

Conference Reader
2nd Annual Conference of
Computational Literary Studies
CCLS 2023 Würzburg
June 2023

Venue: University of Würzburg | Zentrum für Philologie und Digitalität
Campus Hubland Nord | Room 01.001 | Emil-Hilb-Weg 23 | D-97074 Würzburg

Local Organizer: Priority programme [SPP 2207 Computational Literary Studies](#)

Contact: spp2207@uni-wuerzburg.de

Hashtag: #CCLS2023

Conference Programme

Thursday | June 22, 2023

1:00 p.m. to 2:15 p.m. | Session 1

- Opening
- Human Depiction in Portuguese – Cláudia Freitas*, Diana Santos
- A Novel Approach for Identification and Linking of Short Quotations in Scholarly Texts and Literary Works – Frederik Arnold*, Robert Jäschke

2:45 p.m. to 4:15 p.m. | Session 2

- Automatic Topic-Guided Segmentation of Holocaust Survivor Testimonies – Eitan Wagner, Renana Keydar*, Amit Pinchevski, Omri Abend
- InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline – Kai Kugler*, Simon Munker, Johannes Höhmann, Achim Rettinger
- What do characters do? – Andrew Piper

4:45 p.m. to 6:15 p.m. Session 3

- Need a Good Book about Privacy? – Erik Ketzan*, Jennifer Edmond
- The Authorship of Stephen King's Books Written Under the Pseudonym "Richard Bachman": A Stylometric Analysis – Vincent Neyt, Mike Kestemont, Dorothy Henriette Modrall Sperling*
- Extracting Geographical References from Finnish Literature. Fully Automated Processing of Plain-Text Corpora – Harri Kiiskinen*, Asko Nivala, Jasmine Westerlund, Juhana Saarelainen

6:30 p.m. | Keynote

- Jan Rybicki: Reading too many books: first results on 10,005 Polish original and translated texts.

8:00 p.m. | Joint Dinner

Friday | June 23, 2023

9:15 a.m. to 10:45 a.m. | Session 4

- Stylistic History of the Hungarian Novel Based on Sentence Structures – Botond Szemes
- Why the Daisy sisters are different. a stylometric study on the oeuvre of Swedish author Henning Mankell and the Dutch translations of his work – Martje Wijers
- Translation-based connotation visualization for classical poetic Japanese vocabulary of the Kokin Wakashū ca. 905 – Xudong Chen*, Hilofumi Yamamoto, Bor Hodošček

11:15 a.m. to 12:15 p.m. | Session 5

- What's that Scary Sound? – Svenja Guhr*, Mark Algee-Hewitt
- Connecting the Dots – Leonard Konle*, Merten Kröncke, Simone Winko, Fotis Jannidis
- Computational approaches to opera libretti – Luca Giovannini*, Daniil Skorinkin

12:15 p.m.: Closing

Gender Depiction in Portuguese Distant reading Brazilian and Portuguese literature

Cláudia Freitas¹ 
Diana Santos² 

1. Department of Letters, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, COUNTRY.
2. Department of Literature, University of Oslo, Oslo, COUNTRY.

Citation

Cláudia Freitas and Diana Santos (2023). "Human Depiction in Portuguese. Distant reading Brazilian and Portuguese literature". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-21

Date received 2023-01-19

Keywords

distant reading, annotation, Brazilian literature, Portuguese literature

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. In this paper we look at how masculine and feminine characters are described in literature in Portuguese, using a publicly available literary corpus: *Literateca*. We investigate the words used to characterise human beings, after classifying them in four broad categories, namely those related to the social, appearance, character and emotional axes. We study the influence of genre, literary school, author gender, and time, among others.

1. Introduction

The way people are described is a rich source of information about societies and cultures, revealing values and beliefs of those who describe. In addition to proper names, there are many other ways of human designation, such as the use of human general nouns like *man*, *woman*, *person*, *gentleman*, *lady*, and designation by traits or functions of the people mentioned (using places of origin, professions, family ties, etc, such as *Brazilian*, *doctor*, *mother*, *foreigner*).

In this paper, we look into how human beings are characterised in literature in Portuguese – also called lusophone literature – using a distant reading approach. In particular, we want to investigate the influence of features such as authorship, geographical origin, historical period and gender (both character gender, and authorial gender).

Inspired by Moretti and Sobchuk 2019's warning, we try to go beyond simple visualisations by date or author, and add other ways to look at the data. Following their "dissecting table" analogy, our aim is to find which pieces are able to provide pertinent analysis, triggering meaningful readings. So, we search for "creative cuttings", – such as the "volume" of speech verbs in Katsma 2018 – to give us new insight. Specifically, we add the class *human depiction* to our data; still, we aim for consensual and understandable categories, like "century" in history.

1.1 Gender in literature

The theme of gender roles in fiction texts has received increasing attention in the digital humanities community, as the following works testify.

Underwood et al. 2018, looking at English literature (104 thousand works, from 1703 to 2009), found that the gender difference between characters became less pronounced from the middle of the nineteenth century to the present day: actions and attributes of

characters became less defined by gender categories. In other words, gender roles tend to become more flexible. At the same time, they also found a decrease in the number of feminine characters as the volume of fiction written by women from 1850 to 1950 drops by a half.

Exploring the *Black Drama* collection, which contains plays written between 1950-2006, Argamon et al. 2009 reports poor results when trying to automatically distinguish the gender of the author and/or character. However, they found differences in the way masculine and feminine authors and characters use language. Feminine playwrights allocate more than half (52.1%) of speeches to feminine characters, while 34.7% of speeches in plays by masculine authors belong to feminine characters.

Working with present-day Dutch literary fiction (170 novels published in one sample year), Smeets 2021 found the same imbalance between masculine and feminine characters. However, the author questions what he describes as a “perhaps naive mimetic assumption” according to which the relative absence of feminine characters is a result of their unequal status in society. From the results of his investigation, feminine characters, although fewer in number, occupy a relatively central position in their fictional social networks – they display more relations, both more relations in general and more relations with important characters.

Hoyle et al. 2019, using 3.5 million digitized books in English, analysed the lexical choice (adjectives and verbs) associated to feminine gendered nouns and found that positive adjectives used to describe women were more often related to their bodies than adjectives used to describe men. Following the same trend, Schulz and Bahník 2019 explores the depiction of male and female characters using the Google Books Ngram corpus, focusing on twentieth-century English-language fiction. The study analyses adjective-noun bigrams associated with the words *man*, *woman*, *boy* and *girl*, and reports that adjectives associated with *men* are more positive (“honest”, “wise”, “honorable”, and “able”) than those associated with *women* (“vulgar”, “foolish”). As to preferences, “charming”, “fashionable” and “warm” were relatively feminine words, while “lazy” and “mean” were relatively masculine words. Men were described in decreasingly masculine terms throughout the beginning and end of the 20th century; on the other hand, the masculinity of adjectives used to describe women started to slightly increase from 1968 to 2000.

Weingart and Jorgensen 2013 performed a computational analysis of gendered bodies in ca 200 European fairy tales (German, French and Italian folklore texts translated into English). They show that feminine characters are described more than masculine characters with appearance-evaluative words, suggesting that men are associated with the mind and women with the body.

Cermáková and Mahlberg 2022 explores linguistic descriptions of gendered body language and compare 19th-century British children’s literature (ChiLit Corpus) with contemporary fiction for children (the OCC2000+ corpus, a subcorpus of the Oxford Children’s Corpus). Using a corpus linguistic approach, the authors study sequences of 5 words which contain at least one body part noun and a marker of gender. They found fewer clusters for feminine characters in the 19th century. The contemporary data suggests, on the other hand, a trend for feminine and masculine clusters to become more

similar, and an increasing range of options for the description of feminine characters and their interactional spaces. Using the same ChiLit corpus, Cermáková and Mahlberg 2021 focused on nouns – excluding proper names – frequently used to label people, and found that *Mothers* are the most frequent occurring feminine character in the corpus.

It is also worth noting the existence of studies such as Cao and Daumé 2021 and Lucy and Bamman 2021. The first one explores the consequences of gender bias for machine learning. The paper investigates how different aspects of linguistic notions of gender impact an annotator’s judgements of anaphora, and points out that a significant possible source of bias comes from the annotations themselves – from underspecified annotation guidelines and the human annotators. The authors emphasise that both humans and systems should not over-rely on cues such as names, semantically gendered nouns and terms of address, relying on “relatively safe” cues like syntax instead. At the other pole of the machine learning approach, the study conducted by Lucy and Bamman 2021 raises questions on how to avoid unintended social biases when using large language models for storytelling. Focusing on how GPT-3 may perceive a character’s gender based on textual features such as personal pronouns (*he/she/her* etc), the work finds that stories generated by GPT-3 place masculine and feminine characters in different topics and exhibit many gender stereotypes: for example, feminine characters are more associated with family and appearance than masculine characters.

In this paper, we also try to contribute to the investigation of gender roles, using works written in Portuguese. As a crossover between corpus linguistics and digital humanities, we use morpho-syntactic and semantic information automatically provided by the PALAVRAS parser Bick 2014, and we add extra semantic annotation, which will be described below.

With Larson 2017, we recognize that using gender as a variable in Natural Language Processing is an ethical issue, and that we need to explicitly explain what “gender” means along this work. As Larson 2017 points out, there are many views of how gender functions as a social construct. In this study, we treat gender as binary, since in the vast majority of works in our corpus gender was mainly constructed in terms of the binary distinction femininity/masculinity. But we acknowledge that the category “gender” can be more complex than this binary distinction, and that these kinds of studies, which describe the cultural apparatus around gender for an extended period of time do not in any way purpose to assert what gender is, but only how it was/is perceived. So they should not be used for reinforcing gender stereotypes, as warned against by Mandell 2019.

1.2 Previous work for Portuguese

For distant reading of Portuguese, we are aware of some works dealing with characters in literature Santos and Freitas 2019, as well as of the DIP challenge for automatic character identification in Portuguese Santos et al. 2022b, to which we come back later.

Our point of departure is the work by Freitas et al. 2022¹ – and later extended in Silva 2021’s master thesis – who have suggested a fourfold classification for human characterisation. Human attributes were organised in social, appearance, character and emotional

1. Although published in 2022, the work was conducted in 2018

characteristics. 111

Using OBRas, a corpus of Brazilian literature in the public domain Santos et al. 2018, 112
they studied 223 works by 25 Brazilian authors, two of them women (authoring 3 novels 113
altogether), and observed the following trends: 114

- Men were more frequently described than women (60%-40%), something which 115
may be related to the fact that there were more masculine characters than feminine 116
ones, roughly in the same proportion. 117
- The most frequent masculine characterising words were *bom* (good), *sério* (honest), 118
rico (rich) and *alto* (tall), while *bonita* (beautiful) was by far the top characteristic 119
for women 120
- Almost 50% of women depicting words were about beauty (namely *bonita* and 121
bela) 122
- Character and social predication were most frequent for men; for women, social 123
characterisation reduces to *married* and *rich*. 124
- Emotional characterisations like *feliz* ('happy') were (almost) exclusively used for 125
women. 126

We wanted to check whether these observations held for a wider collection, including 127
Portuguese literature as well. 128

1.3 A brief comparison with DIP 129

It is useful to compare and contrast our study with the recent DIP challenge for Por- 130
tuguese, an evaluation contest for identifying literary characters and some information 131
about them in Brazilian and Portuguese works Santos et al. 2022a, 2023. By describing 132
it and pointing out the differences, we throw some light on different ways to look at 133
(roughly) the same data. 134

For DIP the unit is the literary character, and so the challenge looked at their gender, 135
their profession, occupation and/or social status, and their family relations with other 136
characters. But the unit is the character. In addition, "literary character" in DIP does not 137
include all people. In the present study, we try to look at all mentions of characterisation 138
of people in the works, so our numbers are not per characters, but per mentions of 139
people. 140

We will discuss and compare the findings about character gender in section 4.7. 141

1.4 The importance of studying literature in Portuguese 142

Portuguese has a rich literary tradition, but unfortunately the digitisation efforts are 143
lagging behind other languages. This has for example been discussed in Schöch et al. 144
2021. 145

Also, major actors in the big data landscape, no matter the high number of Portuguese 146
speakers in the world, have not endowed Portuguese with the "current" tools that 147
are available for other languages, even with much fewer speakers/readers/writers, 148

like Hebrew or Italian: there is, for example, no Google Book N-grams² service for Portuguese. 149 150

Likewise, recent reviews of the computational literature landscape, because they do not find enough internationally published DH papers on Portuguese, have decided not to review or include them, therefore contributing actively to the lack of information on lusophone materials and studies. For example, Schöch et al. 2022, page 4 state: 151 152 153 154

As several languages, however, were represented only with relatively low numbers of articles or papers, and in order not to misrepresent the research communities these publications stem from, we decided not to take the materials in several languages into account: (...) 155 156 157 158

This is one of the reasons why we are writing this paper for an international audience. Maybe the results are not so different than the ones our English-speaking or English-studying colleagues obtained, but they are novel because they are obtained from completely different data. 159 160 161 162

2. The material 163

We provide here an overview of the data used, also with the purpose of making it known, and hopefully, useful, for other researchers. And not the least because it shows the methodological problems it invites. 164 165 166

Attempting to complement close readings of canonical authors with a wider material, following Moretti 2000, 2013 and Underwood 2019, we use as many books whose full text is currently publicly available in Portuguese to investigate properties of literary text which can be identified in an automatic way. 167 168 169 170

In order for these data to be shareable and studies replicable, we restrain our data (mostly³) to books in the public domain. We are aware that many more electronic texts exist in electronic form, but by using them we would incur either on law infringement, or at least we would risk creating materials only for our own study, not shareable by others. 171 172 173 174 175

Also, it is important to stress that we are referring to textual versions of the works, not simply images. Optical character recognition for Portuguese, especially for old books, is not good enough yet, so all books have been revised by humans, if not born digital. 176 177 178

2.1 Corpus 179

We used Literateca version 11.1, created on 26 May 2023, comprising ca 32 million tokens of (original) prose (excluding drama) from 1700 on. 180 181

A quantitative overview of the material is in Table 1. 182

Figure 1 shows the distribution of the material in time, by size in words. 183

Literateca is the merge of several literary corpora written in Portuguese, and thus has 184

2. <https://books.google.com/ngrams/>

3. Exceptions are excerpts of books existing in parallel corpora, or texts whose authors gave us permission to use.

Literature	no. of tokens	no. works	no. authors
Total	32,718,621	669	200
Portuguese	20,639,007	306	127
Brazilian	12,079,614	355	73

Table 1: Size of the material, prose from 1700 to the present

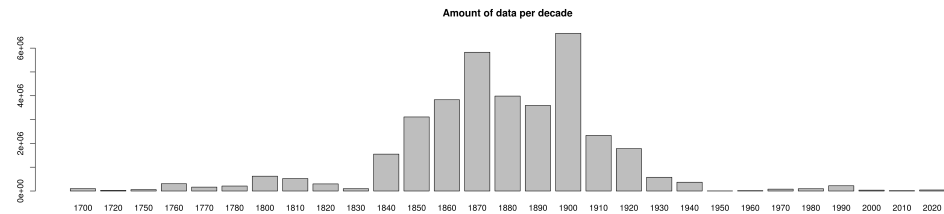


Figure 1: Distribution of words per decade

some particularities:

- It includes literary works by canonical authors, but also other works by those canonical writers which are not usually or necessary deemed literary, such as newspapers chronics, letters, memoirs, and even scholarly works such as history books or ethnographic studies, and travel reports. For previous centuries, even sermons are included. However, these genres are only included for canonical writers.⁴
- It includes drama, poetry and prose.
- Some of the works have updated orthography, others keep the original orthography. Given that there have been several norms of Portuguese spelling across the centuries, this means that there can be a variety of forms for the same word.
- While some authors have all their works included, others have only a few, or just one. Especially for non-canonical writers, there is no claim to completeness.
- Given that the works have been digitised by different bodies and with different tools and for different purposes, there is no claim to homogeneity: works can come from the first or the last paper version, they may keep their prefaces or not, they have different ways of describing chapters, etc.
- All works are marked with author, author gender, date of publication, variety of Portuguese, genre, and whether they are original or translated. Some texts are also classified by the literary school they belong to.

We tried to use as much as this material as we could, but we removed poetry and drama. Poetry is probably a natural choice to be removed, because of syntactic idiosyncrasies – and therefore a worse parser performance –, and because we believe that poetry has not so many mentions of fictional characters. We removed drama, also in prose, because it was heavily unbalanced, given that most of the plays were from Portugal.

4. By this we mean that established authors who belong to the Portuguese and Brazilian canons were fully digitized, that is, everything they published is available. This is in strong contrast with the works of non canonical authors, which may have had some of their (mainly) novels digitized in the context of other projects.

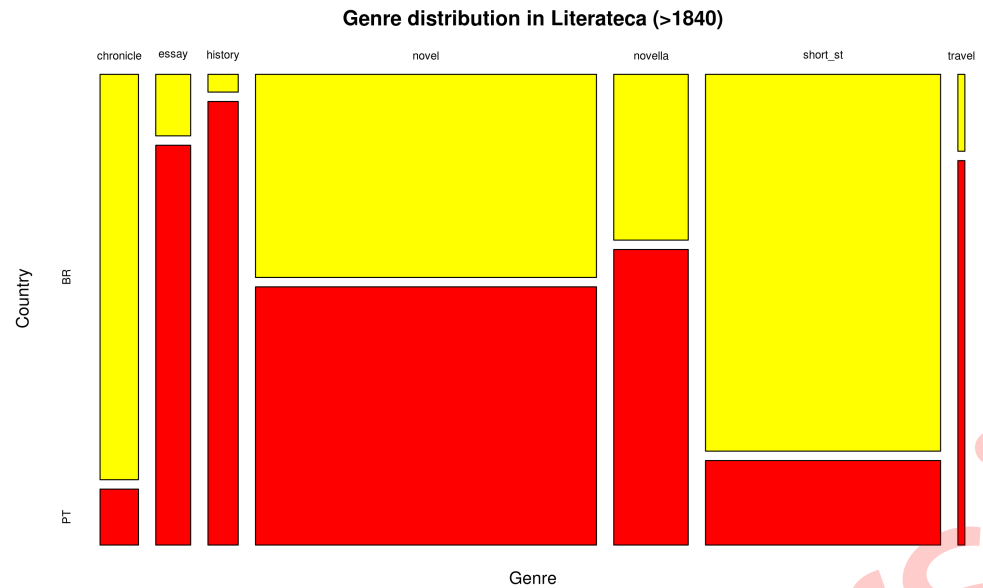


Figure 2: Genre in the full corpus. The unit is the work.

As to prose, we started to use everything published since 1700. It is, anyway, important to recognise that we do not have a balanced corpus, and the lion's part is fiction. We then selected different subsets for different research questions.

- Just fiction, and just non-fiction, to see whether depiction was different across the fiction divide
- Just works published after 1840, to be able to compare Brazilian and Portuguese authors
- Just fiction published after 1840, to be able to compare Brazilian and Portuguese literature

See figures 2 and 3 for a bird's eye view of the genre distributions in total and in fiction.

Only in Figure 3 do we include the variable author gender, since it is only in fiction that we have text written by women.

In Table 2 we give the numbers of words involved for the material published after 1840.

	Fiction	Non fiction	Total
Brazil	10,547,327	1,532,287	12,079,614
Portugal	15,280,938	5,358,069	20,639,007
Total	25,828,265	6,890,356	32,718,621

Table 2: Size in words of the different materials, after 1840.

2.2 Gender attribution

We explore the influence of gender both in characters' description and authorship. Masculine and feminine gender labels were manually ascribed to writers, for our corpus contains works written by canonical authors that have been widely discussed in literary studies. For the non-canonical authors, gender was attributed either based on

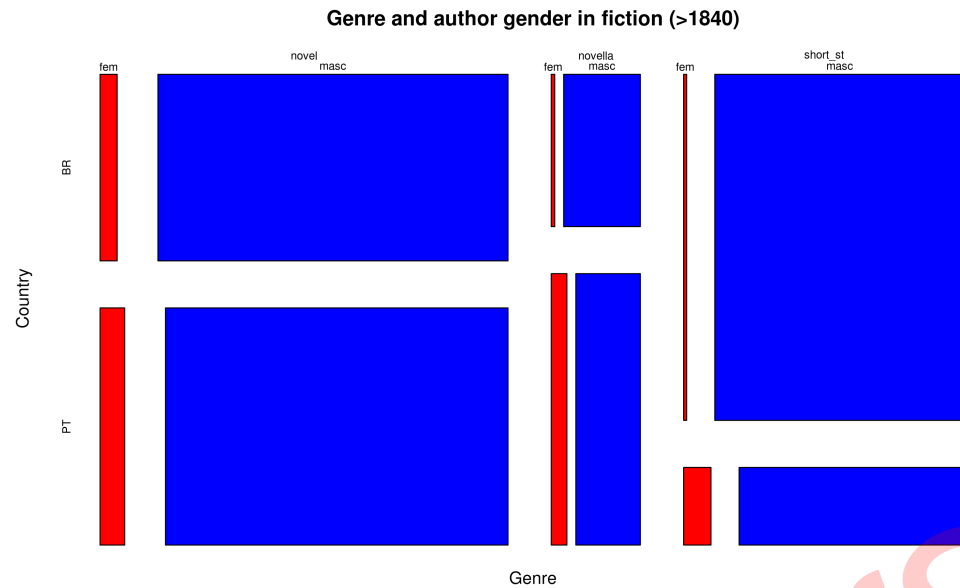


Figure 3: Genre in the fiction corpus. The unit is the work.

adjective/inflected forms used in prefaces or based on their proper names. As to the characters, gender labels were automatically assigned by PALAVRAS parser, and then manually revised by linguists (Rocha et al. 2019; Silva 2021). The linguistic clues followed on attributing and revising gender were syntactic agreement and morphological features.

Portuguese is a Romance language that forces the speakers to specify the gender of nouns (both common and proper nouns) and adjectives. The main formal clue to distinguish masculine and feminine forms is the word's ending: masculine forms tend to end in -o, feminine ones tend to end in -a, and those ending in -e can be both feminine and masculine – *ponte* ('bridge') is feminine, and *pente* ('comb') is masculine. However, there is no perfect equivalence between the ending in -o or -a and the masculine or feminine gender, respectively – *planeta* ('planet') is masculine, and *tribo* ('tribe') is feminine. Therefore, to observe syntactic agreement between the head noun and its modifiers is the most reliable way to assign morphological gender.

When calculating the gender of depicting words, we take into account the gender of the nominal head (noun, proper noun or pronoun) being characterised, not the gender of the words (modifiers) associated with it. This choice is due to the fact that, although adjectives can be inflected for gender in most of the cases, the search patterns we used also retrieve nouns, which do not admit inflection. Thus, nouns like *anjo* ('angel') will always be masculine, even if the mentioned angels are feminine. When considering the gender of nominal heads, *anjo*, although a masculine common noun, is classified as a feminine classifier if it modifies a feminine character.

3. The process

250

We wanted to identify all cases where human beings were mentioned to find out how they were described, or depicted. We extended the search patterns used by Silva 2021⁵ in two ways: (i) we enriched the lexicon of general human nouns, including names of professions as targets, and (ii) having extended the amount of works analysed to include works written by Portuguese authors, we broadened the lexicon of characterising words⁶, based on prose of the eighteenth, nineteenth and twentieth centuries in Literateca. Along the process of data analysis, we were forced to discuss previous classification, which lead to precise classification guidelines and to a reclassification of a few words.

We start from the idea that specific linguistic patterns indicate certain (semantic) relationships. So, we have used a set of patterns – relying on the automatic morpho-syntactic annotation – to search the material for instances of describing human beings. Some examples of what the patterns yielded follow (the patterns are publicly available).

- (1) – Ouviste? – perguntou **ela inquieta**. [– Did you hear? **she** asked **restlessly**.] 263
- (2) ...acudiu logo o **padre**, muito **arisco**. [... came the **priest**, very **skittish**.] 264
- (3) Uma **mulher honesta** não tem segredos para seu marido! [A **honest woman** has no secrets from her husband!] 265
- (4) **D. Joana Tecla** era **idiota**. [–Mrs. **Joan Tecla** was an **idiot**.] 267
- (5) Em todo o caso era uma bela **mulher**, **alta** e forte sem ser gorda... [In any case, she was a beautiful **woman**, **tall** and strong without being fat...] 269
- (6) ...calado como a tarde triste, um **homem**, ainda **moço**, vestido como os essênios taciturnos, caminhava... [...silent as the sad afternoon, a **man**, still **young**, dressed like..] 272

Then we proceeded to classify each word of the aforementioned list – which are the words associated with human beings in the examples –, in four (non-mutually exclusive) classes, according to type of characterisation: social, emotional, physical (appearance) and character. In order to group these idiosyncratic data and provide a better view from afar, we analysed the most frequent words and came up with the four classes. We also used the class other if none of the four could hold, and one or more of the four otherwise. As to the assignment of the categories proper, follows their scope and the major decisions associated:

social In addition to professions, occupations and social status, we also included absence of profession like *mendigo* ('beggar'), nationality, civil status, family relations, political opinions like *monárquico* ('monarchist'), and cases which are a consequence of social intercourse, like *ignorante* ('ignorant') or *educado* ('civil' or 'knowledgeable').

