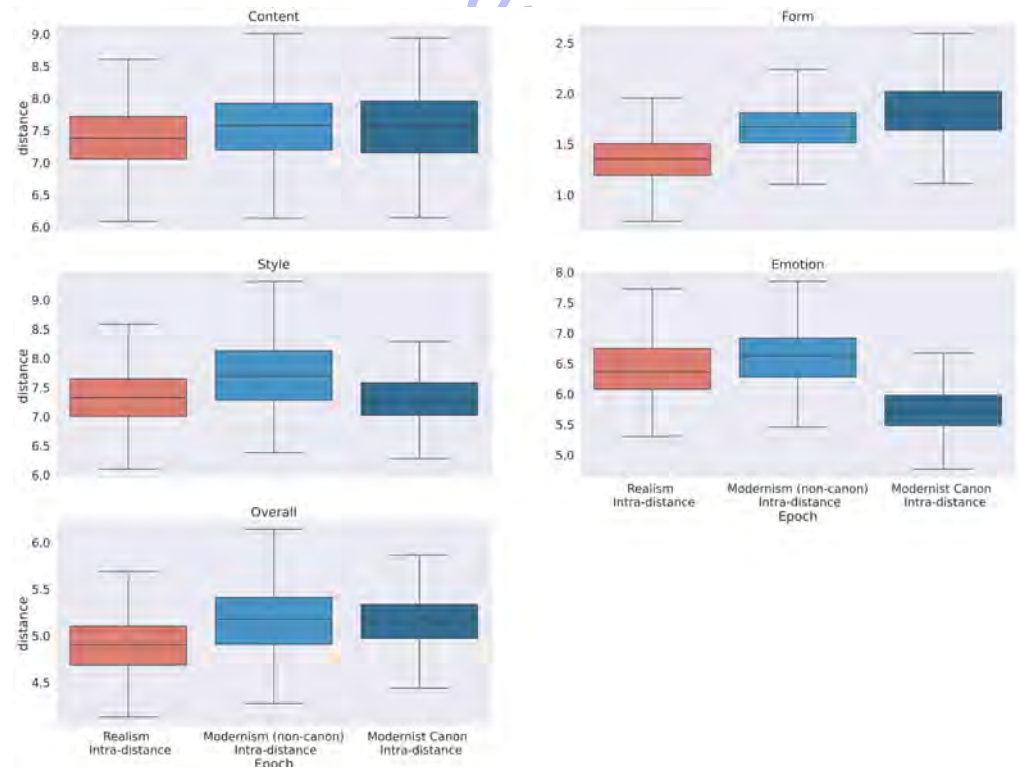


Figure 8: Distances within poems from Realism, Modernism and canonic Modernism. Distances in ‘content’, ‘style’, ‘emotion’ and ‘overall’ are measured in the space of *paraphrase-XLM-R* embeddings transformed via similarity learning (see section 4.4). Distances in ‘form’ are measured in the Feature-Form embedding space (see section 4.1). Each boxplot represents pairwise euclidean distances of 2000 samples with a size of 20 poems.

15. In our study, in accordance with German literary history, Stefan George (6 poems), Hugo von Hofmannsthal (6 poems), Arno Holz (19 poems), Else Lasker-Schüler (3 poems), and Rainer Maria Rilke (24 poems) represent canonical modernism.

To test hypothesis 2, we compare the distances within realism with those within modernism (Fig. 8). In all dimensions, the distances within modernism are greater than in realism, most clearly in the dimension ‘form’. Thus, the hypothesis that modernist poetry is more heterogeneous than realist poetry can be confirmed by our data. However, the differences in heterogeneity are mostly small and should not be overemphasized. Modernist poems by canonical authors are slightly more heterogeneous than non-canonical poems regarding the dimension ‘form’. Otherwise, the canonical poems are not characterized by greater distances among themselves than non-canonical modernist poems. On the contrary, the distances for the dimensions of style and especially emotion are much smaller within the canonical texts than within the non-canonical modernist poems. All in all, the canonical texts are no more heterogeneous than the non-canonical ones. This is surprising, since one might have expected a particularly high degree of individuality and thus heterogeneity in the canon. In any case, it must be kept in mind that the subcorpus of canonical modernist poems is very small (58 poems, 5 authors), which limits the validity of the results. Further research is needed here.

Hypothesis 3: There is a fundamental “rupture” between modernist poetry and earlier, more traditional poetry. It is difficult for us to say, based on our data, whether the distance between realist and modernist poetry is particularly large, since we do not know the distances between other literary periods with which we could compare our results to. But we can compare the distance between realism and modernism with distances *within* periods, for example with those within realism. If the distances between realism and modernism are greater than within realism, it can at least be said that modernism is different from realism.

In all dimensions, the distances between realism and modernism are larger than the distances within realism. However, these differences in distance are not enormous. Moreover, the two-dimensional plot above (Fig. 6) shows that modernist poems appear not only outside realism, but often within the realist spectrum as well. All in all, to speak of a fundamental ‘rupture’ between the periods seems exaggerated, at least for our data.

One might assume that researchers use the metaphor of ‘rupture’ because they focus on other, namely canonical texts. The distance between realism and canonical modernism is indeed larger than the distance between realism and non-canonical modernism regarding the form, and at least a tiny bit larger for the dimensions ‘content’ and ‘overall’. But in terms of style, canonical modernism is no further from realism than non-canonical modernism, and in regards to emotion, the distance between realism and canonical modernism is even smaller than between realism and non-canonical modernism. Thus, our results do not show that the distances between canonical modernism and realism are systematically larger than between non-canonical modernism and realism. The idea that the canonical texts set a trend that the non-canonical texts follow, just not as decisively, cannot be confirmed.

One might expect the canonical modernist poems to be at least closer to the non-canonical modernist texts than to the realist poems, but this is not true either, according to our

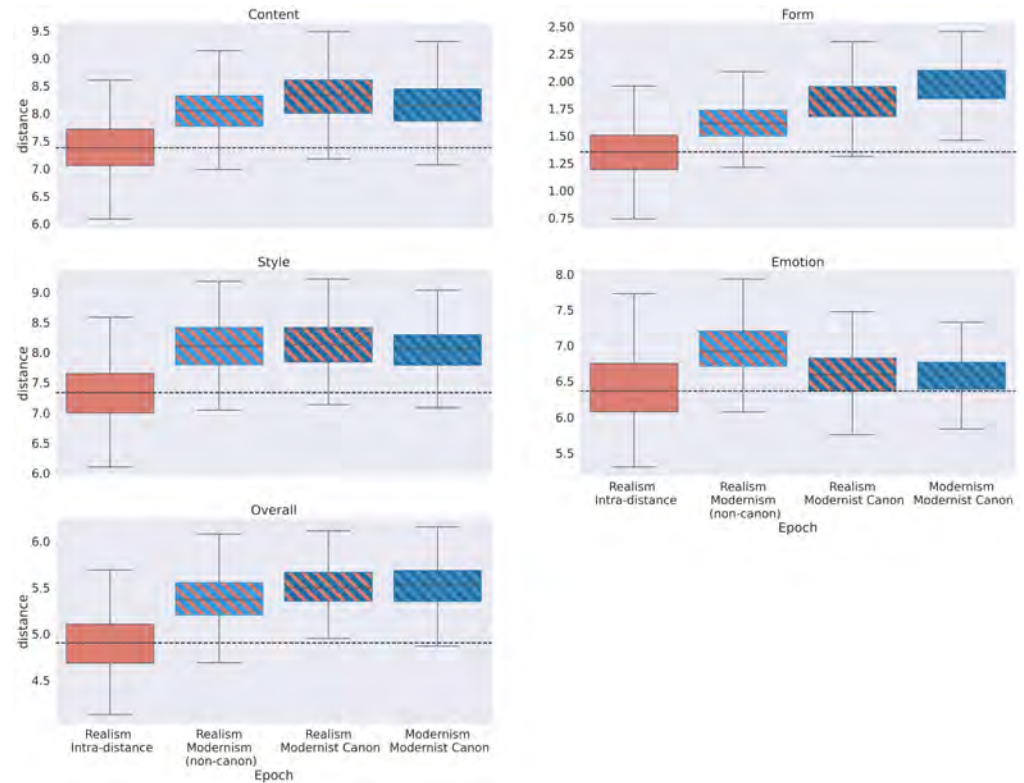


Figure 9: Distances within poems from Realism and between Realism/non-canon Modernism, Realism/canonic Modernism and non-canon Modernism/canonic Modernism. Distances in ‘content’, ‘style’, ‘emotion’ and ‘overall’ are measured in the space of *paraphrase-XLM-R* embeddings transformed via similarity learning (see section 4.4). Distances in ‘form’ are measured in the Feature-Form embedding space (see section 4.1). Each boxplot represents pairwise euclidean distances of 2000 samples with a size of 20 poems.

data: The distances from canonical modernist poems to realist texts on the one hand 547
and to non-canonical modernist texts on the other hand do not differ significantly. In 548
the case of the dimension ‘form’, the canonical modernist poems are even closer to the 549
realist ones than to the non-canonical modernist ones. 550

The results for the canon are counter-intuitive and call for further research. Again, our 551
observations may have something to do with the fact that our subcorpus of canonical 552
texts is very small and that we only analyze short poems. 553

To further explore the differences between modernist and realist poetry in our vector 554
space, we constructed a timeline from a graph network. The network was created using 555
all pairwise distances (or similarities more precisely) between the document vectors. For 556
all dimensions except ‘form’, the distances are based on the vectors of *paraphrase-XLM-R*, 557
after the adaptation with similarity learning. For ‘form’, only the formal feature vector 558
similarities were used. All distances were standardized per dimension to lie between 0 559
and 1 (due to the different metrics used to determine the vector distances). 560

Each node in the graphs represents a span of 5 years (i.e. 1865 for the span 1863-1867). 561
The edge between two year slices is depicted by the mean distances of a sample of 30 562
poems - if less than 30 poems were available, poems were drawn multiple times. The 563

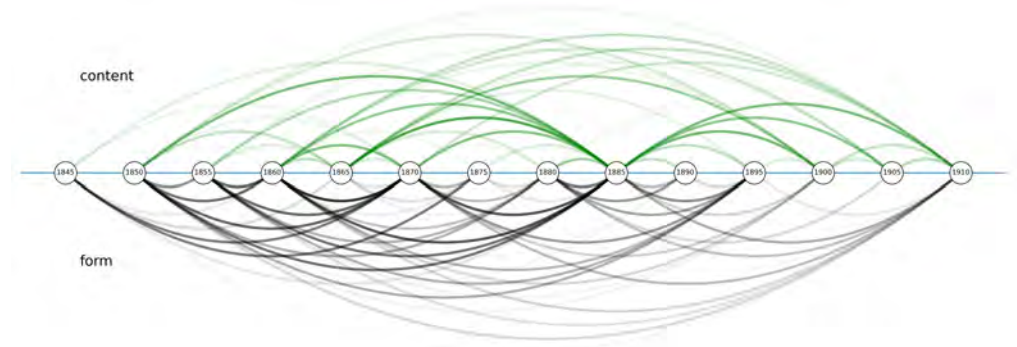


Figure 10: Graph timelines for the ‘content’ and ‘form’ dimensions based on the mean pairwise similarities of 30 poems, sampled for each 5-year time span, based on the similarity-adapted vectors of *paraphrase-XLM-R* (content) and the formal feature vectors (form). See appendix for a larger version of this figure.

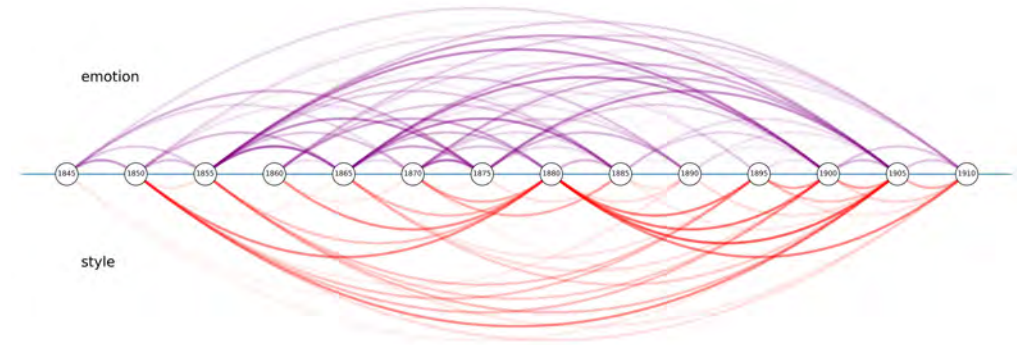


Figure 11: Graph timelines for the ‘emotion’ and ‘style’ dimensions based on the mean pairwise similarities of 30 poems, sampled for each 5-year time span, based on the similarity-adapted vectors of *paraphrase-XLM-R*. See appendix for a larger version of this figure.

alpha of one edge between two years visualizes the degree of their similarity based on 564
the chosen poems. We only used poems where the corresponding years were manually 565
checked and corrected by us if necessary. This amounted to 321 poems between 1845 566
and 1911 specifically. 567

From this visualization which is based not on the assignment of the poems to a period 568
by the editors of the anthologies, but on the publication date of the poems, we can 569
make some observations. In terms of form, we can surmise from the timeline that realist 570
poems are more similar to each other and thus more homogenous than modernist poems 571
are (coinciding with our findings from hypothesis 2). Additionally, the further the 572
nodes are away from realism, the weaker the similarity becomes, implying that later 573
modernist poems become even more estranged from the form of realist poems. The 574
networks for content and style seem similar: both suggest a kind of split between the 575
epochs, hinting at the possibility that modernist and realist poetry have a higher inter- 576
than extra-epochal similarity (coinciding with our findings from hypothesis 3). The 577
timelines could potentially not only help with identifying whether a rupture between 578
the epochs exists or not but also when exactly such a rupture occurs. While ‘style’ shows 579

its split around 1880, the split for ‘content’ appears to be at around 1885, implying
that the change from realism to modernism first became apparent in style and then in
content. For ‘emotion’, we cannot discover any kind of pattern in the timeline, suggesting
that emotions thematized or expressed in the poems might contribute to a continuity
between the two epochs.

In summary, we were able to confirm some important hypotheses from literary studies,
while differentiating or relativizing others. Our data supports the view that naturalist
poetry is closer to realism than to modernism; however, simply equating naturalist and
realist poetry would not be appropriate. We showed that modernist poetry is indeed
more heterogeneous than realist poetry, even though the differences are limited. Finally,
our findings suggest that the change from realism to modernism was an evolutionary
transition rather than a revolutionary disruption. The results encourage increased at-
tention in literary history to processes of gradual, limited change, rather than thinking
only in terms of either stasis or rupture.

The assumptions made in this section are still only based on exploratory visualizations
and comparatively little data. Subsequent research could expand this subcorpus of
year-annotated poems (most importantly including longer poems as already mentioned)
while further research questions could investigate these assumptions, e.g. whether the
rupture between the epochs could have happened at slightly different points in time for
different dimensions or whether ‘form’ really is the most suitable dimension to measure
homogeneity and heterogeneity within realism and modernism for example.

In a recent article (Underwood and So 2021) discuss the question of whether the map-
ping of cultural artifacts to some spatial representation is not ‘distorting’ them, whether
cultural relationships obey a spatial logic at all. Their experiments show that even
if we have some seemingly convincing arguments against this kind of mapping, we
accumulate more and more empirical evidence that it works very often astonishingly
well. Our paper adds to this evidence: Textual representations in high-dimensional
space seem well-suited to express even complex text models though more empirical
work may expose its shortcomings in the future. In the meantime, we hope our approach
can be used to reevaluate our understanding of the fundamental concept of similarity,
not only in Computational Literary Studies.

6. Data availability 611

Data can be found here: <https://github.com/cophi-wue/jcls2022-poem-similarity> 612
613

7. Software availability 614

Software can be found here: <https://github.com/cophi-wue/jcls2022-poem-similarity> 615
616

8. Acknowledgements 617

This work was funded by the Deutsche Forschungsgemeinschaft as part of the SPP 2207 618
Computational Literary Studies in the project *The beginnings of modern poetry - Modeling* 619
literary history with text similarities. 620

9. Author contributions 621

References 622

- Andreotti, Mario (2014). *Die Struktur der modernen Literatur. Neue Formen und Techniken des Schreibens: Erzählprosa und Lyrik*. ger. 5., stark erweiterte und aktualisierte Auflage. 623
UTB 1127. Bern: Haupt Verlag. ISBN: 978-3-8252-4077-6. 624
625
- Anz, Thomas (2007). "Thesen zur expressionistischen Moderne". de. In: *Literarische* 626
Moderne. Begriff und Phänomen. Ed. by Sabina Becker and Helmuth Kiesel. Berlin/New 627
York: De Gruyter, pp. 329–346. ISBN: 978-3-11-092661-3. URL: [https://www.degruyter](https://www.degruyter.com/document/doi/10.1515/9783110926613.329/html) 628
[.com/document/doi/10.1515/9783110926613.329/html](https://www.degruyter.com/document/doi/10.1515/9783110926613.329/html) (visited on 12/14/2021). 629
- Arent, Wilhelm, ed. (1885). *Moderne Dichter-Charaktere*. Berlin: Kanzlah. 630
- Arora, Sanjeev, Yingyu Liang, and Tengyu Ma (2017). "A Simple but Tough-to-Beat 631
Baseline for Sentence Embeddings". In: *ICLR*. 632
- Austermühl, Elke (2000). "Lyrik der Jahrhundertwende". In: *Naturalismus, Fin de siècle,* 633
Expressionismus, 1890–1918. Ed. by York-Gothart Mix. Hansers Sozialgeschichte der 634
deutschen Literatur vom 16. Jahrhundert bis zur Gegenwart 7. München/Wien: Carl 635
Hanser, pp. 350–366. 636
- Avenarius, Ferdinand, ed. (1882). *Deutsche Lyrik der Gegenwart seit 1850. Eine Anthologie* 637
mit biographischen und bibliographischen Notizen. Aus den Quellen. Dresden: Ehlermann. 638
- Bär, Daniel, Torsten Zesch, and Iryna Gurevych (2011). "A Reflective View on Text 639
Similarity". en. In: *Proceedings of Recent Advances in Natural Language Processing*, 640
pp. 515–520. 641
- (Jan. 2015). "Composing Measures for Computing Text Similarity". en. In: p. 31. 642

- Becker, Sabina and Helmuth Kiesel (2007). "Literarische Moderne. Begriff und Phänomen" de. In: *Literarische Moderne. Begriff und Phänomen*. Ed. by Sabina Becker and Helmuth Kiesel. Berlin/New York: De Gruyter, pp. 9–36. ISBN: 978-3-11-092661-3. URL: <https://www.degruyter.com/document/doi/10.1515/9783110926613.9/html> (visited on 12/14/2021).
- Benzmann, Hans, ed. (1904). *Moderne deutsche Lyrik*. Leipzig: Reclam.
- Bern, Maximilian, ed. (1877). *Deutsche Lyrik seit Goethes Tode*. Leipzig: Reclam.
- Bethge, Hans, ed. (1905). *Deutsche Lyrik seit Liliencron*. Leipzig: Hesse.
- Bierbaum, Otto Julius, ed. (1893). *Moderner Musenalmanach auf das Jahr 1893*. München: Albert.
- ed. (1894). *Moderner Musenalmanach auf das Jahr 1894*. München: Albert.
- Bonsels, Waldemar, Hans Brandenburg, Bernd Isemann, and Will Vesper, eds. (June 1905). *Die Erde*. München: Bonsels.
- Bromley, James, Isabelle Guyon, Yann LeCun, Eduard Sackinger, and Roopak Shah (1993). "Signature Verification Using a "Siamese" Time Delay Neural Network". In: *Proceedings of the 6th International Conference on Neural Information Processing Systems*. San Francisco, CA, USA, pp. 737–744.
- Chan, Branden, Stefan Schweter, and Timo Möller (Dec. 2020). "German's Next Language Model". In: *arXiv:2010.10906 [cs]*. arXiv: 2010.10906. URL: <http://arxiv.org/abs/2010.10906> (visited on 07/15/2021).
- Conradi, Hermann (1885). "Unser Credo". In: *Moderne Dichter-Charaktere*. Ed. by Wilhelm Arent. Berlin: Wilhelm Friedrich, pp. I–IV.
- Corbineau-Hoffmann, Angelika (2013). *Einführung in die Komparatistik*. 3. neu bearbeitete Aufl. Berlin: Erich Schmidt.
- Ehrmanntraut, Anton, Thora Hagen, Leonard Konle, and Fotis Jannidis (2021). "Type- and Token-based Word Embeddings in the Digital Humanities". In: *CHR*.
- Fähnders, Walter (1998). *Avantgarde und Moderne 1890-1933*. Lehrbuch Germanistik. Stuttgart/Weimar: J. B. Metzler.
- Federmann, Herta, ed. (1908). *Der Schatzbehälter*. München: Steinicke & Lehmkuhl.
- Frick, Werner (2007). "Avantgarde und longue durée. Überlegungen zum Traditionsverbrauch der klassischen Moderne". de. In: *Literarische Moderne. Begriff und Phänomen*. Ed. by Sabine Becker and Helmuth Kiesel. Berlin/New York: De Gruyter, pp. 97–112. ISBN: 978-3-11-092661-3. URL: <https://www.degruyter.com/document/doi/10.1515/9783110926613.97/html> (visited on 12/16/2021).
- Friedrich, Hugo (1992). *Die Struktur der modernen Lyrik. Von der Mitte des neunzehnten bis zur Mitte des zwanzigsten Jahrhunderts*. rowohlts enzyklopädie. Reinbek: Rowohlt.
- Friedrich, Paul, ed. (1911). *Neuland. Ein Buch jüngstdeutscher Lyrik*. Berlin: Borngräber.
- Gemmel, Ludwig, ed. (1898). *Die Perlenschnur. Eine Anthologie moderner Lyrik*. Berlin, Leipzig: Schuster & Loeffler.
- Goltschnigg, Dietmar (2007). "Traditionszusammenhänge der österreichischen Moderne (am Beispiel der Heine- und Büchner-Rezeption)". de. In: *Literarische Moderne. Begriff und Phänomen*. Ed. by Sabine Becker and Helmuth Kiesel. Berlin/New York: De Gruyter, pp. 169–180. ISBN: 978-3-11-092661-3. URL: <https://www.degruyter.com/document/doi/10.1515/9783110926613.169/html> (visited on 12/16/2021).