5. Which, in turn, are an improvement of the patterns used in Freitas et al. 2022.

6. The list comprises not only adjectives and nouns, but also verbs (for past participles), given that it is a feature of PALAVRAS that most participles are analysed as verbs even though in an adjectival context.

appearance Physical appearance, including clothing or lack of it, as well as those features associated with time, as *jovem* ('young') or *velho* ('old').

emotional Feelings, emotions and emotional tendencies.

character Personality traits, also including cognitive properties, such as intelligence or lack of it. It also includes evaluations according to social conduct, such as *honesto* ('honest'), *malcriado* ('rude') or *pretensioso* ('snob').

It is important to mention that each category works as a label, which in turn encodes 4 perspectives on people: 'appearance' refers to what is visible; 'social' refers to the various roles someone can play in society; 'character' refers to internal/cognitive characteristics; and 'emotion' refers to emotional traits. We could also, and more broadly, consider two large classes: internal characteristics ('character' + 'emotion') and external characteristics ('appearance' + 'social'). We note that the words classified can often refer to non-human entities, as is the case of the next example (7). But if they could modify a human person, they were classified accordingly. However, the results presented in the next sections refer only to those cases where the characterisation was assigned to human beings, such as example (8), since only they are retrieved by the patterns applied.

(7) – Que **triste** pensamento!... [What a **sad** thought!]

(8) – Mas a **triste** senhora continuava a choramingar. [But the **sad** woman kept weeping.]

We classified the retrieved words out of context, except in those rare cases where we had to check whether the adjective had been used as characterising at all in the corpus⁷. For example, initially we wanted to discard the words *granítico* ('made of granite') and *triumfal* ('of triumph'), but we checked the corpus and there were instances where both were applied to human characters, so they were retained in our list.

(9) – Sim, o velho Afonso é **granítico**... [– Yes, old Afonso is **granitic**...]

(10) Nunca as mulheres **triumfais** me fizeram bater o coração... [**Triumphal** women never made my heart beat...]

The classification was done manually by the authors of this paper, and divergences were heartily discussed. We dismissed mistakes, either because (i) they were not characterisation words, (ii) they resulted from wrong parsing, or (iii) we decided they were not relevant to our goals. As to the exclusion:

- We did not take into account "complex adjectives" in the sense of having more than one word, like *bem intencionado* ('having good intentions'), *mal intencionado* ('having bad intentions'), *bem educado*⁸ ('polite'), etc.

7. Actually, there was one case where we consistently considered the context: in Portuguese, the word *grande* can mean either *big* or *great*. Since each meaning corresponds, in general, to a different syntactic position – *grande homem* ('great man'); *homem grande* ('big man'), we used this information to correctly classify each of the occurrences: *character* or *appearance*, respectively.

8. But note that *educado* and *bem-educado*, as words of size one, were included.

- We did not classify relational adjectives, such as *partidário* (*de...*) ('partisan'), *apologista* (*de...*) ('in favour of'), *comparável* (*a ...*) ('comparable to'), *emparelhado com* ('pairing with'), *semelhante a* ('similar to'), since a precise characterization would require a close reading of each sentence.
- We threw away misspellings, except for lack of diacritics.⁹ Our rationale being that, in future improved versions of the corpus, the corrected words would be correctly annotated.

Following the annotation approach adopted in the AC/DC project Santos 2014, underlying Literateca, we used multiple classification when two or more categories/senses could be assigned to a characterising word (vague or ambiguous words). References to madness, for instance, were considered both social and character. The same for habits like *madrugador* ('early riser') and *bêbado* ('drunkard' or 'drunk'), which can be either due to biology or social bringing up. The word *acanhado* (shy), can be interpreted as a not-expansive person (character) or as someone fearful (emotion), and the same applies to *impaciente* (impatient), which can be interpreted as anxious (emotion) or restless (character).

Finally, cases such as *maravilhoso* ('wonderful'), *incomparável* ('incomparable'), *ideal* ('ideal') or *horrível* ('horrible'), where it is not clear to which axis they apply out of context, were classified as referring simultaneously to 'character', 'social' and 'appearance'.

To verify the degree of reliability of the classifications and the adequacy of the classes, Silva 2021 carried out a study on the inter-annotator agreement of 15 people in the classification of occurrences considered especially difficult. The degree of agreement was 80%. We have not carried any further studies on this matter.

After this classification, we ended up with a list of 4481 words¹⁰ which might be employed in depicting human beings, see Table 3. Due to the vagaries of the parser, we list the lemmas which can be verb infinitives for past participle forms, because we use the lemmas in our patterns.

type	size
social	1391
appearance	672
emotional	514
character	1578
other	326
total	4481

Table 3: Depicting words, by category. Recall that words can belong to more than one category.

In order to provide a richer description of this list, we show in Table 4 how often depicting words are vague or ambiguous.

We then annotated the corpus with this new classification¹¹ and computed how often

9. That is, we considered missing accents something that could be present in the original paper edition, but not OCR mistakes.

10. Available from <https://www.linguatca.pt/Gramateca/ListaPredicadoresClassificados.txt>.

11. The classification is encoded in the following tags `pred:carater`, `pred:aparencia`, `pred:social` and `pred:emo`. To find them in Literateca, search for `[sema=".*pred:social.*"]`, etc.

type	size
appearance character	88
appearance emo	12
appearance social	8
appearance character social	10
character emo	107
character emo social	1
character social	80
emo social	9
total with more than one class	315

Table 4: Words belonging to several categories.

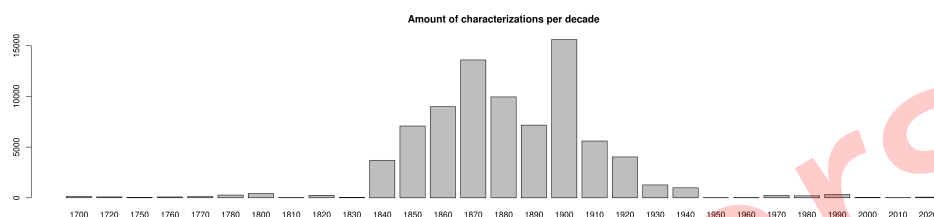


Figure 4: Characterised people in the corpus per decade.

and when the words were used to describe human beings.

We start by providing a picture of the distribution of human characters in time, in Figure 4, as well as how many depicting events we were able to identify, in Figure 5.

A comment is in place: the decade of 1830 is a clear outlier, because it contains one short text only, of 19,334 words, a political pamphlet by Alexandre Herculano, in the whole decade. The same happens with 1950, which in the material is only represented by 4,777 words of Jorge Amado's *Gabriela, Cravo e Canela*.

4. Analysis

The first thing we report is the proportion of these subclasses in our material. Table 5 shows the raw numbers, and also those referring to masculine and to feminine characters.¹² Figure 6 displays the overall distribution of characterisation words.

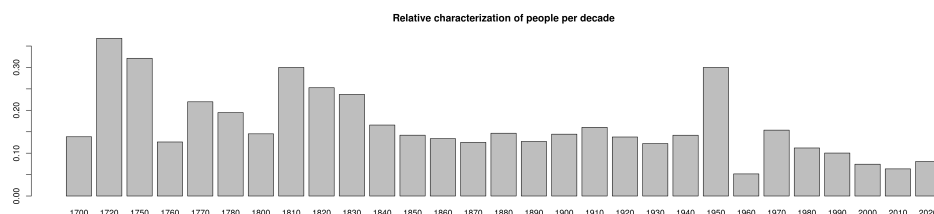


Figure 5: Relative characterisation per person, per decade

¹² It should be noted that the numbers do not add up because in some cases the parser is not able to assign a morphological gender, marking them as M/F. Also, remember that by "character" here we mean mentions to people, not distinct characters.

	Total	Masc. characters	Fem. characters
People	578,815	352,851	173,370
Characterised people	80,415	52,252	24,664
Social	11793	7813	3534
Appearance	15394	9099	5862
Emotion	9670	5562	3895
Character	23880	16542	6394

Table 5: Different depiction classes, in general and per gender of the characterised person, using the subject's gender.

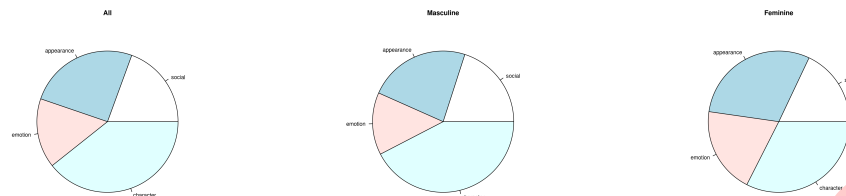


Figure 6: Distribution of characterisation words among the 4 classes, for all, masculine and feminine depictions

The first observation is that there are way more masculine than feminine characters in the material (ca. twice as many). Feminine characters are, however, almost as often characterised as the masculine ones: 14.2% against 14.8%.

The second remark is that the by far most frequent subclass deals with character (most frequent words: *bom* ('good'), *grande* ('great'), *honrado* ('honourable, honest'), *simples* ('simple'), *digno* ('with dignity'), *excelente* ('excellent')), followed by appearance (most frequent words: *velho* ('old'), *novo* ('young')¹³, *antigo* ('old-fashioned'), *jovem* ('young'), *belo* ('beautiful'), *formoso* ('beautiful'), *bonito* ('pretty')).

Social characterisation comes third (most frequent words: *rico* ('rich'), *ilustre* ('illustrious'), *nobre* ('noble'), *casado* ('married'), *célebre* ('famous'), *pobre* ('poor'), *livre* ('free'), *famoso* ('famous')). while emotional characterisation is the least frequent (*pobre* ('poor'), *infeliz* ('unhappy'), *valente* ('brave'), *feliz* ('happy'), *triste* ('sad'), *desgraçado* ('miserable'), *alegre* ('joyful'), *humilde* ('humble')).

Thirdly, feminine characters have a higher chance of being characterised by their appearance compared to masculine ones (23.8% vs. 17.4%), something that corroborates the findings in previous studies, and to which we return in subsection 4.2.

4.1 Does textual genre matter?

Does it make more sense to look only at literary texts, removing travel writing, essays, history and political writings?

On the one hand, we had left all material because we wanted to look at the way people described people in Portuguese, but then it is also conceivable that the kinds of information about people are rather different when you write the history of the Inquisition, an essay about your fellow writers, or a report of how you crossing Africa, compared with

13. It may seem surprising at first to include age as appearance, but it is something that we visually assess.

a narrative where you introduce fictional characters.

So, we reproduced our queries removing all texts not classified as novels, novellas or short stories, and the new numbers are in Table 6.

	Total	Masculine	Feminine
Words	25,828,265		
People	490,892	291,403	159,216
Characterised people	47,450	30,036	16,620
Social	8968	5720	2979
Appearance	12,951	7401	5226
Emotion	8767	4922	3665
Character	19,002	12587	5773

Table 6: Different depiction classes, in general and per gender of the characterised person, using the subject's gender, only in novels, novellas and short stories.

It is interesting to see that removing the non-fictional prose genres does not change the relative order of the subcategories, but increases the percentage of feminine characters, that raises from 30.0% to 32.4/%, and characterised feminine characters, from 33.2% to 35.0/%.

As to the characterisation of masculine and feminine characters, we have trends similar to the ones we present for the full material, as shown in Figure 7: masculine targets are characterised, by far, by their character, while feminine targets are (almost) equally characterised by their appearance and character.

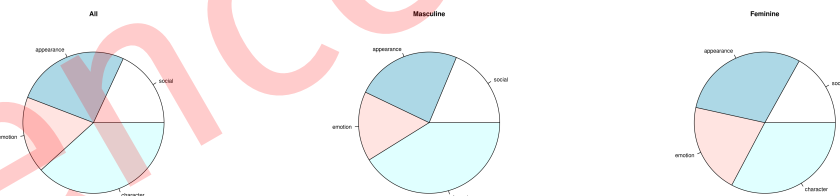


Figure 7: Relative characterisation per gender in novels, novellas and short stories.

Let us see now for the non-fiction part, whether the picture is different. In Table 7 we describe the masculine and feminine characterisations in the (considerably smaller) non-fiction part.

	Total	Masculine	Feminine
Words	6,890,356		
People	87,923	61,448	14,154
Characterised people	10,537	8033	1899
Social	2825	2093	555
Appearance	2443	1698	636
Emotion	966	688	245
Character	4878	3955	621

Table 7: Different depiction classes, in general and per gender of the characterised person, using the subject's gender, only in non fiction.

The percentage of feminine characters, and feminine characterisations shrunk consid-

erably: 16% and 18%, confirming that women are even less important in the public sphere.

Now social characteristics are – globally – more frequent than appearance. Character remains the most important form of describing people, and emotion the least.

In Figure 8 we present the distribution of the four kinds of features, and see that the few women that are mentioned have a large proportion of appearance descriptions, even more in non-fiction than in fiction.

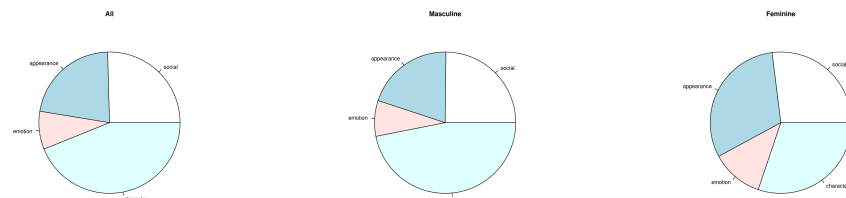


Figure 8: Relative characterisation per gender in non-fiction.

4.2 Differences when describing masculine and feminine characters

The previous figures have shown that appearance is more frequent when describing feminine characters. This can also be appreciated in the barplot of figure 9

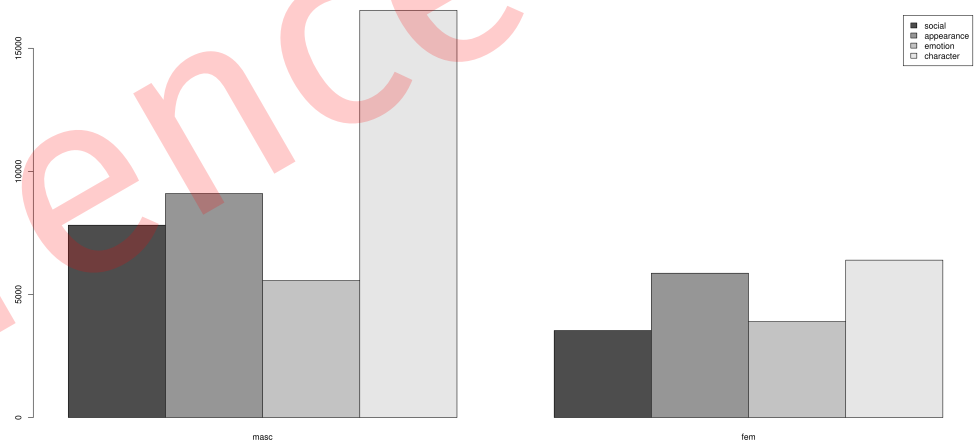


Figure 9: Relative characterisation per gender in the whole material, as a barplot.

However, this is just the tip of the iceberg. The analysis of depictive words preferentially used with masculine and feminine characters can be more revealing than the general analysis we presented in figure 9, which takes into account the whole bunch of depictive words. In order to be evaluated as 'preferred', a word must (i) be used for masculine targets at least for 80% of the occurrences, or for feminine targets more than 60% of the occurrences; (ii) have a total frequency of 4 or more.

In cases where different lexical items correspond to gendered male/female pairs (*mãe/pai* ('mother/father'); *rainha/rei* ('queen/king'); *namorada/namorado* ('girlfriend/boyfriend') etc), we manually grouped the elements of the pair as if they shared the same lemma

so they could be included in the preference count.

419

The new data are in figure 10, which shows a slightly different picture, in which (i) words of the *emotional* axis are almost not seen at all, and almost disappear in the feminine characters, (ii) the balance between *appearance* and *character* in feminine depiction gives way to a characterisation based mainly on *appearance*, which accounts for half of all preferred feminine characterisations, and (iii) *appearance*, the second most frequent characterisation (of both masculine and feminine characters), drops to the third position when associated with masculine characters, and up to the first position, when associated with feminine characters.

427

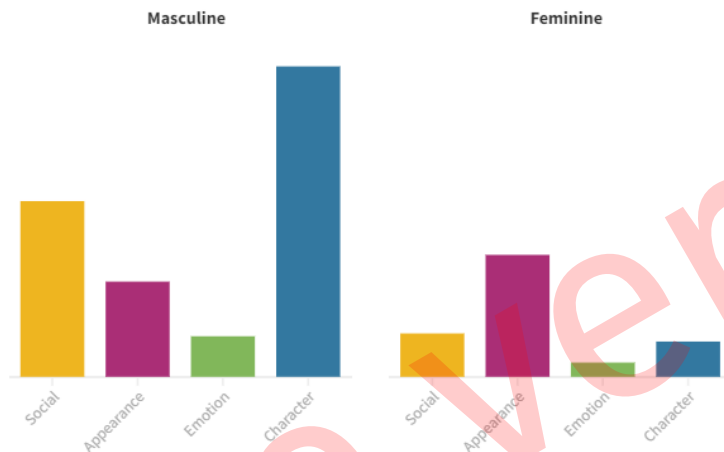


Figure 10: Preferred characterisation per gender

The appearance axis is the second most common for both genders, but figures 11 and 12, complementary to figure 10, provide a few details that enrich the analyses.¹⁴

428

429

As noted in previous studies, typically feminine *social* characterisations relate to the familiar environment (*mãe* ('mother'), *prima* ('cousin')), but mentions to marital status are the highlight (*casada* ('married') and *viúva* ('widow') are the most frequent words, but *adúltera* ('adulteress') is frequent as well). Marital status, in turn, is absent as typical masculine social characterisation. These are related to (positive) social recognition such as *ilustre* ('illustrious'), *célebre* ('famous'), *famoso* (another word for 'famous') and *poderoso* ('powerful').

436

On the feminine emotional axis, words associated with love and sweetness (*adored* and *sweet*) stand out, but also words associated with sadness and insecurity (*poor*, *tearful*, *jealous*, *offended*) and fear (*terrified*). On the other hand, bravery is the masculine highlight: *valente* ('brave'), is, by far, the most frequent word, and *atrevido* ('cheeky/audacious') is in the 6th. Anxiety also appears (*desesperado* ('desperate') is the third most frequent).

442

443

444

Finally, masculine characters seem to be taken by surprise more often than feminine ones, being *maravilhado* ('marveled'), *assombrado* ('haunted') and *surpreso* ('surprised'), which might be due to their roles in narrative events.

Appearance, although highly typical for feminine targets, varies relatively little as to

445

14. In figures 11 and 12, words such as *beautiful_1* and *pretty_2* relate to different Portuguese words that could be translated into the same English word, such as *bonita e formosa*, which could be both translated as *pretty*.

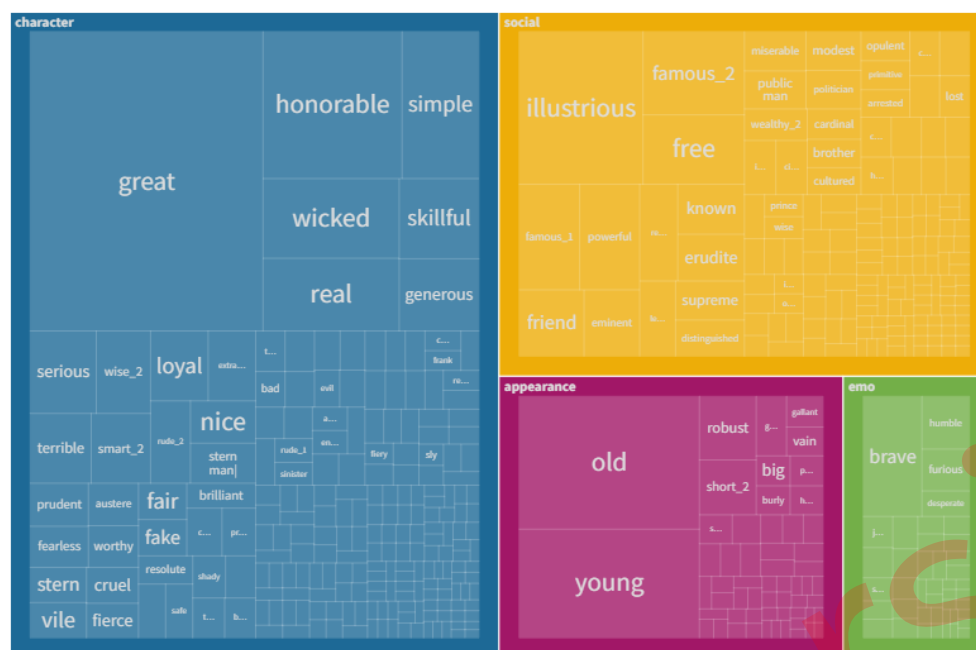


Figure 11: Preferred characterisation of masculine characters.

the most mentioned attributes: beauty (*bonita; formosa, bela, linda*, Portuguese words for 446
beautiful) or the lack of it (*feia* ('ugly')) is the most frequent feature. In the masculine 447
appearance axis, age and size, instead of beauty, are the most mentioned attributes: 448
velho ('old') and *jovem* ('young'); *robust, big* and *short*. 449

In the character axis, typically masculine, stands out *grande* ('great'), *honrado* ('hon- 450
ourable'). Other highly mentioned positive traits are *generoso* ('generous'), *habilidoso* 451
('skillful'), *real, sério* ('serious') and *leal* ('loyal'). For the feminine targets, the highlights 452
are *virtuosa* ('virtuous'), *inocente* ('innocent') and *meiga* ('sweet'). 453

4.3 Does the gender of the author matter? 454

Do these findings change according to the author's gender? In our material, see Table 8, 455
feminine authors use more appearance descriptions than masculine ones, as shown in 456
Figure 13. 457

However, there is a huge difference in the size of the material compared: there are only 458
1.2 million words written by women compared to almost 32 million words by men. In 459
fact, this is an inescapable problem, given the reduced number of writings by women in 460
our corpus: only 19 authors¹⁵ who wrote 33 works in prose. 461

Even though the material is heavily unbalanced, we tried to discern any interesting trend 462
in works written by women as far as whose appearance was more described – could 463
it be that they would emphasise or concentrate more on the appearance of masculine 464
characters? 465

15. Júlia Lopes de Almeida, Virgínia de Castro e Almeida, Ana Plácido, Teresa Margarida da Silva e Orta, Maria Amália Vaz de Carvalho, Maria O'Neill, Maria Firmina dos Reis, Florbela Espanca, M.M.S.A. e Vasconcelos, Cláudia Campos, Maurícia C. de Figueiredo, Maria Luísa Marques da Silva, Matilde Isabel de Santana e Vasconcelos Moniz Bettencourt, Ana de Castro Osório, Alice Moderno, Maria Peregrina de Sousa, Paulina Filadélfia, Clarice Lispector and Sónia Coutinho, by decreasing number of words in the corpus

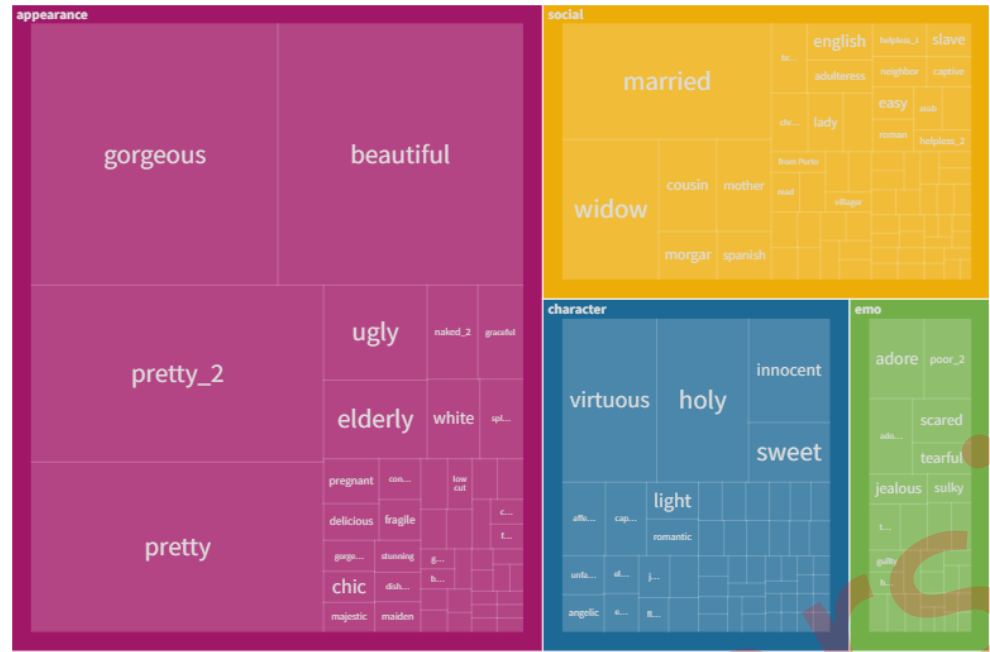


Figure 12: Preferred characterisation of feminine characters.