- Gomaa, Wael and Aly Fahmy (Apr. 2013). "A Survey of Text Similarity Approaches". 687
In: *international journal of Computer Applications* 68. doi: [10.5120/11638-7118](https://doi.org/10.5120/11638-7118). 688
- Grandini, Margherita, Enrico Bagli, and Giorgio Visani (2020). *Metrics for Multi-Class* 689
Classification: an Overview. arXiv: [2008.05756](https://arxiv.org/abs/2008.05756) [stat.ML]. 690
- Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug 691
Downey, and Noah A. Smith (May 2020). "Don't Stop Pretraining: Adapt Language 692
Models to Domains and Tasks". In: *arXiv:2004.10964* [cs]. arXiv: 2004.10964. URL: 693
<http://arxiv.org/abs/2004.10964> (visited on 07/24/2020). 694
- Häntzschel, Günter (1991). *Bibliographie der deutschsprachigen Lyrikanthologien 1840-1914*. 695
München: K. G. Saur. 696
- Hiebel, Hans H. (2005). *Das Spektrum der modernen Poesie. 1: 1900 - 1945*. ger. Würzburg: 697
Königshausen & Neumann. ISBN: 978-3-8260-3200-4. 698
- Huch, Margarethe, ed. (1911). *Frauenlyrik der Gegenwart*. Leipzig: Eckardt. 699
- Jacobowski, Ludwig, ed. (1899). *Neue Lieder der besten neueren Dichter für's Volk*. Berlin: 700
Liemann. 701
- Kiesel, Helmuth (2004). *Geschichte der literarischen Moderne. Sprache – Ästhetik – Dichtung* 702
im zwanzigsten Jahrhundert. München: C.H. Beck. ISBN: 978-3-406-51145-5. 703
- Klambauer, Günter, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter (Sept. 704
2017). "Self-Normalizing Neural Networks". In: *arXiv:1706.02515* [cs, stat]. arXiv: 705
1706.02515. URL: <http://arxiv.org/abs/1706.02515> (visited on 12/21/2021). 706
- Klinger, Cornelia (2002). "Modern/Moderne/Modernismus". de. In: *Ästhetische Grund-* 707
begriffe. Ed. by Karlheinz Barck, Martin Fontius, Dieter Schlenstedt, and Friedrich 708
Wolfzettel. Vol. 4. Stuttgart/Weimar: J.B. Metzler, pp. 121–167. ISBN: 978-3-476-02357-5 709
978-3-476-00533-5. doi: [10.1007/978-3-476-00533-5_6](https://doi.org/10.1007/978-3-476-00533-5_6). URL: http://link.springer.com/10.1007/978-3-476-00533-5_6 (visited on 12/16/2021). 711
- Kneschke, Emil, ed. (1865). *Anthologie deutscher Lyriker seit 1850*. Leipzig: Lorck. 712
- Krippendorff, Klaus (2011). *Computing Krippendorff's Alpha-Reliability*. URL: [https://re](https://repository.upenn.edu/asc_papers/43) 713
[pository.upenn.edu/asc_papers/43](https://repository.upenn.edu/asc_papers/43). 714
- Lamping, Dieter (2000). *Das lyrische Gedicht. Definitionen zu Theorie und Geschichte der* 715
Gattung. 3rd ed. Göttingen: Vandenhoeck & Ruprecht. ISBN: 978-3-525-20778-9. 716
- (2008). *Moderne Lyrik*. Göttingen: Vandenhoeck & Ruprecht. 717
- (July 2012). *Klassiker der Moderne. Über die Kanonisierung moderner Literatur*. de-DE. 718
URL: <https://literaturkritik.de/id/16853> (visited on 12/16/2021). 719
- Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, 720
and Roberto Zamparelli (2014). "A SICK cure for the evaluation of compositional 721
distributional semantic models". In: *Proceedings of the Ninth International Conference on* 722
Language Resources and Evaluation (LREC'14). Reykjavik, Iceland: European Language 723
Resources Association (ELRA), pp. 216–223. URL: [http://www.lrec-conf.org/pr](http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf) 724
[oceedings/lrec2014/pdf/363_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf) (visited on 12/19/2021). 725
- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier (Sept. 2015). "The Unified 726
and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and 727
Alignment". en. In: *Computational Linguistics* 41.3, pp. 437–479. ISSN: 0891-2017, 1530- 728
9312. doi: [10.1162/COLI_a_00227](https://doi.org/10.1162/COLI_a_00227). URL: [https://direct.mit.edu/coli/article](https://direct.mit.edu/coli/article/41/3/437-479/1524) 729
[/41/3/437-479/1524](https://direct.mit.edu/coli/article/41/3/437-479/1524) (visited on 12/20/2021). 730

- McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Großberger (2018). "UMAP: Uniform Manifold Approximation and Projection". In: *The Journal of Open Source Software* 3.29. 731-733
- Moltke, Max, ed. (1882). *Neuer deutscher Parnass. Silberblicke aus der Lyrik unserer Tage*. Leipzig: Rühle. 734-735
- Nöth, Winfried (2008). "Stil als Zeichen". In: *Rhetoric and Stylistics. Handbooks of Linguistics and Communication Science*. Ed. by Ulla Fix, Andreas Gardt, and Joachim Knappe. Berlin, New York: de Gruyter, pp. 1178-1196. 736-738
- Polko, Elise, ed. (1860). *Dichtergrüße. Neuere deutsche Lyrik*. Leipzig: Amelang. 739
- Prakoso, Dimas Wibisono, Asad Abdi, and Chintan Amrit (Mar. 2021). "Short text similarity measurement methods: a review". en. In: *Soft Computing* 25.6, pp. 4699-4723. ISSN: 1433-7479. DOI: [10.1007/s00500-020-05479-2](https://doi.org/10.1007/s00500-020-05479-2). URL: <https://doi.org/10.1007/s00500-020-05479-2> (visited on 12/21/2021). 740-743
- Prutz, Robert, ed. (1859). *Deutsche Dichter der Gegenwart. Ein lyrisches Album*. Prag: Kober & Markgraf. 744-745
- Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". en. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3980-3990. DOI: [10.18653/v1/D19-1410](https://www.aclweb.org/anthology/D19-1410). URL: <https://www.aclweb.org/anthology/D19-1410> (visited on 12/20/2021). 746-751
- (2020). "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation". en. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4512-4525. DOI: [10.18653/v1/2020.emnlp-main.365](https://www.aclweb.org/anthology/2020.emnlp-main.365). URL: <https://www.aclweb.org/anthology/2020.emnlp-main.365> (visited on 12/20/2021). 752-756
- Renner, August, ed. (1899). *Das lyrische Wien. Eine moderne Lese*. Wien, Berlin, Leipzig: Georg Szelinski. 757-758
- Salton, Gerard and Michael J. McGill (1983). *Introduction to Modern Information Retrieval*. New York a.o.: McGraw-Hill. URL: https://books.google.de/books/about/Introduction_to_Modern_Information_Retrieval.html?id=7f5TAAAMAAJ&redir_esc=y (visited on 12/16/2021). 759-762
- Sandig, Barbara (2006). *Textstilistik des Deutschen*. Berlin, New York: de Gruyter. 763
- Selbmann, Rolf (1999). *Die simulierte Wirklichkeit. Zur Lyrik des Realismus*. OCLC: ocm42430273. Bielefeld: Aisthesis. ISBN: 978-3-89528-238-6. 764-765
- (2007). "Die Lyrik des Realismus". In: *Realismus. Epoche – Autoren – Werke*. Ed. by Christian Begemann. Darmstadt: Wissenschaftliche Buchgesellschaft, pp. 189-206. 766-767
- Shaver, Phillip, Judith Schwartz, Donald Kirson, and Cary O'Connor (1987). "Emotion Knowledge. Further Exploration of a Prototype Approach". en. In: *Journal of Personality and Social Psychology* 52.6, pp. 1061-1086. ISSN: 1939-1315, 0022-3514. DOI: [10.1037/0022-3514.52.6.1061](https://doi.org/10.1037/0022-3514.52.6.1061). URL: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.52.6.1061> (visited on 12/21/2021). 768-772

- Sprenkel, Peter (1998). *Geschichte der deutschsprachigen Literatur 1870-1900. Von der Re- 773*
ichsgründung bis zur Jahrhundertwende. Geschichte der deutschen Literatur von den 774
 Anfängen bis zur Gegenwart 9.1. München: C.H. Beck. ISBN: 978-3-406-44104-2. 775
- Stockinger, Claudia (2010). *Das 19. Jahrhundert. Zeitalter des Realismus*. de. Akademie 776
 Studienbücher Literaturwissenschaft. Berlin: Akademie Verlag. ISBN: 978-3-05-005290- 777
 8. URL: [tml \(visited on 12/14/2021\). 779](https://www.degruyter.com/document/doi/10.1524/9783050052908/h 778

<a href=)
- Szubert, Benjamin, Jennifer E. Cole, Claudia Monaco, and Ignat Drozdov (Dec. 2019). 780
 "Structure-preserving visualisation of high dimensional single-cell datasets". en. In: 781
Scientific Reports 9.1, p. 8914. ISSN: 2045-2322. DOI: [10.1038/s41598-019-45301-0](https://doi.org/10.1038/s41598-019-45301-0). 782
 URL: <http://www.nature.com/articles/s41598-019-45301-0> (visited on 783
 12/21/2021). 784
- Tille, Alexander, ed. (1896). *Deutsche Lyrik von Heute und Morgen*. Leipzig: Neumann. 785
- Underwood, Ted and Richard Jean So (June 2021). "Can We Map Culture?" en. In: 786
Journal of Cultural Analytics 6.3. Publisher: Department of Languages, Literatures, 787
 and Cultures, p. 24911. DOI: [10.22148/001c.24911](https://doi.org/10.22148/001c.24911). URL: [tics.org/article/24911-can-we-map-culture](https://culturalanaly 788

<a href=) (visited on 12/21/2021). 789
- Vietta, Silvio (1992). *Die literarische Moderne. Eine problemgeschichtliche Darstellung der 790*
deutschsprachigen Literatur von Hölderlin bis Thomas Bernhard. Stuttgart: J.B. Metzler. 791
 ISBN: 978-3-476-00790-2. 792
- Wieland, Klaus (2019). "Die deutschsprachige Lyrik der Frühen Moderne (1890-1930)". 793
 de. In: *Recherches Germaniques* 14, pp. 5-27. ISSN: 0399-1989. DOI: [10.4000/rg.976](https://doi.org/10.4000/rg.976). 794
 URL: <https://journals.openedition.org/rg/976> (visited on 12/16/2021). 795
- Willatzen, Peter Johann, ed. (1875). *Blüthenzweige deutscher Lyrik nach Goethe. Eine An- 796*
thologie. Bremen: Kühtmann. 797
- Winko, Simone (2003). *Kodierte Gefühle. Zu einer Poetik der Emotionen in lyrischen und 798*
poetologischen Texten um 1900. Berlin: Erich Schmidt. 799
- Zelle, Carsten (2005). "Komparatistik und 'comparatio' - der Vergleich in der Vergle- 800
 ichenden Literaturwissenschaft : Skizze einer Bestandsaufnahme". de. In: *Kompara- 801*
tistik. Jahrbuch der Deutschen Gesellschaft für Allgemeine und Vergleichende Literaturwis- 802
senschaft, pp. 13-33. URL: [or/index/index/year/2017/docId/43475](http://publikationen.ub.uni-frankfurt.de/frontdo 803

<a href=) (visited on 12/16/2021). 804

Almas Abdibayev  ¹

Yohei Igarashi  ²

Allen Riddell  ³

Daniel Rockmore  ¹

1. Department of Computer Science, Dartmouth College, Hanover.

2. Department of English, University of Connecticut, Storrs.

3. Department of Information and Library Science, Indiana University, Bloomington.

Keywords:

limericks, computational poetics,
minimal pairs, evaluation, language
models

Licenses:

This article is licensed under:



Abstract. Computational poetics encompasses the wide range of challenges implicit in analyzing and generating poetry – in all of its many forms – through computational techniques and frameworks. In this paper we build on a nascent body of work that has proposed the use of the limerick as a “model organism” for computational poetics, and in particular the use of Benchmarked Poetic Minimal Pairs (BPoMP) as an investigative framework, especially for the evaluation of the poetic abilities of deep learning language models. To that end, we include results for two new BPoMP tasks of interest for limerick analysis – the word deletion task and the limerick completion tasks. We include a release of a data set for the deletion task. We also offer up a suite of an additional ten BPoMP challenges whose precise formulations still require detail.

1. Introduction

1

Much less would I care to try sliding [limericks] through the ... apertures
of a calculating machine, in order to discover the leading “traits” or themes
with which they are concerned, even assuming that anything meaningful
could be learned in such a way. — Gershon Legman, *The Limerick* (1969)

2

3

4

5

What do computers “know” or recognize about poetic form? And can they “learn”
about poetry? This paper explores such questions under the heading of “computational
poetics,” using limericks as a paradigmatic case or a model organism, first introduced
in (Abdibayev, Igarashi, et al. 2021) and elaborated on below. We use an experimental
framework called “minimal pairs” to examine the extent to which language models
(Jurafsky and Martin 2021) can discern elements of poetic language and form as well as
the poem’s overall integrity.

6

7

8

9

10

11

12

Nearly fifty years ago, the folklorist and limerick historian, Gershon Legman, expressed

13

his distaste at the very idea of the computational analysis of limericks, a particularly folk poetic kind (see epigraph). But despite these admonitions, recent work (Abdibayev, Riddell, and Rockmore 2021; Abdibayev, Igarashi, et al. 2021) has focused attention on the limerick for several reasons, viewing the limerick as – borrowing from the life sciences – a “model organism” for computational poetics.

In the life sciences, many disciplines rely on model organisms: a handful of organisms are studied for their “representational scope,” that is, their ability to stand in for many other organisms and phenomena and thereby “create knowledge that can be projected beyond the immediate domain in which it was produced.” For example, the weed known as thale cress is a key model organism for the broader study of the genetics, evolution, and development of many plant species (Ankeny and Leonelli 2020). Other familiar model organisms include the fruitfly (*Drosophila*), the roundworm (the nematode *C. elegans*), and the mouse (*Mus musculus*). Each has the property of simplicity – at least relative to their larger research environment – as well as some degree of pliability and clarity vis-a-vis interrogative pathways. That is to say, model organisms are generally chosen both for the ease with which a potentially influential parameter can be isolated and then tweaked as well as the ability to understand the effect of that modulation on a phenomenon of interest. *C. elegans* has only 1000-3000 cells (depending on how you count), a few hundred neurons, and about 20,000 genes. *Drosophila* turn over a generation every week. Questions of evolution, genetic engineering, and neuroscience have the potential of being answered at these scales of time, space, and components, and with that, provide a solid platform for broader speculation. In general, model organisms have been critical for the important advances that have been made over at least the past half-century (including several Nobel Prizes) in human genetics, neuroscience, reproductive science, botany, and biology.

Poetry – the complex interplay of sounds, rhythm, words, meanings, narrative, visual formatting, and more – is manifested across a wide range of forms. Such diversity can prove challenging in the search for general principles that might apply across computational approaches. Hence isolating a particular form like the limerick provides a good place to start. The limerick is a relatively short and simple form that happens to have a high density of poetic features: five verse lines with an *aabba* rhyme scheme and a 3-3-2-2-3 accentual-metrical arrangement; the presence of trisyllabic feet, i.e., anapests, dactyls, amphibrachs, depending on how one recites or hears the poem; and usually a condensed, humorous narrative structure (for a fuller discussion of the limerick form, see Preminger, Brogan, and Warnke (1993)). These features can be manipulated, as we have done in our various experiments to date. Limericks also serve as a valuable model because language models in widespread use today tend to require short texts, and the limerick form has high linguistic-formal interest relative to its brevity (Liu et al. 2019). The notion of a model organism for literary study was first popularized in the seminal paper of Mary Poovey (Poovey 2001), who argued that lyric poetry served as the model organism for literary criticism. We hew somewhat more closely to the analogy and inspiration from the sciences. The limerick form is an experimental and analytic environment where progress is highly likely, and our method and findings may

be generalizable to other short poetic forms (for example, epigrams, haiku, clerihews, quintains, and even sonnets) – and, beyond that, to longer poetic forms and potentially literary language generally.

We therefore join existing work in computational approaches to poems in English, both those engaged in machine reading and machine writing. We contribute to work that seeks to automate the detection and analysis of poetic features, language, or kinds (for example, see (Anttila and Heuser 2016; Houston 2014; H. Long and So 2016)) and work where computers are trained to output or compose poetry (for example, see (Ghazvininejad et al. 2017; Lau, Cohn, et al. 2018)). In particular, Long & So’s work in “literary pattern recognition” and their stylistic taxonomy of the haiku inform our work with a similarly short poetic form. More generally, we also take our cue from foundational applications of machine learning to literary texts (Bode 2018; Algee-Hewitt 2017; H. J. Long 2021; Piper 2018; So 2020; Underwood 2019). The model organism of the limerick also promises to contribute to formalist investigations of English poetry. Our project complements work like the Princeton Prosody Archive and other endeavors that, in concert with the “New Formalism” and “Historical Poetics,” have brought sustained attention to poetic form in literary study in recent years.¹ Finally, we build here on other recent work in computational poetics and language modeling: the minimal pairs method was introduced for limericks (Abdibayev, Riddell, and Rockmore 2021) and then slightly expanded in (Abdibayev, Igarashi, et al. 2021) with a system for detecting some of the main features of the limerick form while also producing a publicly available data set of limericks.² We make use of that data set herein.

Our main contribution in this paper is to continue the expansion of the testbed of “minimal pairs” challenges for poetry, the “benchmark of poetic minimal pairs” (BPoMP) (Abdibayev, Riddell, and Rockmore 2021), which are inspired by the “benchmark of linguistic minimal pairs” (BLiMP) framework (Warstadt et al. 2020). We describe BLiMP and BPoMP in greater detail below. The first poetic minimal pair tests evaluated the degree to which language models could detect limerick end rhymes from non-rhymes and the overall limerick structure (Abdibayev, Riddell, and Rockmore 2021). In this paper we report on a test set and results pertaining to new BPoMP challenges: word deletion and a synthetic fifth line. The former tests if a language model can distinguish a given limerick from a version of it with missing words (in the sense of identifying the former as more limerick-like). The latter creates the challenge of distinguishing an original limerick from a version of it where the original fifth line has been replaced by a computer-generated one. Both of these challenges recall various aspects of literary and textual practice, from erasure poetry to popular limerick completion contests held during the early twentieth-century “great limerick boom” (McInerney 2001).

We release a new BPoMP data set concurrently with this paper, freely and publicly

1. ‘See, e.g., The Princeton Prosody Archive (<https://prosody.princeton.edu/>) and the essays deriving from it.

2. The collection of limericks used therein is available at Zenodo: <https://zenodo.org/record/5722527>. These limericks comprise a cleaned subset of a larger corpus, also filtered as best as possible to adhere to formal limerick structure as well as to exclude offensive language. See the documentation at the site as well as the paper referenced in text.

available, thereby enabling reproducibility of results in computational poetics. The combination of a curated and publicly available corpus of material with open source models produces a “standard package” (in the sense of Fujimura (1992)) for deep learning in computational poetics, and creates new opportunities for other computational literary studies scholars to engage with the machine learning techniques and tools for critical and creative work in poetry and literature. This research enhances literary scholarship by providing a testbed for evaluating the extent to which computers can analyze and compose short verse. Furthermore, a set of “benchmarked” computational poetic tasks creates a familiar setting for computer scientists and especially the deep learning community, by articulating measurable targets for interrogating the poetic capabilities of current and future language models. In addition to the deep exploration of deletion and completion tests explored below, we include a suite of new tests, whose design – which can be subtle – is still underway. We hope that by introducing this next set of tests herein we are able to foster interest and collaboration in the broader community in the BPoMP schema. In the next section we give some more background on language models, BPoMP in general, and our two new BPoMP challenges. In Section 3, we explain in detail the deletion tests and the completion test and our results. Section 4 is a discussion of the tests, including implications for poetics. We close in Section 5 with discussions of future work.

2. Background

In this section we give a brief overview of the language models that we are evaluating using our minimal pairs method. We then give more detail on the BPoMP construct as well as some discussion of the word deletion and last line completion minimal pairs. We also describe the corpus (OEDILF) from which we source our limerick data set and the filtering process that produces the data sets for these experiments.

2.1. Language Models

The renaissance of neural network models (often marketed under the heading of “deep learning”) has greatly advanced expectations for a machine’s ability to perform machine reading and machine writing. Applied to language modeling (Jurafsky and Martin 2021; Goldberg 2017), these models present exciting opportunities for research in literary studies and computer-supported creative work.

Our work explores the power of the GPT-2, BERT, TransformerXL, and (causal) XLNet language models. Each is based on the computationally efficient “Transformer” architecture (Vaswani et al. 2017), a basic mathematical model that, given some text, predicts with varying probabilities the surrounding text in any human-produced text instance. This is an encoding of each word in the vocabulary as a *vector*, which is effectively a list of numbers. This model depends on a range of numbers – parameters – that have been set according to the likelihoods of various text strings occurring in a large body of text, such as all the writing in Wikipedia. Pre-training is the algorithmic setting of these

parameters based on the example text corpora. 134

The Transformer architecture derives from the better known ideas and architectures 135
based on and inspired by “neural networks” (see e.g., Gurney (1997)), which are them- 136
selves loosely modeled on the “wet” network of neurons in our own brains: connected 137
collections of billions of simple cells (neurons) whose signaling patterns underlie the 138
abilities of all animals to encode learning and learned behaviors. Early neural networks 139
with a relatively small number of (mathematical) neurons were but one of a large 140
number of basic “classifiers,” mathematical models for segmenting data and predictive 141
algorithms.³ It was something of a surprise that when the model sizes were dramatically 142
increased – going from tens of parameters to orders of magnitude larger – that neural 143
networks showed great and in many ways unforeseen abilities to do data discrimination 144
and prediction. Modern machine learning continues to ride the wave of the strength of 145
these architectures and modern computing has enabled the ability to continuously fit 146
more and more parameters in models of increasing size and complexity. 147

“GPT” stands for Generative Pre-trained Transformer. GPT, GPT-2, and now GPT-3 148
(Radford et al. 2019; Brown et al. 2020) are the three generations of a basic – GPT – 149
architecture specially designed to produce human-level text. They come in “sizes” 150
(small, medium, large, etc.,) and the successive generations are largely distinguished by 151
the expansion of the number of parameters embedded in the models and the requisite 152
sizes of the (pre-) training sets needed to tune these models – tens of billions in the case 153
in GPT-2 and hundreds of billions in the case of GPT-3. 154

“BERT” stands for Bidirectional Encoder Representations from Transformer. The “bidi- 155
rectional” modifier reflects a model that looks at word sequences both backwards and 156
forwards for training (Devlin et al. 2018). 157

“TransformerXL” (Dai et al. 2019) is a causal language model, meaning that unlike 158
BERT it only uses preceding words to predict the next word. Its standout ability is the 159
addition of a recurrence mechanism, which is a form of machine memory. Information 160
from previous word sequence segments processed by the model (a fixed-length context 161
window of words that the model uses to predict the next word) is carried over to the 162
next segment, which theoretically allows Transformer-XL to look farther behind in text 163
to make a good prediction. 164

“XLNet” is an evolution of aforementioned TransformerXL model that uses a clever 165
mathematical trick to approximate all possible orders of words in a sentence, where 166
instead of a fixed-order (left to right) context, the model is exposed to a randomly per- 167
muted sequence of all words that both precede and succeed the word we are predicting, 168
while predicting the last word (or last few words) of this sequence. Since the new 169
context includes tokens both from the left and right of the original context of the target 170
word, the model is bidirectional like BERT. At the same, time it can also be used for 171
left-to-right decoding (i.e., generation), like GPT-2 (Yang et al. 2019). 172

3. The origin story goes back to McCulloch and Pitts’s original modeling of neural activity (McCulloch and Pitts 1958) and later, Rosenblatt’s invention of the “perceptron,” a simple mathematical model of a neuron (Rosenblatt 1943).