	Total	Feminine author	Masculine author
Words	25,828,265	1,206,744	24,621,521
People	490,892	24,271	466,621
Characterised people	57,680	2235	55,445
Social	8968	355	8613
Appearance	12951	595	12356
Emotion	8704	533	8171
Character	19002	887	18115

Table 8: Different depiction classes, for masculine and feminine authors, in novels, novellas and short stories.

We get 265 appearance descriptions of feminine characters, and 319 of masculine characters, in 985 characterisations of feminine characters and 1195 characterisations of masculine characters. In other words, 26.9% of feminine characterisations and 26.7% of masculine characterisations involve their appearance. But we acknowledge that numbers are too small to be conclusive. In any case it is conspicuous that both genders have roughly the same characterisation frequency in literature written by women.

Despite the imbalanced data, figure 14 shows preferential characterisation regarding gender of both characters and writers. In what follows we sketch some differences between human depiction in works written by men and women. The main difference is the increase of appearance in masculine characterisation in works written by woman.

Beginning with feminine characters and focusing on women writers only, we found that *married* is no longer among the most frequent social depictions, but *widow* and *single* remain. Despite still being frequent, less space is devoted to beauty in works written by women. On the other hand, age is more present: *young* and *old*. As to emotional characterisation, *happy* and *adorable* are the highlights, and none of the preferred emotional words relate to sadness. As to character, the highlights of feminine depiction words are

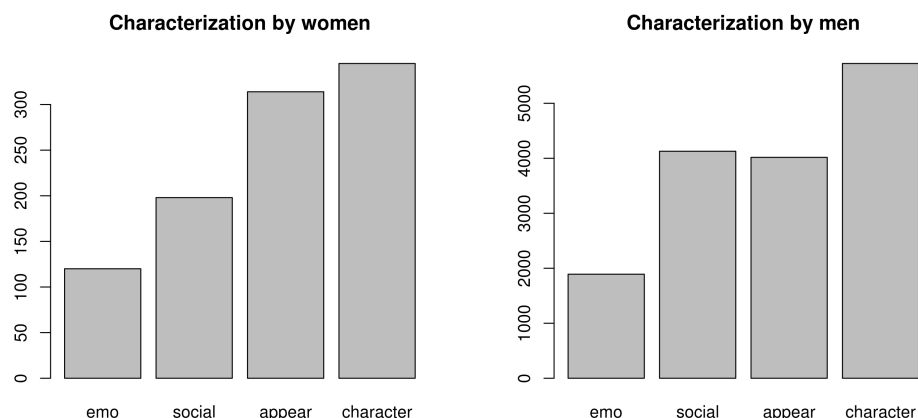


Figure 13: Characterisation by masculine and feminine authors. E PRECISO REFAZER

honest, infamous, crazy, refined and dangerous. In the social axe, masculine characters are mainly *married* and *noble*. Positive emotions are present for masculine characters as well (*happy/pleased, enthusiastic*), but bravery (*brave*) has only one occurrence. Masculine appearance follows the general trend, and masculine character are mainly *kind* and *honourable*.

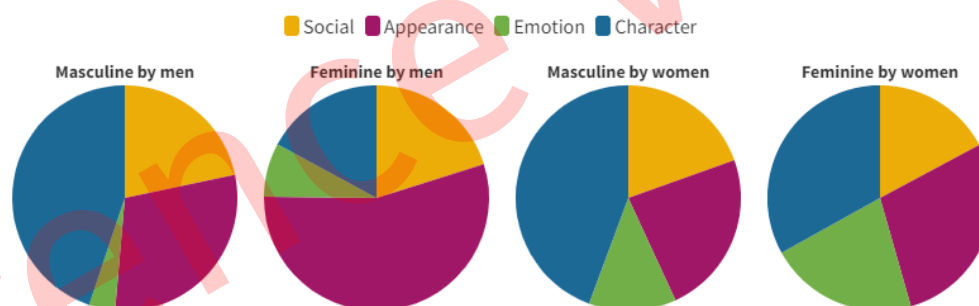


Figure 14: Preferred characterisation by masculine and feminine authors.

4.4 Difference between Brazil and Portugal

Are there differences between the two countries as regards people's characterisation?

We compared the works from 1840 to the present (Brazil became independent in 1822, and, as already mentioned, for the 1830 decade we only have one work by a Portuguese author).

We decided to compare only novels, novellas and short stories between the two countries, because the non-fiction parts differ widely: While we have a large body of texts on history in the Portuguese side, we have almost only short essays in newspapers on the Brazilian side. The results are presented in Table 9.

We see that *character* and *social* characterisation are somewhat higher in Portuguese literature, while the other categories – especially emotion – are more pronounced in Brazilian literature. One may wonder if this is due to a more socially rigid society in Portugal, or whether the cause lies with the historical novels (almost absent in the

	Total	Brazil	Portugal
People	486,575	209,283	277,292
Characterised people	46,704	19,642	27,062
Social	8887	3545	5342
Appearance	12877	6199	6678
Emotion	8704	4874	3650
Character	18782	7649	11133

Table 9: Different depiction classes in novels, novellas and short stories, in general and per author nationality, after 1840.

Brazilian material, and quite frequent in the Portuguese material).

We also investigated whether the differences among genders are more obvious in the Brazilian material, or different from the ones in the Portuguese material. For this, we created Table 10, where we can see that Brazilian literature has a higher proportion of mentions of feminine characters (36.5%) than the Portuguese (29.7%). This may again be due to the historical novels, but will have to be investigated closer.

	Br total	Br fem.	Br masc.	Pt total	Pt fem.	Pt masc.
People	202,829	74,020	118,088	275,301	81,847	165,796
Characterised	17453	6381	10591	24548	8452	15372
Social	3545	1216	2217	5342	1753	3434
Appearance	6199	2579	3472	6678	2618	3885
Emotion	3474	1444	1949	5230	2206	2925
Character	7649	2446	4955	11133	3292	7452

Table 10: Different depiction classes in novels, novellas and short stories after 1840, per author nationality and per gender of the characterised.

Here we see that the social status of male characters is more important in Portuguese literature.

If we now compare the distribution by country by gender, presented in 15, masculine characters seem to be similarly depicted. But for feminine characters, there are significantly relatively less mentions of their social status and more mentions of their appearance in Brazilian authored works.

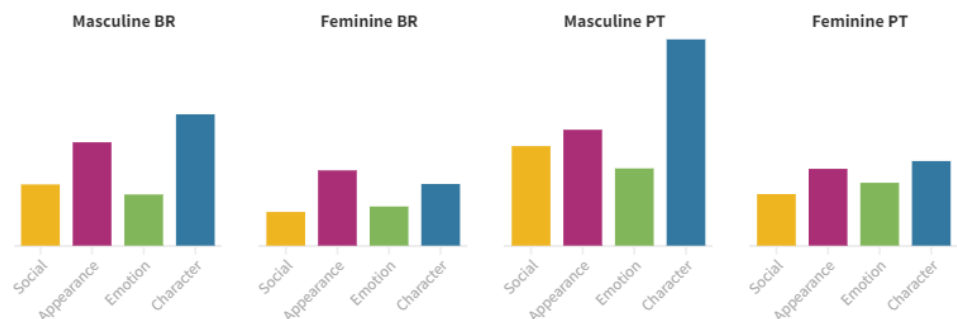


Figure 15: Characterisation by country

4.5 Differences among authors

In table 11, we show the distribution of the kinds of characterisation for 12 canonical authors, 6 Brazilian and 6 Portuguese.

Author	land	nr	total	char	soc	app	emo	mfreq
Camilo Castelo Branco	PT	42	4045	1781	938	845	481	pobre
Machado de Assis	BR	140	1864	793	219	643	209	bom
Eça de Queirós	PT	16	2487	1019	420	923	125	bom
JM de Macedo	BR	7	1325	411	232	515	167	velho
Aluísio Azevedo	BR	13	1307	513	191	374	229	pobre
José d'Alencar	BR	15	887	331	154	370	32	velho
Coelho Neto	BR	17	966	369	81	440	76	velho
Humberto de Campos	BR	6	766	169	193	368	36	velho
Júlio Dinis	PT	9	1038	430	127	302	179	pobre
Teófilo Braga	PT	4	419	144	82	112	81	pobre
Alexandre Herculano	PT	8	809	321	201	228	59	velho
Raul Brandão	PT	5	206	73	24	102	7	grande

Table 11: Different depiction classes per authors, ordered by number of characterisations. "nr" shows the number of different fiction works by that author in Literateca, and "mfreq" the most frequent characterising word.

We see there are some differences among these authors. They agree in that none emphasises an explicitly emotional description, and several authors follow the "general" pattern in fiction: first *character*, then *appearance*, then *social*, and last *emotion*: Machado de Assis, Eça de Queirós, Aluísio de Azevedo, José de Alencar, Júlio Dinis, Teófilo Braga and Alexandre Herculano.

But in José Manuel de Macedo, Coelho Neto and Raul Brandão *appearance* is the most frequent characterisation, and *character* is the second most frequent.

As to the relative order of *character* and *social* characterisation, Humberto de Campos is the only one who reverts the "canonical" order, using more *social* characterisations than those reflecting *character*, while Camilo Castelo Branco (incidentally the author with more works in Literateca) is the only that describes more *social* than *appearance*.

In any case, there are also differences in the amount of characterisation provided by each author: Figure 16 illustrates how much each author depicts, i.e. how many characterisations he uses per number of words.

In Figure 17, we represent each author in a plane formed by internal and external characteristics.

4.6 The influence of literary school

For a subset of the works of Literateca we have metadata about the literary school they belong, as has been described in Santos et al. 2020.

We selected all works marked as romantic in one group (11,850,395 words, 175 books), and those marked as realist or naturalistic (7,616,384 words, 121 different books) in another group,¹⁶ to see whether one could identify differences as to people's depictions,

16. Note that the groups are not mutually exclusive: there are a few books classified as both romantic and realist, which correspond to the transition between the two schools.

Relative characterization per author

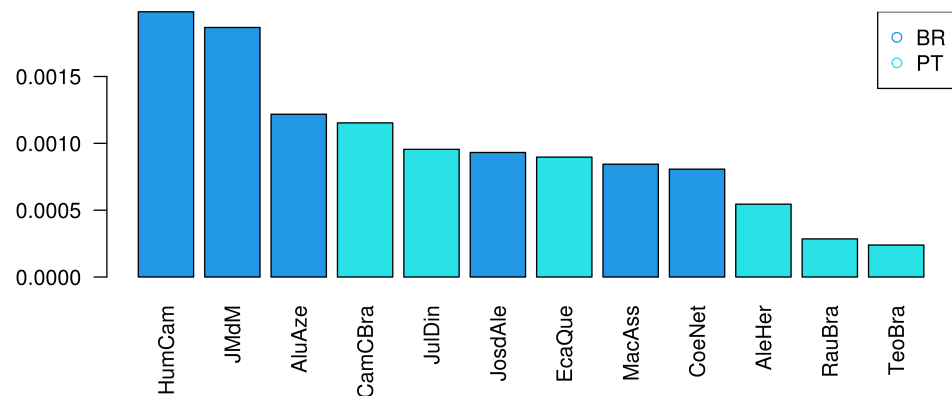


Figure 16: Characterisation by author.

just based on this fourfold sub-classification, and also according to the gender of who gets characterised. The results are presented in Table 12 and in Figure 18.

	Romantic	fem	masc	Realist	fem	masc
People	238,338	74,991	142,245	149,699	52,771	86,861
Characterised	22,733	8140	14041	13834	5187	8244
Social	4629	1510	3002	2516	946	1501
Appearance	5573	2279	3179	3944	1678	2147
Emotion	4370	1932	2350	2635	1112	1464
Character	9389	2899	6237	5649	1819	3650

Table 12: Different depiction classes in novels, novellas and short stories, per literary school and per gender of the characterised.

The first interesting remark is that there are (relatively) more mentions of feminine characters in realist works than in romanticism. However 10.9% of the feminine occurrences are characterized in romantic books (and 9.9% of masculine occurrences), but only 9.8% in realist ones (compared to 9.5% for masculine).

We see that in romanticism there are far more *character* characterisations of masculine characters than in realism, where the relationship across all kinds of characterizations is stable across genres. In addition, realism describes physical appearance of both genders, while romanticism prefers feminine appearance.

4.7 Going back to DIP

DIP has clearly demonstrated that there are fewer feminine characters in lusophone literature.

But in this study we see that those feminine characters are relatively more characterised, at least for appearance, than the masculine ones.

Ideally, and for the near future, we would like to connect the two studies/activities/forms of distantly looking at literature and provide, for each literary work, not only their description in terms of characters (as DIP does), but also how each character is

Authors by relative characterization

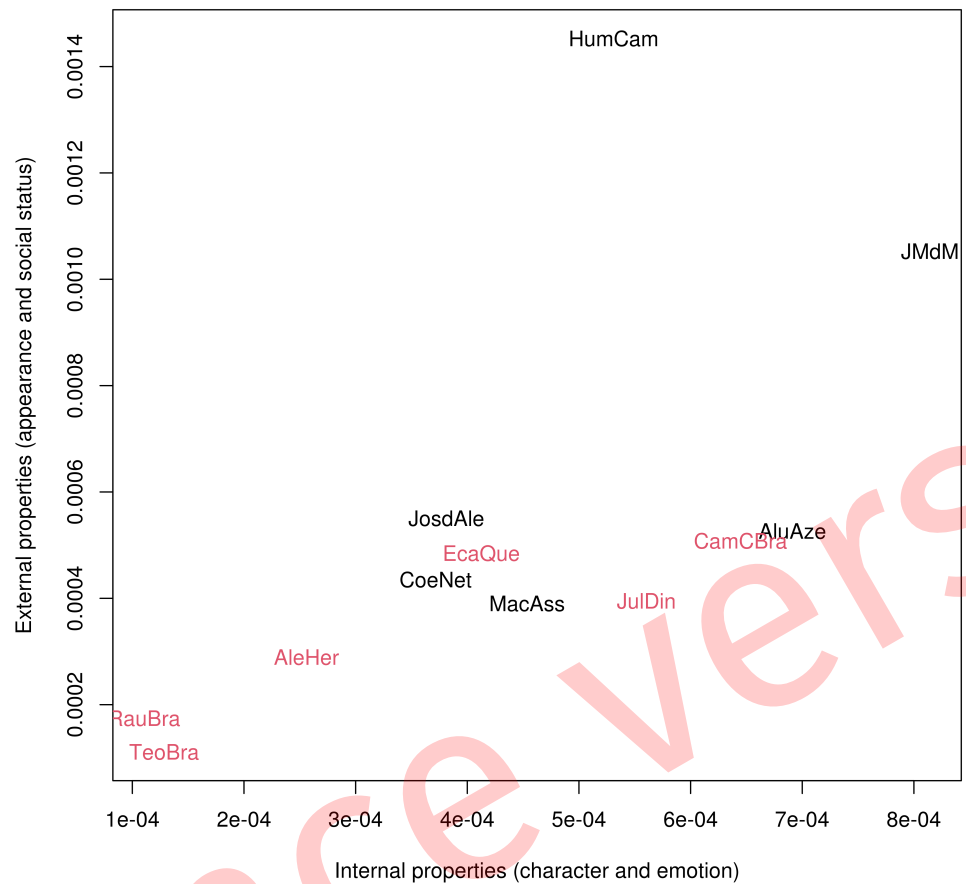


Figure 17: Characterisation by author as far as kind and relative weight of characterisation.

characterised, using the present work and some form of anaphoric resolution of the non-proper name depictions and of those cases where human subjects (whether or not proper names) are omitted (Freitas and Souza 2021 found omitted subjects in 41% of clauses in Brazilian literature material).

We might therefore link kinds of characters with particular clusters of properties, like the beautiful rich woman and the poor honest lad and the evil old priest.

5. Concluding remarks

In this paper, we offered some insights into human depiction based on distant reading literature in Portuguese. We can summarise our results as follows: human depiction seems to obey the pattern *character, social, appearance* and *emotion* for masculine characters, and *character* and *appearance, social* and *emotion* for feminine characters. If we consider only preferred depiction words, differences between feminine and masculine characters become more pronounced, and changing lens – from distant to close reading – reveals that features associated with characters are related to their genders. The results also suggest an impact of the author's gender in the types of characterisation used, but the limited number of works written by women hinders a more definite conclusion.

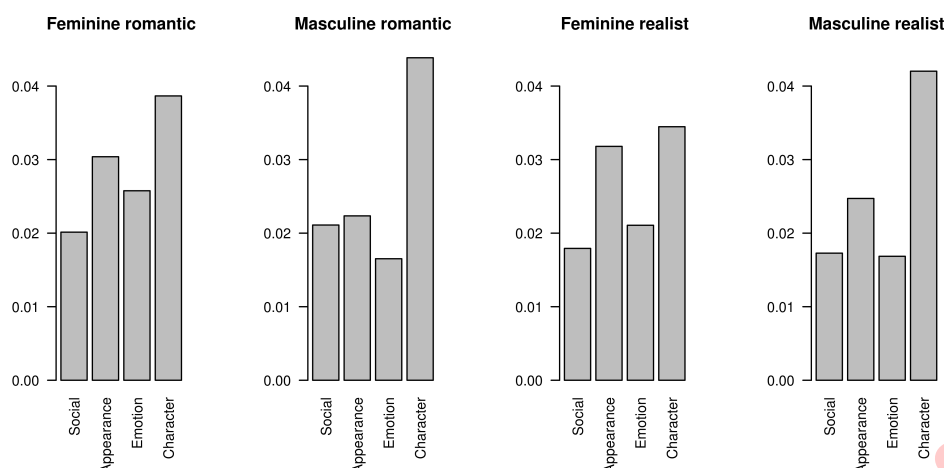


Figure 18: Characterisation per literary school and per gender

We acknowledge that the material we used (works and words) is smaller than those used in other studies conducted under the umbrella of Digital Humanities. However, our findings show that an advantage of annotated data is the opportunity to see trends and patterns even in moderate-sized collections. On the other hand, we stress that another intention with this work is to convince (the Portuguese-speaking community, mainly) to enlarge literary Portuguese-language collections with machine readable texts.

In the near future, we would like to assess the precision of each rule used, and to correct the detected mistakes, as well as to widen the scope of characterisation. We are aware that human depiction is not restricted to the lexical-syntactic patterns we used, and to detect other ways Portuguese language manifests characterisation is, therefore, a natural route to continue the investigation.

We are also aware that our study reflects mainly the vision of male authors of nineteenth and early twentieth century, and that therefore it is by no means an unbiased description of gender.

Other studies that we may still do on this material is to add an evaluation view: of these ways of depicting, which ones are positive, negative, or neutral? This is more straightforward for character and emotional words, but also possible for appearance and even social descriptions.

We could also separate age from appearance, and check what this dimension may bring.

Anyway, all material is open for inspection, from the lists of the characterising words to the patterns used, and the annotated works themselves, which allow interested researchers to redo our searches and even refine them.

6. Data Availability

We make available in Zenodo:

- the list of characterising words, classified in five classes: [10.5281/zenodo.7979566](https://zenodo.org/record/7979566)
- the patterns to find them in the corpus, together with the commands to create the

tables and/or figures used in the paper: [10.5281/zenodo.7979619](https://zenodo.org/record/7979619) 597

7. Software Availability 598

Not relevant 599

8. Acknowledgements 600

Funding, Funding and thanks! 601

9. Author Contributions 602

Cláudia Freitas: Conceptualization, Writing – original draft, review & editing 603

Diana Santos: Conceptualization, Writing – original draft, review & editing 604

References 605


- Argamon, Shlomo, Charles Cooney, Russell Horton, Mark Olsen, Sterling Stuart Stein, and Robert Voyer (2009). "Gender, Race, and Nationality in Black Drama, 1950-2006: Mining Differences in Language Use in Authors and their Characters". In: *Digit. Humanit. Q.* 3.2. <http://www.digitalhumanities.org/dhq/vol/3/2/000043/000043.html>. 606
- Bick, Eckhard (2014). "PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese". In: *Working with Portuguese Corpora*. Ed. by Tony Berber Sardinha and Thelma de Lurdes São Bento Ferreira. Bloomsbury Academic, 279–302. 607
- Cao, Yang Trista and III Daumé Hal (Nov. 2021). "Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*". In: *Computational Linguistics* 47.3, 615–661. ISSN: 0891-2017. [10.1162/coli_a_00413](https://direct.mit.edu/coli/article-pdf/47/3/615/1971880/coli_a_00413.pdf). eprint: https://direct.mit.edu/coli/article-pdf/47/3/615/1971880/coli_a_00413.pdf. https://doi.org/10.1162/coli%5C_a%5C_00413. 608
- Cermáková, Anna and Michaela Mahlberg (2021). "The representation of mothers and the gendered social structure of nineteenth-century children's literature". In: *English Text Construction* 14.2, 119–149. ISSN: 1874-8767. <https://doi.org/10.1075/etc.00044.cer>. <https://www.jbe-platform.com/content/journals/10.1075/etc.00044.cer>. 609
- (2022). "Gendered body language in children's literature over time". In: *Language and Literature* 31.1, 11–40. [10.1177/09639470211072154](https://doi.org/10.1177/09639470211072154). eprint: <https://doi.org/10.1177/09639470211072154>. <https://doi.org/10.1177/09639470211072154>. 610
- Freitas, Cláudia, Flávia Martins, and Liana Biar (Feb. 2022). "Um 'olhar discursivo' sobre predicação e gênero: aproximações metodológicas entre corpus e discurso". In: *Texto Livre* 15, e36213. [10.35699/1983-3652.2022.36213](https://doi.org/10.35699/1983-3652.2022.36213). 611

- Freitas, Cláudia and Elvis Souza (2021). “Sujeito oculto às claras: uma abordagem descritivo-computacional / Omitted subjects revealed: a quantitative-descriptive approach”. In: *Revista de Estudos da Linguagem* 29.2, 1033–1058. ISSN: 2237-2083. 10.17851/2237-2083.29.2.1033-1058. <http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/17439>.
- Hoyle, Alexander Miserlis, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell (July 2019). “Unsupervised Discovery of Gendered Language through Latent-Variable Modeling”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 1706–1716. 10.18653/v1/P19-1167. <https://aclanthology.org/P19-1167>.
- Katsma, Holst (2018). *Loudness in the novel*. workingpaper. <https://litlab.stanford.edu/LiteraryLabPamphlet7.pdf>.
- Larson, Brian (Apr. 2017). “Gender as a Variable in Natural-Language Processing: Ethical Considerations”. In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, 1–11. 10.18653/v1/W17-1601. <https://aclanthology.org/W17-1601>.
- Lucy, Li and David Bamman (June 2021). “Gender and Representation Bias in GPT-3 Generated Stories”. In: *Proceedings of the Third Workshop on Narrative Understanding*. Virtual: Association for Computational Linguistics, 48–55. 10.18653/v1/2021.nuse-1.5. <https://aclanthology.org/2021.nuse-1.5>.
- Mandell, Laura (2019). “Gender and Cultural Analytics: Finding of Making Stereotypes?” In: *Debates in the Digital Humanities*. Ed. by Matthew K. Gold and Lauren F. Klein. Manifold Scholarship.
- Moretti, Franco (2000). “The slaughterhouse of literature”. In: *Modern Language Quarterly* 61.1.
- (2013). *Distant Reading*. Verso.
- Moretti, Franco and Oleg Sobchuk (2019). “Hidden in Plain Sight: Data Visualization in the Humanities”. In: *New Left Review* 118, 86–115.
- Rocha, Luísa, Cláudia Freitas, and Diana Santos (Oct. 2019). “Preparação para Leitura Distante em português: diálogos entre PLN e Humanidades Digitais”. In: *Anais do TILic 2019*. <https://www.linguateca.pt/Diana/download/Rochaetal2019.pdf>.
- Santos, Diana (2014). “Corpora at Linguateca: Vision and roads taken”. In: *Working with Portuguese Corpora*. Ed. by Tony Berber Sardinha and Thelma de Lurdes São Bento Ferreira. Bloomsbury Academic, 219–236.
- Santos, Diana and Cláudia Freitas (Oct. 2019). “Estudando personagens na literatura lusófona”. In: *STIL 2019 - Symposium in Information and Human Language Technology and Collocates Events, October 15-18, 2019, Salvador, BA, Proceedings of conference*, 48–52.
- Santos, Diana, Cláudia Freitas, and Eckhard Bick (Sept. 2018). “OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in public domain”. In: *CorLex*. <http://www.linguateca.pt/Diana/download/CorLex.pdf>.
- Santos, Diana, Cristina Mota, Emanuel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão, and Roberto Willrich (2022a). *Introduction to DIP: goal, setup, resources and results*. Encontro do DIP. https://www.linguateca.pt/aval_conjunta/dip/apr_encontro/DIPpresentation.pdf.
- (2023). “DIP - Desafio de Identificação de Personagens: objectivo, organização, recursos e resultados”. In: *Linguamática*. to appear.

- Santos, Diana, Emanuel Pires, Cláudia Freitas, Rebeca Schumacher Fuão, and João Marques Lopes (June 2020). "Periodização automática: Estudos linguístico-estatísticos de literatura lusófona". In: *Linguamática* 12.1, 81–95. 677 678 679
- Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanuel Pires, Rebeca Schumacher, and Paulo Silva Pereira (2022b). "Identifying literary characters in Portuguese: Challenges of an international shared task". In: *Computational processing of the Portuguese language, 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022 Proceedings*. Ed. by Vlória Pinheiro, Pablo Gamallo, Raquel Amaro, Caroline Scarton, Fernando Batista, Diego Silva, Catarina Magro, and Hugo Pinto. Springer, 413–419. 680 681 682 683 684 685 686
- Schöch, Christof, Tomaz Erjavec, Roxana Patras, and Diana Santos (2021). "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives". In: *Modern Languages Open* 1, 1–19. <https://doi.org/10.3828/mlo.v0i0.364>. 687 688 689
- Schöch, Christof, Evgeniia Fileva, and Julia Dudar (Feb. 2022). *CLS INFRA D3.1 Baseline Methodological User Needs Analysis*. 10.5281/zenodo.6389333. <https://doi.org/10.5281/zenodo.6389333>. 690 691 692
- Schulz, Daniel and Štěpán Bahník (2019). "Gender associations in the twentieth-century English-language literature". In: *Journal of Research in Personality* 81, 88–97. ISSN: 0092-6566. <https://doi.org/10.1016/j.jrp.2019.05.010>. <https://www.sciencedirect.com/science/article/pii/S0092656619300510>. 693 694 695 696
- Silva, Flávia Martins da Rosa Pereira da (2021). *Diferenciações de gênero na caracterização de personagens: uma proposta metodológica e primeiros resultados*. 697 698
- Smeets, Roel (2021). *Character Constellations: Representations of Social Groups in Present-Day Dutch Literary Fiction*. Leuven University Press. ISBN: 9789462702950. <http://www.jstor.org/stable/j.ctv21wj5cb> (visited on 12/17/2022). 699 700 701
- Underwood, Ted (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press. 702 703
- Underwood, Ted, David Bamman, and Sabrina Lee (2018). "The transformation of gender in English-language fiction". In: *Journal of Cultural Analytics* 3.2, 11035. 704 705
- Weingart, Scott and Jeana Jorgensen (May 2013). "Computational analysis of the body in European fairy tales". In: *Literary and Linguistic Computing* 28.3, 404–416. ISSN: 0268-1145. 10.1093/llc/fqs015. eprint: <https://academic.oup.com/dsh/article-pdf/28/3/404/2784264/fqs015.pdf>. <https://doi.org/10.1093/llc/fqs015>. 706 707 708 709

A Novel Approach for Identification and Linking of Short Quotations in Scholarly Texts and Literary Works

Frederik Arnold¹ 
Robert Jäschke¹ 

1. Berlin School for Library and Information Science, Humboldt-Universität zu Berlin , Berlin, Germany.

Citation

Frederik Arnold and Robert Jäschke (2023). "A Novel Approach for Identification and Linking of Short Quotations in Scholarly Texts and Literary Works". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-12

Date received 2023-01-31

Keywords

quotation linking, literary works, scholarly works, machine learning, language models

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. We present two approaches for the identification and linking of short quotations between scholarly works and literary works: *ProQuo*, a specialized pipeline, and *ProQuoLM*, a more general language model based approach. Our evaluation shows that both approaches outperform a strong baseline and the overall performance is on the same level. We compare the performance of ProQuoLM on texts with and without (page) reference information and find that reference information is not used. Based on our findings, we propose the following steps for future improvements: further analysis of the influence of a bigger context window for better handling of long distance references and the introduction of positional information of the literary work so that reference information can be (better) utilized.

1. Introduction

Scholarly and literary texts do not exist in a vacuum but rather interact in various ways: literary scholars quote literary works, scholarly works and other sources, to support the reasoning of their interpretations or to build on earlier publications. Although (literary) interpretations usually only concretely refer to certain passages of a (literary) text, they often claim to interpret the entire work. We know little about the inner workings of this skillful selection, the corresponding attentional behavior, as well as the canonization of passages, that, for various reasons, lend themselves to support the interpretation. The intertextual relationships between interpretations and objects of interpretation vary in nature, ranging from relatively vague references to renderings of clearly identifiable passages of text in the interpreter's own words to direct quotations.

Long quotations, that is, quotations of a length of five words or more, can be identified using text reuse detection methods (cf. Arnold and Jäschke 2021). Shorter quotations are a major challenge for reasons we will explain in a moment. They are important, however, either because they apply to particularly weighty words or because they are indicative of references to passages. Other uses include intertextuality research, for example, in the analysis of quotations from Hamlet (Hohl Trillini and Quassdorf 2010) or Shakespeare in general (Molz 2020), argument mining in scholarly texts where the context in the literary work is relevant to understand how texts are analyzed (cf. Descher and Petraschka 2018, Winko and Jannidis 2015), or the identification of key passages, that

Literary work	Scholarly work
<p>Wo ist die Hand so zart, daß ohne Irren Sie sondern mag beschränkten Hirnes Wirren, So fest, daß ohne Zittern sie den Stein Mag schleudern auf ein arm verkümmert Sein? Wer wagt es, eitlen Blutes Drang zu messen, Zu wägen jedes Wort, das unvergessen In junge Brust die zähen Wurzeln trieb, Des Vorurteils geheimen Seelendieb? Du Glücklicher, geboren und gehegt Im lichten Raum, von frommer Hand gepflegt, Leg hin die Waagschal, nimmer dir erlaubt! Laß ruhn den Stein – er trifft dein eignes Haupt!</p>	<p>Hier laufen wir Gefahr, uns zu jenem selbstbezogenen Messen und Wägen verführen zu lassen, das uns "Glücklichen", die "geboren und gehegt/Im lichten Raum, von frommer Hand gepflegt" (882) wurden, nicht erlaubt ist.</p>

Figure 1: Example shows an excerpt of a scholarly work (Schaum 2004) which quotes from a literary work (excerpt from Droste-Hülshoff 1979). A single word quotation is shown in green, a long quote in dark blue and a (page) reference in light blue.

is, passages that are particularly important to expert readers (cf. Arnold and Fiechter 2022).

As already mentioned, quotations can be of varying length from single words to whole paragraphs. Bibliographic references, often in footnotes or a dedicated reference section, identify the work a quotation is taken from. Page references, either in footnotes or in parentheses in the running text, are often used to indicate specific pages.¹ Despite this information, identifying the exact source location of a quotation is a hard task. Existing tools for the identification of quotations, for example, Copyfind (Bloomfield 2016), Passim (Smith et al. 2014), TextMatcher (Reeve 2020) or Quid (Arnold and Jäschke 2021), are not suitable for unambiguously identifying instances which are shorter than at least a couple of words, as they often rely on text reuse detection methods.

For these shorter quotations, especially for quotations consisting of just one word, a number of challenges need to be solved. Firstly, short quotations are much more likely to have multiple possible sources in the literary work, which makes it more difficult to link a quotation to its source. Secondly, quotation from other sources, for example, other scholarly works or quotations from the Bible, are much more likely to also occur in the literary work just by chance.

In this paper, we present and compare two tools for the identification and linking of short quotations between scholarly works and literary works: *ProQuo* and *ProQuoLM*. Quotations, long and short, are often accompanied by citation information, for example, page or line numbers, either in the running text in parentheses or in footnotes (cf. Figure 1). Our main idea behind *ProQuo* is to use the references corresponding to long quotations as examples to distinguish references corresponding to short quotations from other text in parentheses and other references, for example, Bible references or references to other literary works. We then extract relations between short quotations and references and use that information and the position of long quotations as anchors to link short quotations to the literary work.

1. In this work, whenever we talk about *references*, we refer to the second type of reference, the one used to indicate specific pages.

We compare this specialized pipeline with its explainable steps to a more general, state-of-the-art neural language model approach which we named ProQuoLM. For this second approach, we first extract candidates for short quotations and then use a fine-tuned language model to filter the candidates. The comparison allows us to investigate and illustrate the advantages and disadvantages of a pipeline with explainable steps and a blackbox neural language model approach. This is especially relevant in light of recent discussions about computational approaches in digital humanities (cf. Da 2019).

This paper is organized as follows: In Section 2, we provide an overview on related work. In Section 3 we describe our approaches. In Section 4 we present our dataset and experimental setup, followed by Section 5 where we present the results.

2. Related Work

Our task is related to reference extraction and segmentation, quotation detection and quotation attribution. Existing tools for reference extraction and segmentation (*GROBID* 2008–2022; Prasad et al. 2018) focus on STEM fields (science, technology, engineering, medicine) where references appear in a dedicated reference section and are referenced in the running text in some form, for example, author-year mentions. The focus is on the identification of these reference sections, linking author-year mentions in the running text to entries in the reference section and the segmentation of references into individual fields, for example, author, title, year etc.

The next related task, quotation detection, aims to identify reported speech, thought and writing in text (Papay and Padó 2019; Pareti et al. 2013; Scheible et al. 2016). This task is normally constrained to individual texts and a focus on speech. For our task on the other hand, we are interested in the detection of quotations as a type of scholarly citation.

Quotation attribution is the task of identifying the source of a quotation (Almeida et al. 2014; Elson and McKeown 2010). Existing approaches are often focused on speaker attribution in fiction or news paper articles. For the task at hand our goal is different. We want to distinguish between quotations from a given primary literary work and other sources and identify a specific occurrence in the case of multiple occurrences.

We aim to combine aspects of these three tasks into the new task of identifying quotations in one text and linking those quotations to their source in another text by using page references.

Arnold and Jäschke 2021 presented Quid: a tool for the identification of text re-use with a focus on quotations with a length of at least five words between literary and scholarly works. Five words is not a hard limit but they determined that shorter quotations generate too many ambiguous matches without more advanced methods. Quid outperformed other approaches which led us to the decision to use it in this work. We will also use Quid for the extraction of candidates for quotations shorter than five words which we then filter further.

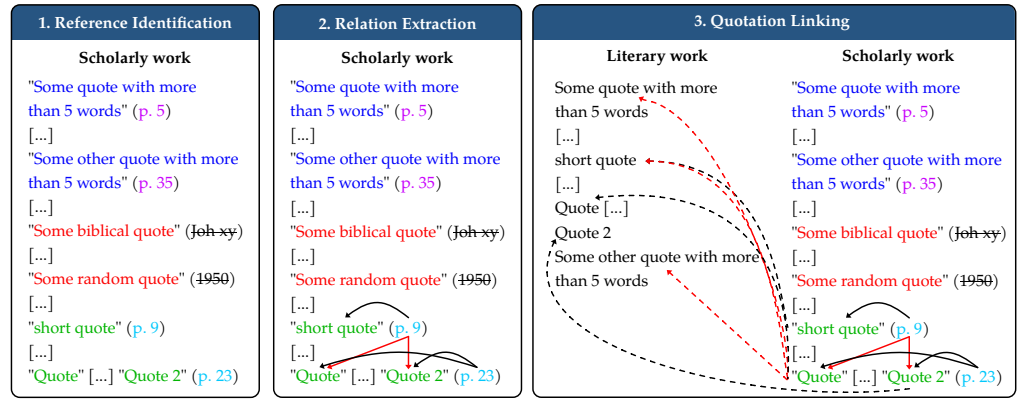


Figure 2: Visualization of quotation identification and linking in three steps.

3. Methods

In this section, we first define the task and then present two approaches to solve it. The first approach is a specialized pipeline and the second approach is a more general approach based on a neural network language model.

3.1 Task

Our overall goal is to identify short quotations in the scholarly work and link the quotations to their source text in the literary work. For this task, we make the following assumptions. Firstly, we work with a corpus of scholarly works for which we know that their main focus is on the primary literary work which we are interested in. Secondly, we assume that all quotations appear in quotation marks and that the texts do not contain errors, for instance, due to OCR. Handling texts with such issues is out of scope of this work and there are other efforts solving this task (Brunner et al. 2020).

We focus on scholarly works with references in parentheses in the running text. This decision was made based on the number of scholarly works in our corpus with references in the running text (cf. Section 4.1) and due to the high variance in structure of references in footnotes.

3.2 ProQuo

Figure 2 shows the building blocks of our first approach which we named *ProQuo*. This approach is divided into three steps: *Reference Identification*, *Relation Extraction*, and *Quotation Linking*. In the first step, we use long quotations (dark blue), extracted using Quid, and their references (pink) as anchors. We use these known references as examples to identify other references to the literary work (light blue) and distinguish them from other text in parentheses (strikethrough). In the second step, we then link the identified references to their corresponding short quotations (green). In the final step, the identified short quotations are linked to their source in the literary work (black dashed arrows).

3.2.1 Step 1: Reference Identification

113

The goal of this step is to distinguish between true references to the literary work (Figure 2, light blue) and other text in parentheses. References are written in a number of ways, as Table 1 shows.² Often references only contain a page number (Ex. 1, 3) but they can also contain line numbers (Ex. 2) or information on the cited edition (Ex. 4). In this work, we are only interested in page numbers and ignore the other information. To extract the page number from a reference string, we perform the following searches until we get a match:

- A number which immediately follows the string “S. ”
- A number which is not preceded by a letter.

A scholarly work can use any of the variants from Table 1 to reference the literary work and at the same time use some other variant to point to other (literary) works or even use a similar looking variant to reference the same source in case of citations from collected works. At the same time, we need to distinguish true references from other text which appears in parentheses. This includes dates (Ex. 8), other citations, for example, Bible citations (Ex. 9), and text in general.

No.	Example	Citation Target
1	(12)	Literary work
2	(12, 12-14)	
3	(S. 12)	
4	(HKA V,1. S.12)	
5	(I, S.12)	
6	(Jb, 12)	
7	(SW9 II, 12)	
8	(1987)	Other
9	(Johannes 8, 11)	
10	(other text)	

Table 1: Examples for references.

To overcome these challenges, we use the following approach. We first identify the best example for a reference to the literary work in the scholarly work. We use quotations longer than five words (Figure 2, dark blue) to extract up to n_{ref} examples of the type of reference (Figure 2, pink) for a specific scholarly work. The examples are extracted starting with the longest quotation with a maximum distance of d_{ref} characters between reference and quotation and a maximum reference length of l_{ref} . If less than three examples could be found, we use the one from the longest quotation. Otherwise, all examples are clustered with spectral clustering into two clusters. We use the probability that two references are similar (the model to determine similarity is described below) as the similarity in the affinity matrix for the clustering. From the bigger cluster, we then select the reference example which belongs to the longest quotation. This clustering procedure is necessary to reduce the probability of selecting an incorrect reference example, which could happen in cases where Quid made a mistake or when the long quotation is not followed by a reference but some other text in parentheses.

². Examples taken from real texts are shown in the original language. Translations: “S.” → “p.”, “Johannes” → “John”. Other translations are given in the text in brackets.

To classify whether two references are similar, we trained a siamese network (Bromley et al. 1993) for binary classification. The network is made up of two sub-networks, each a character-level BiLSTM (Hochreiter and Schmidhuber 1997) on top of an embedding layer. The outputs of the sub-networks are compared using Manhattan distance. Two references are classified as similar if the probability given by the model is over a threshold t_{ref} . Using this model, all text in parentheses is compared against the selected example to distinguish between true references and other text occurring in parentheses.

3.2.2 Step 2: Relation Extraction

The goal of this step is to identify relations between quotations, that is, text in quotation marks, and the references identified in the previous step. First, we extract all quotations and create all possible combinations of quotations and references where the quotation and reference are within a distance of d_{rel} tokens. We determine tokens by white space tokenization. We surround the quotation which we are interested in with a start and end tag, replace the reference text with a special tag and also replace all other references with another special tag. Then, we use a machine learning model to classify each pair as belonging together (Figure 2, solid black arrows) or not (for example, *Quote/Quote 2* and *p. 9* (solid red arrows)). We classify a quotation and reference as belonging together if the probability given by the model is over the threshold t_{rel} . For quotations with multiple reference candidates, we take the relation with the highest probability.

For the classification we compare two machine learning models: a token-level BiLSTM with a classification layer with sigmoid activation and a fine-tuned German uncased BERT model³ (Devlin et al. 2019) with a linear layer on top of the pooled output.

3.2.3 Step 3: Quotation Linking

The goal of this step is to link quotations from the scholarly work to their source in the literary work (Figure 2, dashed black arrows) and exclude other possible candidates (dashed red arrows). The main idea is to use long quotations with known links and references as anchors. We then link short quotations relative to these known positions.

Scholarly works cite different editions of the literary work. Since automatic identification of the cited edition is out of the scope of this work, we decided to map all citations to one edition. To achieve this, we estimate a *virtual page* size by using the references from the long quotations:

$$page_size = \frac{last_quote_end - first_quote_start}{last_page - first_page} \quad (1)$$

Here *page_size* is the estimated page length (the number of characters) of the literary work, *first_quote_start* and *last_quote_end* are character positions of the first and last quotation in the scholarly work, respectively, and *first_page* and *last_page* are the corresponding page numbers in the literary work, respectively.

Using this virtual page size we can approximate the character position of short quotations in the literary work. It should be noted that short quotations can appear without a reference. We distinguish between short quotations with and without a reference and

3. <https://huggingface.co/dbmdz/bert-base-german-uncased>

the approach differs:

$$page_diff = quote_page - first_page \quad (2)$$

$$quote_pos = first_quote_start + (page_diff \times page_size) \quad (3)$$

For *quotations with a reference*, we approximate the character position of the quotation in the literary work by using Equations 2 and 3, where *quote_page* is the page number of the quotation we want to link and *page_diff* is the distance in number of pages between the quotation and the first known page number.

For *quotations without a reference*, we first try to find the closest quotation in the scholarly work from the already linked quotations within a certain distance d_{link} . We use the midpoint of that quotation as the approximate position.

If an approximate position could be determined, we use this position to define a search range r_{link} . For single word quotations, we then first perform exact string matching and if that does not lead to any matches, we perform fuzzy matching. In case of multiple matches, we take the match closest to the approximated quote position.

For longer quotations, we first try to find an exact match in the determined range. If that leads to exactly one match, that match is used. If there are no matches, the whole text is searched. If that does not lead to a single exact match, we use the matches from Quid as candidates. If there is a single candidate in the given range with an overlap of at least o_{link} %, that candidate is used. If there are no matches, the whole text is searched for a single unambiguous result.

If no approximate position could be determined, the whole text is searched for a single exact match and if there are no matches, we perform fuzzy matching and only use a single unambiguous result.

In our corpus, 11 scholarly works cite an edition of *Michael Kohlhaas* in parallel print. These texts were manually identified and this information is passed to the algorithm to adjust the calculations to only count every other page.⁴

3.3 ProQuoLM

Having seen and appreciated the complexity of the aforementioned bespoke approach, we want to analyze, how state-of-the-art neural language models can solve the task when it is formulated in a very simple way such that they can be fine-tuned and applied.

For the second approach, we first extract all text in quotation marks and for each quotation we determine all candidates in the literary work. For determining the candidates we use the same (fuzzy) matching approach as in Section 3.2.3. We then fine-tune the same German uncased BERT model as before for binary classification between a quotation and a candidate, both with a context window. Both text fragments, that is, quotation with context and candidate with context, have a maximum length of l_{lm} tokens each. We also surround the quotation which we are interested in with a start and end tag in both fragments. From all candidates, we select the one with the highest probability over a threshold t_{lm} .

4. This is just a very rough approximation. The topic of parallel editions is much more complex and beyond the scope of this work.

4. Experiments

In this section, we first give an overview of the dataset and our annotations. We then present the experiments to evaluate both approaches on texts with references in the running text. Finally, we evaluate ProQuoLM on texts with all reference information removed.

4.1 Dataset and Annotation

We assess our methods by analyzing two literary texts, *Die Judenbuche* by Annette von Droste-Hülshoff (1979) and *Michael Kohlhaas* by Heinrich von Kleist (1978). For each text our corpus contains 44 and 49 interpretive scholarly articles, respectively, which were previously annotated in the ArguLIT project (Winko 2017–2020) using TEI/XML (TEI Consortium, eds. 2022).⁵ The annotations include quotations of different types, such as those from the primary literary work, other literary works or scholarly works. The original annotations were limited to clearly marked quotations, that is, with quotation marks. In this evaluation, we only focus on quotations coming from the primary literary work. The 93 scholarly works use references either in parentheses in the running text or in footnotes. For this work, we focus on scholarly works with references in the running text and ignore footnotes in all experiments, this includes quotations in footnotes. This decision was made mainly due to the varying structure of quotations in footnotes and to keep the focus on quotations with references in the running text. For *Die Judenbuche* 24 scholarly works and for *Michael Kohlhaas* 33 scholarly works have references in the running text.

We extended the original annotations of these scholarly works in two annotation tasks. In the *reference annotation task*, three persons annotated reference strings (cf. Table 1, p. 5) and links between reference strings and quotations. Five of the texts were annotated by all three annotators with F_1 -score inter-annotator agreements between pairs of annotators of 0.88, 0.93, and 0.90. Table 2 shows statistics for the number of (short) quotations from the primary literary work with and without references. We also show the number of quotations in footnotes which only account for around 10 % of short quotations.

Literary work	Die Judenbuche	Michael Kohlhaas
All quotations (primary work)	1 736	1 788
Quotations with a reference	1 467	1 547
Short quotations	817	862
Short quotations with a reference	672	736
Quotations in footnotes	94	80

Table 2: Statistics for *Die Judenbuche* and *Michael Kohlhaas*.

In a *linking annotation task*, two persons annotated the origin of quotations from scholarly texts in the literary text. In this task, not only the literary works with references in the running text were annotated but for *Die Judenbuche*, all 44 scholarly works were annotated. The additional annotated texts contain 270 short quotations. For consistency, we also ignore footnotes in the additionally annotated texts where references appear in

5. For the sake of brevity, we will reference *Die Judenbuche* and *Michael Kohlhaas* with J and K, respectively.

footnotes, effectively resulting in texts without any reference information. This data is used to evaluate the performance of ProQuoLM, which does not rely on explicit reference information, on texts without references. To evaluate inter-annotator agreement, again, the same five texts as before were annotated by both annotators which resulted in an F_1 -score inter-annotator agreement of 0.90.

4.2 References in Running Text

For the experiments in this section, we perform 5-fold cross validation. We calculate precision and recall following Arnold and Jäschke (2021). We optimized the hyperparameters once on the validation data from the first split of our cross validation and use the hyperparameters for all evaluations.

4.2.1 Reference Identification

To evaluate the performance of our model, we compare it against a baseline that classifies texts in parentheses as a reference if there is at least one number contained and the text is not longer than the maximum reference length l_{ref} , which we set to 25 characters. This value was chosen as it is in the 99 percentile of lengths in our corpus

The output dimension of the embedding layer and the BiLSTM hidden state are both 32. A dropout of 0.2 is applied. The batch size was set to 512 and the network was trained for 10 epochs with binary crossentropy loss and Adam optimizer with a learning rate of 0.001. The number of examples n_{ref} is set to 5. This worked well in our tests and leaves some room for incorrect examples. For the maximum distance d_{ref} , we determined 20 characters to work well. The inputs are padded/truncated to the maximum reference length. The classification threshold t_{ref} is set to 0.7.

4.2.2 Relation Extraction

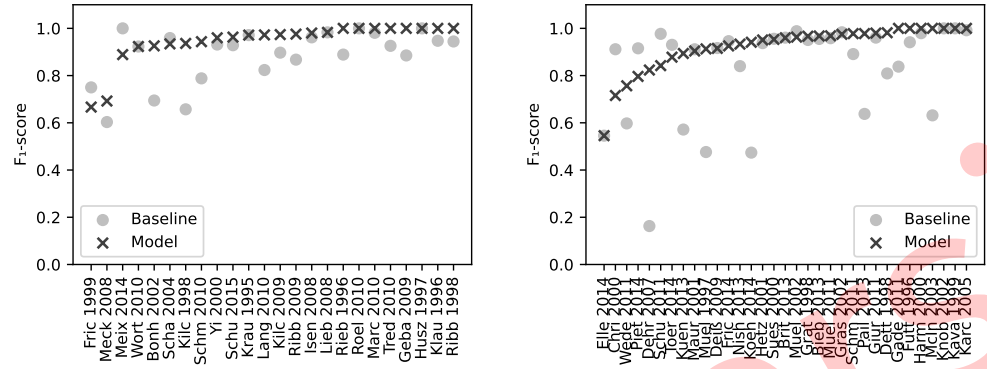
To evaluate the performance of our two models, we compare them against three baselines. The first baseline (*Ref After*) always takes the closest reference after the quotation. The second baseline (*Ref Before*) works the same way but takes the closest references before the quotation and the last baseline (*Ref Closest*) takes the closest reference before or after the quotation.