2.2. BPoMP

173

The BPoMP method evaluates a language model by giving it a choice between a right 174
and wrong instance, where the wrong instance is a “minimally” doctored version of the 175
correct – original – instance. For example, one might have Poem A, but then replace a 176
single word in Poem A with a random word to create Poem B. The more capacious a 177
language model, the more reliably it is able to distinguish between the original and the 178
corruption. The design of minimal pairs to isolate a phenomenon of interest can be 179
rather subtle. 180

There are three reasons to use BPoMP challenges when comparing language models’ 181
ability to model poetry. First, the BPoMP challenges provide a useful “second opinion” 182
when considering model performance. Traditional measures of model “fit” such as per- 183
plexity (see e.g., Chen, Beeferman, and Rosenfeld (1998)) are often unreliable or difficult 184
to calculate in settings where the observed data is high dimensional. Performance on 185
BPoMP challenges, by contrast, is easy to calculate and to interpret. Second, the BPoMP 186
challenges can be used in settings where traditional evaluations such as perplexity are 187
unavailable. As examples of this, BPoMP can be used to compare two language models 188
which use different tokenization strategies and to compare a bidirectional language 189
model with a traditional (“causal”) language model. Third, using the BPoMP challenges 190
to evaluate models can yield insights into the strengths and weaknesses of particular 191
models. It is, for example, easy to imagine one language model performing better on 192
BPoMP challenges involving rhyme but worse on all other challenges. Such a result may 193
indicate that some component of the model is doing well at capturing regularities in 194
language that relate to rhyme. If this component can be isolated, it could be borrowed 195
by other models. While one might balk at leaving such evaluations of models to the 196
machine, recent work supports the use of machine – rather than human – evaluation 197
(Clark et al. 2021). 198

The preceding arguments in favor of the BPoMP challenges closely resemble those 199
offered in favor of the BLiMP challenge set (Warstadt et al. 2020). Whereas BLiMP helps 200
bring into focus the strengths and weaknesses of language models’ ability to adequately 201
model differences between grammatical and ungrammatical sentences, BPoMP helps 202
researchers characterize models’ capacities to capture differences between poetic and 203
non-poetic language. 204

In Abdibayev, Riddell, and Rockmore (2021) a dataset of 10,000 minimal pairs of limer- 205
ick/corrupted limericks were used in a task of “choosing” between the two options to 206
determine the original limerick. Transformer-based models were used. The minimal 207
corruptions were (1) shuffling two rhyming end-of-the-line words, (2) shuffling two 208
rhyming lines, (3) replacing an end-of-the-line word by a non-rhyming synonym. While 209
the models identified the original limerick at rates better than chance, there was a good 210
deal of room for improvement. It is fair to say that the models have yet to demonstrate 211
that they have developed an ear for poetry. 212

This work on detecting formal elements (poetic analysis) complements, but takes a 213

fundamentally different approach from, useful, existing prosodic parsers and pedagogical tools – e.g., “Prosodic” and “For Better For Verse.”⁴ Whereas such tools are focused on discerning or teaching meter, our standardized package is concerned with investigating more fundamentally – via our adaptation of the minimal pairs method – what “poetic knowledge” a language model possesses or can learn, and therefore has a more comprehensive scope which includes not only accentual patterning but other formal elements (rhyme, musical devices like alliteration, and so on).

2.3. Word Deletion

Word deletion is the focus of one of the new minimal pair benchmarks in this paper. In short, a language model is presented with a limerick and a near twin, corrupted by the removal of one, two, or three words. The model then “chooses” between the two by calculating the likelihood of both poems. The text with the higher likelihood wins. This textual challenge probes the models’ ability to discern the semantic, syntactical, and grammatical integrity and form of the limerick genre. Word deletion should disturb these qualities and a language model should know this in the probabilistic sense in which it “knows.”

The word deletion tasks also recalls certain textual and literary practices. Presenting our models with an original text and a version of the same text with missing words mimics the longstanding problem of omissions in the historical transmission of texts. Missing words are an inevitability in the manuscript documentary record, and the scholarly practice of textual criticism has codified various kinds of deletions. For example, a manuscript might be missing words because of *saut du même au même*: when the same word is repeated in close proximity, the scribe “copies the text as far as its first occurrence ... then looking back at the exemplar to see what he must copy next he inadvertently fixes his eye on the second occurrence of the word and proceeds from that point.” The result is that “the intervening words are omitted from his copy” (Reynolds and Wilson 1991). Missing words and other such common errors degrade the transcription, but are crucial for textual scholars in positing the relationships between different manuscripts (that is, the practice of stemmatics or stemmatology): missing words can help to establish how related or unrelated a given manuscript is to other manuscript copies of the “original” (the archetype) (Reynolds and Wilson 1991). The deletion-based BPoMP arguably resembles the first and most fundamental step in such textual scholarship – to identify the more “correct” or more likely text, which is not missing words – and measures how well models can automatically carry out this task.

In a different literary context, the minimal pairs challenge also evokes erasure poetry. An erasure poem is a type of found poem, where an existing composition written by another is manipulated to create a new work through the blacking out or omission of some words or letters (*Erasure | Academy of American Poets* 2021). Another way to look at the deletion task, then, is that our algorithm is taking an original limerick and creating

4. On “Prosodic,” see <https://litlab.stanford.edu/hooddistance>. The longstanding digital humanities tool, “For Better For Verse,” can be found at <https://prosody.lib.virginia.edu/>

an erasure poem out of it. This task raises all sorts of questions about the human and machine discernment of erasure poems. What exactly might distinguish “good” erasure poetry from the randomized omission of a composition’s original words? Although we are assuming for our purposes that the erasure poem is the “incorrect” choice, how might models learn to recognize legitimate or artistically compelling instances of erasure poems? The task also raises interesting ontological questions about the nature of found and manipulated poetry like erasure poems.

2.4. (Poor) Last Line Completion

Our second minimal pair experiment tests models on their ability to distinguish an existing limerick from a similar limerick in which the fifth line of the original limerick has been overwritten by a computer-generated fifth line. The first four lines of each poem in the pair are identical. This experiment involves, then, a stage where a natural language generation technique—whatever technique one uses—completes a four-line fragment of a limerick with a plausible fifth line that fulfills some of the requirements of the limerick form: primarily the *a* end rhyme (that rhymes with lines 1 and 2 in the *aabba* rhyme scheme) and a line length that also generally conforms to those of lines 1 and 2 (typically the longer lines of the limerick in terms of the number of words, compared to lines 3 and 4). (In our experiment, we did not require the synthetic fifth line to meet the rhythm requirement of three stresses and triplet meter.)

This challenge asks the language model being evaluated to discern the original limerick from the synthetic version. This is testing whether or not the model has some minimal sense of how limericks typically end. The machine-generated last line is not a typical line, but it is a good fake, so a model purporting to model narrative and/or semantics cannot “cheat” by just looking to see if it has the right number of syllables and rhyme scheme. Thus, the last line completion minimal pairs also probe a language model’s ability to encode coherence in the limerick. Note that is possible that a language model performs well on word deletion and not on line completion or vice versa.

This test pushes the definition of a “minimal” alteration to its limit in that we replace an entire line of a limerick. Yet this aspect of the experiment recalls the history of the limerick form. Although the origins of the limerick form are obscure, we do know that it was initially popularized in the nineteenth century by Edward Lear’s *A Book of Nonsense* (1846) and then reached another peak of popularity around the turn of the twentieth century. During this latter peak – called by one limerick historian “the great limerick boom” – several promotional competitions elicited the public to submit a final line for a limerick fragment, for a chance at winning substantial prize money (McInerney 2001). The most well-documented example in the limerick literature is a contest put on in 1907 by the cigarette company, J. Samuda & Co. (Legman 1969). Offering a prize of 3£ per week for the rest of the winner’s life, Samuda’s contest asked the public to complete the following limerick fragment, which advertised the company’s product:

That the “Traylee’s” the Best Cigarette,
Is a tip that you cannot regret:

And in buying, I'll mention
There's a three-pound-a-week pension, ...

294

295

September 20, 1907. T.P. WEEKLY. 875

THE GREATEST OF ALL LIMERICKS

£3 per Week for Life. **£3 per Week for Life.**

Think What it Means to You!

£3 PER WEEK FOR LIFE

Mr. J. Samuda's Great Limerick Competition may make you independent as long as you live.

NO ENTRANCE FEE. NO RISK. GUARANTEED PENSION.

ABSOLUTE FAIRNESS ASSURED.

The Competition.
This is certainly the greatest of all the Limerick Competitions. Mr. Samuda will pay £3 per week for life to the actual winner—no Commission! Prizes for £25 each to the runners-up, and third prizes of £10 each, and so on, until all the £100,000 of the fund is paid out. And the £100,000 is paid out in 21 days.

What £3 per Week Means.
Try and realize what £3 per week would mean to you. Every week you would receive £3—it would keep you in comfort or health. If you go on working it would double your income at a stroke. It is no chance that will provide you with every luxury you now enjoy, a better home, better food, better clothes, a better car. You would not have to fear sickness or accident, old age, poverty, or death.

The Object.
Mr. Samuda is the well-known manufacturer of the celebrated "Aston" Cigarettes, and he has recently placed on the market a new Virginia Cigarette, the "Traylee" Cigarette, which is sold in boxes of 100 in a tin. As all readers will undoubtedly be anxious to successfully place any new Cigarette on the market, Mr. Samuda has decided to establish this Limerick Competition for the express purpose of bringing the "Traylee" Cigarette prominently before the notice of the British public.

The Conditions.
The conditions are simple. Below we give a list of them. All you have to do is to send the line and fill in the last line of the limerick, sending same to us with your order for 100 "Traylee" Cigarettes, together with a postal order for 2s. 6d. and 3d. for postage. In 48 to 60 days your money Cigarettes will come, and you will pay 2s. 6d. for 100, even if they buy them in bulk of 10 for 10, or their successors. As the "Traylee" Virginia Cigarette is far superior to any other Cigarette sold at this price, each competitor gets

full value for his money, and without additional cost, or risk, has an opportunity of winning a life long independence. Every man who smokes Cigarettes should try as a matter of course. It costs him nothing.

The Judges. Absolute Fairness Guaranteed.
As this is a special competition, and not a weekly feature, every line submitted will be carefully judged by an independent committee including the Editor of a well-known literary journal. In this way no unfairness can rest assured that the best line will win. There will be no attempt of change about the award, for the judging and consideration of the solutions rest entirely with the independent committee, who have not the slightest interest in this contest.

Solutions Should be Sent in at Once.
As the committee do not wish their work to be impeded in the last few days of the competition, competitors are earnestly requested to post their entries and solutions at the earliest possible moment.

Important.
Any number of solutions may be sent in, providing each solution is accompanied by a separate order for 100 "Traylee" Virginia Cigarettes, and the price including postage, 2s. 6d. Additional lines should be written clearly on a plain sheet of paper, and placed in the left margin. Competitors are especially requested not to write letters, or the names of friends, on the left margin, as the committee will not have time to read or answer same.

Your Money Refunded.
All orders will be executed in strict rotation, and the cigarettes will be dispatched at the earliest possible moment. If any competitor after having paid, say, five of the cigarettes, does not consider them better in quality than any other cigarette he has purchased at the same price, he may return the balance and have his money refunded; but in such cases it is absolutely understood that the solution sent in is thereby cancelled and annulled. Therefore only hold good to which seven days of the closing of the competition.

A Final Word.
Readers should remember that in sending their order for 100 "Traylee" Virginia Cigarettes, they secure full value for their contribution, as their chance to win the large prize of £3 per week for life, or one of the many smaller prizes, does not cost them a single penny. Mr. Samuda fully realizes that the quality of the "Traylee" Virginia Cigarette, which will surely equal any other and surpasses any recommendation to smokers for whom it is made, is the greatest recommendation of the cigarette, and he has decided to make the contest of a few years' time the original prize will thus be secured.

To Wives, Mothers, and Sisters.
Alone every day your husband, son, brother, or sister smokes cigarettes. Why not make upon his purchasing 100 "Traylee" Cigarettes and let your share competition? Think it over! It may mean your being independent for the rest of your life!

Points to Remember.
In the interests of competitors the following points are written for their guidance:
1. An independent committee of judges will decide.
2. Every Limerick sent in will be read and considered.
3. The money will be sent in the "Daily Mail," the "Daily Express," and the "Daily Telegraph" on Thursday, October 10th, 1907.
4. Cigarettes will be sent to prize-winners within a fortnight of the closing of the contest.
5. The winner's weekly income will be remitted from the new fine competition fund, and cigarettes will be forwarded on the first of each month afterwards.
6. No member of the firm, employee, or friend of any member or employee, will be eligible to compete.
Any number of solutions may be sent, providing each is accompanied by a separate order for 100 "Traylee" Cigarettes, and the price, 2s. 6d.

CUT THIS COUPON OUT AND FILL UP CAREFULLY—

ENTRY COUPON.

To the Limerick Committee,
Mr. J. SAMUDA & CO.,
21, Leadenhall Street, London, E.C.

I enclose—Kindly send me post paid a box of 100 "Traylee" Virginia Cigarettes, for which I enclose P.O. 2s. 6d. and 3d. for postage.
If I receive my cigarettes and cannot receive this order you are at liberty to return my solution and money, and by so doing I release you from all liability in connection with the competition.

Full Name _____
Postal Address _____
Town _____

N.B.—All orders will be dispatched in strict rotation.

Here is an unfinished Limerick, to which the last line should be added:
That the "Traylee" the Best Cigarette,
Is a tip that you cannot regret;
And in buying, I'll mention
There's a three-pound-a-week pension.

I agree to accept the Committee's decision as final, and enter the Coupon on this subject and enclosing.

Please sign here _____
Address Envelope _____
The Limerick Committee,
Mr. J. SAMUDA & CO.,
21, Leadenhall Street, London, E.C.

If more than one solution is sent write same on a plain piece of paper and pin to this entry form (see Rules).

Figure 1: J. Samuda & Co. advertisement

J. Samuda & Co.'s advertisement proclaimed that the competition would identify "the 296
greatest of all limericks," but really, as the ad itself reveals, the competition was a 297
gimmick aimed at promoting a new product, Traylee Cigarettes (see Figure 1). Hence 298
contestants needed to mail in an order for the cigarettes in order to submit a fifth line 299
for the limerick. Evidently, J. Samuda & Co. went on to hold several such promotional 300
competitions involving limerick completion, alongside other similar competitions put 301
on by others. In 1907 alone, there were more than 11 million postal orders for such 302
limerick contests. The winning line for J. Samuda & Co.'s 1907 contest was, "Two good 303
'lines' – one you give, one you get," punning on "line" (the poetic line, the cigarette, and 304
the financial life line) (McInerney 2001). 305

In any case, what matters here is less the history of limerick contests *per se* than the 306

fascinating resonances between the poetic challenges we have set for language models 307
and such historical practices and precedents. These resonances – between manuscript 308
omissions and deletions tasks, between popular poetry contests and the autoregressive 309
generation of final limerick lines – suggest a rich direction for future media archaeology- 310
and historical poetics-inflected inquiries into the history of poetry (and texts generally) 311
and the methodologies of contemporary computational literary studies. 312

2.5. Dataset 313

The limericks used for the research in this paper originated as a subset of the content from 314
the website *The Omnificent English Dictionary in Limerick Form* ("OEDILF"). Established 315
in 2004, it is an amateur, crowd-sourced project whose goal is to have at least one limerick 316
for every meaning of every word found in the *Oxford English Dictionary*. User-submitted 317
limericks are subject to approval by moderators, and, if approved, are published on the 318
website.⁵ Among the benefits of OEDILF is that it comprises a large number of limericks 319
which can be sorted according to different categories of metadata.⁶ On the website, the 320
limericks are organized according to different categories: (a) authors, (b) topics. Last 321
but not least, many printed anthologies of limericks highlight particularly misogynistic 322
and/or racist limericks. While OEDILF has its share of ribald poems, it skews in such a 323
way that it provides a large archive of poems from which we can create a good corpus. 324

We work from a subset of OEDILF originally gathered in (Abdibayev, Riddell, and 325
Rockmore 2021). Therein, two levels of filtering reduced the original 110,610 published 326
limericks to a set of 65,000. The first level was based on simple structural criteria 327
(limericks must have 5 lines and must use words – rather than symbols, like emojis or 328
formulae). A next level kept only those limericks that verifiably – by machine – satisfied 329
basic structural properties. Specifically, only limericks where all end-of-the-line words 330
could be verified by machine as satisfying the rhyme scheme were kept. For our fifth 331
line completion task we only picked limericks whose end rhyme words for the first and 332
second lines could be located within our rhyming dictionary. We then used this set for 333
our beam-search algorithm to generate synthetic fifth lines (see section 3.4). 334

For the deleted words minimal pairs, we further filtered limericks to keep only those 335
where at least three end-of-line words are found in Merriam-Webster's Collegiate The- 336
saurus. This produces 34,699 limericks from which we sampled 10,000 limericks (to 337
reduce computational burden). Later we used this smaller set as a testing ground for 338
the delayed beam-search task. 339

3. Experiments and results 340

In this section we set up our experimental procedure and present the results. We first 341
flesh out our experimental design that serves as a framework for the BPoMP tests. 342

5. <http://http://www.oedilf.com/db/Lim.php?View=About>

6. <http://www.oedilf.com/db/Lim.php>

3.1. General structure of all BPoMP tasks 343

3.1.1. Probability of a sequence 344

The BPoMP challenges are probabilistic in nature, in that the final “judgement” of the machine is a comparison of probabilities derived from a pair of inputs. 345 346

In the general schema, a language model G will take as input a variable length sequence of words L , and for each word in the sequence output a probability distribution over the possibilities of a word that it “thinks” should come next in the sequence. The sample space W in our case is some predefined and finite set of words. The probability distribution refers to the collection of probabilities P for all words in the set of words known to the model, which we call its *vocabulary*. 347 348 349 350 351 352

A sequence L can be as short as a sentence or as long as a paragraph. The model produces a probability for each possible next word in a set of words. The word with the highest probability serves as the model’s best “guess” as to what comes next in the sequence, based on what it has witnessed so far. During training we expose the model to gigabytes of human written text, divided into chunks called training examples, and correct the model’s predictions of each word in these training examples – in the sense of modifying an underlying algorithm to give more appropriate probabilities based on the truth – based on words that precede the word in question. 353 354 355 356 357 358 359 360

3.1.2. Tokens 361

The machine works at a slightly finer level of granularity — tokens. In some cases tokens correspond to whole words, while in others (such as ours) they might refer to subword segments, such as astronomical being tokenized (that is split into tokens) as astro and nomical. The vocabulary is then defined as all tokens that we have established through the pre-processing of a – usually very large – training text into subword units by means of a count-based compression algorithm (Sennrich, Haddow, and Birch 2015). 362 363 364 365 366 367

3.2. Formal definition of BPoMP 368

Having established these concepts we now can explain BPoMP’s structure in finer detail. A language model G takes as an input two sequences S and S^* , which correspond to tokenized limericks L and L^* and processes each independently, in no particular order. L^* is a transformation of L , or more formally $L^* = f(L)$, where f is a function that alters (“minimally corrupts”) the original limerick L . The alterations are aimed at singling out particular linguistic phenomenon associated with limericks. 369 370 371 372 373 374

By processing L , we mean outputting $G(S)$ which will provide a probability P for each token in S , the tokenized version of L . We denote all tokens in the vocabulary of the model as V . 375 376 377

Once we compute the probability for each token in both S and S^* , we compute the total probability of sequences S and S^* , via 378 379

$$P(S) = \prod_{i=1}^{|S|} P(w_i | w_{<i})$$

where w_i is the word at a position i and $w_{<i}$ are all words before within the limerick. 380

Or, rewritten for clarity: 381

$$P(S) = \prod_{i=1}^{|S|} G(w_1, \dots, w_{i-1}).$$

$G(w_1, \dots, w_{i-1})$ is the language model's *estimate* of an abstract "true" probability of encountering token w_i given words that come before it. 382 383

Note that some models, such as BERT, violate this formulation by considering the "score" of each word by looking both at preceding and succeeding words. There are workarounds to their more exotic ways of computing (Lau, Armendariz, et al. 2020) what we can instead call the *pseudo-likelihood* of a word (as these formulations do not satisfy the formal properties of a probability distribution). We will not go into the details of this as it does not contribute to the understanding of our experiments. Bigger models (e.g., GPT2-medium v. GPT2) generally give estimates closer to the aforementioned "true" probability. 384 385 386 387 388 389 390 391

In practice we work with $\log P(S) = \sum_{i=1}^{|S|} \log P(w_i)$ instead of $P(S)$ as it simplifies the computation. However, for simplicity of exposition in the examples we will just use probability P . The beauty of this approach lies in its simplicity and universality across models: unlike explicit classification it requires no pretraining and no additional computational units on top of existing model (adding which may introduce more discrepancies in the comparison between models). Moreover, it opens up possibility of studying how much "poetic knowledge" a model can learn without being explicitly constructed to do so. 392 393 394 395 396 397 398 399

We now can delineate the general structure of any BPoMP task. Given a tokenized, human-written limerick S and its tokenized, automatically generated alteration S^* , we ask a model G to compute $\log P(S)$ and $\log P(S^*)$. Comparing the two numbers tells us whether the model finds the original or the alteration to be more likely. When it deems the original more likely it scores a point. Our overall metric is then a simple accuracy measure: divide the total points the model scored by the total number of test examples used in the experiment. An example of a BPoMP test point is presented in Figure 2. 400 401 402 403 404 405 406

3.3. New BPoMPs 407

We now present two new BPoMPs that further refine our understanding of capabilities of these large models to encode poetic concepts.. 408 409

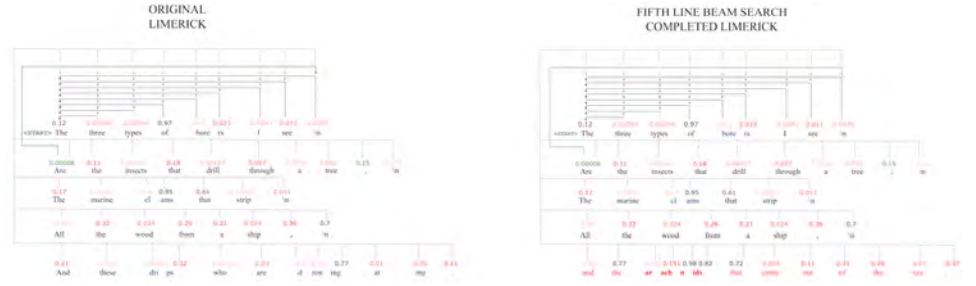


Figure 2: Example of a BPoMP task, specifically, 5th line completion task. To the left is an original, tokenized limerick with each token characterized by its own probability. To the right is an altered limerick, where the 5th line was replaced by a machine-generated one. All probabilities were produced by GPT-2 medium. The colors of probabilities correspond to magnitude. The arrows represent what words were used by the model when predicting the outputted probability. For simplicity, we didn't include all arrows from the second line onward (with the exception of the first word for explanatory purposes). Every model receives a start token at the beginning of every sequence, for which the probability is not computed (presumed to be 1).