For the BiLSTM model, the output dimension of the embedding layer is 64, the hidden state is 64, and a dropout of 0.3 is applied. The batch size was set to 128 and the network was trained for 5 epochs with binary crossentropy loss and Adam optimizer with a learning rate of 0.01. We use WordPiece embeddings (Wu et al. 2016) with a 8000 token vocabulary. The classification threshold t_{rel} is set to 0.4. The BERT model was fine-tuned for 3 epochs with a batch size of 12 and a learning rate of 10^{-5} . The classification threshold t_{rel} is set to 0.5. The maximum distance d_{rel} between a quotation and reference to still be considered is 100 tokens which is in the 93 percentile of distances in our corpus. We tried to increase the maximum distance but got overall worse results as false positives increased. The input is padded/truncated to a length of 200 tokens.

4.2.3 Quotation Linking	291
To evaluate the performance of our algorithm, we compare it against a baseline which always links a quotation to the first matching instance.	292
We determined a search range r_{link} of one page before and after the approximate position to work best. For quotations without a reference, the maximum distance d_{link} is 500 tokens. The minimum candidate overlap o_{link} is 70 %.	293
4.2.4 The Complete Pipeline and Language Model Approach	297
In this experiment, we perform two evaluations of our two approaches and compare the results against the same baseline as for the quotation linking task. We first perform the same 5-fold cross validation as before and then a second evaluation where we split the scholarly works by the literary work they interpret and train on the texts from one literary work and evaluate on the other. This is relevant as it indicates how well the approaches can generalize and perform on a completely new literary work.	298
For ProQuoLM, the model was fine-tuned for 3 epochs with a batch size of 4 and a learning rate of 10^{-5} . The classification threshold t_{lm} is set to 0.5 and the maximum length l_{lm} to 200 tokens.	299
4.3 References in Footnotes	307
Our second approach ProQuoLM does not rely on explicit reference information for quotations. With this experiment, we investigate whether reference information is needed at all or if our second approach can also handle texts with references in footnotes. We do this by evaluating how well ProQuoLM performs on texts where all reference information is removed including footnotes.	300
5. Results	313
We first present the results for the experiments of the individual steps of ProQuo (Sections 5.1 to 5.3), followed by the results of the complete pipeline ProQuo compared to ProQuoLM (Section 5.4). Finally, we present how ProQuoLM performs on texts without any reference information (Section 5.5).	314
5.1 Reference Extraction	318
Table 3 shows the results for our baseline and model for reference extraction. Our model outperforms a strong baseline for both literary works. The baseline only misses cases where the reference is not in parentheses or does not contain a number, for example, <i>ibd.</i> [ibid.] False positives include dates, Bible quotations, or quotations from other scholarly texts. Our model misses less <i>ibd.</i> references but all cases not in parentheses and some other special cases. This includes instances where the reference style differs from all other references, for example, references to a specific verse (V. 8) and not a page. Other false negatives include references that consist of two references (S. 47 and S. 50) and references which differ from the rest as they are followed by additional information (Jb, 35, Herv. durch Autor [author's emphasis]). False positives include instances where	319

Approach	Die Judenbuche			Michael Kohlhaas		
	Precision	Recall	F ₁	Precision	Recall	F ₁
Baseline	0.86	0.95	0.90	0.80	0.95	0.87
Model	0.95	0.96	0.95	0.97	0.90	0.93

Table 3: Precision, recall, and F₁-score for *Die Judenbuche* and *Michael Kohlhaas* for reference classification.



(a) *Die Judenbuche*

(b) *Michael Kohlhaas*

Figure 3: F₁-score comparison for reference extraction.

numbers appear in parentheses with the same style as true references but are used to structure the text (e. g., in enumerations) or reference other scholarly works.

Figures 3a and 3b enable a more fine grained analysis.⁶ For *Die Judenbuche*, we can see that our model outperforms or is on par with the baseline for all texts except three. We get similar results for *Michael Kohlhaas*, except that for seven texts the baseline performs better than the model. The results illustrate the importance of our model. Texts for which the baseline struggles often have a high number of quotations with references from sources other than the primary literary work.

5.2 Relation Extraction

Table 4 shows the results of our two models and three baselines. *Ref Closest* is the best performing baseline with an F₁-score of 0.65 (J) and 0.75 (K). *Ref Closest* has the highest recall but lacks precision. This is to be expected as the baseline does not distinguish between quotations from the primary literary work and quotations from other sources. The poor performance of *Ref Before* confirms that references typically follow a citation.

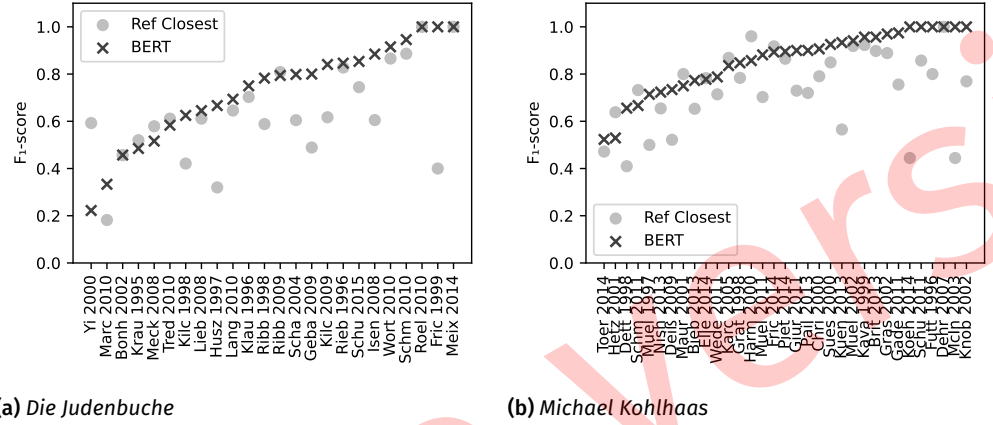
The LSTM-based model outperforms all three baselines. The BERT model performs best overall but worse for *Die Judenbuche* than *Michael Kohlhaas*.

For *Die Judenbuche*, there are 213 false negatives; 98 of those are the result of long distances, that is, the distance between quotation and references is larger than 100 tokens. Another 67 are instances where the reference appears before the quotation. We get a similar result for *Michael Kohlhaas* with 178 false negatives, 69 long distance and 76 reference before quotation. The instances where the reference appears before the

6. The horizontal axes are labeled with the first (up to four) letters of the first author’s name followed by the year of publication. The labels can be used to identify the texts on: [Anonymized for review].

Approach	Die Judenbuche			Michael Kohlhaas		
	Precision	Recall	F_1	Precision	Recall	F_1
Ref After	0.59	0.63	0.61	0.72	0.78	0.75
Ref Before	0.25	0.21	0.23	0.14	0.12	0.13
Ref Closest	0.57	0.76	0.65	0.66	0.86	0.75
LSTM	0.83	0.59	0.69	0.85	0.69	0.76
BERT	0.83	0.68	0.74	0.93	0.81	0.86

Table 4: Precision, recall, and F_1 -score for *Die Judenbuche* and *Michael Kohlhaas* for relation extraction.



(a) *Die Judenbuche* (b) *Michael Kohlhaas*

Figure 4: F_1 -score comparison for relation extraction.

quotation are problematic due to the fact that a reference before a quotation is a lot less likely and our training data is limited in that regard.

Figures 4a and 4b show a comparison of the best baseline and the BERT model. These results illustrate the importance of the model for the difficult texts where the difference in performance between the baseline and model is largest. But they also show that the model struggles with some texts. In the case of *Yi 2000*, for example, all false positives are instances where the reference appears before the quotation.

5.3 Quotation Linking

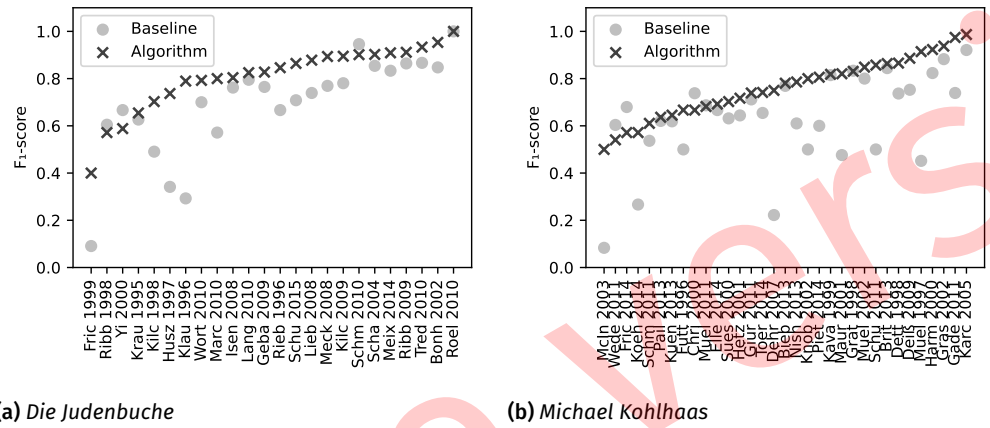
Table 5 shows the results for the quotation linking step. We compare our algorithm against one baseline.

The algorithm outperforms the baseline for both literary works and achieves a high precision. The baseline struggles with texts with a low percentage of quotations from the primary literary work which still appear in the literary work.

Approach	Die Judenbuche			Michael Kohlhaas		
	Precision	Recall	F_1	Precision	Recall	F_1
Baseline	0.65	0.78	0.71	0.59	0.75	0.66
Algorithm	0.85	0.78	0.81	0.87	0.69	0.77

Table 5: Precision, recall, and F_1 -score for *Die Judenbuche* and *Michael Kohlhaas* for quotation linking.

Our algorithm generates 154 false negatives for *Die Judenbuche*. 102 of those are single word quotations and 107 have a reference in our annotations. For *Michael Kohlhaas*, we get 271 false negatives of which 180 are single word quotations and 215 have a reference. These results indicate that for further improvements better handling of single word quotations is necessary. The results for *Die Judenbuche* would also indicate that an improvement in the relation extraction step should improve the overall results of the pipeline. At first glance, the overall worse results for *Michael Kohlhaas* in combination with the better results in the relation extraction step do not support this theory. But *Michael Kohlhaas* is roughly twice as long as *Die Judenbuche* which makes the linking step considerably harder which could counteract the better relation extraction performance.



(a) *Die Judenbuche* (b) *Michael Kohlhaas*

Figure 5: F_1 -score comparison for quotation linking.

5.4 The Complete Pipeline and Language Model Approach

The results in Table 6 demonstrate that both approaches – ProQuo and ProQuoLM – perform on the same level. Compared to the baseline, the pipeline is a big improvement in precision, but recall is lower than the baseline for both literary works. Overall, ProQuoLM works best, with improvements in recall over ProQuo. ProQuoLM produces 166 false negatives for *Die Judenbuche*, 104 are single word quotations and 130 have a reference in our annotations. Similarly for *Michael Kohlhaas*, the results contain 267 false negatives, 177 are single word quotations and 222 have a reference.

Approach	Die Judenbuche			Michael Kohlhaas		
	Precision	Recall	F_1	Precision	Recall	F_1
Baseline	0.65	0.78	0.71 [0.60, 0.78]	0.59	0.75	0.66 [0.57, 0.70]
ProQuo	0.87	0.72	0.79 [0.73, 0.83]	0.87	0.66	0.75 [0.70, 0.79]
ProQuoLM	0.88	0.75	0.81 [0.74, 0.86]	0.87	0.69	0.77 [0.70, 0.81]
Split by literary work						
ProQuo	0.87	0.72	0.79 [0.72, 0.83]	0.86	0.63	0.73 [0.67, 0.77]
ProQuoLM	0.82	0.74	0.78 [0.70, 0.82]	0.76	0.70	0.73 [0.66, 0.77]

Table 6: Precision, recall, and F_1 -score for *Die Judenbuche* and *Michael Kohlhaas* for the full pipeline. For each F_1 -score, the upper and lower bound of the 95 % confidence interval is reported.

The second evaluation shows the performance of ProQuo and ProQuoLM for training

and evaluation split by literary work. This is relevant as it indicates what the performance will be on a completely new literary work. We can see that the difference in performance is larger when the scholarly works from *Die Judenbuche* are used as training data. This is not surprising as there are less scholarly works for *Die Judenbuche* and therefore less training data in that case.

The results show that the performance of both tools is on the same level but from a usability perspective ProQuoLM is superior to ProQuo. The approach is less complex, the creation of training data is a lot less time consuming and there is no need for specific handling of parallel print editions.

5.5 References in Footnotes

In this final experiment we evaluate the performance of ProQuoLM trained solely on scholarly texts from *Michael Kohlhaas* and tested on scholarly texts from *Die Judenbuche* with references in footnotes. But, as before, we exclude footnotes, effectively resulting in scholarly works without any reference information. We compare ProQuoLM against the same baseline as before.

Approach	Precision	Recall	F ₁
Baseline	0.53	0.87	0.66
ProQuoLM	0.81	0.83	0.82

Table 7: Precision, recall, and F₁-score for *Die Judenbuche* for texts with references in footnotes.

Table 7 shows that the performance is similar to the other results. This means that even without reference information ProQuoLM performs on the same level as ProQuo which further highlights the advantages as it is more versatile. It also leads us to the conclusion that ProQuoLM currently cannot make use of the information contained in references. Considering that there is no information available to the model from where in the literary work a candidate is taken, this is not surprising. Another reason could be that BERT can struggle with capturing numeracy (Wallace et al. 2019).

6. Discussion

We presented two approaches for the identification and linking of short quotations between scholarly works and literary works. ProQuo is a pipeline consisting of three steps. We evaluated each step individually and the complete pipeline. ProQuo outperforms a strong baseline, which lacks precision, especially in cases with quotations from different sources. Our results illustrate that the simple approach of just performing text matching is not sufficient for the task at hand.

The second approach, ProQuoLM, performs on the same level as the pipeline but is superior from a usability perspective as it is less complex, more versatile and the creation of training data is less time consuming. We therefore consider ProQuoLM to be a better starting point for future improvements. However, it should be noted that, depending on the overall goal, ProQuo has the advantage that the idea behind the overall approach and the individual steps can be explained which makes it easier to identify specific

issues. The following observations might not have been made without the pipeline and the possibility to investigate individual steps. The development of two approaches is more time and resource consuming but can be beneficial.

From our experiments, we can observe a number of things. Firstly, the distance between a quotation and corresponding reference information can be quite large but our context window is limited due to limitations of current language models. Secondly, the quotation linking step struggles with single word quotations even if they come with a reference. Lastly, ProQuoLM performs on the same level with and without reference information. Based on these observations alone it is not possible to determine the exact source of the remaining issues without further experiments. As a first step, we propose to test ProQuoLM with positional information from where a candidate is taken in the literary work to see if ProQuoLM can make use of reference information at all in the current version. Additionally, it might also be the case that more training would already improve this approach. Also, explicit usage of reference information from the first step of the pipeline in combination with ProQuoLM could be promising but, again, is limited by the fact that reference information can be scattered throughout the text.

Other areas for improvement include the resolution of references which point to other references, for example *ibid.*, and references with multiple page numbers, page ranges or line numbers which are currently not properly handled. We also do not handle quotations with multiple occurrences in the literary work. In the current approach, quotations are never linked to more than one occurrence.

For the presented approaches, we assume a corpus of scholarly works for which we know that the main source of quotations is a certain literary work. Arnold and Jäschke (2022) have found that existing approaches for automatic extraction of bibliographic information do not work for scholarly works in literary studies. This led us to conclude that advances in the extraction of literature references are needed before we can make use of bibliographic information to automatically match scholarly works with the main literary work in focus. Advances in this area would also allow for proper handling of citations from different editions of the literary work.

Another assumption we made for this work is that all quotations appear in quotation marks and that the texts do not contain errors, for instance, due to OCR or mistakes made by the authors. We did not analyze how such errors influence the results as it is beyond the scope of this work. Based on our findings, it seems likely that these errors would have a bigger impact on ProQuo compared to ProQuoLM considering that the former relies more on the availability of specific information. But a deeper analysis is needed to come up with quantifiable results.

7. Data Availability

The annotated scholarly works can currently not be made available due to copyright restrictions. All data, which can be made available, can be found here: <https://hu.berlin/proquo-data>.

8. Software Availability 457

Software can be found here: <https://hu.berlin/proquo> 458

9. Acknowledgements 459

Parts of this research were funded by the German Research Foundation (DFG) priority programme (SPP) 2207 *Computational Literary Studies* project *What matters? Key passages in literary works* (grant no. 424207720). We would like to thank the project's student assistants Gregor Sanzenbacher and Nathalie Burkowski and our colleague Benjamin Fiechter for their annotation work and Steffen Martus for giving feedback on the manuscript. 460 461 462 463 464 465

10. Author Contributions 466

Frederik Arnold: Software, Experiments, Conceptualization, Writing – original draft 467

Robert Jäschke: Conceptualization, Writing – original draft 468





References 469




- Almeida, Mariana S. C., Miguel B. Almeida, and André F. T. Martins (Apr. 2014). “A Joint Model for Quotation Attribution and Coreference Resolution”. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, 39–48. 10.3115/v1/E14-1005. <https://aclanthology.org/E14-1005>. 470 471 472 473 474
- Lesen, was wirklich wichtig ist - Die Identifikation von Schlüsselstellen durch ein neues Instrument zur Zitatanalyse* (2022). Potsdam, Germany. 10.5281/zenodo.6327917. <https://doi.org/10.5281/zenodo.6327917>. 475 476 477
- Arnold, Frederik and Robert Jäschke (2021). “Lotte and Annette: A Framework for Finding and Exploring Key Passages in Literary Works”. In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. NIT Silchar, India: NLP Association of India (NLP AI), 55–63. <https://aclanthology.org/2021.nlp4dh-1.7>. 478 479 480 481
- (2022). “A Game with Complex Rules: Literature References in Literary Studies”. In: *Proceedings of the Workshop on Understanding Literature references in academic full TEXT*. Cologne, Germany: CEUR Workshop Proceedings, 7–15. <https://ceur-ws.org/Vol-3220/paper1.pdf>. 482 483 484 485
- Bloomfield, Lou (2016). *Copyfind*. <https://plagiarism.bloomfieldmedia.com/software/copyfind/>. 486 487
- Bromley, Jane, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah (1993). “Signature Verification Using a Siamese Time Delay Neural Network”. In: *Advances in Neural Information Processing Systems*. Vol. 6. 488 489 490
- Brunner, Annalen, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis (2020). “To BERT or not to BERT-Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of four Types of Speech, Thought and Writing Representation.” In: *SwissText / KONVENS*. 491 492 493 494

- Da, Nan Z. (2019). "The Computational Case against Computational Literary Studies". 495
In: *Critical Inquiry* 45.3, 601–639. [10.1086/702594](https://doi.org/10.1086/702594). 496
- Descher, Stefan and Thomas Petraschka (2018). "Die Explizierung des Impliziten". In: 497
Scientia Poetica 22.1, 180–208. [doi:10.1515/scipo-2018-007](https://doi.org/10.1515/scipo-2018-007). 498
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: 499
Pre-training of Deep Bidirectional Transformers for Language Understanding". In: 500
Proceedings of the 2019 Conference of the North American Chapter of the Association for 501
Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short 502
Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 4171– 503
4186. [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). <https://aclanthology.org/N19-1423>. 504
- Droste-Hülshoff, Annette von (1979). *Die Judenbuche*. Frankfurt am Main: Insel Verlag. 505
ISBN: 3-458-32099-7. <https://www.projekt-gutenberg.org/droste/judenbch/index.html>. 506
[x.html](https://www.projekt-gutenberg.org/droste/judenbch/index.html). 507
- Elson, David K. and Kathleen R. McKeown (2010). "Automatic Attribution of Quoted 508
Speech in Literary Narrative". In: *Proceedings of the Twenty-Fourth AAAI Conference on* 509
Artificial Intelligence. AAAI'10. Atlanta, Georgia: AAAI Press, 1013–1019. 510
- GROBID (2008–2022). <https://github.com/kermitt2/grobid>. swb: 1:dir:dab86b296 511
e3c3216e2241968f0d63b68e8209d3c. 512
- Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: 513
Neural Computation 9.8, 1735–1780. ISSN: 0899-7667. [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735). 514
eprint: [https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997](https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf) 515
[.9.8.1735.pdf](https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf). <https://doi.org/10.1162/neco.1997.9.8.1735>. 516
- Hohl Trillini, Regula and Sixta Quassdorf (2010). "A 'key to all quotations'? A corpus- 517
based parameter model of intertextuality". In: *Literary and Linguistic Computing* 518
25.3, 269–286. [10.1093/lc/fqq003](https://doi.org/10.1093/lc/fqq003). 519
- Kleist, Heinrich von (1978). "Michael Kohlhaas". In: *Werke und Briefe in vier Bänden*. 520
Ed. by Michael Holzinger. CreateSpace Independent Publishing Platform, 7–113. 521
<http://www.zeno.org/nid/2000516902X>. 522
- Molz, Johannes (2020). *A Close and Distant Reading of Shakespearean Intertextuality: Towards 523
a Mixed Method Approach for Literary Studies*. Open Publishing in the Humanities. 524
Universitätsbibliothek Ludwig-Maximilians-Universität. [10.5282/oph.4](https://doi.org/10.5282/oph.4). 525
- Papay, Sean and Sebastian Padó (2019). "Quotation Detection and Classification with 526
a Corpus-Agnostic Model". In: *Proceedings of the International Conference on Recent* 527
Advances in Natural Language Processing (RANLP 2019). Varna, Bulgaria: INCOMA 528
Ltd., 888–894. [10.26615/978-954-452-056-4_103](https://doi.org/10.26615/978-954-452-056-4_103). <https://aclanthology.org/R19> 529
[-1103](https://aclanthology.org/R19-1103). 530
- Pareti, Silvia, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska 531
(2013). "Automatically Detecting and Attributing Indirect Quotations". In: *Proceed-* 532
ings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, 533
Washington, USA: Association for Computational Linguistics, 989–999. [https://ac](https://aclanthology.org/D13-1101) 534
[lanthology.org/D13-1101](https://aclanthology.org/D13-1101). 535
- Prasad, Animesh, Manpreet Kaur, and Min-Yen Kan (2018). "Neural ParsCit: a deep 536
learning-based reference string parser". In: *International Journal on Digital Libraries* 537
19.4, 323–337. ISSN: 1432-5012, 1432-1300. [10.1007/s00799-018-0242-1](https://doi.org/10.1007/s00799-018-0242-1). (Visited on 538
03/11/2021). 539
- Reeve, Jonathan (July 2020). *JonathanReeve/text-matcher: First Zenodo release*. Zenodo. 540
version 0.1.6. [10.5281/zenodo.3937738](https://doi.org/10.5281/zenodo.3937738). 541

- Schaum, Konrad (2004). "Ironie und Ethik in Annette von Droste-Hülshoffs Juden- 542
buche". In: Beiträge zur neueren Literaturgeschichte; [Folge 3], Bd. 204. Winter. 543
Chap. Die Judenbuche als Sittengemälde, 99–194. 544
- Scheible, Christian, Roman Klinger, and Sebastian Padó (2016). "Model Architectures 545
for Quotation Detection". In: *Proceedings of the 54th Annual Meeting of the Association 546
for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association 547
for Computational Linguistics, 1736–1745. 10.18653/v1/P16-1164. <https://aclanthology.org/P16-1164>. 548
549
- Smith, David A., Ryan Cordell, Elizabeth Maddock Dillon, Nick Stramp, and John 550
Wilkerson (2014). "Detecting and Modeling Local Text Reuse". In: *Proceedings of the 551
14th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '14. London, United 552
Kingdom: IEEE Press, 183–192. ISBN: 9781479955695. 553
- TEI Consortium, eds. (2022). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 554
Version 4.4.0. <https://www.tei-c.org/Guidelines/P5/>. 555
- Wallace, Eric, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner (2019). "Do 556
NLP Models Know Numbers? Probing Numeracy in Embeddings". In: *Proceedings 557
of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th 558
International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong 559
Kong, China: Association for Computational Linguistics, 5307–5315. 10.18653/v1 560
/D19-1534. <https://aclanthology.org/D19-1534>. 561
- Winko, Simone (2017–2020). *The making of plausibility in interpretive texts. Analyses of 562
argumentative practices in literary studies*. DFG-funded research project (grant no. 563
372804438). Georg-August-Universität Göttingen. [https://gepris.dfg.de/gepris 564
/projekt/372804438?language=en](https://gepris.dfg.de/gepris/projekt/372804438?language=en). 565
- Winko, Simone and Fotis Jannidis (2015). "Wissen und Inferenz – Zum Verstehen 566
und Interpretieren literarischer Texte am Beispiel von Hans Magnus Enzensbergers 567
Gedicht Frühschriften". In: *Literatur interpretieren: Interdisziplinäre Beiträge zur Theorie 568
und Praxis*. Ed. by Jan Borkowski, Stefan Descher, Felicitas Ferder, and Philipp David 569
Heine. Leiden, Niederlande: Brill | mentis, 221–250. ISBN: 978-3-95743-897-3. <https://doi.org/10.30965/9783957438973>. 570
571
- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang 572
Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva 573
Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, 574
Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, 575
Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, 576
Macduff Hughes, and Jeffrey Dean (2016). *Google's Neural Machine Translation System: 577
Bridging the Gap between Human and Machine Translation*. 10.48550/ARXIV.1609.0814 578
4. <https://arxiv.org/abs/1609.08144>. 579

Automatic Topic-Guided Segmentation of Holocaust Survivor Testimonies

Eitan Wagner¹ 
Renana Keydar² 
Amit Pinchevski³ 
Omri Abend¹ 

1. School of Computer Science and Engineering, Hebrew University of Jerusalem , Jerusalem, Israel.
2. Department of Law and Digital Humanities, Hebrew University of Jerusalem , Jerusalem, Israel.
3. Department of Communications, Hebrew University of Jerusalem , Jerusalem, Israel.

Citation

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend (2023). "Automatic Topic-Guided Segmentation of Holocaust Survivor Testimonies". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-21

Date received 2023-02-17

Keywords

computational, literary, studies

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. In recent decades, efforts have been made to gather and digitize the testimonies of living Holocaust survivors. The challenge we now face is attending to those thousands of human stories, which while safely stored in archives, may nevertheless disappear into oblivion. Despite recent advances in narrative analysis in the fields of Computational Literature (CL) and Natural Language Processing (NLP), existing language model technology still faces challenges in analyzing elaborate narratives and long texts. One such challenge is text segmentation – a long-standing issue in the area of CL and NLP. In our work, we propose a computational method to approach this problem. Our research draws on testimony transcripts from the Shoah Foundation (SF) Holocaust archive for supervised topic classification, which is then used as topic guidance for automatic segmentation.

1. Introduction

Major efforts are being devoted to improving digital access to Holocaust testimonies across the world for safeguarding, cataloging, and disseminating survivors' personal accounts. The imminent passing of the last remaining survivors coincides with the transformation from analog platforms (such as film, video, and television) to digital platforms (big data, online access, social media), which introduces great challenges—and great opportunities—to the future of Holocaust memory. As the phase of survivors' testimony collection reaches its inevitable conclusion, pressing questions emerge: how can we approach and make sense of the enormous quantity of materials collected, which by now exceeds the capacity of human reception? How can we study and analyze the multitude of testimonies in a systematic yet ethical manner, one that respects the integrity of each personal testimony? How can new technology help us cope with the gap between mass atrocity and mass testimony (Keydar 2019)?

Advances in the fields of Computational Literature (CL) and Natural Language Processing (NLP) hold the potential of opening new avenues for the computational analysis of testimony narratives at scale. However, notwithstanding recent developments in the field, existing language modeling technology still faces challenges in analyzing elaborate narratives and long texts in general.

One intuitive approach to dealing with long texts is through an intermediate step of segmentation. Although seemingly simple, the notion of segmentation is not easy to define, as there are many different considerations that may favor placing certain boundary points, rather than others.

In our work, we propose to perform segmentation under the guidance of topics, resulting in the task known as *topical segmentation* (Marti A Hearst 1997; Kazantseva and Szpakowicz 2012). For topics, we use a set of predefined topical categories, created by domain experts. The task of topical segmentation is well studied, but previous work has mostly addressed it in the context of structured, well-defined segments, such as segmentation into paragraphs, chapters, or segmenting text that originated from multiple sources. We tackle the task of segmenting running (spoken) narratives, which poses hitherto unaddressed challenges. As a test case, we address Holocaust survivor testimonies, given in English. Other than the importance of studying these testimonies for Holocaust research, we argue that they provide an interesting test case for topical segmentation, due to their unstructured surface level, relative abundance (tens of thousands of such testimonies were collected), and the relatively confined domain that they cover. Our work leverages the annotations from the Shoah Foundation (SF) Holocaust testimony archive for supervised topic classification and uses these topics as guidance for the topical segmentation of the testimonies.

In this contribution, we discuss the importance and challenges of narrative segmentation as the basis for the analysis of more complex narratological phenomena and as a method for representing the narrative flow. We follow Reiter’s (Reiter 2015) notion of narrative segments as a pragmatic intermediate layer, which is a first step towards the annotation of more complex narratological phenomena. We expand Reiter’s work in two main directions: 1) We hypothesize that boundary points between segments correspond to low mutual information between the sentences proceeding and following the boundary. Based on this hypothesis, we develop a computational model of segmentation for unstructured texts that has the prospect of being identifiable automatically and theoretically sound. We propose a simple computational method for automatic segmentation. 2) We focus on topic-based segmentation rather than on event-based segmentation (Gius and Vauth 2022). This allows the use of topic models and classifiers which are easier to obtain compared to event models. We provide a set of expert-created Holocaust-related topics and train a classifier with it. We consider it an important addition to recent efforts to operationalize narrative theory in CL and NLP.

Our work also contributes to current Holocaust research, seeking methods to better access testimonies (Artstein et al. 2016; Fogu et al. 2016). We expect our methods to promote theoretical exploration and analysis of testimonies, enabling better access, research, and understanding of the past.

This work presents the following results:

- We define segments in a theoretically and sound manner, building upon information-theory measures.
- We propose a simple and effective algorithm for segmentation, independent of topics.

- We evaluate the model and argue for the necessity of guidance for segmentation, especially in unstructured texts. 62 63
- We construct data and models for topic classification. We propose models to infer topics and segments as a combined task. 64 65
- We show that giving a reference (“gold”) segmentation leads to better topics, but it seems difficult to design a joint model that gives a segmentation that benefits the topics. 66 67 68
- We discuss future directions to address this difficulty. 69

We note that the technical and algorithmic part of the paper was adopted with minimal changes from Wagner et al. 2022. 70 71

2. Approaches to Narrative Segmentation 72

2.1 Narrative Analysis 73

Proper representation of narratives in long texts remains an open problem in computational narratology and NLP (Castricato et al. 2021; Mikhalkova et al. 2020; Piper et al. 2021; Reiter et al. 2022). High-quality representations for long texts seem crucial to the development of document-level text understanding technology, which is currently unsatisfactory (Shaham et al. 2022). One possible avenue for representing long texts is to cast them as a sequence of shorter segments, with inter-relations between them. 74 75 76 77 78 79

This direction has deep conceptual roots. Beginning with Aristotle’s theory of drama, narratological analysis has relied on the identification of and distinction between one event and the next. But what guides the segmentation – what constitutes the divide between events – has remained obscure. While some genres offer structural cues for segmentation, such as diary entries, scenes in a dramatic play, stanzas in a poem, or chapters in a novel, other forms of narratives do not always present clear units or boundaries (Gius and Vauth 2022; Zehe et al. 2021). 80 81 82 83 84 85 86

From the computational perspective, much work has been done in the direction of probabilistic schema inference, focusing on either event schemas (Chambers 2013; Chambers and Jurafsky 2009; M. Li et al. 2020) or persona schemas (Bamman et al. 2013, 2014). 87 88 89 90

A common modern approach for modeling narratives is as a sequence of neural states (Rashkin et al. 2020; Wilmot and Keller 2020, 2021). Wilmot and Keller 2020 presented a neural GPT2-based model for suspense in short stories. This work follows an information-theoretic framework, modeling the reader’s suspense by different types of predictability. Wilmot and Keller 2021 present another neural architecture for story modeling. Due to their strong performance in text generation, neural models are also commonly used for story generation, with numerous structural variations (Alhussain and Azmi 2021; Rashkin et al. 2020; Zhai et al. 2019). However, a general drawback of the neural approach is the lack of interpretability, which is specifically crucial in the context of drawing qualitative conclusions from experiments. 91 92 93 94 95 96 97 98 99 100

A different approach represents and visualizes a narrative as a sequence of interpretable 101

topics. Min and Park 2019 visualized plot progressions in stories in various ways, including the progression of character relations. Antoniak et al. 2019 analyzed birth stories, but used a simplistic, uniform segmentation, conjoined with topic modeling, to visualize the frequent topic paths.

Inspired by this approach, we seek to model the narrative of a text using topic segmentation, dividing long texts into topically coherent segments and labeling them, thus creating a global topical structure in the form of a chain of topics. An NLP task of narrative representation of a given document may benefit from knowing something about the document’s high-level structure. Topical segmentation is a lightweight form of such structural analysis: given a sequence of sentences or paragraphs, split it into a sequence of topical segments, each characterized by a certain degree of topical unity (Kazantseva and Szpakowicz 2014). This is particularly useful for texts with little structure imposed by the author, such as speech transcripts, meeting notes, or, in our case, oral testimonies. Topic segmentation can be useful for the indexing of a large number of testimonies (tens of thousands of testimonies have been collected thus far) and as an intermediate or auxiliary step in tasks such as summarization (Jeff Wu et al. 2021) and event detection (Wang et al. 2021).

Unlike recent supervised segmentation models that focus on structured written text, such as Wikipedia sections (Arnold et al. 2019; Lukasik et al. 2020) or book chapters (Petthe et al. 2020), we address the hitherto mostly unaddressed task of segmenting and labeling unstructured (transcribed) spoken language. For these texts, we do not have large datasets of divided text. Moreover, there may not be any obvious boundaries that can be derived based on local properties. This makes the task more challenging and hampers the possibility of taking a completely supervised approach.

To adapt the model to jointly segment and classify, we incorporate into the model a supervised topic classifier, trained over manually indexed one-minute testimony segments, provided by the USC Shoah Foundation (SF).¹ Inspired by Misra et al. 2011, we also incorporate the topical coherence based on the topic classifier into the segmentation model.

2.2 Topic Classification

The *topic* of a text segment is the subject or theme guiding the segment. Latent Dirichlet Allocation (LDA; Blei et al. 2003) is a popular method to extract latent topics in an unsupervised fashion. In LDA, the definition of a topic is a distribution over a given vocabulary. This definition is very flexible, with one of the results being computationally heavy at inference time.

In recent years, attempts have been made, in various degrees of success, to apply topic models for the purpose of narrative analysis prose (Jockers and Mimno 2013; Uglanova and Gius 2020) and drama (Schöch 2021). Topic models can also be useful in the analysis of complex narratives, such as oral testimonies and other free-form narrative texts (Keydar et al. 2022). Despite its popularity, we found LDA to be too heavy computationally for inference on texts with many segments. Therefore we did not apply LDA in our research. Instead, we used supervised text classification over domain-specific

1. <https://sfi.usc.edu/>

predefined topics.

144

In supervised classification, we have a list of predetermined topics and a set of texts, each assigned a topic from the list. A classifier is trained to predict the topic for a given text. Since the introduction of BERT (Devlin et al. 2018), the common practice in NLP is to use a neural model that was pretrained on general language tasks and finetune it for the downstream task of classification. This method achieves impressive results even without a vast amount of labeled data, thus proving a natural choice for many domains.

145

146

147

148

149

150

2.3 Text Segmentation

151

Considerable previous work addressed the task of text segmentation, using both supervised and unsupervised approaches. Proposed methods for unsupervised text segmentation can be divided into linear segmentation algorithms and dynamic graph-based segmentation algorithms.

152

153

154

155

Linear segmentation, i.e., segmentation that is performed on the fly, dates back to the TextTiling algorithm (Marti A Hearst 1997), which detects boundaries using window-based vocabulary changes. Recently, He et al. 2020 proposed an improvement to the algorithm, which, unlike TextTiling, uses the vocabulary of the entire dataset and not only of the currently considered segment. TopicTiling (Riedl and Biemann 2012) uses a similar approach, using LDA-based topical coherence instead of vocabulary only. This method produces topics as well as segments. Another linear model, BATS (Q. Wu et al. 2020), uses combined spectral and agglomerative clustering for topics and segments.

156

157

158

159

160

161

162

163

In contrast to the linear approach, several models follow a Bayesian sequence modeling approach, using dynamic programming for inference. This approach allows making a global prediction of the segmentation, at the expense of higher complexity. Implementation details vary, and include using pretrained LDA models (Misra et al. 2011), online topic estimation (Eisenstein and Barzilay 2008; Mota et al. 2019), shared topics (Jeong and Titov 2010), ordering-based topics (Du et al. 2015), and context-aware LDA (W. Li et al. 2020).

164

165

166

167

168

169

170

Following recent advances in neural models, these models have been used for the task of supervised text segmentation. Pethe et al. 2020 presented ChapterCaptor which relies on two methods. The first method performs chapter break prediction based on Next Sentence Prediction (NSP) scores. The second method uses dynamic programming to regularize the segment lengths toward the average. The models use supervision for finetuning the model for boundary scores, but can also be used in a completely unsupervised fashion. They experiment with segmenting books into chapters, which offers natural incidental supervision.

171

172

173

174

175

176

177

178

Another approach performs the segmentation task in a completely supervised manner, similar to supervised labeled span extraction tasks. At first, the models were LSTM-based (Arnold et al. 2019; Koshorek et al. 2018), and later on, Transformer-based (Lukasik et al. 2020; Somasundaran et al. 2020). Unlike finetuning, this approach requires large amounts of segmented data.

179

180

181

182

183

All of these works were designed and evaluated with structured written text, such as book chapters, Wikipedia pages, or artificially stitched segments, where supervised

184

185

data is abundant. In this work, we address the segmentation of texts of which we have little supervised data regarding segment boundaries. We, therefore, adopt elements from the unsupervised approaches combined with supervised components and design a model for a novel segmentation task of unstructured spoken narratives.

3. Spoken Narratives: Holocaust Testimonies as a case study

3.1 Corpus

Our data consists of Holocaust survivor testimonies. We received 1000 testimonies from the Shoah Foundation. All testimonies were conducted orally with an interviewer, recorded on video, and transcribed as text. The lengths of the testimonies range from 2609 to 88105 words, with mean and median lengths of 23536 and 21424 words, respectively.

The testimonies were transcribed as time-stamped text. In addition, each testimony recording was divided into segments, typically a segment for each minute. Each segment was indexed with labels, possibly multiple. The labels are all taken from the SF thesaurus.² The thesaurus is highly detailed, containing ~ 8000 unique labels across the segments. It's worth noting that the division into segments was done purely by length and does not take the labels into consideration.

3.2 Motivation for Topical Segmentation

Typically, narrative research faces a trade-off between the number of narrative texts, which is important for computational methods, and the specificity of the narrative context, which is essential for qualitative narrative research (Sultana et al. 2022). Holocaust testimonies provide a unique case of a large corpus with a specific context.

The large and specific corpus provides motivation for schema alignment. Assuming that there are common thematic units across testimonies, it should be possible to extract parallel segments. For example, many testimonies address a common event or experience (e.g., "deportation" or "physical hardship") and we might want to compare various aspects of the different reports. To do this we first need to align the testimonies according to these topics.

As opposed to some work that focuses on events in a narrative (Gius and Vauth 2022), we choose to focus on topics. This is for multiple reasons. First, our data is not completely event-based. There are parts that are not events (e.g., "family life", "reflection") and the focus is on the personal experience and not on historical facts. Second, event-complex extraction is a highly complex computational task and has very scarce annotated data (Ning et al. 2018).

As opposed to traditional topic modeling, which is completely unsupervised, we decided to make use of the large annotated testimony corpus. We create a supervised dataset for topic classification and use it in a similar manner as a topic model.

2. <https://sfi.usc.edu/content/keyword-thesaurus>

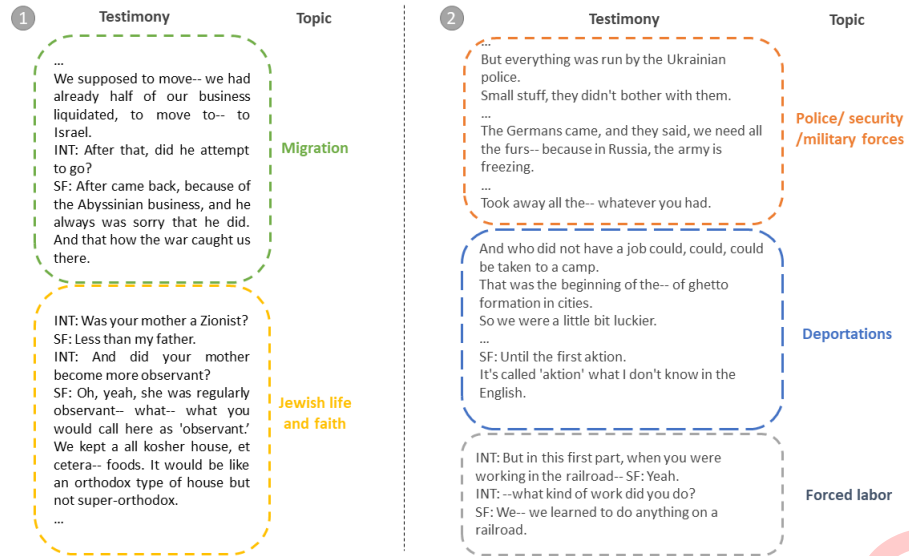


Figure 1: Examples of topic-related segment changes. Both examples are from SF testimony no. 43109.

3.3 Challenges

224

In spoken texts, there are no obvious boundaries. This is unlike common data used for segmentation, such as stitched-up texts (Misra et al. 2011), book chapters (Pethe et al. 2020), or Wikipedia sections (Arnold et al. 2019). Manual inspection of human-annotated segments in the testimonies shows that in many cases the segment boundary highly depends on the expected topic (see figure). In some cases, there is clearly a topic change, but it is not clear where the transition took place. Even in cases where the boundary is in accordance to surface cues, the topics still play a role as they decide what should be read together.

225
226
227
228
229
230
231
232

We argue that proper segmentation cannot truly be separated from the topical structure. When there are no surface-form cues, the segments must depend on a higher-level structure. We claim that there is an essential difference between the prediction of existing segments, in which case the textual cues are given and play a heavy role in the creation of the segmentation, and the generation of a new segmentation for unstructured text, in which case the topical structure is more dominant.

233
234
235
236
237
238

Our experiments (see below) show that given annotated segments with topics, knowledge of the segment boundaries is highly beneficial for predicting the sequence of topics. Put differently, in the topical segmentation task, the segmentation boundaries are predictive of the topic sequence. This finding motivates the exploration of the task of topical segmentation even if the goal of the user is to extract a sequence of topics (and their localization in the text is of less utility).

239
240
241
242
243
244

The SF annotations use a very large set of labels. This hurts the uniformity of the annotation and raises computational issues. Therefore, it is necessary to reduce the size of the label set as we describe in the following section.

245
246
247

3.4 Building the Datasets

248

Classification. As some of the labels in the SF annotations are very rare, and given the noise in the data, using the full SF label set directly as classification labels is dispreferred. Instead, we reduced the number of labels through an iterative process of manual expert annotation and clustering. The SF thesaurus uses a hierarchical system of labels, ranging from high-level topics (e.g., “politics”, “religion and philosophy”), through mid-level (e.g., “camp experiences”, “ghetto experiences”), to low-level labels (e.g., “refugee camp injuries”, “forced march barter”). For the purpose of compiling the list of topics, we focused on mid-level labels. Then, with the help of domain experts from the field of Holocaust studies, we created a list of 29 topics that were deemed sufficiently informative, yet still generalizable across the testimonies. We added the label *NO-TOPIC*, which was used for segments that address technical details of the testimony-giving event (e.g., changing the tape), and do not include Holocaust-related content. For the full topic list see figure 2.

We filtered out testimonies that were not annotated in the same fashion as the others, for example, testimonies that did not have one-minute segments or ones that skipped segments altogether. We used these testimonies for development and testing. We also filtered out all segments that had more than one label after the label conversion. We ended up with a text classification dataset of 20722 segments with 29 possible labels.

We added to the input texts an extra token to indicate the location within the testimony. We divided each testimony into 10 bins with equal segment counts and added the bin index to the input text.

Segmentation. The testimonies in the SF were divided based on pure length. Segment boundaries can appear in the middle of a topic or even a sentence. Therefore, the SF segments cannot be used for the evaluation of topical segmentation. Instead, for evaluation and test sets, we manually segmented and annotated 20 full testimonies. In this set of testimonies, the segmentation was performed based on a given list of topics. We used testimonies from SF that were not annotated in the same manner as the others, and therefore not seen for the classifier. The annotation was carried out by two trained annotators, highly proficient in English. We note that this process was done independently from the SF indexing and therefore the number of testimonies is relatively small.

An initial pilot study to segment testimonies without any prior requirements and no topic list yielded an approximate segment length (the results of these attempts were not included in the training or test data). The approximate length was not used as a strict constraint, but rather as a weak guideline, so as to align our expectations with the annotators.

The approximate desired average segment length was given to the annotators as well as the final topic list. The first annotator annotated all 20 testimonies, which were used for development and testing. The second annotator annotated 7 documents, used for measuring the inter-annotator agreement. The full annotation guidelines can be found in the Supplementary Material (§12.2).

Altogether, for our test data, we obtained 20 testimonies composed of 1179 segments

Topic	Description
1 Adaptation and survival	Any act of finding ways to adapt to the war and persecution and to survive in Ghetto, camps, etc.
2 After the war	Not liberation, but post-war life
3 Aid	Either giving or receiving aid
4 Antisemitism and persecutions	This mostly refers to pre-war episodes, before the ghetto or camps
5 Before the war	This mostly refers to the opening parts relating the pre-war life in the hometown, family, friends, school, etc.
6 Betrayals	Any betrayal by friends, neighbors, locals, etc.
7 Brutality	Any acts of brutality, physical or mental during the war - intended and performed by someone. To be distinguished from hardship which can describe of a certain condition of hardship
8 Camp	Any events that take place in the concentration or death camps
9 Deportations	Deportation from the city/village to the ghetto, and from the ghetto to the camps. This includes any forced transport to an undesired destination.
10 Enemy collaboration	Either jews or locals collaborating with the Nazi regime or their representatives
11 Escape	Any escape from hometown, from the ghetto, from prison or camps
12 Extermination/execution/ death march	Any event of violent intended killing
13 Extreme	killing of a child, suicide, surviving a massacre
14 Family and friendships	Stories involving family members, friends, loved ones
15 Forced labor	Any events taking place in labor camps or as part of forced labor
16 Ghetto	Any event taking place in the ghetto
17 Hardship	Any description of physical or mental hardship
18 Hiding	Hiding places, woods, homes while running away or stories of being hidden by others (farms, monasteries, etc.),
19 Jewish life and faith	Any event relating to jewish life and its practices - school, prayer, shabbat, synagogue, before, during and after the war
20 Liberation	Events relating to allies liberation of camps
21 Migration	Either pre or post-war migration to other countries
22 Non Jewish faith	Any mention of non-jewish beliefs, practices etc.
23 Police/ security /military forces	Events relating to soldiers and police, either enemy or allies
24 Political activity	Protests, political parties, either for or against Nazis
25 Prison	Captivity in prison - to be distinguished from camps
26 Reflection/memory/trauma	
27 Refugees	Mostly the post-war episodes in refugee/displaced persons camps
28 Resistance and partisans	Any act or resistance, organized or individual
29 Stills	Presentation of pictures

Figure 2: List of topics with their description.

with topics. The segment length ranges from 13 to 8772 words, with a mean length of ~ 485 . We randomly selected 5 testimonies for parameter estimation, and the remaining 15 were used as a test set.

4. Topical Segmentation

We have a document X consisting of n sentences $x_1 \dots x_n$, which we consider atomic units. Our task is to find $k - 1$ boundary points, defining k segments, and k topics, where every consecutive pair of topics is different.

4.1 General Considerations

Designing a model for topical segmentation involves multiple, possibly independent, considerations which we present here. For more technical details, see Wagner et al. 2022.

Our general approach to segmentation requires a scoring method that can be applied to each possible segment. Given these scores and the desired number of segments, we can then select the segmentation with the highest score.

We compose a segment score based on both local and non-local properties. For local scores, we propose to use Point-wise Mutual Information (PMI). Given a language model (LM), we hypothesize that the mutual information between two adjacent sentences can predict how likely the two sentences are to be in the same segment. These scores need additional supervision beyond the LM pretraining. Given these scores, the extraction of a segmentation for a given text is equivalent to maximizing the LM likelihood of text, under the assumption that each sentence depends on one previous sentence and that each segment depends on no previous sentences. A formal proof can be found in the supplementary material (§12.1).

This is opposed to recent work in this direction that uses the Next Sentence Prediction (NSP) scores (Pethe et al. 2020). We argue that the pretrained NSP scores do not capture the probability of two given consecutive sentences being in the same segment, since even if the second sentence is in another segment, it still is the next sentence.