3.4. New BPoMP Challenge 1: Random word deletion task

410

In this task we alter the original limerick by deleting M words, for $M \in \{1, 2, 3\}$ (each 411
choice of M is a separate BPoMP task). By “word” we mean any string that is at least 412
2 characters long, separated from other words by spaces. Whenever a word is deleted 413
(whether one or several) its surrounding punctuation is preserved. As noted above, 414
this task somewhat resembles the protocols of erasure poetry. 415

An important consideration is sequence length difference. In this case the deletion 416
will create a twin that is shorter in word length. This in turn will trivially increase 417
its total probability because the summation of log-probabilities is equivalent to the 418
multiplication of regular probabilities, which in itself is lower for longer sequences on 419
average since we are multiplying numbers between 0 and 1. Thus, to correct for this 420
effect we rescale the total probability of each sequence by length. In other words, we 421
compare 422

$$P(S_{\text{original limerick}}) = \frac{1}{|S|} \sum_{i=0}^{|S|} \log P(w_i | w_{<i})$$

against

423

$$P(S_{\text{altered limerick}}) = \frac{1}{|S_{\text{original limerick}}| - M} \sum_{j=0}^{|S_{\text{original limerick}}| - M} \log P(w_j | w_{<j}),$$

This formulation will be used throughout the paper.

424

The results (presented in Table 1) show that most models have no difficulty distin- 425
guishing between original limericks and limericks whose semantics were altered by 426
word removals. However, one model remains an outlier: XLNet performs very poorly 427

Model	Accuracy		
	Delete 1 word	Delete 2 words	Delete 3 words
GPT2	0.89	0.96	0.97
BERT	0.88	0.95	0.97
TransformerXL	0.772	0.8417	0.8816
XLNet	0.4589	0.4412	0.402
Human	0.95	0.975	0.925

Table 1: The BPoMP word deletion test. We delete a varying number (either, 1, 2, or 3) of words (a word is at least two letters long) from a limerick. The task is to pick the original limerick – i.e., to label the text that is most likely to be a limerick. The second line shows the average (2 subjects) human (baseline) performance. All tested models easily solve the task.

compared to others. This can be attributed to the somewhat unnatural causal nature of the task.

A second important consideration is the possibility that the three models that perform well on this task (GPT2, BERT, and TransformerXL) are merely picking up on the grammatical and syntactical conventions they have learned from their training data. In other words, the models are basically functioning as grammar checkers on sentences with missing words. That said, the training data for these models likely include poetic language, and so these models are not only detecting standard prose usage. Still, in order to begin to address this, in the second BPoMP task described below, we test how well models can detect a real limerick ending from a synthetic one – and, in some of those minimal pairs, the limericks exhibit comparable degrees of conformity to what be considered as general and standard English language usage.

Human test subjects were tasked with a slightly different task: to mark if limericks were altered or not. This is due to ease with which one can solve the machine task if presented with both limericks. Their results are uniformly high: performance dips on 3 deleted words task, but that can largely be attributed to small sample size, and perhaps, the greater allowance judges made for what passes as a legitimate limerick.

Below is an example of a random deletion task with 3 words removed from a limerick:

Low to the ground as it goes,	Low to the ground as goes,
The centipede uses its nose	centipede uses its nose
To find insects to eat,	To find to eat,
While an army of feet	While an army of feet
Moves what looks like a flexible hose	Moves what looks like a flexible hose.

3.4.1. Deleting rhyming vs non-rhyming words

Here we present an analysis of the effect of removal of rhyming EOL (end-of-line) words as opposed to any other word. We perform the exact same test on all models as in the previous subsection. In the control group we exclude rhyming EOL words from removal. If rhyming words carry informational importance for models then we expect to see increase in accuracy. Table 2 summarizes the results.

Model	Accuracy					
	1-r	1-nr	2-r	2-nr	3-r	3-nr
GPT2	0.9133	0.8991	0.9146	0.9556	0.9163	0.9819
BERT	0.8754	0.8888	0.8727	0.9505	0.8817	0.9801
TransformerXL	0.7464	0.7769	0.7488	0.841	0.751	0.8818
XLNet	0.2897	0.4978	0.2909	0.4931	0.2843	0.4712

Table 2: The BPoMP rhyming importance test. We delete a varying number (either, 1, 2, or 3) of either rhyming or non-rhyming words (a word is at least two letters long) from a limerick. Rhyming word deletion is denoted as {1, 2, 3}-r in the table. Non-rhyming deletion is denoted {1, 2, 3}-nr. The task is to pick the original limerick – i.e., to label the text that is most likely to be a limerick. By looking at the difference in performance across ten thousand examples, we can theorize on the importance of rhyming words to the models.

With the exception of GPT-2 in a single scenario we can conclude that rhyming words 453
are not heavily utilized by the models and in fact tend to have lower importance in terms 454
of determining original limericks. This is in line with results obtained in (Abdibayev, 455
Riddell, and Rockmore 2021). 456

3.5. Transformer completion task: Beam Search and delayed Beam Search 457

In the Transformer *completion task* we compute the probabilities comparing an original 458
limerick and its corruption obtained by replacing the fifth line with a line generated 459
by *another* Transformer-based language model (a smaller version of GPT-2 – not one of 460
the models that is being tested) given the first four lines. We ensure that the machine 461
completion necessarily rhymes with the first two lines according to the limerick rhyming 462
scheme. In human evaluation, participants picked out the machine-completed line in 463
all 40 of the test pairs. 464

3.5.1. Exact Search 465

To complete the 5th line using a causal model of language (such as GPT-2) we use an 466
algorithm called *search*, but more specifically, a variation on it called *beam search* that we 467
will outline in the next subsection. 468

To understand this algorithm we need to explain its purest form first — *exact search*. 469
Exact search works by considering every possible combination of tokens and computing 470
the sequence probability score for each. We then can choose the most likely sequence of a 471
desired length N using our trained models. The caveat is that we cannot exactly compute 472
the most likely sequence because the size of the search space grows exponentially at 473
each step. 474

As illustration consider generating a most likely sequence of length 5, such as "Jane vis- 475
ited during the snowstorm." For a vocabulary of size 50,265, the exact search for the most 476
likely sequence has the following total cost: the first step would take 50,265 searches, 477
the second 2.5 ($50,265^2$) billion searches, the third 126 trillion ($50,265^3$) searches, the 478
fourth $6.38e^{18}$ ($50,265^4$) searches and the fifth $3.21e^{23}$ ($50,265^5$), which is one order of 479

magnitude less than a trillion of trillions of steps. 480

On average, the 5th line of a limerick in our dataset is 10 tokens long (standard deviation 2.14). Thus, we had to consider different, less computationally intensive methods. 481 482

3.5.2. Beam Search 483

Beam search is an approximation technique that restricts the search terms to only a fixed number (k) of sequences at a time, rather than all of them at once. 484 485

The first step is the simplest: on input of an initial sequence – the first four lines of the limerick – the model produces a probability distribution (for the next word) over all words in the vocabulary, V . We rank them, with the k highest probability words making it through the first round. In the second step, for each of the k words we compute the probability for every possible next word in the vocabulary. With size of vocabulary $|V|$ (50,256 word fragments for GPT2) and k words the total cost is $k|V|$. Similarly, in every subsequent step, we generate a probability distribution over the entire possible vocabulary for each of the k surviving words (or more generally, word sequences) and then we re-rank the resulting $k|V|$ sequences to leave only k most likely ones once again. This is beam search. In practice, previous work (Shaham and Levy 2021) has found that beam search performs surprisingly well despite its limited “field of view”.

Nevertheless, beam search can suffer from a tendency to produce k non-diverse sequences (Holtzman et al. 2019): that is, all of the highly probable outputs are very similar, differing only in one or two words. It rarely ever explores the defined vocabulary, preferring to generate articles (“the,” “a,” etc.), as they are very frequently encountered in any text, meaning that it almost never ends up producing a line that ends with a rhyming word for many limericks. We find that we need to use k of size 500-700 to produce rhyming completions. Thus, to get as many limerick 5th line completions as possible we ran this process for a set of 64,872 limericks which only yielded us 5330 (8%) completions for original limericks. The number is substantially higher (14,155) if we count all completions of the same limerick that vary by end-of-line rhyme word. We tested models with a sample of 10,000 pairs of original-completion pairs. Separately, we sampled 20 pairs of distinct limericks (that is there were no same limericks in the sample) and presented them to our human test subjects.

We accept the generated line as final if (1) we produce at most N tokens (a number slightly higher than number of tokens in the original 5th line, typically by 2), (2) it ends with a sentence stopping symbol, such as a period or an exclamation mark. The second requirement serves as a partial stoppage in the stream of generation, but does not necessarily mean that the model does not intend to keep going. This can be verified if a model produces its own generation stop symbol (different from grammatical sentence stoppage symbols) that in turn tells us that the model does not intend to continue a sequence.

As Table 3 shows, models have a very strong preference for machine-generated completions, while humans do not. If completions were perfectly comparable to human written

Model	GPT-2 medium	BERT	TransformerXL	XLNet (causal)	Human
Accuracy	0.0736	0.1478	0.0253	0.415	1

Table 3: Accuracy for the “5th line replacement BPoMP test”. In this test a new 5th line for a limerick is generated by a neural network model (base GPT2, 142M parameters) given the original four lines. We generated the completions using the base GPT-2 model after using beam search with results that end with a rhyming word. To preserve the quality of our task, these completions have not been selected for a high probability of the sequence but tested on our models regardless of their absolute probability as determined by base GPT-2 that generated them. The limerick with the highest score as per the calculation is then “picked” by the machine as the original. The last column shows the average (2 subjects) human performance on the task of distinguishing the original limerick from the corrupted limerick. The results demonstrate that models do not perform well at picking out machine completed limericks, while humans have no trouble with the task.

ones, we’d expect parity between human and machine performance (floating around 0.5 mark). However, XLNet remains an outlier showing somewhat strong performance. The fact that humans perform well on this task while machines struggle suggests that humans have higher tolerance for what’s considered “valid” language. This further suggests that the underlying model may need to incorporate a more expansive view of what it means to “learn language” in order to “understand” or even identify poetry.

Below are two examples of originals and minimally altered limericks:⁷

Example #1:

In my favourite recipe book	In my favourite recipe book	
Every dish has a photo. I look	Every dish has a photo. I look	
At the words (on the page	At the words (on the page	
The pic faces) to gauge	The pic faces) to gauge	528
How to roast, boil or fry what I cook.	the amount of time it takes for a dish to cook.	

Example #2:

Said a guy whose divorce just went through,	Said a guy whose divorce just went through,	
“I’m so lucky to bid you adieu.	“I’m so lucky to bid you adieu.	
Best of all is I won	Best of all is I won	
At the lotto, and hon,	At the lotto, and hon,	530
I don’t need to share any with you.”	I’m so delighted to have you back.”“Thank you,”	

3.5.3. Delayed Beam Search

To remedy the problem of certain limericks never yielding results using beam search and specifically to produce diverse solutions, we utilize *Delayed Beam Search* (DBS) (Massarelli et al. 2019). DBS samples the first few words (a number we choose empirically; in our experiments the number is 3) using top-p sampling (Holtzman et al. 2019) and then switches to a regular beam search. Top-p sampling works by reducing the probabilities of all words whose value falls outside of a cumulative range of probability

7. The altered limericks only appear to have an extra line, and do not actually have an extra line.

p (a hyperparameter we set), normalizing the rest to sum to 1, and then simply sampling non-uniformly using these (newly normalized) probabilities. After generating the top k sequences using this algorithm we check if the last word rhymes with the end rhyme words of the first and second lines. If so, we accept this completion as valid. Note that in this approach several differing completions of the same limerick can make it into our test set. We run this algorithm on average 1000 times for each limerick and add any valid completions generated during this process, before proceeding to the next limerick. Similarly, in this task we are not guaranteed to produce a completion since it is initially stochastic. When we sample the words we simply pick them from a set using probabilities that the model provides. Due to the costs involved in running this process, we restrict the completion generation to a smaller (compared to beam search) set of limericks. Importantly, in our experiments the beam search part uses $k = 5$.

The motivation behind the DBS approach is two-fold: (1) As mentioned above, regular beam search tends to produce non-diverse text and thus, rarely generates a sequence that contains a valid end rhyme; and (2) top- p sampling tends to produce text that often seems unrelated to its preceding context. That said, combining the two helps to alleviate the drawbacks that we see when either is used alone (Massarelli et al. 2019). We additionally cleaned completions of some garbage symbols generated by the model (such as a newline) and then put these limericks into a test set. After using this approach on 10,000 limericks we generated 4014 test examples derived from 2595 original limericks. Below is one example:⁸

The Absolute: what do we feel
From the Absolute? Not a great deal.
Our emotional scenes
Are directed by genes,
and we've worked hard to make them feel like
they're real.

The Absolute: what do we feel
From the Absolute? Not a great deal.
Our emotional scenes
Are directed by genes,
and such things are not theirs to reveal.

Another example:

Uncle Ed had repaired to his bed
With a terrible pain in his head,
And by noon he was dead—
So the coroner said—
'Cause his cerebral artery bled.

Uncle Ed had repaired to his bed
With a terrible pain in his head,
And by noon he was dead—
So the coroner said—
That would be I, if he had not been
dead.

The results for this task are presented in Table 4. The language models perform poorly, but not as much as previous task, going from 16% average performance (across all models) in the previous task to 19%. In particular, GPT-2-medium correctly identifies 17% of the test cases, as opposed to 7% it did prior.

We believe this is a byproduct of the beam search procedure. By sampling the first three tokens we do not overfit to the model's preferences. Moreover, since DBS partially

8. Note that on the lefthand side, the last line extends to include those last two words – i.e., it is not a six-line poem. The righthand side is the original, correct version!

Model	GPT-2 medium	BERT	TransformerXL	XLNet (causal)	Human
Accuracy	0.17	0.42	0.07	0.12	1

Table 4: Accuracy for the “5th line replacement BPoMP test using Delayed Beam search”. In this test a new 5th line for a limerick is generated by a neural network model (base GPT2, 142M parameters) given the original four lines. The last column shows the average (3 subjects) human performance on the task of distinguishing the original limerick from the corrupted limerick.

samples the completions, the resulting lines were notably poorer in terms of proper grammar and did not follow the logic of the previous lines as closely as the lines generated by a regular beam search. We stress that we never compared these to human written 5th lines during generation, so they were never filtered to beat (i.e., be more probable under the model) the original completions. In turn, a possible explanation is that the models’ statistical “preferences” (those decoded by the beam search procedure) differ from linguistic preferences of humans, at least when not explicitly trained on poetry.

We presented our human test subjects with 20 samples from the generated set of completed limericks accompanied by their originals. All human judges scored perfectly on the test. Our explanation for this is straightforward: the completions generated by the model tend to be visibly longer than the typical completions written by a human (average of 11 words with std of 2.73 compared to 7 words with std 1.23). This suggests that the poetic knowledge of language models still have some way to go in terms of sensing coherence, a punchline, poetic closure, and meaning generally.

4. Future Work

4.1. Rhyme Probability and Artistry

The second BPoMP challenge necessitated generating synthetic fifth lines, only a percentage of which had correct end rhymes to match lines 1 and 2 of the source limerick (given the *aabba* rhyme scheme). In future work, we hope to explore more minutely how language models fare in generating different kinds of rhyme words given a certain initial rhyme, and the broader implications of *improbable* or difficult rhyme words for poetic artistry.

4.2. Poetic Minimal Pairs Examples

The investigation into the poetic knowledge of language models approaches poeticity or literariness using a novel approach. We anticipate future work expanding the BPoMP framework to other kinds of poems beyond the limerick and to other poetic features. Below are a next “level” of examples of minimal pairs. Put aside for now is the issue of preparing such sets, a necessary sub-step that surfaces its own interesting set of challenges in computational poetics.

Ballad or Common Meter 598

Ballad or Common Meter (four-line stanza, with two pairs of a line of iambic tetrameter followed by a line of iambic trimeter. 599
600

Emily Dickinson original vs. minimally flawed example (syllable count). 601

Original: 602

Great streets of silence led away 603

To neighborhoods of pause — 604

Here was no notice — no dissent — 605

No universe — no laws. 606

607

Minimally flawed example: 608

Great streets of silence led away 609

To neighborhoods of pause — 610

Here was no notice — no **resistance** — 611

No universe — no laws. 612

Strong Stress (aka Accentual Meter) 613

Each line has the same number of stresses regardless of the total number of syllables per line. The example is from Samuel Taylor Coleridge's "Christabel" [1816]), where every line in the poem has four accents (with a variable number of total syllables per line): 614
615
616

Original: 617

The **night** is **chill**, the **cloud** is **gray**: 618

'Tis a **month** before the **month** of **May**... 619

Minimally flawed (has extra stress in the second line) 620

The **night** is **chill**, the **cloud** is **gray**: 621

'Tis **many** **months** before the **month** of **May**... 622

Iambic Pentameter 623

Iambic pentameter (from Tennyson, "Ulysses"): 624

Original: 625

...Made **weak** by **time** and **fate**, but **strong** in **will** 626

To **strive**, to **seek**, to **find**, and **not** to **yield**. 627

Minimally flawed (final foot of line 2 is a trochee) 628

...Made weak by time and fate, but strong in will 629

To strive, to seek, to find, and not perish. 630

Rhyme (from Thomas Gray, "Elegy Written in a Country Churchyard") **631**

Original: 632

Full many a gem of purest ray serene, 633
 The dark unfathom'd caves of ocean bear: 634
 Full many a flow'r is born to blush unseen, 635
 And waste its sweetness on the desert air. 636
 637

Minimally flawed example (the fourth line's end rhyme has been altered with a non-rhyme): 638
 639

Full many a gem of purest ray serene, 640
 The dark unfathom'd caves of ocean bear: 641
 Full many a flow'r is born to blush unseen, 642
 And waste its sweetness on the desert **sand**. 643

Rhyme in a Limerick (Edward Lear, "There Was an Old Man with a Beard") **644**

Original: 645

There was an Old Man with a beard, 646
 Who said, "It is just as I feared!— 647
 Two Owls and a Hen, four Larks and a Wren, 648
 Have all built their nests in my beard!" 649

Minimally flawed example (the third line's internal "Hen"- "Wren" rhyme has been disrupted by "Crow"):

There was an Old Man with a beard, 652
 Who said, "It is just as I feared!— 653
 Two Owls and a Hen, four Larks and a **Crow**, 654
 Have all built their nests in my beard!" 655

Assonance (from John Keats, "Ode on a Grecian Urn") **656**

Original (recurring long "i"s) 657

Thou still unravished bride of quietness, 658
 Thou foster-child of silence and slow time... 659

Minimally flawed example: 660

Thou still unravished bride of quietness, 661
 Thou foster-child of **muteness** and slow time... 662

Alliteration (from Shakespeare's Sonnet #30) **663**

Original (recurring sibilants) 664

When to the sessions of sweet silent thought	665
I summon up remembrance of things past	666
Minimally flawed example (in the second line, “summon” is replaced by “conjure”):	667
When to the sessions of sweet silent thought	668
I conjure up remembrance of things past	669
Consonance (from W.H. Auden, “That night when joy began”)	670
Original (consonance in “flush” and “flash”)	671
That night when joy began	672
Our narrowest veins to flush ,	673
We waited for the flash	674
Of morning’s levelled gun.	675
Minimally flawed example (the “flush”-“flash” consonance is disrupted):	676
That night when joy began	677
Our narrowest veins to flush,	678
We waited for the blaze	679
Of morning’s levelled gun.	680
Imagery and Meaning (from Elizabeth Bishop, “Pink Dog”)	681
Original:	682
Oh, never have I seen a dog so bare!	683
Naked and pink, without a single hair...	684
Startled, the passersby draw back and stare.	685
Minimally flawed example (in the second line, the imagery is made less consistent by replacing “hair” with “care”):	686
Oh, never have I seen a dog so bare!	688
Naked and pink, without a single care ...	689
Startled, the passersby draw back and stare.	690
Chiasmus and Meaning (from Emily Dickinson, “Much Madness is divinest sense”)	691
Original:	692
Much Madness is divinest Sense -	693
To a discerning Eye -	694
Much Sense - the starkest Madness...	695
Minimally flawed example (in the third line, the parallelism and meaning are disrupted by replacing “Sense” with “Nonsense”):	696
	697

Much Madness is divinest Sense - 698
 To a discerning Eye - 699
 Much **Nonsense** - the starkest Madness... 700

5. Conclusion 701

In this paper, we reported on our experiments in computational poetics with the limerick, thereby continuing its use as a “model organism” for the discipline. Namely, we presented the formulation of and outcome of two tests constructed using the “minimal pairs” experimental method for poetry (BPoMP), which are designed to probe the extent to which language models can classify good limericks from slightly altered ones. The language models performed quite well in the first challenge, where an original limerick was compared with its “corrupted twin,” the same but with a few words omitted (which had the effect of disrupting the poem’s grammar, syntax, and meaning). In the second challenge, we gave language models a choice between an original limerick and the same limerick except the latter’s fifth line now given by a plausible machine-generated replacement for the original final line. On this task, models demonstrate much room for improvement. 702 703 704 705 706 707 708 709 710 711 712 713

Both BPoMP challenges raise all manner of interesting questions about models and their ability to detect human-generated verse from computer-generated verse; resemblances between these tasks and methods of textual criticism, erasure poetry, and the history of the limerick form; and more. Our experiments also point us to future avenues of inquiry, including additional minimal pair challenges that isolate different features of poetry, rhyming artistry, and other unexpected resonances and challenges at the intersection of language models, textual criticism and literary history and analysis. 714 715 716 717 718 719 720

6. Data availability 721

Data can be found here: <https://anonymous.4open.science/r/BPomP-Benchmark-of-Poetic-Minimal-Pairs-6CD7/> 722
723

7. Acknowledgements 724

8. Author contributions 725

Almas Abdibayev: Conceptualization, Data Analysis and Preparation, Programming, Writing – original draft 726
727

Yohei Igarashi: Poetics, Literary Criticism, Data Analysis, Writing - original draft 728

Allen Riddell: Methodology, Conceptualization, Data Analysis and Preparation, Writing – original draft 729
730

Daniel Rockmore: Methodology, Conceptualization, Data Analysis, Writing – original draft 731
732

References 733

- Abdibayev, Almas, Yohei Igarashi, Allen Riddell, and Daniel Rockmore (Nov. 2021). “Automating the Detection of Poetic Features: The Limerick as Model Organism”. In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, pp. 80–90. URL: <https://aclanthology.org/2021.latechclfl-1.9>. 734
735
736
737
738
739
- Abdibayev, Almas, Allen Riddell, and Daniel Rockmore (Sept. 2021). “BPoMP: The Benchmark of Poetic Minimal Pairs – Limericks, Rhyme, and Narrative Coherence”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd., pp. 1–9. URL: <https://aclanthology.org/2021.ranlp-main.1>. 740
741
742
743
744
- Algee-Hewitt, Mark (2017). *Canon/Archive: studies in quantitative formalism from the Stanford Literary Lab*. Ed. by Franco Moretti. New York: n+1 Foundation. 315 pp. ISBN: 978-0-9970318-7-4. 745
746
747
- Ankeny, Rachel A. and Sabina Leonelli (Nov. 2020). “Model Organisms”. In: *Elements in the Philosophy of Biology*. ISBN: 9781108593014 9781108742320 Publisher: Cambridge University Press. DOI: [10.1017/9781108593014](https://doi.org/10.1017/9781108593014). URL: <https://www.cambridge.org/core/elements/model-organisms/F895B26EAC0373BCA5A138835AC73AEA> (visited on 11/28/2021). 748
749
750
751
752

- Anttila, Arto and Ryan Heuser (June 21, 2016). "Phonological and Metrical Variation across Genres". In: *Proceedings of the Annual Meetings on Phonology* 3.0. ISSN: 2377-3324. DOI: [10.3765/amp.v3i0.3679](https://doi.org/10.3765/amp.v3i0.3679). URL: <http://journals.linguisticsociety.org/proceedings/index.php/amphonology/article/view/3679> (visited on 08/03/2021).
- Bode, Katherine (2018). *A world of fiction: digital collections and the future of literary history*. Ann Arbor, MI: University of Michigan Press. 252 pp. ISBN: 978-0-472-13085-6.
- Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020). "Language models are few-shot learners". In: *arXiv preprint arXiv:2005.14165*.
- Chen, Stanley, Douglas Beeferman, and Ronald Rosenfeld (1998). *Evaluation Metrics For Language Models*.
- Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith (Aug. 2021). "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 7282–7296. DOI: [10.18653/v1/2021.acl-long.565](https://doi.org/10.18653/v1/2021.acl-long.565). URL: <https://aclanthology.org/2021.acl-long.565>.
- Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov (2019). "Transformer-xl: Attentive language models beyond a fixed-length context". In: *arXiv preprint arXiv:1901.02860*.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.
- Erasure | Academy of American Poets (2021). URL: <https://poets.org/glossary/erasure> (visited on 12/13/2021).
- Fujimura, Joan H (1992). "Crafting Science: Standardized Packages, Boundary Objects and 'Translation'". In: *Science as Practice and Culture*. Ed. by A Pickering. Chicago: University of Chicago Press, pp. 168–211.
- Ghazvininejad, Marjan, Xing Shi, Jay Priyadarshi, and Kevin Knight (2017). "Hafez: an Interactive Poetry Generation System". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pp. 43–48.
- Goldberg, Yoav (Apr. 2017). *Neural Network Methods in Natural Language Processing*. Ed. by Graeme Hirst. San Rafael: Morgan & Claypool Publishers. ISBN: 978-1-62705-298-6.
- Gurney, Kevin (1997). *An Introduction to Neural Networks*. UCL Press.
- Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi (2019). "The curious case of neural text degeneration". In: *arXiv preprint arXiv:1904.09751*.
- Houston, Natalie M. (2014). "Toward a Computational Analysis of Victorian Poetics". In: *Victorian Studies* 56.3. Publisher: Indiana University Press, pp. 498–510. ISSN: 0042-5222. DOI: [10.2979/victorianstudies.56.3.498](https://doi.org/10.2979/victorianstudies.56.3.498). URL: <https://www.jstor.org/stable/10.2979/victorianstudies.56.3.498> (visited on 12/11/2021).
- Jurafsky, Dan and James H. Martin (2021). *Speech and Language Processing*. 3rd ed. draft. Prentice Hall.