Based on previous work, we also consider the non-local properties of segment length (Pethe et al. 2020), and topical coherence Misra et al. 2011. Given a domain-specific multi-label classifier, we use the classification log probabilities as the coherence scores.

Given the desired number of segments, we have a structured prediction task that requires dynamic programming in order to be executed in polynomial time, where the degree of the polynomial is decided by the order of dependency. Inference of the optimal topic assignment according to a given classifier also requires a dynamic algorithm to avoid identical adjacent topics.

4.2 Topical Segmentation Models

We propose various models and baselines for the task of topical segmentation. Some models perform segmentation and topic assignment separately (“pipeline”) and some jointly.

Topic-Modeling Based. Misra et al. 2011 performed topical segmentation based on topic modeling, where the selected segmentation is that with the highest likelihood, based on the Latent Dirichlet Allocation model (LDA, Blei et al. 2003). The topic model gives a likelihood score to each segment and the segmentation that maximizes the product of likelihoods is selected. Inference is equivalent to finding the shortest path in a graph with n^2 nodes.

NSP-based Segmentation. The approach in the first ChapterCaptor model is to perform linear segmentation based on Next Sentence Prediction (NSP) scores. Using a model that was pretrained for NSP, they further finetune the model with segmented data, where a positive label is given to two subsequent spans in one segment, and a negative label is given to two spans that are in different segments.

The second ChapterCaptor model leverages the assumption that segments tend to have similar lengths. Given data, they compute the expected average length, L , and add regularization towards average-length segments. We denote this model with NSP+L.

LMPMI-based Segmentation.. Adapting the NSP scores for segmentation seems sub-optimal in domains for which we do not have enough segmented data. We propose to replace the NSP scores with language-modeling (LM) and Point-wise Mutual Information (PMI) scores. Specifically, for each possible boundary index i , we define:

$$LMPMI_i = \log \frac{P_{LM}(x_i, x_{i+1})}{P_{LM}(x_i) \cdot P_{LM}(x_{i+1})} \quad (1)$$

where the probabilities are the LM probabilities for the sentences together or alone.

These scores can be computed by any pretrained language model, and the log-scores replace the NSP scores in both previous methods. We denote these models with PMI and PMI + L.

Pipeline Topic Assignment. Given a segmentation for the document and a topic classifier, we infer a list of topics. We need to find the optimal topic sequence under the constraint of no identical adjacent elements. This can be formalized as an HMM inference task, which can easily be found using dynamic programming.

Joint Segmentation and Topic Assignment. In another method, we take into account the segment classification scores in addition to a length penalty. We jointly infer a segmentation and topic assignment using the following dynamic formula:

$$\begin{aligned} cost(n, k, t) = \min_{\substack{1 \leq i \leq n-1 \\ t' \in T}} (cost(i, k-1, t') + \alpha \cdot \frac{|n-i-L|}{L} + \beta \cdot \log P(t'|X_i \dots X_n)) \\ + (1 - \alpha - \beta) \cdot PMI_n \quad (2) \end{aligned}$$

where $cost(n, k, t)$ represents the cost of a boundary at index n with $k-1$ previous bound-

aries and topic t as the last topic. α, β are hyperparameters controlling the components. We denote this model with PMI + T.

Baseline Models. As a point of comparison, we also implemented simple baseline models for segmentation and topic selection. These models can be used in a pipeline.

For segmentation, we divide a text into equally lengthed segments, given a predetermined number of segments. This method was used by Antoniak et al. 2019 and, with slight modifications, by Jeff Wu et al. 2021, as it is extremely simple and efficient.

For topic assignment, we sequentially sample topics from a uniform distribution over the set of given topics. We avoid repeating topics by giving probability 0 to the previous topic.

We denote these baselines with UNIFORM.

Implementation Specifics. The classifier was selected by fine-tuning various Transformer-based models with a classification head. Base models were pretrained by HuggingFace. We experimented with Distilbert, Distilroberta, Electra, RoBERTa, XLNet, and DeBERTa in various sizes. For our experiments we chose to use Distilroberta, which showed an accuracy score of ~ 0.55 , which was close to that of the larger models, doing this with way faster training and inference. We trained with a random 80-20 data split on 2 GPUs for ~ 10 minutes with the Adam optimizer for 5 epochs with $batch\text{-}size=16$, $label\text{-}smoothing=0.01$ and other settings set as default. We selected this classifier for our final segmentation experiments.

From the 20 manually segmented testimonies, we randomly took 5 testimonies a development set for hyperparameter tuning. Based on the results on this set, we chose $\alpha = 0.8$ for the PMI + L model and $\alpha = \beta = 0.2$ for the PMI + T model.

The LDA topic model was pretrained on the same training data as the classifier's (§3.4), before running the segmentation algorithm. We trained the LDA model with 15 topics using the Gensim package,⁴ which we also used for the likelihood estimation of text spans given an LDA model.

We used HuggingFace's pretrained transformer models for the NSP scores and LM probabilities. We used FNET (Lee-Thorp et al. 2021) for NSP and GPT2 (Radford et al. 2019) for LM probabilities. We tuned the context size parameter on the development set, resulting in $C = 3$.

4.3 Evaluation Methods

Here we discuss appropriate metrics for the segmentation and topic assignments.

Segmentation. Measuring the quality of text segmentation is tricky. We want to give partial scores to segmentations that are close to the manually annotated ones, so simple Exact-Match evaluation is overly strict. This is heightened in cases like ours, where there is often no clear boundary for the topic changes. For example, in one place the

3. <https://pypi.org/project/transformers/>

4. <https://radimrehurek.com/gensim/>

witness says “he helped us later when we planned the escape”. This sentence comes between 397
getting help (the *Aid* topic) and escaping (the *Escape* topic). We would like to give at 398
least partial scores for boundaries either before or after this sentence. 399

Various attempts have been made to alleviate this problem and propose more relaxed 400
measures. Since the notion of “closeness” strongly depends on underlying assumptions, 401
it seems hard to pinpoint one specific measure that will perfectly fit our expectations. 402
Following this rationale, we report a few different measures. 403

The first measure we report is the average F1 score, which counts overlaps in the exact 404
boundary predictions. Another measure we used is average WindowDiff (WD; Pevzner 405
and Marti A. Hearst 2002), which compares the number of reference boundaries that fall 406
in an interval with the number of boundaries that were produced by the algorithm. We 407
also measured the average Segmentation Similarity (S-sim; Fournier and Inkpen 2012) 408
and Boundary Similarity (B-sim; Fournier 2013) scores. These scores are based on the 409
number of edits required between a proposed segmentation and the reference, where 410
Boundary Similarity assigns different weights to different types of edits. In F1, B-sim, 411
and B-sim a higher score is better and in WindowDiff a lower score is better. We used 412
the segeval python package⁵ with the default settings to compute all of these measures. 413
Notably, the window size was set to be the average segment length (in the reference 414
segmentation for the particular testimony) divided by 2. 415

Topic Assignment. To measure the similarity between a predicted topic assignment and 416
the reference assignment we used two different measures. One measure was python’s 417
difflib SequenceMatcher (SM) scores, which are based on the *gestalt pattern matching* 418
metric Ratcliff and Metzener 1988, that considers common substrings. In this metric, a 419
higher score means stronger similarity. 420

Another measure we used is the Damerau–Levenshtein edit distance (Edit, Damerau 421
1964), which measures distance by the number of actions needed to get from one 422
sequence to another. We normalized the edits by the number of topics in the reference 423
data. For the Edit distance, lower is better. 424

5. Results and Discussion 425

We evaluate our models for both the segmentation and the resulting topic sequence. 426

We do not report scores for the LDA-based model since it did not produce a reasonable 427
number of segments, and its runtime was prohibitively long (in previous work, it was 428
run on much shorter text). We also implemented the models with different sizes of 429
GPT2. Observing that the size had no significant effect, we report the results with the 430
base model (“gpt2”) only. 431

5.1 Annotator Agreement. 432

Evaluating on the 7 documents that were annotated by both annotators, we achieve 433
Boundary score = 0.324, *Sequence Matching* = 0.4 and *Edit distance* = 0.73. 434

5. <https://pypi.org/project/segeval/>

Model	F1	WD	S-sim	B-sim
UNIFORM	0.052	0.568	0.958	0.026
NSP + L	0.04	0.584	0.958	0.02
PMI	0.172	0.537	0.963	0.094
PMI + L	0.173	0.535	0.964	0.095
PMI + T	0.165	0.54	0.962	0.09

Table 1: Segmentation scores. We evaluate PMI-score models with and without length penalties (PMI and PMI + L, respectively). We also evaluate a joint model for segmentation with topics (PMI + T), a uniform length segmentor (UNIFORM) and a Next Sentence Prediction segmentor with length penalties (NSP + L). For F1, S-SIM and B-SIM, higher is better and for WD lower is better. The number of segments is decided using the expected segment length.

In complex structured tasks, the global agreement score is expected to be low. Agreement in these cases is therefore often computed in terms of sub-structures (e.g., attachment score or PARSEVAL F-score in parsing instead of exact-match). Since no local scores are common in segmentation tasks, we report only the global scores despite their relative strictness. Compared to the boundary score of uniform-length segmentation (which is much better than random), we can see that the annotator agreement was larger by an order of magnitude. Eyeballing the differences between the annotators also revealed that their annotations are similar.

We note that the annotators did not always mark the same number of segments (and topics), and this can highly influence the scores. We also note that the annotators worked completely independently and did not adjudicate.

5.2 Model Performance

Segmentation. Table 1 presents the results for the segmentation task. We see that PMI-based models are significantly better than the uniform length segmentation and the NSP-based model. Among the PMI-based models, there is no clear advantage for a specific setting, as the local PMI model is slightly better than the models with global scores.

We note that due to the nature of the metrics, specifically how they normalize the values to be between 0 and 1, the different measures vary in the significance of the gaps.

In figure 3 we present two examples of outputs of the PMI model. In the first case, the human annotator did not put boundaries where the model did, but the model’s predictions seem plausible. In the second example, the model predicted a boundary in the same place as the annotator.

Topic Assignment. Table 2 presents our results for the topic assignments produced by our models and the baselines. For comparison, it also presents the scores for topic creation based on the classifier when the real annotated segments are given.

Here we see that the pipeline methods with uniform or NSP segmentation provide slightly better topics than the joint inference model or the simple PMI model. All models based on the classifier perform significantly better than the baselines.

1	Testimony	2	Testimony	Annotator's Topic
	<p>INT: What was your father's occupation? HP: Well, what was my father occupation? My family owned a shop on the main square in a place called Rakovnik...</p> <p>INT: What was sold in the shop? HP: ... Toys... And thinking back, what was really interesting that the toys were already, at that time, imported some of them from Germany, and some of them still I have seen a New Zealand being imported. [LAUGHS] Yes.</p> <p>INT: What kind of man was your father? HP: What kind of man? ... Of course, very nice ... And I think what had a big influence on my father, when the ... And he became very involved, also politically, in the Czech or Czechoslovak Republic. ...</p> <p>INT: What was the name of that organization in Czech? HP: ... I think it was Czech [NON-ENGLISH]. ... but I think that was. Yeah.</p> <p>INT: What was your father's role in the home? HP: What was the father's role in my home? I mean, he was working. You know, he was in a shop. ... because my mother also worked in a shop... INT: What was your mother's name? ...</p> <p>INT: And her maiden name? HP: ... My mother was born not long-- I would say a small village not far away from Rakovnik. And I think it was a very large family. And I think that I knew my grandfather. At that time, when I was a kid, he was well over 80. I never knew my grandmother. Yes.</p> <p>INT: Do you know how your parents met? ...</p>		<p>... It was an old house, but for Europe the houses are not old if they're 1850 or whatever. ...</p> <p>INT: What languages did you speak at home? HP: We spoke Czech. You can hear my Czech accent. Yes. But I went to school in Prague, later on, and there was German school to learn German, and I also learned French.</p> <p>INT: Was your family religious? HP: Semi-religious. ... Compared what I can see here or especially in Melbourne, I could say not so-- not as much religious. But we-- the strange thing was we didn't have a Friday night, but my father usually, for Friday evening, bought something which was nice....</p>	<p>Before the war</p> <p>Jewish life and faith</p>

Figure 3: Examples of outputs from the PMI segmentation model. In 1 the predicted boundaries were not marked by the annotators. In 2 the model and the annotators agreed. Both examples are from SF testimony 43109.

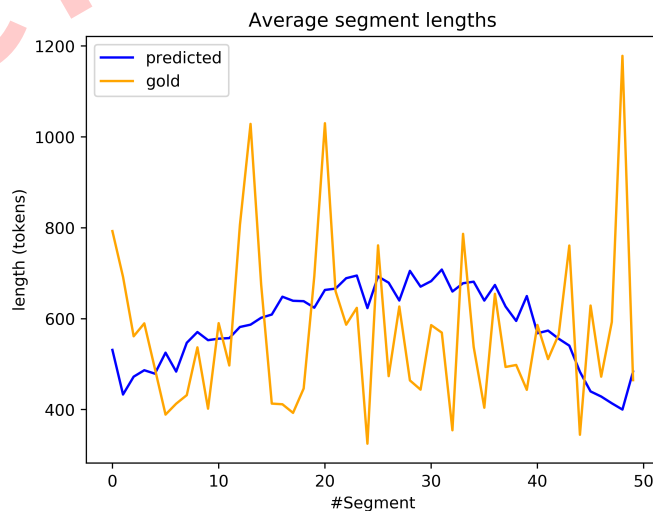


Figure 4: Segment length (in tokens) as a function of the segment index in the testimony. The predicted segments are decided by the PMI model for 50 segments. The gold segmentation was normalized to 50 segments and then averaged.

Model	SM	Edit
UNIFORM	0.138	1.13
UNIFORM + CL	0.378	0.872
NSP + CL	0.369	0.875
PMI + CL	0.36	0.892
PMI + T	0.375	0.872
GOLD + CL	0.478	0.5

Table 2: Performance of the various models for topic lists. In Sequence Matching (SM) higher is better and for the Edit Distance, lower is better. In all cases, the number of topics was set as the length divided by the expected segment length rounded. The models we evaluate are uniform segmentation, NSP segmentation with length penalties, and PMI segmentation, all with dynamic topic assignment based on the classifier (UNIFORM + CL, NSP + CL and PMI + CL, respectively), and the joint segmentation and classification model (PMI + T). The baseline model is uniform topic generation (UNIFORM), which samples topics independently of the given text, and avoids repeating the previous topic.

5.3 Discussion

Our experiments show that topic assignment given the real segmentation (GOLD + CL) yields better topics than all other models. This suggests that a high-quality segmentation does contribute to the topic assignment, which motivates work on segmentation, even if the desired product is (only) the sequence of topics, without their localization. The GOLD + CL model in fact achieves higher topic similarity than the inter-annotator agreement. This might be explained by the fact that the GOLD + CL model was given the exact number of segments, while this was not specified for the annotators.

Regarding the segmentation models, our results show that the PMI methods present better performance for the segmentation task, compared to previous methods. However, we find that the automatic segmentation results do not contribute to the topic assignment. Also, within our different PMI models, we see that additional length and topic scores do not yield substantial improvements, neither for the segmentation nor for the topics. This is somewhat surprising and might mean that the sensitivity of our classifier to exact boundaries is low, or that the produced segments did not yet cross a usefulness threshold for topic classification.

Another seemingly surprising result is that larger sizes and domain fine-tuning of the GPT2 model do not improve performance, sometimes actually hurting it.

Delving more into the outputted segments created by the PMI models (see 3), it seems that these models do produce meaningful segments with good boundaries, but they do not always match the manual boundaries, as the exact segmentation also depends on the given set of topics. we hypothesize that there is a gap between the “surface level topic changes”, that are reflected in clear textual cues (e.g., a new question by the interviewer) and “high-level topic changes”, that highly depend on our prior domain-specific topical interests.

If the hypothesis is correct we would expect to see more PMI boundaries in parts of the testimony that are more structured, compared to the reference boundaries that would be more prevalent where there are more Holocaust-related themes.

In figure 4 we plot the average lengths (in tokens) of the segments as a function of the location within the testimony. Since the number of segments in the reference data varies,

we normalize the lengths as if there were 50 segments in each testimony. We can see that the PMI segments tend to be longer in the middle of the testimony and shorter at the beginning and end. The reference segmentation has less of a pattern.⁶ This result supports our hypothesis. The SF testimonies have a relatively uniform structure at the beginning and the end, so it should be easier to detect surface-level changes there.

This is an important argument regarding contemporary segmentation models. It is common to test the model with highly structured cases, like book sections (Pethe et al. 2020), Wikipedia sections (Arnold et al. 2019), and randomly stitched stories (Misra et al. 2011). In these cases, it is important to assert that the high performance is not due to surface-level cues, in which case the model only predicts traces of previously generated sections and does not actually segment a long document.

We note that the data for the classification and segmentation are restricted to a specific domain, specifically Holocaust testimonies. This limits the generalization of our models to other domains. The general ideas are domain-independent, and some models can be readily used, the application of the models that use the classifier will require adaptation to a new domain.

We also note that the use of single-label topic classification has its limitations. It seems that in some cases the topics are not mutually exclusive (e.g., a segment can involve both “Family” and “Ghetto”). This makes the topical segmentation task less conclusive. In future work we intend to model the temporal and spatial paths of testimonies, allowing segmentation and alignment in a more robust manner.

6. Conclusion

We presented models for combined segmentation and topic extraction for narratives. We found that: (1) segmentation boundaries can be indicative of the sequence of topics (as demonstrated by using the gold standard segmentation; however, (2) topic lists inferred dynamically given a classifier are not very sensitive to the actual segmentation, allowing the extraction of high-quality topic lists even with uniform segmentation. In addition, we find that (3) local PMI scores are sufficient to infer a segmentation with better quality than previous models; (4) additional features such as segment lengths and topics seem to have limited influence on the quality of the segmentation;

Our work addresses the segmentation and topic labeling of text in a naturalistic domain, involving unstructured, transcribed text. Our model can segment noisy texts where textual cues are sparse.

In addition to the technical contribution of this work, it also makes important first steps in analyzing spoken testimonies in a systematic, yet ethical manner. With the imminent passing of the last remaining Holocaust survivors, it is increasingly important to design methods of exploration and analysis of these testimonies, so as to enable us to use the wealth of materials collected in the archives for studying and remembering their stories.

6. We averaged over 500 predicted testimonies and only 20 reference ones, so it is expected that the reference lengths will be noisier.

7. Data Availability 532

Data will be given upon permission from the Shoah Foundation. 533

8. Software Availability 534

Software can be found here: <https://github.com/eitanwagner/holocaust-segmentation> 535
536

9. Acknowledgements 537

The authors acknowledge the USC Shoah Foundation - The Institute for Visual History and Education for its support of this research. We thank Prof. Gal Elidan, Prof. Todd Presner, Dr. Gabriel Stanovsky, Gal Patel and Itamar Trainin for their valuable insights and Nicole Gruber, Yelena Lizuk, Noam Maeir and Noam Shlomai for research assistance. This research was supported by grants from the Israeli Ministry of Science and Technology and the Council for Higher Education and the Alfred Landecker Foundation. 538
539
540
541
542
543
544

10. Author Contributions 545

Eitan Wagner: Conceptualization, Investigation, Computational analysis, Visualization, Experimentation, Implementation, Writing – original draft, Writing – review and editing 546
547

Renana Keydar: Conceptualization, Data Curation, Supervision, Writing – original draft, Writing – review and editing 548
549

Amit Pinchevski: Conceptualization, Writing – original draft, Writing – review and editing 550
551

Omri Abend: Conceptualization, Supervision, Writing – original draft, Writing – review and editing 552
553

11. Ethical Statement 554

We abided by the instructions provided by each of the archives. We note that the witnesses identified themselves by name, and so the testimonies are not anonymous. Still, we do not present in the analysis here any details that may disclose the identity of the witnesses. We intend to release our codebase and scripts, but those will not include any of the data received from the archives; the data and trained models used in this work will not be given to a third party without the consent of the relevant archives. We note that we did not edit the testimony texts in any way. The compilation of the Holocaust related topics was done by Holocaust experts based on the SF thesaurus hierarchy. We did not apply automatic generation at any point. 555
556
557
558
559
560
561
562
563

12. Supplementary Material

12.1 Equivalence of PMI and Likelihood

We have a document $X = x_1, x_2 \dots, x_n$ which we want to divide into k segments.

We assume that the LM probability for each sentence depends only on the previous sentence and that in the case of a boundary at index i , sentence i is independent of all previous sentences. Under these assumptions, the segmentation that places boundaries at the places with minimal PMI is the same segmentation that maximized the LM likelihood.

Proof: Assume we have a boundary set $B = (i_1, i_2, \dots, i_k)$.

For any $i \in B$ we have:

$$PMI(x_i, x_{i-1}) = \frac{P(x_i | x_{i-1})}{P(x_i)} = 1$$

Therefore we get:

$$\begin{aligned} {}_B P(X) &= {}_B P(X) \cdot \prod_{i=1}^n \frac{1}{P(x_i)} = {}_B \prod_{i \notin B} \frac{P(x_i | x_{i-1})}{P(x_i)} \prod_{i \in B} \frac{P(x_i)}{P(x_i)} \\ &= {}_B \sum_{i \notin B} \log PMI(x_i, x_{i-1}) = {}_B \sum_{i=1}^n \log PMI(x_i, x_{i-1}) \quad (3) \end{aligned}$$

12.2 Annotation Guidelines

Annotation Guidelines for Topical Segmentation

In this task, we divide Holocaust testimonies into topically coherent segments. The topics for the testimonies were predetermined. We have 29 content topics and a NULL topic. The full list is attached. Each segment has one topic (multi-class, not multi-label), and a change of topic is equivalent to a change of segments.

The segmentation annotation will be as follows:

- The testimonies are already divided into sentences. A segment change can only be between sentences.
- Our goal is to annotate segmentations. For this, we will assign a topic for each sentence. Since the main focus is the segment, the topic should be given based on a segment and not a single sentence.
- The changing of a topic, if it does not include further information, should not be marked as a separate topic, rather it should be combined with the surrounding topics. If there is a change of topics there then the Overlap should be marked as True over these sentences.

- Regarding the number of requested segments, we want an approximate average segment length of 30 sentences. This is a global attribute, as the actual Segment lengths can (and should) vary, depending on the topics. Any single segment should be decided mainly by content and not by constraints regarding the segment lengths.
- After deciding the segment scope, all sentences can be marked at once. No need to mark them one by one.
- No sentence should be left without a topic (NULL is also a topic). If the topic is unclear then one should be chosen. It should not be left empty.
- A “thumb rule” in cases of multiple options is to choose a topic that is more Holocaust-specific. For example, a hiding story about a family member should be assigned to “Hiding” and not to “Family and Friendships”.

13. Data Availability

Data will be given upon permission from the Shoah Foundation.

14. Software Availability

Software can be found here: <https://github.com/eitanwagner/holocaust-segmentation>

15. Acknowledgements

The authors acknowledge the USC Shoah Foundation - The Institute for Visual History and Education for its support of this research. We thank Prof. Gal Elidan, Prof. Todd Presner, Dr. Gabriel Stanovsky, Gal Patel and Itamar Trainin for their valuable insights and Nicole Gruber, Yelena Lizuk, Noam Maeir and Noam Shlomei for research assistance. This research was supported by grants from the Israeli Ministry of Science and Technology and the Council for Higher Education and the Alfred Landecker Foundation.