- Lau, Jey Han, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu (2020). "How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context". In: *Transactions of the Association for Computational Linguistics* 8, pp. 296–310. In: *Transactions of the Association for Computational Linguistics* 8, pp. 296–310.
- Lau, Jey Han, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond (July 2018). "Deep-speare: A joint neural model of poetic language, meter and rhyme". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1948–1958. DOI: [10.18653/v1/P18-1181](https://doi.org/10.18653/v1/P18-1181). URL: <https://aclanthology.org/P18-1181>.
- Legman, G[ershon] (1969). *The Limerick : 1700 examples, with notes, variants, and index*. eng. New York: Bell Pub. Co.
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv: [1907.11692](https://arxiv.org/abs/1907.11692) [cs.CL].
- Long, Hoyt and Richard Jean So (2016). "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning". eng. In: *Critical inquiry* 42.2, pp. 235–267. ISSN: 0093-1896.
- Long, Hoyt J. (2021). *The values in numbers: reading Japanese literature in a global information age*. New York: Columbia University Press. 1 p. ISBN: 978-0-231-55034-5.
- Massarelli, Luca, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel (2019). "How decoding strategies affect the verifiability of generated text". In: *arXiv preprint arXiv:1911.03587*.
- McCulloch, Warren S. and Walter Pitts (1958). "A logical calculus of the ideas immanent in nervous activity". In: *Bulletin of Mathematical Biophysics* 5.4, pp. 115–133.
- McInerney, Vincent (2001). *Writing for radio*. Manchester ; New York : New York, NY: Manchester University Press ; Distributed exclusively in the USA by Palgrave. 276 pp. ISBN: 978-0-7190-5842-4 978-0-7190-5843-1.
- Piper, Andrew (2018). *Enumerations: data and literary study*. Chicago ; London: The University of Chicago Press. 243 pp. ISBN: 978-0-226-56861-4 978-0-226-56875-1.
- Poovey, Mary (2001). "The model system of contemporary literary criticism". In: *Critical Inquiry* 27.3, pp. 408–438.
- Preminger, Alex, Terry V.F. Brogan, and Frank J. Warnke, eds. (1993). *New Princeton Encyclopedia of Poetry and Poetics*. Princeton University Press.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9.
- Reynolds, L. D. and N. G. Wilson (1991). *Scribes and scholars: a guide to the transmission of Greek and Latin literature*. 3rd ed. Oxford : New York: Clarendon Press ; Oxford University Press. 321 pp. ISBN: 978-0-19-872145-1 978-0-19-872146-8.
- Rosenblatt, Frank (1943). "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". In: *Psychological Review* 65.6, pp. 386–408.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch (2015). "Neural machine translation of rare words with subword units". In: *arXiv preprint arXiv:1508.07909*.

- Shaham, Uri and Omer Levy (2021). "What Do You Get When You Cross Beam Search with Nucleus Sampling?" In: *arXiv preprint arXiv:2107.09729*. 840
841
- So, Richard Jean (2020). *Redlining culture: a data history of racial inequality and postwar fiction*. New York: Columbia University Press. ISBN: 978-0-231-19772-4 978-0-231-19773-1. 842
843
844
- Underwood, Ted (2019). *Distant horizons: digital evidence and literary change*. Chicago: The University of Chicago Press. 206 pp. ISBN: 978-0-226-61266-9 978-0-226-61283-6. 845
846
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008. 847
849
- Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman (2020). "BLiMP: The Benchmark of Linguistic Minimal Pairs for English". In: *Transactions of the Association for Computational Linguistics* 8, pp. 377–392. DOI: [10.1162/tac1_a_00321](https://doi.org/10.1162/tac1_a_00321). eprint: https://doi.org/10.1162/tac1_a_00321. URL: https://doi.org/10.1162/tac1_a_00321. 850
851
852
853
854
- Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le (2019). "Xlnet: Generalized autoregressive pretraining for language understanding". In: *arXiv preprint arXiv:1906.08237*. 855
856
857

Yvonne Völkl  ¹

Sanja Sarić  ²

Martina Scholger  ²

1. Institute of Interactive Systems and Data Science (ISDS), Graz University of Technology, Graz.

2. Institute Centre for Information Modelling - Austrian Centre for Digital Humanities, University of Graz, Graz.

Keywords:

topic modeling, French, Spanish,
Spectator press, 18th century,
Literary Gender Studies

Licenses:

This article is licensed under:



Abstract. Gender-specific knowledge – just like knowledge in general – is generated through discourses that are disseminated through (mass) media. Among the first mass media is the Spectator press (*Moralische Wochenschriften*), which spread all over Europe throughout the 18th century. With their gender-specific discourses, analyzed in *Spectatoriale Geschlechterkonstruktionen* (Völkl 2022), they decisively promote the development of a (bourgeois) gender model, shaping the social perception of gender until today. Against this background, the present article examines the gender-specific discourses in the French and Spanish Spectator periodicals by means of topic modeling, which detects semantically related words. The study, which originates from the project *Distant Spectators. Distant Reading for Periodicals of the Enlightenment* (Scholger et al. 2019–2021), shows that topic modeling reinforces previous findings on gender-specific discourses in the Spectator periodicals. Moreover, it offers new perspectives concerning this research corpus.

1. Introduction

1

The Spectator periodicals are a popular journalistic genre of the 18th century which (co-)constructs and preserves the cultural knowledge of its time in general and gender-specific knowledge in particular, propagating a heteronormative society. As a broadly effective medium circulating from England throughout the Western world, the Spectator periodicals also promote the transcultural dissemination of a transforming understanding of gender,¹ in conjunction with the changing values, norms and practices among

2

3

4

5

6

7

1. In the course of the 18th century, the perception of female and male bodies and their genitalia changed. The so far gradually assumed difference between women and men is increasingly interpreted qualitatively and a complementary understanding of two genders can assert itself (cf. Laqueur 2003[1990]). The new perception of women and men also leads to a cultural redefinition of their gender relations (e.g. woman as the 'moral gender', cf. Steinbrügge 1987) and to a major shift in the conception of virtue, which was originally

the constituting middle classes. 8

The quantitative-statistical as well as the discourse-analytical and interpretative study 9
on gender-specific ways of worldmaking in the French- and Spanish-language Spectator 10
periodicals (Völkl 2022) reveals that the French-language periodicals of the first half 11
of the 18th century contribute to the dissemination of the notion of a ‘natural’ gender 12
difference, which primarily appears together with a discourse of character and/or 13
physical differences. From the middle of the century onwards, in which the periodicals 14
are also published in Spain,² the discourse of difference is expanded to include the aspect 15
of complementarity, finally recognizing woman and man as a mutually complementary 16
entity. Due to her alleged closeness to nature, in this discourse of complementarity, the 17
woman is hierarchically placed under the authority of man, whose assumed higher 18
ability to reason is considered superior. 19

In order to disseminate the discourses of difference and complementarity, the French- 20
and Spanish-language periodicals draw particularly on the notion of virtue (French: 21
vertu; Spanish: *virtud*). According to research on the Enlightenment period, this term 22
generally functions as a gender-specific key concept (‘geschlechtsspezifischer Leitbegriff’ 23
according to Pabst 2007) and stands in opposition to the notion of vice (French: *vice*; 24
Spanish: *vicio*) (cf. Bolufer Peruga 1998, Kilian 2002, Schaufler 2002, Steinbrügge 1987). 25
Furthermore, the discourse on virtues and vices is combined with positive and negative 26
(character) traits and behavioral patterns, which are hierarchized and assessed as worthy 27
or not worthy of emulation. Among the ignoble vices on the one hand, one can find, for 28
example, hypocrisy, idleness, vanity, or jealousy, which should be avoided by women and 29
men alike (and thus remain gender-*unspecific*). The virtues worthy of emulation on the 30
other hand, are constructed in a gender-specific way, with the ‘female’ virtues revolving 31
around concepts such as decency, modesty, kindness, shamefulness, beauty or (a specific 32
female) education, while the ‘male’ virtues only include (a specific male) education, 33
honesty, and reason. In order to make the large number of virtues and vices known to 34
the Spectator audience – which decidedly also included women – they are incorporated 35
into gender-stereotypical models, illustrating ideal images or warning examples. E.g. 36
the characteristics of egoism and vanity, which are considered vicious, are linked to the 37
stereotypical models of the coquette or the fop and contrasted with virtuous models of 38
women and men. The gender-stereotypical models with their manifold virtues and vices 39
are enveloped into countless (exemplary) stories from everyday life and in (character) 40
portraits, which are narratively woven into the plot (cf. Völkl 2022, 282–286). 41

To quantitatively verify these observations on the (entire) Spectator corpus, a topic 42
modeling analysis was carried out in the course of the project *Distant Spectators. Distant* 43
Reading for Periodicals of the Enlightenment (DiSpecs) (Scholger et al. 2019–2021), after 44
which special attention was given to the interpretation of those topics that stand out from 45
a gender-theoretical perspective. The computed values and their visual representation 46

connotated to the meritorious properties and qualities of men (Latin: *vir*) and was feminized as of the end of 47
the 17th century only (cf. Pabst 2007, 25ff.).

2. For a description of the Spanish Spectator periodicals and an in-depth analysis of the use and function of 48
the letter as mode of communication with the public, see Hobisch 2017.

were intended to provide a new perspective on the corpus and create new theories and questions. The following chapters first describe the related work, the research material, and the methodology, before presenting the results and findings of the topic modeling analysis with regard to gender-specific discourse in the Spectator periodicals.

2. Related work

Topic modeling has become an integral part of the range of methods used in digital humanities, and more specifically in computational literary studies. According to the survey of Du, it has been increasingly used since 2011 (cf. Du 2019). In the field of historical newspapers and periodicals, topic modeling was conducted for analyzing the social and political life of Civil War Richmond based on the *Richmond Daily Dispatch* (cf. Nelson 2020) and for investigating the discourse dynamics in historical newspapers published in Finland between 1854 and 1917 (cf. Marjanen et al. 2020). Regarding the Enlightenment period, Schöch applied topic modeling on French Classical and Enlightenment drama for sub-genre classification (cf. Schöch 2017), and Roe et al. analyzed the discursive structure in the *Encyclopédie* of Denis Diderot and Jean le Rond d'Alembert (cf. Roe, Gladstone, and Morrissey 2016).

A persistent point of criticism in the application of topic modeling is the lack of explainability and comprehensibility of the results (cf. Hu et al. 2014, 424–425, Liu et al. 2017, 1–2). This is very much related to the lack of documentation of single working steps and parameters applied in the topic modeling process, as Du pointed out (cf. Du 2019): In order to guarantee the reproducibility of the results, it is crucial to have details on the number of documents, the conducted pre-processing steps, the number of topics and iterations selected in the actual modeling process, etc. To address this criticism, this contribution aims to not only provide the results of our topic modeling analyses, but also to transparently document the workflow that led to them.

3. The research corpus

While the Spectators have previously been studied through close reading as a work-centred approach, there have been no previous activities that explore this genre from a distant reading perspective. For this reason, the project DiSpecs engaged in text mining of the collection of 3,863 periodical issues in six languages,³ assembled and edited during the digital scholarly edition project *The 'Spectators' in the International Context* (Ertler et al. 2011–2021). In the DiSpecs project, topic modeling was used for investigating the semantic and stylistic structure.

What proved to be very useful for the analysis was the fact that the documents were already available in XML/TEI format (TEI Consortium 2021). This includes not only

3. The corpus contains periodicals in English, French, German, Italian, Portuguese, and Spanish, but due to the rather small corpus size of English, German, and Portuguese, these languages were not considered in the topic modeling analysis. Therefore, we analyzed 1,658 French-, 1,344 Italian-, and 690 Spanish-language issues.

the annotation of metadata and structural elements (e.g. paragraphs and pagination),
but also narrative forms (e.g. self-portrait, letter/letter to the editor, fable) and narrative
levels of representation,⁴ as well as subjects (e.g. 'Idea of man', 'Nature', 'Economy',
'Theatre Literature Arts'), mentioned places, person names, and intellectual works. The
annotation format provided through the application of the Text Encoding Initiative
(TEI) standard enables easier extraction of certain structures of the data for the analysis
(e.g. metadata, headings, footnotes, editorial comments) with the possibility to separate
issues into paragraphs and to exclude parts or whole issues during the pre-processing
of the data, which will be explained closer in subsection 4.2.

4. Topic modeling workflow

The unsupervised probabilistic topic modeling method aims to identify hidden thematic
structures in large text collections (cf. Blei 2020, 8), which means that the algorithm
recognizes patterns in the data without having a training subset or a desired output
(cf. Alloghani et al. 2020, 4). The resulting topics usually consist of thematically related
words, i.e. tokens. However, some topics have structural rather than thematic signifi-
cance. They can provide insight into the writing style of the author, terms typical for
a genre, adjectives describing a matter, repeatedly mentioned places or persons. This
is due to the fact that the method's algorithms measure the co-occurrence of words,
following statistical assumptions, meaning that if the same words often occur together
in documents, they are most likely thematically related (cf. Blei 2020, 9).

Multiple algorithms were developed for topic modeling, but one of the most prominent,
and the one we used in our analysis, is Latent Dirichlet Allocation (LDA). We owe
this choice to the DARIAH-DE team, who developed a Jupyter Notebook embedding
the *dariah_topics* Python library for topic modeling (DARIAH-DE 2019) with MALLET
(McCallum 2002–2018), a toolkit that builds on LDA. We adapted and expanded these
notebooks to incorporate them into our topic modeling analysis workflow,⁵ which can be
divided into four main parts: data evaluation, pre-processing of the data, topic modeling
creation and post-processing of the results (Figure 1). As we demonstrate in this chart,
individual steps of the workflow have to be repeated to optimize the results. Further on
in this chapter, we will describe how we conducted these steps and what decisions were
important for quality results.

4. The Spectator periodicals stand out for their multi-layered system of communication consisting of various narrative levels of representation, which are embedded in various narrative forms. Further, narrative forms are also intertwined within each other when e.g. the fictitious editor includes a supposedly authentic reader's letter in the periodical, which, in turn, narrates a story about a woman, who then enters into a dialogue with another woman about a letter to a man (cf. Fischer 2014, 81–83). This epistolary correspondence between editor and readers has been considered a major element for the success of the Spectator periodicals in the course of the 18th century (cf. Hobisch 2018).

5. The Jupyter Notebooks with the Python code are provided via GitHub: <https://github.com/distantस्पेक्टатор/DiSpecs>.

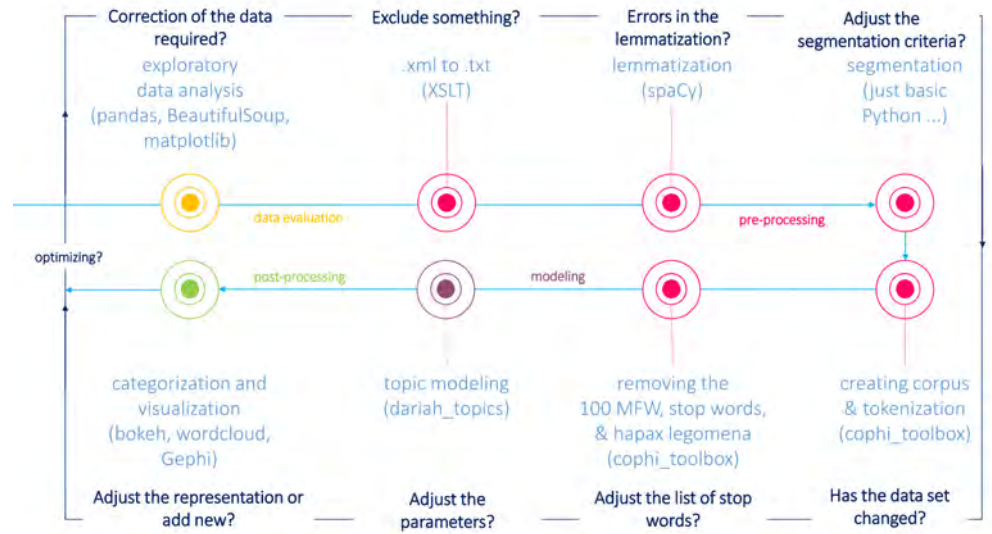


Figure 1: The topic modeling workflow.

4.1. Data evaluation

113

To get an overview of the French- and Spanish-language research data, we conducted a number of exploratory data analysis steps. This included evaluating and visualizing the size of the corpus, the number of issues per periodical and per author, the number of tokens per issue, as well as the distribution of manually assigned keywords and narrative forms. This simple statistical analysis allows insight into the corpus, which can be relevant for interpreting and evaluating the results. For example, comparing manually assigned keywords with topics identified through topic modeling is used for cross-evaluation of these two approaches, i.e. finding out how similar the human- and the machine-assigned topics are. In addition, discrepancies in the metadata could be detected and corrected. Getting this insight was possible thanks to the TEI annotation of the Spectator corpus, which allows extracting all the relevant data structures from the documents, either with Python libraries like *Beautiful Soup* or with XSLT while transforming the XML/TEI to plain text files.

4.2. Pre-processing

127

Our workflow, building on DARIAH-DE, required plain text files as input for topic modeling. As part of this transformation, we extracted metadata from the TEI files and used it to build file names: publication year, periodical name, author of the periodical, volume, issue, and persistent identifier of the file. This way, we had easy access to specific parts of the files' metadata even when using plain text files. We also filtered the text material. On the one hand, we divided the collection into separate corpora

according to their language and excluded files that did not contain manually assigned keywords (e.g. tables of contents). On the other hand, we removed titles and subtitles, since they were repetitive and therefore had a disproportionate impact on the result.

These plain text files were already fulfilling formal requirements to proceed with topic modeling, meaning they were classified per language, in the desired format, and containing metadata in the file names. But LDA treats inflected forms (e.g. Span. *mujer* – ‘woman’, and *mujeres* – ‘women’, or *muger*/*mugeres* in 18th century orthography) as different concepts. A topic can therefore include multiple forms of the same concept, which can result in semantically poor topics. To avoid this, we decided to lemmatize the Spectator texts before modeling the topics, using natural language processing with *spaCy* to replace each inflected word (e.g. *mugeres*) with its lexical base form (*mujer*), i.e. lemma. This step was one of the most challenging, since *spaCy* was not trained on historical language. Wrongly lemmatized tokens had to be replaced with the correct lemma through a dictionary. Although still not without errors, the decision to lemmatize brought much cleaner and semantically richer results than the preliminary experiments with non-lemmatized texts.

Since topic modeling measures the co-occurrence frequency of tokens in the same document, another pre-processing step was to define what will be treated as a document. We decided to segment the issues in paragraphs with a minimum of 500 tokens, whereby longer paragraphs were avoided by cutting off a paragraph after the first following sentence’s end, if the number of included tokens had surpassed 600. Remaining paragraphs with less than 200 tokens were appended to the preceding paragraphs of the same issue, to avoid very short paragraphs. Although there were still a few outlier paragraphs left, this method resulted in a larger quantity of documents with a similar token number instead of a smaller quantity of documents with more strongly varying token numbers. Since there is no state-of-the-art consensus on the optimal number of tokens in a document, the selection was based on preliminary experiments with different values.

With this set of resized and lemmatized documents, we continued with the workflow as provided by DARIAH-DE, with some practical adjustments. From the imported and tokenized documents, we removed redundant tokens as a last pre-processing step, since some tokens do not have semantic significance or are simply irrelevant and therefore not desired to be part of the final result. These tokens are a) the 100 most frequent words (MFW), because they tend to be functional words, like pronouns, articles, prepositions etc., b) the hapax legomena (tokens occurring only once in the corpus), and c) a project specific stop word list. To create the stop word lists, we adjusted the *Stopwords ISO* (Diaz 2016) lists and expanded them after each of our topic modeling cycles with new resulting topic keywords we identified as irrelevant.⁶

6. Besides functional words, such as the Spanish *inmediatamente* (immediately) or *entonces* (therefore), or some frequently used adjectives and modal verbs like the French *grand* (great) and *devoit* (should), we also excluded some nouns from the analysis, e.g. the Spanish *número*, as it is often used as an issue title.

4.3. Topic model creation

172

Since topic modeling is an unsupervised machine learning method, the researcher cannot impact the result by assigning categories in advance. There are, however, a couple of factors that do influence the results. One of them is, as previously mentioned, the pre-processing of the data. Another one is the choice of the input parameters: the number of topics, the number of iterations and the hyperparameter optimization interval. [Table 1](#) gives an overview of relevant parameters in our topic modeling analysis.

	French	Spanish
Number of periodicals	25	18
Number of issues	1,658	690
Extracted segments	6,752	3,190
Lemmatization	yes	yes
Removed features	100 MFW, 801 stop words, hapax legomena	100 MFW, 823 stop words, hapax legomena
Number of topics	25	18
Iterations	2,000	2,000
Alfa hyperparameter	5.0 (MALLET default)	5.0 (MALLET default)
Beta hyperparameter	0.01 (MALLET default)	0.01 (MALLET default)
Hyperparameter optimization	20	20

Table 1: Parameters used in the topic modeling analysis of French- and Spanish-language periodicals.

The number of topics is thus decided by the researcher. The reasonable number of topics in a text collection depends on the text scope, but also the genre and the thematic richness. Our approach was to experiment with different numbers of topics and evaluate the results to decide the optimal number of topics for each text corpus. Eventually, we determined 25 French and 18 Spanish topics. The number of topic keywords, on the other hand, is not a matter of the researcher's decision: each topic consists of all tokens from the text collection, whereas the distribution of these tokens varies in the individual topics. So, each token from the treated text collection can be found in each resulting topic, but with a varying probability which is never equal to 0% (cf. S. Bock et al. 2016, 13). The researcher familiar with the analyzed content then decides on how many of the topic tokens i.e. keywords they see as significant to represent in the results. For each analyzed group, we chose to output the first 20 tokens.

The researcher also sets the number of iterations. More iterations lead to a longer processing time but can lead to more reliable and stable results until a limit is reached after which the quality stagnates (cf. Jockers 2014, 147). Choosing an optimization interval is optional and depends on the desire to observe the difference in topic weight, by "allowing some topics to be more prominent than others" (McCallum 2002–2018).⁷ In our analysis, we conducted 2,000 iterations with an optimization in every 20 iterations.

But even with the same data and the same parameters, the output of two modeling cycles is never exactly the same in terms of the topics per document and the words per topic distribution, due to the probabilistic and unsupervised nature of the method. Nevertheless, using our data and parameters, the comparison of multiple results showed a re-emergence of the same topics, with rather insubstantial differences in the sequence

7. Schöch gave a more detailed reflection on hyperparameter optimization in his scientific blog (cf. Schöch 2016).

of the most frequent topic keywords, as well as the probability of the topics, which suggests a sufficient stability of the model.

4.4. Post-processing

The last step in the topic modeling workflow is the post-processing of the results. The probability of topics is being computed for each individual document (which, as explained in subsection 4.2, is a segment of a periodical's issue). Using these values, we computed the probability of topics per periodical and represented the results from different perspectives, utilizing multiple visualization techniques.⁸

A common way to represent topics are heat maps. The heat map (Figure 2) is a visual representation of the data frame matrix (Table 2) resulting from the topic modeling and the computed results per periodical, consisting of periodicals (X axis), topics (Y axis) and the probabilities of each topic per periodical as values, where darker color represents higher probability.

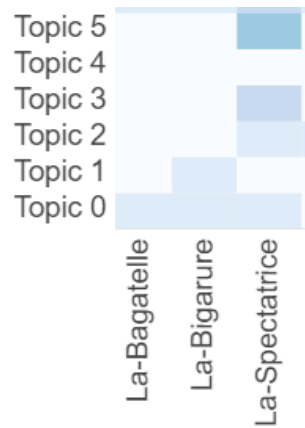


Figure 2: Heat map detail of 3 periodicals.

	La Bagatelle	La Bigarure	La Spectatrice
...	...	...	...
Topic 5	0.0196085	0.01663	0.126381
Topic 4	0.011347	0.0313603	0.03801
Topic 3	0.004227	0.00814	0.08031
Topic 2	0.03328	0.0153783	0.05184
Topic 1	0.02778	0.0434557	0.01843
Topic 0	0.064768	0.0685915	0.04833

Table 2: Data frame matrix detail corresponding to the heat map detail in Figure 2.

For each periodical, as well as for each topic, we created a bar chart (e.g. Figure 3). This technique offers a focus on one periodical or topic, whereas the heat map is better

8. The heat map, bar charts and word clouds of the French periodicals can be viewed under the URL <https://gams.uni-graz.at/o:dispeccs.result.tm.fr>. Spanish results are accessible under the URL <https://gams.uni-graz.at/o:dispeccs.result.tm.es>.

suited for getting an overview of the whole corpus. Both the heat map and the bar chart creation are part of the original DARIAH-DE Jupyter Notebook.

Additionally, we decided to use word clouds to visualize the top 100 keywords of the respective topics. A larger font size indicates a keyword with a higher probability inside a topic (Figure 4). We chose this visualization method despite some critics claiming that it is difficult for users to infer the relationship of the words from it (cf. Dobson 2021, §20). While we do agree with this point, a visual overview of all topic keywords is beneficial next to a visualization of the topic distribution, especially when the topics are not labeled.

Another way we used Python libraries to represent topics in selected periodicals is with line diagrams, which show the prevalence of topics over the issues of a single periodical, i.e. over time. This is only a relative prevalence over time, since the time span between the issues was not always constant and is not explicitly available in the metadata. The interactive diagram (created using the library *bokeh*) can be viewed on our project website.⁹ Here, it is possible to zoom in, create sections, activate or deactivate the visibility of individual topics, and save the created versions of the diagram.

Finally, we used the software *Gephi* to create networks of topics, periodicals and manually assigned keywords (Gephi.org 2008–2021). More precisely, we created a force-directed graph using the algorithms Fruchterman Reingold and Force Atlas 2 (Figure 7).¹⁰ The periodical nodes are represented as pie charts showing the distribution of a certain manually assigned keyword throughout the periodical's issues. The web presentations include color legends and numerical data for the pie charts. The size of a periodical node (pie chart) indicates whether the number of analyzed periodical issues from the topic modeling set is larger or smaller in comparison to other periodicals. Note that numerous issues do not mean the same as a large amount of text, since some issues can be very long while others are quite short. The size of the topic nodes indicates whether a topic has a high or low representativity in the analyzed set of texts. The edges are higher weighted (thicker) if the likelihood of a topic in a periodical is higher. Nodes with the same color belong to the same community. This means that the densities of the edges between these nodes are higher than from these nodes towards the rest of the network. But, since this is a small network where all topics occur in all periodicals to some degree, the weighted modularity of this network is low, and the community structure is not perfectly clear. Nevertheless, it is possible to detect topics that often co-occurred in periodicals.

As shown by the visualizations, the topics are non-semantically labeled (Topic 0, Topic 1, Topic 2, ...), and the numbers give no statement about the importance or frequency of the topic but are only used to distinguish the topics. This approach is contrary to the

9. Line diagrams: <https://gams.uni-graz.at/archive/objects/o:dispecs.result.tm.fr/methods/sdef:TEI/get?mode=diachronic> and <https://gams.uni-graz.at/archive/objects/o:dispecs.result.tm.es/methods/sdef:TEI/get?mode=diachronic>.

10. The full visualizations can be viewed on our web page: <https://gams.uni-graz.at/archive/objects/o:dispecs.result.tm.fr/methods/sdef:TEI/get?mode=topic-network> and <https://gams.uni-graz.at/archive/objects/o:dispecs.result.tm.es/methods/sdef:TEI/get?mode=topic-network>.

occasionally seen practice, where researchers either label their topics by interpreting them (e.g. Boyd-Graber, Hu, and Mimno 2017, 40, or Blevins 2010) or by using a few of the most relevant keywords, as proposed by the DARIAH-DE Jupyter Notebook. In recent years, we noticed an increase in the non-semantic labeling approach (e.g. Horstmann and Kleymann 2019, Krautter et al. 2020, or Chehal, Gupta, and Gulati 2021). We also decided to proceed without labels because the interpretation of a topic depends on the reception horizon of the researcher. This further impedes the obtrusion of a certain perspective and leaves room for different interpretations. We did, however, provide our interpretation in textual form. The gender-specific topics will be elaborated in section 5 and section 6.

To ensure transparency and comprehensibility of the visualizations and interpretations, all the underlying raw data can be downloaded by the user, including the topic keywords list and the word weights. Nevertheless, it has to be pointed out that for understanding the results of distant reading, a certain familiarity with the source material through close reading expertise is always required to create meaning from the results and generate added value for related research. As Shadrova also suggests, “[i]t is of crucial importance to make the underlying contextualization, the model, explicit, both through hypothesis-based work and by tying results back to the theoretical and conceptual debates in the field” (Shadrova 2021, 16).

5. Topic modeling in the French-language Spectator periodicals

Among the 25 topics of the Spectator periodicals published in French language, at least six topics stand out from a gender-specific perspective. Topics 4, 22, and 24 directly, topics 9, 18, and, 21 indirectly relate to character, behavior and roles of women and men within the (emerging bourgeois) society in the 18th century (see Table 3).

Topic 4 lists the various French terms for ‘marriage’ and ‘getting married’ (*mariage*, *marier*, *épouser*), ‘family’ (*famille*), ‘child’ (*enfant*), or ‘house’ (*maison*), which are terms that construct the destiny of young (!) women (*filles*, *demoiselle*) within the domestic sphere (in contrast to the public sphere, which is attributed to men).¹¹ In this private sphere, her main duty is to take tender (*tendresse*) care (*soin*) of her husband (*mari*) and children.

Topic 22 refers to the vocabulary used in the translation of the *Female Spectator* (1749–51), *La Spectatrice*, *traduite de l’anglais* (1750–51), as indicated in the bar chart with a probability of over 0.3 within this periodical (Figure 3), which is much higher in relation to other periodicals. *La Spectatrice* is one of the few spectatorial titles specifically directed to (bourgeois) women. This focus on the female readership also reverberates in the

11. Regarding gender discourse in 18th century France, see the articles of G. Bock and Zimmermann 1997, Brink 2008, Honegger 2011, or Sieuzac 2009. As to the presence of women in society and literature, see the essay collection edited by Jacobs et al. 1979. Concerning the theoretical and literary discourse on the woman as the ‘moral gender’ in the 18th century, see Steinbrügge’s monograph (Steinbrügge 1987). As to the representation of women in the French Enlightenment press, see Dijk 1988 and to the history of the ‘presse féminine’ in France, see Sullerot 1966.

Topic 4	Topic 9	Topic 18	Topic 21	Topic 22	Topic 24
filles	vertu	heureux	aimer	dame	air
jeune	mérite	dieu	sentir	quoiqu	bel
pere	vie	doux	bonheur	égard	sexe
mariage	nature	oeil	passion	manière	dame
fil	ame	tendre	lettre	passion	beauté
mere	hommes	tendre	amant	tem	jeune
mari	propre	main	tendre	sexe	visage
famille	bonheur	ciel	moment	mauvais	oeil
père	vice	ame	malheureux	propre	plaire
âge	passion	voix	douleur	convenir	aimable
enfan	heureux	feu	sentiment	peine	mode
marier	conduite	aimable	perdre	conduite	habit
chevalier	action	charme	malheur	obliger	joli
épouser	sage	gloire	heureux	penser	figure
tendresse	honneur	peine	tendresse	montrer	femmes
demoiselle	digne	objet	devenir	dessein	goût
devenir	noble	bel	ame	affection	conversation
maison	mal	terrebeauté	oeil	devenir	grace
soin	fortune	sage	objet	liberté	rire
amant	estime	brillant	étois	avis	compagnie

Table 3: Gender-specific topics in French-language Spectators.

first term of the topic with ‘lady’ (*dame*). The subsequent terms used, such as ‘passion’ (*passion*), ‘bad’ (*mauvais*), ‘suitable’ (*propre*), ‘corresponding’ (*convenir*), ‘conduct’ (*conduite*), or ‘affection’ (*affection*), indicate that this topic is concerned with the behavior of women in public, especially in the company of men.

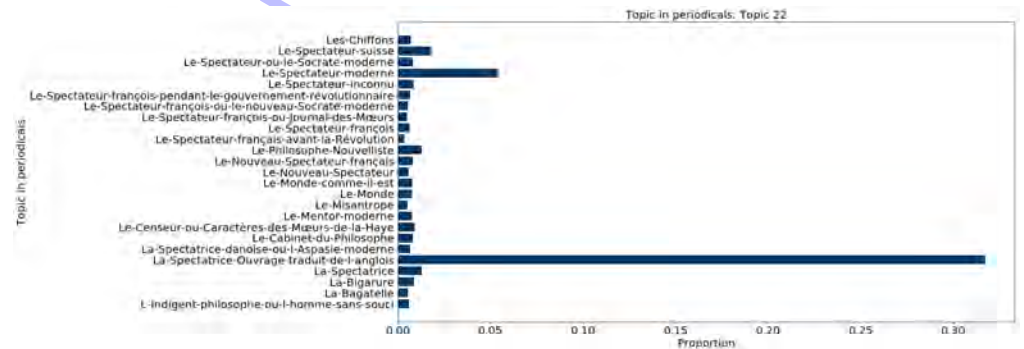


Figure 3: Distribution of topic 22 in French-language periodicals.

Topic 24, visualized as word cloud (Figure 4), also lists attributes associated with the ‘fair sex’ (*beau sexe*).¹² On the one hand, a woman has to ‘please’ (*plaire*) through her inner beauty – expressed by terms such as ‘beautiful’ (*bel*), ‘beauty’ (*beauté*), ‘amiable’ (*aimable*), and ‘grace’ (*grâce*) – and on the other hand through her outer beauty –

12. The French term *beau sexe* for the female part of the population is a compound and has been separated during the topic modeling process. Further, the term *beau* has been lemmatized into *bel*. This is the reason why the terms *bel* and *sexe* appear separately in topic 24. Nonetheless, their immediate position next to each other indicates their connection.

expressed as well by the terms ‘beautiful’ (*bel*) and ‘beauty’ (*beauté*), but also by ‘pretty’ 297
(*joli*) or ‘taste’ (*goût*). Both inner and outer beauty are accentuated by appropriate 298
‘clothing’ (*habit, mode*), good ‘taste’ (*goût*), and ‘conversation practices’ (*conversation*) 299
that are understood as suitable for a woman. Her ‘appearance’ (*air*), i.e. her outward 300
appearance, has the highest priority here, as can be seen from the prominent position 301
of the term in the first place, and is represented in all periodicals (see also Topic 17 302
of the Spanish periodicals, where the orientation on outward appearances manifests 303
through terms such as *moda* – ‘fashion’, *adornar* – ‘to adorn’, *gustar* – ‘to please’, *hermoso* – 304
‘beautiful’, *hermosura* – ‘beauty’). 305



Figure 4: 100 MFW in topic 24 in French-language periodicals.

The discourse on women within the French-language periodicals is further supported 306
by topics 9, 18, and 21. While the first three topics mentioned above explicitly evoke 307
terms for women (e.g. *beau sexe, femme*), and also use self-explaining terms alluding to 308
their status (e.g. *dame* – ‘lady’, *demoiselle* – ‘unmarried young woman’, *mère* – ‘mother’) 309
as well as gender-specific, heteronormative practices (e.g. *marier, épouser* – the act of 310
getting married), the terms used in topics 9, 18, and 21 are more implicit to the extent 311
that they only indirectly allude to the gender-specific discourse and roles of women and 312
men in the (bourgeois) society within the Spectator periodicals. 313

The terms occurring in **topic 9** describe virtuous behavior and practices. The gender-specific key concept of virtue (cf. Pabst 2007) stands at the very beginning of the word sequence. The following terms refer to the fact that virtue leads to (individual and collective) ‘happiness’ (*bonheur*). In general, 18th century philosophers equate ‘virtue’ with ‘happiness’, for only those who lead a virtuous life can contribute to their own

happiness and to the happiness of the community. Virtue is thus seen as a means to achieve the individual and collective goal of happiness (cf. Völkl 2022, 121–122).¹³

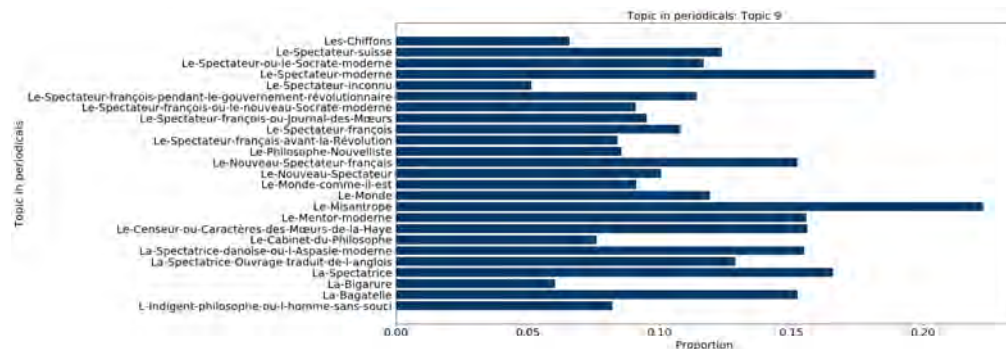


Figure 5: Distribution of topic 9 in French-language periodicals.

Topic 9 can be found in all Spectator periodicals at a median rate of 0.62 (Figure 5), which makes it the most probable in the corpus. This is not surprising because most periodicals explicitly state their goal in their first lines, which is to turn all people into useful members of the society – a society which is becoming increasingly complex and integrated into a nation (cf. Ertler 2010, 100). In terms of women as useful members of society, the role of the (bourgeois) woman is conceptualized in three ways: as spouse, housewife and mother. Outside the domestic sphere, she has no right to exist, which is the reason why, for example, the image of the learned woman was defamed in the Spectator periodicals at the beginning of the 18th century and has subsequently been omitted altogether – according to the motto ‘out of sight, out of mind’ (cf. Völkl 2022, 309–310).

Topic 18 results in terms referring to the virtuous ideal image of both women and men. The terms ‘tender’ (*doux, tendre*), ‘amiable’ (*aimable*), ‘grace’ (*charme*), ‘prudent’ (*sage*), ‘witty’ (*brillant*) here refer to inner virtues, while the terms ‘beautiful’ (*bel*) and ‘beauty’ (*beauté*) can refer to inner and outer virtues at the same time, as explained above. Although this is not a frequent topic, it is consistently present in all Spectator periodicals.

Topic 21 exhibits terms that can be assigned to the discourse field of love. They are associated positively or negatively with love. For example, next to the appositives ‘to love’ (*aimer*), ‘happiness’ (*bonheur*), or ‘tender’ (*tendre*), one can find the pejoratives such as ‘unhappy’ (*malheureux*), ‘pain’ (*douleur*), or ‘to lose’ (*perdre*). The frequency of individual terms will be discussed below.

A look at the distribution of topic 21 within the French-language periodicals (Figure 6) reveals that the three successive periodicals *Le Nouveau Spectateur* (1758–60), *Le Monde comme il est* (1760), and *Le Monde* (1760–61) of Jean-François de Bastide (1724–1788) are particularly endowed with this topic. The literary and cultural studies research carried out by Fischer-Pernkopf et al. and Völkl support the finding that Bastide continuously

13. On the discourse of happiness in the 18th century, cf. Mauzi 1969, on the concept of ‘happiness’ in *The Spectator*, cf. Norton 2015.

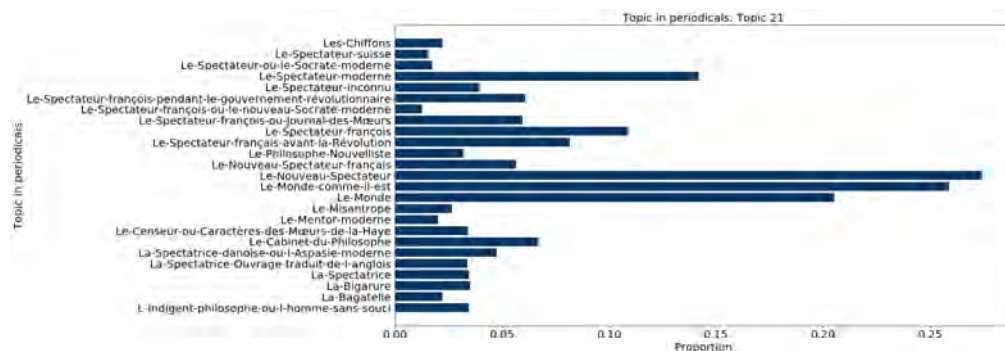


Figure 6: Distribution of topic 21 in French-language periodicals.

narrates exemplary stories of happy and unhappy (heterosexual) lovers (cf. Fischer- 348
Pernkopf, Mussner, and Ertler 2018, Völkl 2022). 349

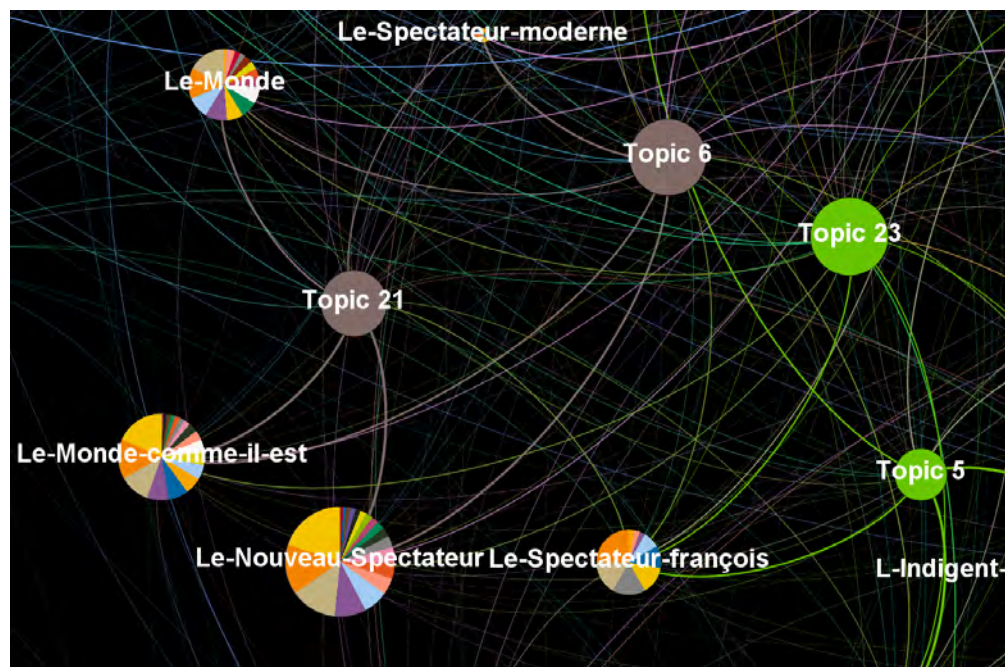


Figure 7: Detail from the network graph of periodicals and topics, showing the prevalence of topic 21 in the three periodicals of Jean-François de Bastide.

The proximity of the nodes and the edges weight in the network analysis graph in 350
Figure 7 also illustrates the prevalence of topic 21 in all of Bastide's periodicals. It further 351
shows that topics 6¹⁴ and 23¹⁵ are also very common in Bastide's periodicals. They 352
include typical narrative vocabulary (e.g. *demander* – 'to ask', *répondre* – 'to answer', *entrer* 353
– 'to enter', *entendre* – 'to listen', *lire* – 'to read') and typical narrative elements (e.g. *ami* – 354
'friend', *maison* – 'house', *chambre* – 'room', *porte* – 'door'). Based on the accumulation 355
of narrative terms, it can be concluded that in Bastide's periodicals, the discourse of 356

14. Topic 6: penser vérité mal vrai honneur ami réflexion juger mauvais défaut répondre caractere sentir droit
convenir lorsqu connoître entendre lire quelquefois

15. Topic 23: maison demander heure chambre paraître jeune tem main peine passer entrer revenir air ami
vouloit porte arriver alloit entendre sortir

love is primarily conveyed through stories and storytelling. This interpretation of the topic modeling results is supported by previous literary and cultural research in this field, which also stress the strong narrative design of Bastide's periodicals (cf. Fischer-Pernkopf, Mussner, and Ertler 2018, Mussner 2016, Völkl 2022).

Additionally, the analysis of the issues manually annotated with the subjects/keywords 'Image of women' and 'Image of men'¹⁶ of the *Nouveau Spectateur*¹⁷ and the *Monde comme il est*,¹⁸ identified that the following five narrative forms (*Erzählformen*) are predominantly used to discuss family life (in particular education) and couple relationships (with a focus on the romantic tender love relationship): general account (*allgemeine Erzählung, AE*), heteroportrait (*Fremdporträt, FP*), metatextuality (*metatextueller Kommentar, MT*), dialogue (*Dialog, D*) and letter/letter to the editor (*Leser*innenbriefe, LB*) (cf. Völkl 2022, 209). Concerning the distribution and arrangement of these text types, it has to be emphasized that they also repeatedly appear intertwined within each other, which leads to the – for the Spectator periodicals – typical multi-layered system of communication (cf. Fischer 2014, 74–83).

Figure 8, which shows a statistical examination of all issues of Bastide's periodicals, further supports the above-mentioned results. It shows that Bastide uses the following narrative forms as predominant communication strategy: metatextual commentaries (MT), letters/letters to the editor (LB), dialogues (D), and general accounts (AE). While the heteroportrait (FD) only stands on fifth position after citation/motto (ZM).

Furthermore, the three bar charts of the topic distribution in Bastide's periodicals (Figure 9, Figure 10, and Figure 11) indicate a wide distribution of topic 9 (virtuous behavior and action) and topic 6 (describing the postulate of enlightened philosophers: '(Self)reflection' (*réflexion*) leading to 'truth' (*vérité*) and knowledge). This focus on virtue and vice is not surprising at all, considering that the Spectator periodicals aim at the moral education of their female and male audience. The readers of the periodicals in general and of Bastide's periodicals in particular are repeatedly exposed to vicious behavior and actions by means of shorter and longer stories in order to guide them to virtuous behavior and actions. A detailed definition or specification of the social norm designated by the term 'virtue', however, is lacking and thus remains undefined; rather, 'being virtuous' is illustrated indirectly through the depiction of its opposite: 'being vicious'. Via the detour of numerous love and relationship stories as well as character portraits, which clearly highlight vicious behavior and vicious character traits, the readers are thus led to the desired social norm (cf. Völkl 2022, 291–292).

16. In total, the list of subjects comprises 37 keywords, which were determined at the beginning of the digital scholarly edition project (cf. Ertler et al. 2011–2021) and which was slightly expanded in the course of the project.

17. From the 108 issues within the *Nouveau Spectateur* 58 issues (44%) are indexed by the subject 'Image of women' and 16 issues (14,8%) by the subject 'Image of men' (cf. Völkl 2022, 206).

18. Within the 60 issues of Bastide's *Monde comme il est*, 38 issues (64%) are indexed by the subject 'Image of women' and 12 issues (20%) by the subject 'Image of men' (cf. Völkl 2022, 206).

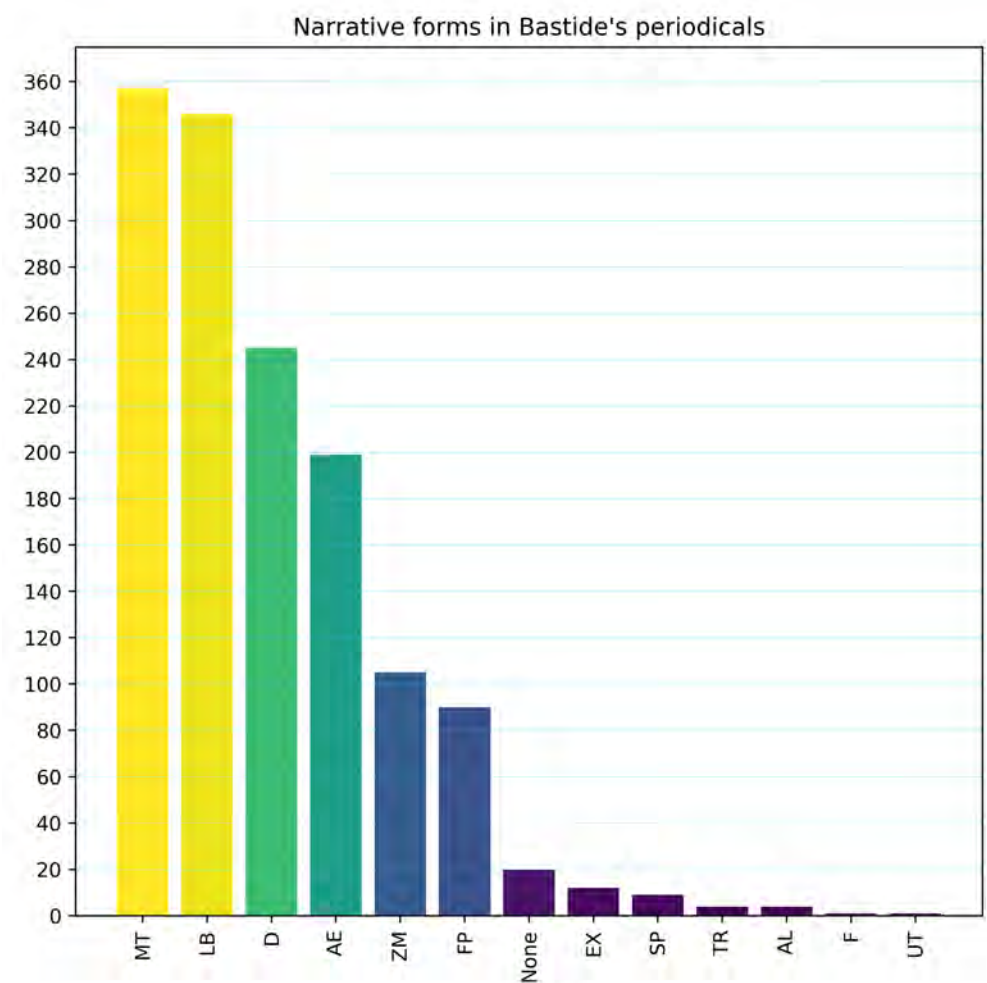


Figure 8: Narrative forms in Bastide's periodicals.

6. Topic modeling in the Spanish-language Spectator periodicals 391

Regarding gender-specific topics, the topic modeling results for the Spanish-language 392 periodicals were similar to those of the French-language Spectators. Within the 18 393 Spanish topics, topics 8, 9, 11 and 17 can be identified as referring to women and men 394 (see Table 4). 395

Topic 8 is headed by the gender-specific key concept of 'virtue' (*virtud*) followed by 396 terms describing elements of a virtuous lifestyle (*vida* – 'life', *amor/amar* – '(to) love', 397 *honor/honrar* – '(to) honor'), thereby showing considerable similarities to topic 9 of 398 the French-language periodicals.¹⁹ This topic similarity is not surprising at all, since 399 the contemporary gender discourse within the French-language periodicals enters the 400 Spanish periodicals – that first appear in Spain from mid-century onward – through 401 numerous translations, imitations, and cultural adaptations. More than in other Euro- 402

19. The Spanish topic 8 and the French topic 9 show the following equivalent terms: *virtud* – *vertu*, *vida* – *vie*, *honor/honrar* – *honneur*, *alma* – *ame*, *viciar* – *vice*, *noble* – *noble*, *felicidad* – *bonheur*.

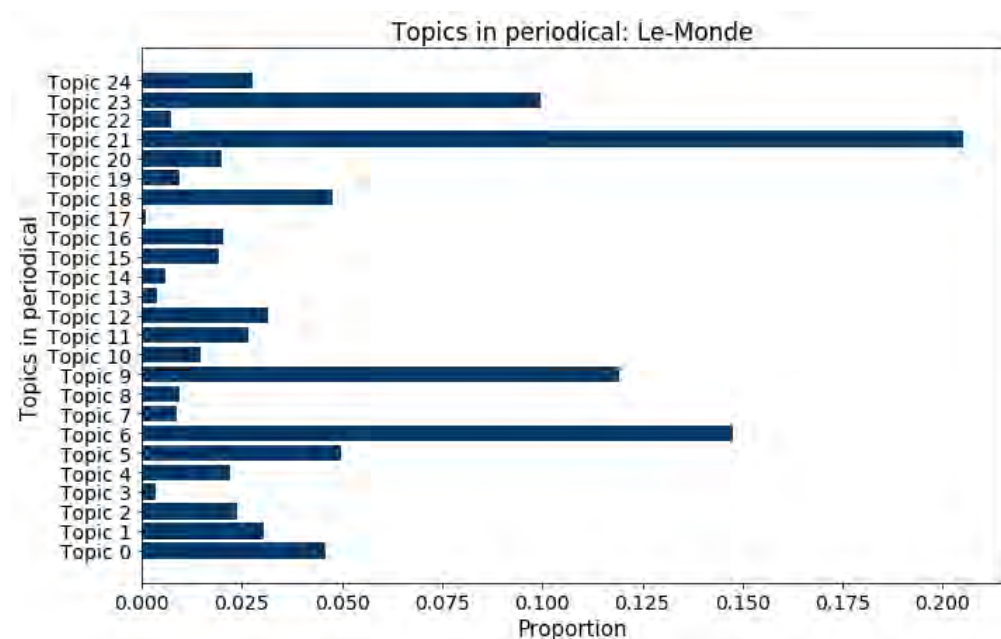


Figure 9: Topics in *Le Monde*.

pean countries, however, women in Spain are excluded from public life and confined 403
to the private sphere which centers on home, family, and motherhood (cf. Völkl 2022, 404
229).²⁰ An abundant presence of topic 8 in all Spanish periodicals is thus an expectable 405
development (Figure 12). 406

Topic 9, with terms such as ‘writing’ (*escribir*), ‘studying’ (*estudiar*), and ‘reading’ (*leer*), 407
alludes to educational activities; the terms ‘science’ (*ciencia*), ‘art’ (*arte*), or ‘history’ 408
(*historia*) of ‘ancient’ (*antiguo*) time to specific study objects. The convergence of these 409
terms suggests that this topic describes the education of a bourgeois man, even though 410
no term referring to a male subject (such as *hombre*) – nor to a female subject (such as 411
mujer) – can be found. In fact, although the Spanish periodicals grant the female gender 412
a certain capacity for education as well, the terms of topic 9 refer to male formation 413
only. Education for young women is conceived differently to education for young men 414
because (as the French-language periodicals) the Spanish Spectators also propagate 415
a complementary gender model, implying that women and men need to be educated 416
specifically for the correct fulfillment of their gender-specific role in society. The Spanish 417
ideal of the virtuous (bourgeois) woman is also praised in her threefold role as spouse, 418
housewife, and mother, through the fulfillment of which she contributes to the common 419
good of society. This image of woman is conceived in a ‘natural complementarity’ to 420
man, whose ideal image is embodied by the ‘hombre de bien’. The latter is characterized 421
by the training of his intellect and subsequently proving useful for his fatherland and 422
the common good. The ‘hombre de bien’ of the 18th century is thus not to be confused 423
with the preceding aristocratic ‘hombre de bien’ of the 17th century, whose idleness 424

20. Regarding gender discourse in 18th century Spain, see e.g. the monographs and articles by Martín Gaité 1972, Hassauer 1997, Bolufer Peruga 1998, Brink 2008, Capel Martínez 2010, or Gronemann 2013; on the gender discourses in the Spanish novels of the ‘siglo de las luces’, see Hertel-Mesenhöller 2001 or Kilian 2002.

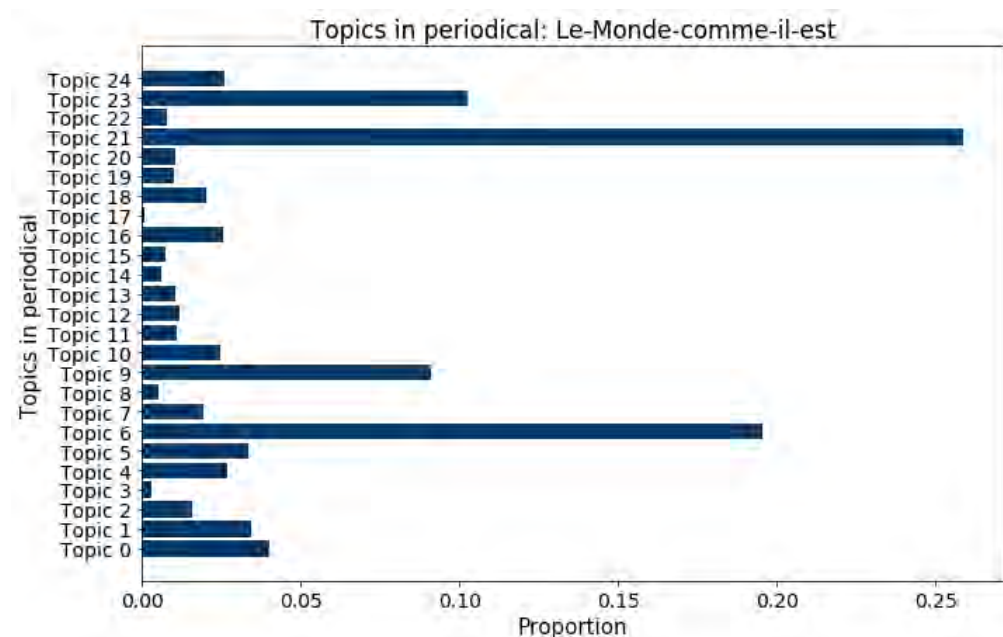


Figure 10: Topics in *Le Monde comme il est*.

causes his reputation to fall below that of an active citizen – regardless of his status (cf. Heße 2008, 113–130).

Very similar to topic 4 in the French Spectators, the content of **topic 11** supports the construction of the heteronormative society, suggesting the ideal role of ‘woman’ (*mujer*) and man in ‘marriage’ (*maridar*, *matrimoniar*), where they become ‘mother’ (*madre*) and ‘father’ (*padre*) of ‘children’ (*hijo*, *niño*). The role of the woman is thus conceived by her ‘husband’s’ (*marido*, *esposo*) side, to whom she is supposed to be a good spouse and housewife. Within the domestic sphere (*familia*), she also receives the role of the ‘caring’ (*cuidar*, *cariño*) mother, who ‘loves’ (*amor*, *amar*) and ‘raises’ (*criar*) her ‘children’ (*hijo*, *niño*). Although a rather infrequent topic (Figure 13), it exists throughout all Spanish Spectators.

Topic 17, represented in Figure 14, points to two discourses associated with the female gender: on the one hand the subject of beauty, on the other hand the then vicious trend of having a relationship with a younger man (*cortejo*). The first eight terms of this topic (*mujer* – ‘woman’, *moda* – ‘fashion’, *dama* – ‘lady’, *gustar* – ‘to please’, *hermoso* – ‘beautiful’, *adornar* – ‘to adorn’, *hermosura* – ‘beauty’) refer to the semantic field of beauty which pervades the spectatorial gender discourses throughout the century and clearly reflects topic 24 of the French-language periodicals (see (Figure 4)). In fact, the Spectator periodicals by and large constantly instruct their readers to cultivate external and, increasingly, internal beauty, because female beauty is perceived as a pledge for marriage (cf. Schaufler 2002, 190) which is seen as the ‘natural’ destiny of the (bourgeois) woman and is thus considered her ultimate goal. At the same time, however, the periodicals warn against falling prey to a cult of beauty that goes hand in hand with the vices of vanity and jealousy. One of these vices, also represented in topic 17, is the

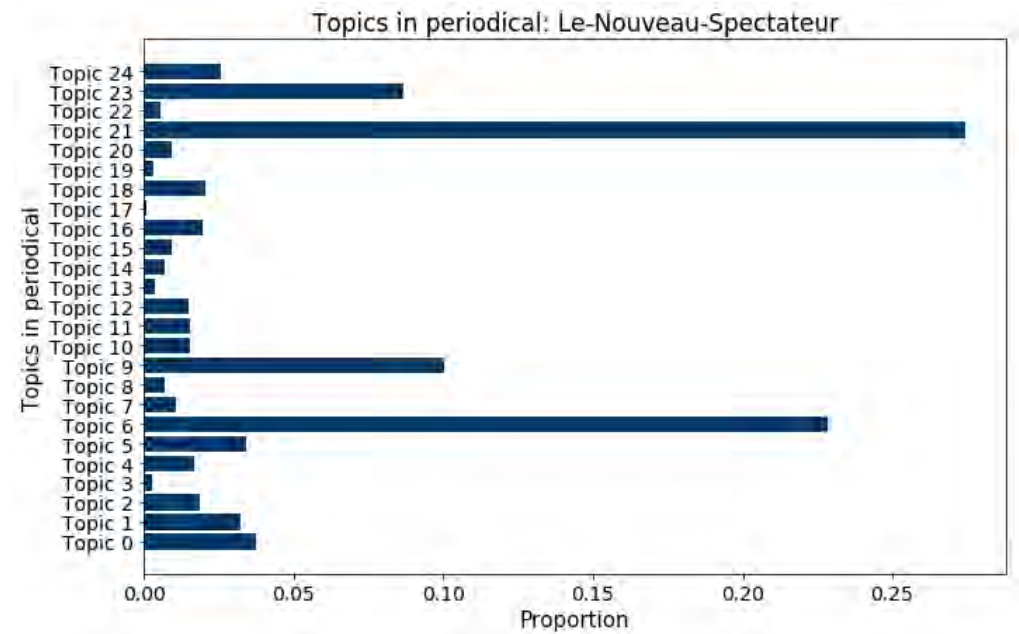


Figure 11: Topics in *Le Nouveau Spectateur*.

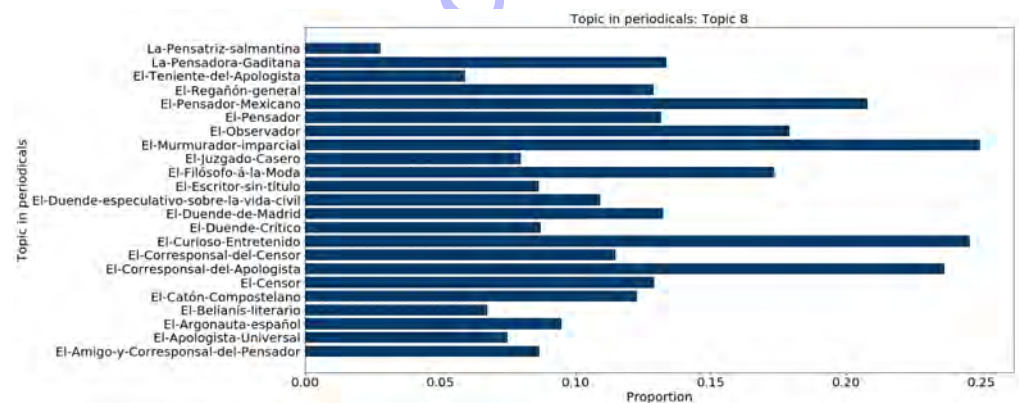


Figure 12: Distribution of topic 8 in Spanish-language periodicals.

gender-stereotypical model of the *cortejo*, i.e. a (younger) man maintaining a very close relationship with a married woman or widow who he ‘visits’ (*visitar*) regularly.²¹ While this (mostly platonic) form of relationship is not a moral offense in aristocratic tradition, it is criticized and stigmatized in the Spectator periodicals.

Regarding the dissemination of the gender-related topics, the Spanish periodicals pursue a similar strategy to their French-language precursors. Likewise, in the Spanish Spectators virtuous and vicious gender-specific values, norms and practices are mostly conveyed through stories and storytelling. Similar to the French topics 6 and 23, the topic modeling process for the Spanish Spectators revealed topics with a high concentration

21. The term ‘cortejo’, which only exists in the masculine form, is not only used to designate the man in this special relationship with a married woman, but also for the woman who allows herself to be courted, and furthermore even to paraphrase the liaison itself (cf. Heße 2008, 135–136).

Topic 8	Topic 9	Topic 11	Topic 17
virtud	españa	hijo	mujer
vida	siglo	mujer	moda
amor	lengua	padre	dama
corazon	ciencia	madre	gustar
honor	escribir	criar	hermoso
vivir	nacion	maridar	adornar
placer	mundo	niño	hermosura
alma	estudiar	familia	personar
amar	historia	amor	gracia
mirar	leer	amar	sexo
mundo	libro	edad	figurar
viciar	arte	tratar	señora
honrar	letra	año	bayle
noble	sabio	marido	cortejo
despreciar	idioma	cuidar	mirar
felicidad	españoles	matrimoniar	cabeza
desear	ciencias	señora	naturaleza
efecto	llamar	esposo	rostro
naturaleza	antiguo	hermano	arte
ojo	naciones	cariño	visitar

Table 4: Gender-specific topics in Spanish-language Spectators.

of narrative vocabulary, such as in topic 2.²² Therein, narrative vocabulary revolves 458
around the semantic fields of movement (e.g. *venir* – ‘to come’, *salir* – ‘to leave’, *llegar* 459
– ‘to arrive’), speech (e.g. *contar* – ‘to narrate’, *entender* – ‘to listen’, *palabra* – ‘word’), 460
and time (e.g. *año* – ‘year’, *hora* – ‘hour’, *noche* – ‘night’), all of which are important 461
components in a story. As can be discerned in Figure 15, topic 2 occurs in all Spanish 462
periodicals. 463

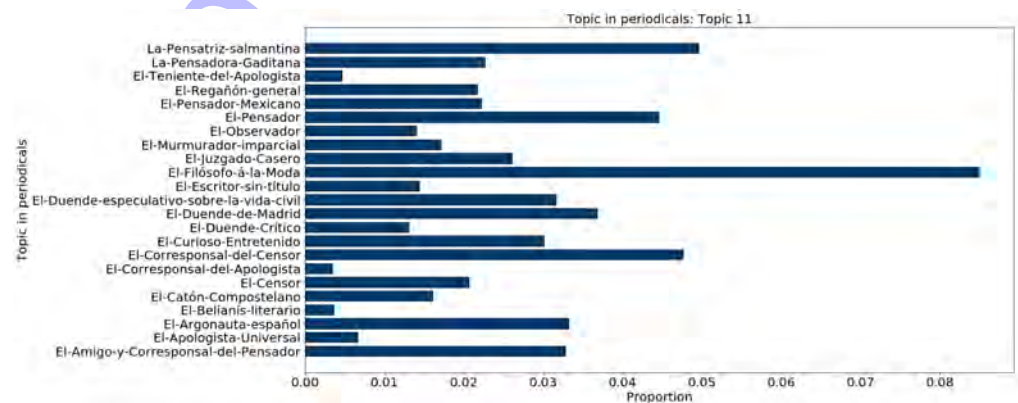


Figure 13: Distribution of topic 11 in Spanish-language periodicals.

22. Topic 2: venir salir pasar tomar mano año llegar llamar mil quedar mundo volver entender contar hora
amigar oír acabar noche palabra

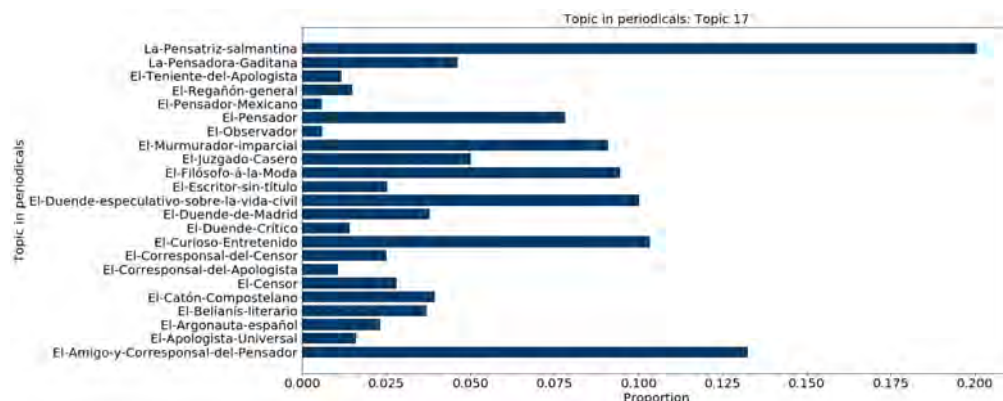


Figure 14: Distribution of topic 17 in Spanish-language periodicals.

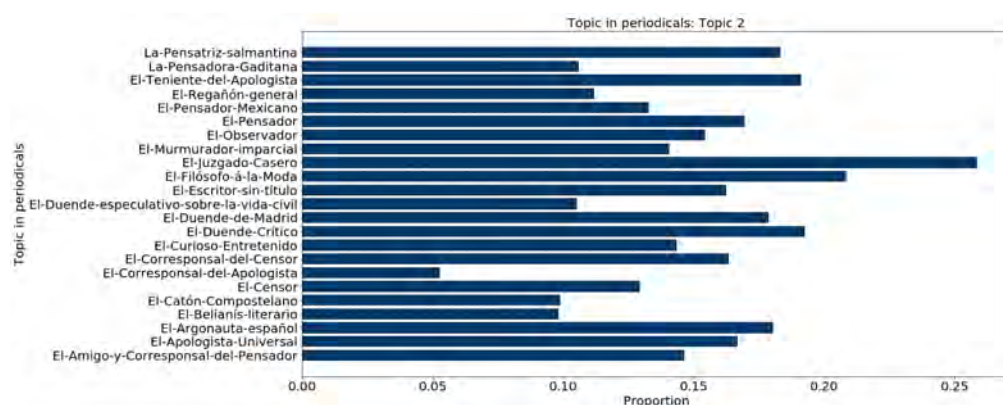


Figure 15: Distribution of topic 2 in Spanish-language periodicals.

7. Conclusion

464

With their gender-specific discourses, the Spectator press (co-)constructed, preserved, and propagated a bourgeois gender model which is still valid in socio-cultural perception today. This contribution investigates 1,658 French- and 690 Spanish-language issues which were analyzed with topic modeling using LDA. The findings with a focus on gender-specific discourse match and reinforce the results from Völkl's study on narrative and media-specific gender construction within the Spectator periodicals (Völkl 2022).

Using topic modeling, gender-specific topics were identified in the Spectator corpus. Additionally, the application of topic modeling also showed that the Spectator press employed a certain narrative vocabulary (French Spectators: Topic 6 and 23; Spanish Spectators: Topic 2). Moreover, the comparison between the French- and Spanish-language periodicals rendered similar results: The Spectator corpus of both languages manifested several topics pertaining to a gender-specific discourse. This discourse can be discerned explicitly in topics which exhibit terms referring to female or male stereotypical models, or implicitly in topics which exhibit terms referring to virtuous and vicious gender-specific values, norms, and practices. These concordances ascertain that topic modeling as the method used for the present analysis can be successfully

employed to question and confirm hypotheses gained through close reading. 481

In addition to our findings on the gender-specific topics in the Spectators, we described 482
the topic modeling workflow used in DiSpecs in [section 4](#). We aim to make our analysis 483
process transparent for other researchers interested in this method. The research com- 484
munity can also benefit from the primary data available in TEI, and the code, which all 485
are publicly available online (Ertler et al. 2011–2021, Scholger et al. 2019–2021, Scholger 486
et al. 2022). 487

The DARIAH-DE Notebooks that implement LDA topic modeling proved to be very 488
useful as a basis in our analysis workflow. With some adaptations and additional pre- 489
processing (especially segmentation and lemmatization) and post-processing steps 490
(e.g. results categorization and additional visualizations) we were able to produce 491
comprehensible and insightful results. Our own experience and the comparison with 492
other topic modeling projects allow us to conclude that pre-processing is a crucial part 493
of the analysis, since it strongly impacts the quality of the results. The decisions on the 494
respective steps depend on the research material and the specific project goals. 495

An advantage of topic modeling is the possibility to analyze more content than with 496
close reading, to illustrate the hypothesis on a broader level than through individual 497
examples, and to present the findings using different types of visualizations. Our topic 498
modeling analysis resulted in measurable data of a large text collection’s semantic 499
structure, which we were able to interpret and comprehensively demonstrate to the 500
Spectators research community. Furthermore, the analysis invoked some new insights 501
into the corpus. Concerning the gender-specific discourse in the Spectators, we saw 502
e.g. with topic 22 that the French translation of the *Female Spectator* is equipped with 503
a specific semantic vocabulary that can almost exclusively be found in this specific 504
periodical. This result can be attributed to the fact that in this case, we are dealing with 505
a translation and not with a genuine French periodical. 506

The primary data and the digital scholarly edition also benefit from the topic modeling 507
analysis. With the resulting topics, it is now possible to revise the keywords manu- 508
ally assigned to the individual issues and to further differentiate them. The list of 37 509
keywords was determined at the beginning of the digital edition project around 2011 510
and was only minimally expanded in the course of the project. Consequently, the list 511
seems somewhat arbitrary: culture- and language-specific topics – such as ‘Apologetic 512
of Spain’ which only apply to a few issues – are on the same level as very broad topics 513
such as ‘Theatre, Literature, Arts’ which combine three areas in one topic. Therefore, 514
the results from topic modeling help to expand and adjust the list of keywords for 515
thematic indexing, thus improving the analysis capabilities within the digital edition, 516
as demonstrated in *LdoD Visual* by Portela and Rito Silva 2017. The identified terms in 517
the topics can be incorporated into the TEI metadata header and subsequently used 518
for a more precise and sophisticated search not only at the document level but also on 519
specific text fragments. 520

Nevertheless, it is necessary to mention certain challenges in using topic modeling. 521

Critics like Dobson argue that the variability of the output depending on the algorithms and set parameters of the method is problematic (cf. Dobson 2021, §20), while Roe, Gladstone and Morrissey also refer to the probabilistic nature of LDA causing variability in individual runs even with the same parameters (cf. Roe, Gladstone, and Morrissey 2016, 4). While we did not compare our LDA results with other algorithms, we agree with Schöch that these variations manifest themselves “in the details of word ranks rather than in the general topics obtained” (Schöch 2017). Parameters have to be tested for individual projects, but once optimized, the method provides relatively stable results.

Furthermore, Murakami et al. as well as Shadrova are skeptical towards methods based on the bag-of-words approach because it ignores the grammatical structures and semantic relations between words (cf. Murakami et al. 2017, 246, Shadrova 2021, 13-14). While we do agree with this statement and believe that every scientific method should be questioned, we also argue that digital methods are not supposed to take on our tasks as humanities experts, but to facilitate research and help us to interpret our data. For these reasons, using a combined approach of topic modeling (and text mining methods in general) and close reading is essential, as well as the understanding of the material itself. As Fechner and Weiß point out, it is not the topics that answer research questions themselves, but the researchers through the interpretations of the topics (cf. Fechner and Weiß 2017, 20).

Besides the contribution to the current state of Spectators research and to practical applications of topic modeling, our work also lays the foundations for future work on 18th century literature. The presented results can be compared with similar research on other genres of that time. In addition to the probabilistic topic modeling approach, we intend to integrate transformer-based models to investigate a new corpus of Spanish epistolary novels, which are considered to have continued propagating gender-specific values, norms, and practices from the Spectators, while also representing an intermediate step towards the 19th century novel.

8. Data availability 549

Data can be found here: <https://gams.uni-graz.at/dispecs> and <https://gams.uni-graz.at/spectators> 550
551

9. Software availability 552

Software can be found here: <https://github.com/distantspectators> 553

10. Acknowledgements 554

We thank the Austrian Academy of Sciences who funded the project DiSpecs, as well as 555
our project team co-members Bernhard Geiger, Christina Glatz, Elisabeth Hobisch, and 556
Philipp Koncar for their cooperation, contribution, and support. 557

11. Author contributions 558

Yvonne Völkl: Results Interpretation, Conceptualization, Writing, Editing 559

Sanja Sarić: Topic Modeling, Conceptualization, Writing, Editing 560

Martina Scholger: Project Coordination, Conceptualization, Writing, Editing 561

References 562

- Alloghani, Mohamed, Dhiya Al-Jumeily, Jamila Mustafina, Abir Hussain, and Ahmed J. 563
Aljaaf (2020). "Chapter 1: A Systematic Review on Supervised and Unsupervised 564
Machine Learning Algorithms for Data Science". In: *Supervised and Unsupervised 565
Learning for Data Science*. Ed. by Michael W. Berry, Azlinah Mohamed, and Bee Wah 566
Yap. Unsupervised and Semi-Supervised Learning. Springer, pp. 3–22. 567
- Blei, David M. (2020). "Topic Modeling and Digital Humanities". In: *Journal of Digital 568
Humanities* 2.1, pp. 8–11. URL: <http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/>. 569
570
- Blevins, Cameron (2010). "Topic Modeling Martha Ballard's Diary". In: URL: <https://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/>. 571
572
- Bock, Gisela and Margarete Zimmermann, eds. (1997). *Jahrbuch für Frauenforschung. 573
Band 2. Die europäische Querelle des Femmes. Geschlechterdebatten seit dem 15. Jahrhundert.* 574
Stuttgart: Metzler. 575
- Bock, Sina, Keli Du, Michael Huber, Stefan Pernes, and Steffen Pielström (2016). "Der 576
Einsatz quantitativer Textanalyse in den Geisteswissenschaften. Bericht über den 577
Stand der Forschung". In: DARIAH-DE Working Papers 18. Ed. by Mirjam Blümm, 578
Thomas Kollatz, Stefan Schmunk, and Christof Schöch. 579
- Bolufer Peruga, Mónica (1998). *Mujeres e Ilustración. La construcción de la feminidad en la 580
Ilustración española*. València: Inst. Alfons el Magnànim. 581

- Boyd-Graber, Jordan, Yuening Hu, and David Mimno (2017). "Applications of Topic Models". In: *Foundations and Trends in Information Retrieval*. URL: https://mimno.infosci.cornell.edu/papers/2017_fntir_tm_applications.pdf.
- Brink, Margot (2008). "Geschlechterstreit und Dialektik der Aufklärung in Spanien und Frankreich. Die ambivalente Rolle von Vernunft und Natur in Egalitäts- und Komplementaritätstheorien des 18. Jahrhunderts". In: *Heißer Streit und kalte Ordnung. Epochen der 'Querelle des femmes' zwischen Mittelalter und Gegenwart*. Ed. by Friederike Hassauer. Göttingen: Wallstein Verlag, pp. 344–364.
- Capel Martínez, Rosa M^a (2010). "Prensa y Escritura Femenina en la España Ilustrada". In: *El Argonauta español* 7. URL: <http://journals.openedition.org/argonauta/431>.
- Chehal, Dimple, Perul Gupta, and Payal Gulati (2021). "Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations". In: *J Ambient Intell Human Comput* 12, pp. 5055–5070. DOI: <http://dx.doi.org/10.1007/s12652-020-01956-6>.
- DARIAH-DE (2019). *Notebook IntroducingMallet.ipynb*. URL: <https://github.com/DARIAH-DE/Topics>.
- Diaz, Gene (2016). *Stopwords ISO*. URL: <https://github.com/stopwords-iso>.
- Dijk, Suzanna van (1988). *Traces de femmes. Présence féminine dans le journalisme français du XVIIIe siècle*. Amsterdam: APA, Holland University Press.
- Dobson, James (2021). "Interpretable Outputs: Criteria for Machine Learning in the Humanities". In: *Digital Humanities Quarterly* 15.2. URL: <http://digitalhumanities.org:8081/dhq/vol/15/2/000555/000555.html>.
- Du, Keli (2019). "A Survey On LDA Topic Modeling In Digital Humanities". In: *Book of Abstracts DH2019*. DOI: <https://doi.org/10.34894/H9UYPI>.
- Ertler, Klaus-Dieter (2010). "Die Moralischen Wochenschriften als Vehikel zur diskursiven Ausdifferenzierung der Nation in Spanien". In: *Beiträge zur Nationalisierung der Kultur im Spanien des aufgeklärten Absolutismus*. Ed. by Jan-Henrik Witthaus. Frankfurt a. M.: Peter Lang, pp. 93–107.
- Ertler, Klaus-Dieter, Alexandra Fuchs, Michaela Fischer-Pernkopf, Elisabeth Hobisch, Martina Scholger, and Yvonne Völkl (2011–2021). *The Spectators in the International Context*. Graz. URL: <https://gams.uni-graz.at/spectators>.
- Fechner, Martin and Andreas Weiß (2017). "Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts". In: *Zeitschrift für digitale Geisteswissenschaften* 2. DOI: http://dx.doi.org/10.17175/2017_005.
- Fischer, Michaela (2014). *Die Figur des Lesers im Kommunikationssystem der Spectateurs*. Frankfurt a. M.: Peter Lang.
- Fischer-Pernkopf, Michaela, Veronika Mussner, and Klaus-Dieter Ertler (2018). *Die «Spectators» in Frankreich. «Le Nouveau Spectateur» und «Le Monde comme il est» von Jean-François de Bastide*. Frankfurt a. M.: Peter Lang.
- Gephi.org (2008–2021). *Gephi V.0.9.2*. URL: <https://gephi.org/>.
- Gronemann, Claudia (2013). *Polyphone Aufklärung. Zur Textualität und Performativität der spanischen Geschlechterdebatten im 18. Jahrhundert*. Frankfurt a. M.: Vervuert.

- Hassauer, Friederike (1997). "Die Seele ist nicht Mann, nicht Weib. Stationen der Querelles des Femmes in Spanien und Lateinamerika vom 16. zum 18. Jahrhundert". In: *Die europäische Querelle des Femmes. Geschlechterdebatten seit dem 15. Jahrhundert*. Ed. by Gisela Bock and Margarete Zimmermann. Vol. 2. Jahrbuch für Frauenforschung. Stuttgart: Metzler, pp. 203–238.
- Hertel-Mesenhöller, Heike (2001). *Das Bild der Frau im spanischen Roman des 18. Jahrhunderts. Im Spannungsfeld von Lebenswirklichkeit und Fiktion*. Frankfurt a. M.: Vervuert.
- Heße, Kristina (2008). *Männlichkeiten im Spanien der Aufklärung. Der Diskurs der moralischen Wochenschriften 'El Pensador', 'La Pensadora gaditana' und 'El Censor'*. Berlin: Logos.
- Hobisch, Elisabeth (2017). *La forma epistolar en los espectadores españoles: Características y tipología de las cartas*. Frankfurt a. M.: Peter Lang.
- (2018). "Les stratégies publicitaires dans les lettres des 'spectateurs' espagnols". In: *Discourses on Economy in the Spectators / Discours sur l'économie dans les spectateurs*. Ed. by Klaus-Dieter Ertler, Samuel Baudry, and Yvonne Völkl. Hamburg: Dr. Kovač, pp. 199–214.
- Honegger, Claudia (2011). "Die kognitiven Prinzipien der neuen Wissenschaften vom Menschen und die Genese einer weiblichen Sonderanthropologie in Frankreich". In: *Die gesellschaftliche Verortung des Geschlechts. Diskurse der Differenz in der deutschen und französischen Soziologie um 1900*. Ed. by Theresa Wobbe. Frankfurt a. M.: Campus, pp. 93–113.
- Horstmann, Jan and Rabea Kleymann (2019). "Alte Fragen, neue Methoden – Philologische und digitale Verfahren im Dialog. Ein Beitrag zum Forschungsdiskurs um Entsagung und Ironie bei Goethe". In: *Zeitschrift für digitale Geisteswissenschaften*. doi: http://dx.doi.org/10.17175/2019_007.
- Hu, Yuening, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith (2014). "Interactive topic modeling". In: *Mach Learn* 95, pp. 423–469. doi: <http://dx.doi.org/10.1007/s10994-013-5413-0>.
- Jacobs, Eva, W. H. Barber, Jean H. Bloch, F. W. Leakey, and Eileen Le Breton (1979). *Woman and society in eighteenth-century France : essays in honour of John Stephenson Spink*. London: Athlone Press.
- Jockers, Matthew Lee (2014). *Text analysis with R for students of literature. Quantitative Methods in the Humanities and Social Sciences*. Heidelberg, New York, Dordrecht, London: Springer, Cham.
- Kilian, Elena (2002). *Bildung, Tugend, Nützlichkeit. Geschlechterentwürfe im spanischen Aufklärungsroman des späten 18. Jahrhunderts*. Würzburg: Königshausen & Neumann.
- Krautter, Benjamin, Janis Pagel, Nils Reiter, and Marcus Willand (2020). "»[E]in Vater, dächte ich, ist doch immer ein Vater«. Figurentypen im Drama und ihre Operationalisierung." In: *Zeitschrift für digitale Geisteswissenschaften*. doi: http://dx.doi.org/10.17175/2020_007.
- Laqueur, Thomas (2003[1990]). *Making Sex. Body and Gender from the Greeks to Freud*. Cambridge et al.: Harvard University Press.
- Liu, Alan, Scott Kleinman, Jeremy Douglass, Thomas Lindsay, Ashley Champagne, and Jamal Russell (2017). "Open, Shareable, Reproducible Workflows for the Digital Humanities: The Case of the 4Humanities.org 'WhatEvery1Says' Project". In: *Digital*

- Humanities* 2017. *Conference Abstracts*. Montréal: McGill University and Université de Montréal, pp. 95–98. URL: <https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf>.
- Marjanen, Jani Marjanen, Elaine Zosa, Simon Hengchen, Lidia Pivovarov, and Mikko Tolonen (2020). “Topic modelling discourse dynamics in historical newspapers”. In: URL: <https://arxiv.org/abs/2011.10428>.
- Martín Gaité, Carmen (1972). *Usos amorosos del dieciocho en España*. Madrid: Siglo XXI de España Editores.
- Mauzi, Robert (1969). *L’idée du bonheur dans la littérature et la pensée françaises au XVIII^e siècle*. Paris: Colin.
- McCallum, Andrew Kachites (2002–2018). *MALLET. A Machine Learning for Language Toolkit. V. 2.0.8*. URL: <https://mimno.github.io/Mallet/index>.
- Murakami, Akira, Paul Thompson, Susan Hunston, and Dominik Vajn (2017). “‘What is this corpus about?’: using topic modelling to explore a specialised corpus”. In: *Corpora* 12.2, pp. 243–277. DOI: <http://dx.doi.org/10.3366/cor.2017.0118>.
- Mussner, Veronika (2016). “Die Moralischen Wochenschriften in Frankreich. ‘Le Monde comme il est’ von Jean-François de Bastide als Spiegel seiner Zeit”. MA thesis. Graz: Universität Graz.
- Nelson, Robert K. (2020). *Mining the Dispatch*. URL: <https://dsl.richmond.edu/dispatch/>.
- Norton, Brian Michael (2015). “The Spectator, Aesthetic Experience and the Modern Idea of Happiness”. In: *English Literature* 1.2, pp. 87–104.
- Pabst, Esther S. (2007). *Die Erfindung der weiblichen Tugend. Kulturelle Sinngebung und Selbstreflexion im französischen Briefroman des 18. Jahrhunderts*. Göttingen: Wallstein.
- Portela, Manuel and António Rito Silva (2017). *Arquivo LdoD: Arquivo Digital Colaborativo do Livro do Desassossego*. Coimbra. URL: <https://ldod.uc.pt/>.
- Roe, Glenn, Clovis Gladstone, and Robert Morrissey (2016). “Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie”. In: *Frontiers in Digital Humanities* 2. DOI: <http://dx.doi.org/10.3389/fdigh.2015.00008>.
- Schaufler, Birgit (2002). *‘Schöne Frauen – starke Männer’. Zur Konstruktion von Leib, Körper und Geschlecht*. Opladen: Leske + Budrich.
- Schöch, Christof (2016). “Topic Modeling with MALLET. Hyperparameter Optimization”. In: URL: <https://dragonfly.hypotheses.org/1051>.
- (2017). “Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama”. In: *Digital Humanities Quarterly* 11.2. URL: <http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html>.
- Scholger, Martina, Bernhard Geiger, Elisabeth Hobisch, Philipp Koncar, Sanja Sarić, Yvonne Völkl, and Christina Glatz (2022). *Distant Spectators. Distant Reading for periodicals of the Enlightenment (DiSpecs)*. GitHub repository. URL: <https://github.com/distant-spectators/DiSpecs>.
- (2019–2021). *Distant Spectators. Distant Reading for Periodicals of the Enlightenment (DiSpecs)*. Graz. URL: <https://gams.uni-graz.at/dispecs>.

- Shadrova, Anna (2021). "Topic models do not model topics: epistemological remarks and steps towards best practices". In: *Journal of Data Mining & Digital Humanities* 2021. DOI: [10.46298/jdmdh.7595](https://doi.org/10.46298/jdmdh.7595). URL: <https://jdmdh.episciences.org/8608>.
711 712 713
- Sieuzac, Laurence (2009). "Éducation et vocation de la femme au Siècle des lumières". In: *Genre & Éducation. Former, se former, être formée au féminin*. Ed. by Paul Pasteur et al. Mont-Saint-Aignan: Presses universitaires de Rouen et du Havre, pp. 271–287.
714 715 716
- Steinbrügge, Lieselotte (1987). *Das moralische Geschlecht. Theorien und literarische Entwürfe über die Natur der Frau in der französischen Aufklärung*. Weinheim: Beltz.
717 718
- Sullerot, Evelyne (1966). *Histoire de la Presse féminine en France des origines à 1848*. Paris: Armand Colin.
719 720
- TEI Consortium (2021). *Guidelines for Electronic Text Encoding and Interchange*. URL: <http://www.tei-c.org/P5>.
721 722
- Vökl, Yvonne (2022). *Spectatoriale Geschlechterkonstruktionen: Geschlechtsspezifische Wissens- und Welterzeugung in den französisch- und spanischsprachigen Moralischen Wochen-schriften des 18. Jahrhunderts*. Bielefeld: transcript.
723 724 725

Conference