16. Author Contributions

Eitan Wagner: Conceptualization, Investigation, Computational analysis, Visualization, Experimentation, Implementation, Writing – original draft, Writing – review and editing

Renana Keydar: Conceptualization, Data Curation, Supervision, Writing – original draft, Writing – review and editing

Amit Pinchevski: Conceptualization, Writing – original draft, Writing – review and editing

Omri Abend: Conceptualization, Supervision, Writing – original draft, Writing – review and editing

References

- Alhussain, Arwa I. and Aqil M. Azmi (May 2021). "Automatic Story Generation: A Survey of Approaches". In: *ACM Comput. Surv.* 54.5. ISSN: 0360-0300. [10.1145/3453156](https://doi.org/10.1145/3453156). <https://doi.org/10.1145/3453156>.
- Antoniak, Maria, David Mimno, and Karen Levy (2019). "Narrative paths and negotiation of power in birth stories". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, 1–27.
- Arnold, Sebastian, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser (Mar. 2019). "SECTOR: A Neural Model for Coherent Topic Segmentation and Classification". In: *Transactions of the Association for Computational Linguistics* 7, 169–184. [10.1162/tacl_a_00261](https://aclanthology.org/Q19-1011). <https://aclanthology.org/Q19-1011>.
- Artstein, Ron, Alesia Gainer, Kallirroi Georgila, Anton Leuski, Ari Shapiro, and David Traum (June 2016). "New Dimensions in Testimony Demonstration". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. San Diego, California: Association for Computational Linguistics, 32–36. [10.18653/v1/N16-3007](https://aclanthology.org/N16-3007). <https://aclanthology.org/N16-3007>.
- Bamman, David, Brendan O'Connor, and Noah A. Smith (Aug. 2013). "Learning Latent Personas of Film Characters". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, 352–361. <https://aclanthology.org/P13-1035>.
- Bamman, David, Ted Underwood, and Noah A. Smith (June 2014). "A Bayesian Mixed Effects Model of Literary Character". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 370–379. [10.3115/v1/P14-1035](https://aclanthology.org/P14-1035). <https://aclanthology.org/P14-1035>.
- Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3, 993–1022.
- Castricato, Louis, Stella Biderman, David Thue, and Rogelio Cardona-Rivera (June 2021). "Towards a Model-Theoretic View of Narratives". In: *Proceedings of the Third Workshop on Narrative Understanding*. Virtual: Association for Computational Linguistics, 95–104. [10.18653/v1/2021.nuse-1.10](https://aclanthology.org/2021.nuse-1.10). <https://aclanthology.org/2021.nuse-1.10>.
- Chambers, Nathanael (Oct. 2013). "Event Schema Induction with a Probabilistic Entity-Driven Model". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, 1797–1807. <https://aclanthology.org/D13-1185>.
- Chambers, Nathanael and Dan Jurafsky (Aug. 2009). "Unsupervised Learning of Narrative Schemas and their Participants". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, 602–610. <https://aclanthology.org/P09-1068>.
- Damerau, Fred J. (Mar. 1964). "A Technique for Computer Detection and Correction of Spelling Errors". In: *Commun. ACM* 7.3, 171–176. ISSN: 0001-0782. [10.1145/363958.363994](https://doi.org/10.1145/363958.363994). <https://doi.org/10.1145/363958.363994>.




- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 10.48550/ARXIV.1810.04805. <https://arxiv.org/abs/1810.04805>.
- Du, Lan, John K Pate, and Mark Johnson (2015). "Topic segmentation with an ordering-based topic model". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Eisenstein, Jacob and Regina Barzilay (Oct. 2008). "Bayesian Unsupervised Topic Segmentation". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, 334–343. <https://aclanthology.org/D08-1035>.
- Fogu, Claudio, Wulf Kansteiner, and Todd Presner (2016). *Probing the ethics of Holocaust culture*. Harvard University Press.
- Fournier, Chris (Aug. 2013). "Evaluating Text Segmentation using Boundary Edit Distance". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, 1702–1712. <https://aclanthology.org/P13-1167>.
- Fournier, Chris and Diana Inkpen (June 2012). "Segmentation Similarity and Agreement". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, 152–161. <https://aclanthology.org/N12-1016>.
- Gius, Evelyn and Michael Vauth (2022). "Towards an Event Based Plot Model. A Computational Narratology Approach". In: *Journal of Computational Literary Studies* 1.1.
- He, Xin, Jian Wang, Quan Zhang, and Xiaoming Ju (2020). "Improvement of Text Segmentation TextTiling Algorithm". In: *Journal of Physics: Conference Series*. Vol. 1453. 1. IOP Publishing, 012008.
- Hearst, Marti A (1997). "Text Tiling: Segmenting text into multi-paragraph subtopic passages". In: *Computational linguistics* 23.1, 33–64.
- Jeong, Minwoo and Ivan Titov (2010). "Multi-document topic segmentation". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*, 1119–1128.
- Jockers, Matthew L and David Mimno (2013). "Significant themes in 19th-century literature". In: *Poetics* 41.6, 750–769.
- Kazantseva, Anna and Stan Szpakowicz (June 2012). "Topical Segmentation: a Study of Human Performance and a New Measure of Quality." In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, 211–220. <https://aclanthology.org/N12-1022>.
- (Aug. 2014). "Hierarchical Topical Segmentation with Affinity Propagation". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 37–47. <https://aclanthology.org/C14-1005>.
- Keydar, Renana (2019). "Mass Atrocity, Mass Testimony, and the Quantitative Turn in International Law". In: *Law & Society Review* 53.2, 554–587.
- Keydar, Renana, Yael Litmanovitz, Badi Hasisi, and Yoav Kan-Tor (2022). "Modeling Repressive Policing: Computational Analysis of Protocols from the Israeli State Commission of Inquiry into the October 2000 Events". In: *Law & Social Inquiry* 47.4, 1075–1105. 10.1017/lsi.2021.63.

- Koshorek, Omri, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant (June 2018). "Text Segmentation as a Supervised Learning Task". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 469–473. 10.18653/v1/N18-2075. <https://aclanthology.org/N18-2075>.
- Lee-Thorp, James, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon (2021). "Fnet: Mixing tokens with fourier transforms". In: *arXiv preprint arXiv:2105.03824*.
- Li, Manling, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss (Nov. 2020). "Connecting the Dots: Event Graph Schema Induction with Path Language Modeling". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 684–695. 10.18653/v1/2020.emnlp-main.50. <https://aclanthology.org/2020.emnlp-main.50>.
- Li, Wenbo, Tetsu Matsukawa, Hiroto Saigo, and Einoshin Suzuki (2020). "Context-Aware Latent Dirichlet Allocation for Topic Segmentation". In: *Advances in Knowledge Discovery and Data Mining* 12084, 475.
- Lukasik, Michal, Boris Dadachev, Kishore Papineni, and Gonalo Simoes (Nov. 2020). "Text Segmentation by Cross Segment Attention". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 4707–4716. 10.18653/v1/2020.emnlp-main.380. <https://aclanthology.org/2020.emnlp-main.380>.
- Mikhalkova, Elena, Timofei Protasov, Polina Sokolova, Anastasiia Bashmakova, and Anastasiia Drozdova (May 2020). "Modelling Narrative Elements in a Short Story: A Study on Annotation Schemes and Guidelines". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 126–132. ISBN: 979-10-95546-34-4. <https://aclanthology.org/2020.lrec-1.16>.
- Min, Semi and Juyong Park (2019). "Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling". In: *PloS one* 14.12, e0226025.
- Misra, Hemant, Franois Yvon, Olivier Capp  , and Joemon Jose (2011). "Text segmentation: A topic modeling perspective". In: *Information Processing and Management* 47.4, 528–544.
- Mota, Pedro, Maxine Eskenazi, and Lu sa Coheur (Nov. 2019). "BeamSeg: A Joint Model for Multi-Document Segmentation and Topic Identification". In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, 582–592. 10.18653/v1/K19-1054. <https://aclanthology.org/K19-1054>.
- Ning, Qiang, Hao Wu, and Dan Roth (July 2018). "A Multi-Axis Annotation Scheme for Event Temporal Relations". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 1318–1328. 10.18653/v1/P18-1122. <https://aclanthology.org/P18-1122>.
- Pethe, Charuta, Allen Kim, and Steve Skiena (Nov. 2020). "Chapter Captor: Text Segmentation in Novels". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Lin-

- guistics, 8373–8383. [10.18653/v1/2020.emnlp-main.672](https://aclanthology.org/2020.emnlp-main.672). <https://aclanthology.org/2020.emnlp-main.672>. 762 763
- Pevzner, Lev and Marti A. Hearst (2002). “A Critique and Improvement of an Evaluation Metric for Text Segmentation”. In: *Computational Linguistics* 28.1, 19–36. [10.1162/089120102317341756](https://aclanthology.org/J02-1002). <https://aclanthology.org/J02-1002>. 764 765 766
- Piper, Andrew, Richard Jean So, and David Bamman (Nov. 2021). “Narrative Theory for Computational Narrative Understanding”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 298–311. <https://aclanthology.org/2021.emnlp-main.26>. 767 768 769 770 771
- 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, 9. 772 773 774
- Rashkin, Hannah, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao (Nov. 2020). “Plot-Machines: Outline-Conditioned Generation with Dynamic Plot State Tracking”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 4274–4295. [10.18653/v1/2020.emnlp-main.349](https://aclanthology.org/2020.emnlp-main.349). <https://aclanthology.org/2020.emnlp-main.349>. 775 776 777 778 779
- Ratcliff, John W and David E Metzener (1988). “Pattern-matching-the gestalt approach”. In: *Dr Dobbs Journal* 13.7, 46. 780 781
- Reiter, Nils (July 2015). “Towards Annotating Narrative Segments”. In: *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Beijing, China: Association for Computational Linguistics, 34–38. [10.18653/v1/W15-3705](https://aclanthology.org/W15-3705). <https://aclanthology.org/W15-3705>. 782 783 784 785
- Reiter, Nils, Judith Sieker, Svenja Guhr, Evelyn Gius, and Sina Zarrieß (June 2022). “Exploring Text Recombination for Automatic Narrative Level Detection”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 3346–3353. <https://aclanthology.org/2022.lrec-1.357>. 786 787 788 789 790
- Riedl, Martin and Chris Biemann (July 2012). “TopicTiling: A Text Segmentation Algorithm based on LDA”. In: *Proceedings of ACL 2012 Student Research Workshop*. Jeju Island, Korea: Association for Computational Linguistics, 37–42. <https://aclanthology.org/W12-3307>. 791 792 793 794
- Schöch, Christof (2021). “Topic modeling genre: An exploration of French classical and enlightenment drama”. In: *arXiv preprint arXiv:2103.13019*. 795 796
- Shaham, Uri, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy (2022). *SCROLLS: Standardized Comparision Over Long Language Sequences*. arXiv: [2201.03533 \[cs.CL\]](https://arxiv.org/abs/2201.03533). 797 798 799
- Somasundaran, Swapna et al. (2020). “Two-Level Transformer and Auxiliary Coherence Modeling for Improved Text Segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, 7797–7804. 800 801 802
- Sultana, Sharifa, Renwen Zhang, Hajin Lim, and Maria Antoniak (2022). “Narrative Datasets through the Lenses of NLP and HCI”. In: https://maria-antoniak.github.io/resources/2022_nlp_hci_narratives.pdf. 803 804 805
- Uglanova, Inna and Evelyn Gius (2020). “The Order of Things. A Study on Topic Modelling of Literary Texts.” In: *CHR* 18-20, 2020. 806 807

- Wagner, Eitan, Renana Keydar, Amit Pinchevski, and Omri Abend (Dec. 2022). “Topical Segmentation of Spoken Narratives: A Test Case on Holocaust Survivor Testimonies”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 6809–6821. <https://aclanthology.org/2022.emnlp-main.457>.
- Wang, Haoyu, Hongming Zhang, Muhao Chen, and Dan Roth (Nov. 2021). “Learning Constraints and Descriptive Segmentation for Subevent Detection”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 5216–5226. <https://aclanthology.org/2021.emnlp-main.423>.
- Wilmot, David and Frank Keller (July 2020). “Modelling Suspense in Short Stories as Uncertainty Reduction over Neural Representation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 1763–1788. 10.18653/v1/2020.acl-main.161. <https://aclanthology.org/2020.acl-main.161>.
- (2021). *A Temporal Variational Model for Story Generation*. 10.48550/ARXIV.2109.06807. <https://arxiv.org/abs/2109.06807>.
- Wu, Jeff, Long Ouyang, Daniel M Ziegler, Nissan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano (2021). “Recursively Summarizing Books with Human Feedback”. In: *arXiv preprint arXiv:2109.10862*.
- Wu, Qiong, Adam Hare, Sirui Wang, Yuwei Tu, Zhenming Liu, Christopher G Brinton, and Yanhua Li (2020). “BATS: A Spectral Biclustering Approach to Single Document Topic Modeling and Segmentation”. In: *arXiv preprint arXiv:2008.02218*.
- Zehe, Albin, Leonard Konle, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, Anneke Schreiber, and Nathalie Wiedmer (Apr. 2021). “Detecting Scenes in Fiction: A new Segmentation Task”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, 3167–3177. <https://aclanthology.org/2021.eacl-main.276>.
- Zhai, Fangzhou, Vera Demberg, Pavel Shkadzko, Wei Shi, and Asad Sayeed (Aug. 2019). “A Hybrid Model for Globally Coherent Story Generation”. In: *Proceedings of the Second Workshop on Storytelling*. Florence, Italy: Association for Computational Linguistics, 34–45. 10.18653/v1/W19-3404. <https://aclanthology.org/W19-3404>.

InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline

Kai Kugler¹ 
Simon Munker¹ 
Johannes Höhmann¹
Achim Rettinger¹ 

1. Computational Linguistics & Digital Humanities, University of Trier , Trier, Germany.

Citation

Kai Kugler, Simon Munker, Johannes Höhmann, and Achim Rettinger (2023). "InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-21

Date received 2023-02-17

Keywords

contextualized word embeddings, derived text formats, text reconstruction, transformer encoder, publication restrictions

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract.

Digital Humanities and Computational Literary Studies apply automated methods to enable studies on large corpora which are not feasible by manual inspection alone.

However, due to copyright restrictions, the availability of relevant digitized literary works is limited. Derived Text Formats (DTFs) have been proposed as a solution. Here, textual materials are transformed in such a way that copyright-critical features are removed, but that the use of certain analytical methods remains possible. Contextualized word embeddings produced by transformer-encoders are promising candidates for DTFs because they allow for state-of-the-art performance on analytical tasks. However, in this paper we demonstrate that under certain conditions the reconstruction of the original text from token representations becomes feasible. Our attempts to invert BERT suggest, that publishing the encoder together with the contextualized embeddings is unsafe, since it allows to generate data to train a decoder with a reconstruction accuracy sufficient to violate copyright laws.

1. Introduction

Due to copyright laws the availability of more recent text material (specifically literary works from the last 100 years) for scientific analyses is quite limited. For disciplines such as Computational Literary Studies (CLS), these legal restrictions make research on contemporary literature difficult because the relevant primary texts may not be published with the research results (e.g. to enable follow-up research on the reusable data), as current principles of scientific data management demand (Wilkinson et al. 2016), however. Depending on national law there might be some degree of freedom to use protected texts for scientific studies and give reviewers access to them, but in most cases they still can't be published fully, making it hard for the research community to reproduce or build on scientific findings. According to German case law, for example, there are now possibilities for making copyright-protected material accessible in the context of scientific research, e.g. for the peer review process¹, but these exceptions are

1. see § 60d UrhG https://www.gesetze-im-internet.de/urhg/_60d.html

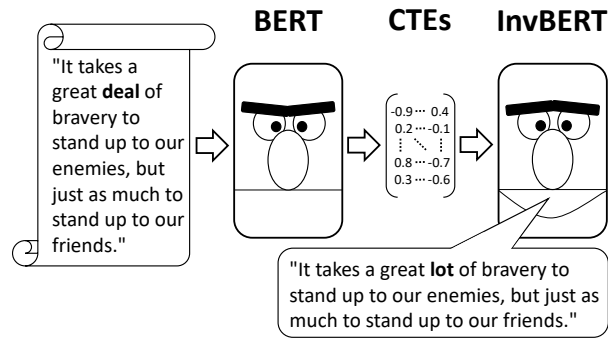


Figure 1: Sample text reconstruction to a Harry Potter quote Rowling (1998) by inverting BERT.

so narrowly defined that the corpora are no longer available for follow-up research. 14

This is a fundamental issue for research in Digital Humanities (DH) and Computational 15
Literary Studies (CLS), but applies also to any analysis of text documents that cannot 16
be made available due to privacy reasons, copyright restrictions or business interests. 17
This, for instance, makes it hard for digital libraries to offer their core service, which 18
is the best possible access to their content. While they provide creative compromise 19
solutions, like *data capsules* or *web-based analysis tools*², such access is always limited and 20
complicates subsequent use and reproducibility. 21

As a consequence, there have been attempts to find a representation formalism which 22
retains as much linguistic information as possible while not disclosing the original text 23
fully. Such text representations have been referred to as Derived Text Formats (DTFs) 24
(Schöch et al. 2020a). While such DTFs are always a compromise between the degree of 25
obfuscation (non-reconstructibility) and degree of analyzability (retained information), 26
there are DTFs with clear advantages over others. In the end, they should always be 27
more informative than not publishing the documents at all. 28

We investigate if Contextualized Token Embeddings (CTE), like the ones obtained 29
from a transformer encoder stack trained on a self-supervised masked language mod- 30
eling (MLM) task (Devlin et al. 2019), are a promising candidate for DTFs. On the 31
one hand, they are the state-of-the-art text representation for most Natural Language 32
Understanding tasks (Wang et al. 2019a,b), including tasks relevant to DH and CLS, 33
like text classification, sentiment analysis, authorship attribution or text re-use (Schöch 34
et al. 2020b). On the other hand, it appears difficult to reconstruct the original text just 35
from its CTEs because, unlike (static) word embeddings, there is no fixed inventory 36
of representations that does not change from sentence to sentence. Thus, we pose the 37
following research question: 38

In which scenarios can protected text documents be released publicly if 39
encoded as contextualized embeddings without the original content being 40
reconstructed to an extent that the publication violates copyright laws? 41

After presenting related work (Sec. 2) we will first formalize different reconstruction 42
scenarios, which allow us to define potential lines of attack that aim at reconstructing 43
the original text (Sec. 3). Next, we will discuss the feasibility of each line of attack. 44

2. see https://www.hathitrust.org/htrc_access_use

In Sec. 4 we focus on the most promising lines of attack by evaluating their feasibility empirically (Sec. 5), before concluding in Sec. 6.

2. Related Work

First, we look at the very recent field of DTFs, before presenting existing work on text reconstruction beyond copyright protected texts.

2.1 Derived Text Formats

DTFs, like n-grams or term-document matrices are an important tool to the Digital Humanities and Computational Linguistics, since they allow the application of quantitative methods to their research objects. However, they have another important advantage: If the publication of an original text is prohibited, DTFs may still enable reproducibility of research (Schöch et al. 2020a,b). This is especially important for CLS, where there is only a small “window of opportunity” of available manuscripts from the year 1800 to 1920 due to technical issues on the lower and copyright restrictions on the upper boundary. Since this is of permanent concern and an obstacle to open science, tools to widen this window are of great importance to the field. Other approaches to tackle this issue, like granting access to protected texts in a closed room setting, come with their own major drawbacks and still do not enable an unhindered exchange of scientific findings. Therefore, in most cases, DTFs like term-document matrices are the best solution available. The aim of these formats is to retain as much information as possible, while minimizing reconstructibility. In reality, however, the latter most often is achieved by compromising on the former. This leads to the variety of feasible analytical down-stream tasks being narrowed. A format that preserves a noticeable amount of information and is already used as a DTF are word embeddings like Word2Vec (Mikolov et al. 2013) or GloVe (Pennington et al. 2014). However, similar to term-document matrices they can only be applied to document-level tasks. Otherwise, there remains considerable doubt regarding their resilience against reconstruction attempts. A promising attempt to alleviate that is by using contextualized word - or more precise token - embeddings (CTEs) generated by pretrained language models instead, since the search space to identify a token grows exponentially with the length of the sequence containing it. Additionally, these embeddings carry even more, especially lexical semantic information (Vulic et al. 2020) and achieve SOTA results on various down-stream tasks.

2.2 Reconstruction of Information from Contextualized Embeddings

Recently, attention was drawn to privacy and security concerns regarding large language models due to prominent voices in ethics in AI (Bender et al. 2021), as well as a collaborative publication of the industry giants Google, OpenAI and Apple (Carlini et al. 2021). In the latter, the authors demonstrated, that these models memorize training data to such an extent, that it is not only possible to test whether the training data contained a given sequence (membership inference, (Shokri et al. 2017)), but also to directly query samples from it (training data extraction). Other recent research supports these findings and agrees that this problem is not simply caused by overfitting (Song and Shmatikov 2019; Thomas et al. 2020). Large language models like GPT-3

(Brown et al. 2020) or T5 (Raffel et al. 2020) were trained on almost the entirety of the available web, which poses a special concern, since sensitive information like social security numbers is unintentionally being included. Hence, a majority of the literature focuses on retrieving information about the training data. However, we argue that such attacks are less successful in the case of literary works, since a) the goal in this scenario would usually be the reconstruction of a specific work, and b) the attacks are not suited to recover more than isolated sequences.

A third prominent type of attack which can be performed quite effectively and reveals some information about training data is attribute inference (Mahloujifar et al. 2021; Melis et al. 2019; Song and Raghunathan 2020). It is also of little relevance, since it aims to infer information like authorship from the embeddings, which is non-confidential in a DTF setting anyways. More so, authorship attribution is actually a relevant field of research in the DHs.

The main threat regarding CTEs as DTFs are embedding inversion attacks, where the goal is the reconstruction of the original textual work they represent. However, research on this topic is still limited and most papers focus on privacy. Therefore, very few go beyond retrieval of isolated sensitive information. E.g. Pan et al. (2020) showed, that it is possible to use pattern-recognition and key-word-inference techniques to identify content with fixed format (e.g. birth dates) or specific keywords (e.g. disease sites) with varying degree of success (up to 62% and above 75% avg. precision respectively). However, this is easier and the search space smaller, than reconstructing full sequences drawn from the whole vocabulary.

To the best of our knowledge, retrieval of the full original text is covered only by Song and Raghunathan (2020). Using an RNN with multi-set prediction loss in a setting with access to the encoding model as a black-box, they were able to achieve an in-domain F1 score of 59.76 on BERT embeddings. However, since privacy was their concern, they did not consider word ordering in their evaluation, which is crucial when dealing with literary works. Therefore, and since they failed to improve on their results using a white-box approach as well, we believe that the security of the usage of CTEs as DTFs still remains an unanswered question.

When dealing with partial-white- or black-box scenarios, a final type of attack should be kept in mind: Inferences about the model itself. Even though not the goal here, successful model extraction attacks (Krishna et al. 2020) may transform a black-box situation into a white-box case. However, critical information can even be revealed by fairly easy procedures like model fingerprinting. This was showcased on eight SOTA models by Song and Raghunathan (2020), who were able to identify the model based on a respective embedding with 100% accuracy.

3. Reconstruction Task and Attack Vectors

This paper is not about improving or applying transformers, but inverting them. To introduce a reconstruction model (cmp. Rigaki and Garcia 2020) we first describe scenarios for possible attacks. Then, we lay out different attack vectors based on the scenarios.

3.1 Reconstruction Scenarios

Formally, the reconstruction scenarios can be defined as follows:

Given: Contextualized token embeddings CTEs of a copyright protected literary document W (typically a book, containing literary works, like poetry, prose or drama) are made available in every scenario. Depending on the scenario additional information is available:

WB - White Box Scenario: The most flexible scenario is given if the encoder $enc()$, including the neural network's architecture and learned parameters, and tokenizer $tok()$ are made openly available in addition to the CTEs. In this case, analytical experiments can be conducted by DH researchers that require to adapt/optimize the encoder $enc()$ and/or the tokenizer $tok()$.

BB - Black Box Scenario: A scenario with little flexibility from the perspective of a DH researcher is given, when the tokenizer $tok()$ and the encoder $enc()$ are made available as one single opaque function and are only accessible for generating mappings from W to CTE. A similar scenario arises if ground truth training data is available (i.e., aligned pairs of W s to CTEs are given). Then the researcher is still able to label his own training data and use it to optimize $enc()$ or embed other data not yet available as CTEs for analysis. However, if provided as a service, the number of queries allowed to be sent to $enc()$ might be limited up to a point where the model is not released at all. Then, existing implementations can be reused in order to perform a standard analytical task if the respective task-specific top layer function is also provided. Note, that BB can be turned into WB by successful model extraction attacks.

GB - Gray Box Scenario: If the encoder-transformer pipeline $tok()$ and $enc()$ used for generating CTEs is available to some degree (e.g., the tokenizer is given) we refer to it as a Gray Box (GB) scenario.

Searched: A function or algorithm $inv(CTE) = \hat{W}$ that inverts the model pipeline or approximate its inverse and outputs reconstructed text \hat{W} from CTEs.

3.2 Inversion Attacks

We consider three lines of attack:

Inverting Functions: Inverting $enc()$ and $tok()$ using calculus requires an attacker to find a closed-form expression for $tok^{-1}()$ and $enc^{-1}()$. Since this requires knowledge of the parameters of the encoder pipeline, this is only applicable to a WB scenario. Even then, this approach would only be feasible if all functions in question are invertible which is not the case for BERT-like transformer-encoder stacks.

Exhaustive Search: Sentence-by-sentence combinatorial testing of automatically generated input sequences to "guess" the contextualized token embeddings would be applicable to WB, GB and BB, as long as an unlimited number of queries to $enc()$ is allowed. However, combinatorial explosion renders this approach infeasible: A sentence of 15 tokens results in $18 \cdot 10^6$ possible combinations, assuming a vocabulary size of 30,522 different tokens, like in the case of BERT_{BASE}.

Machine Learning: Learning an approximation of $tok^{-1}(enc^{-1}())$ can be attempted as soon as training samples are available or can be generated. We assume that an attack is more likely to be successful if components of the embedding generating pipeline are accessible, because in a GB scenario the components can be estimated separately, reducing the complexity compared to an end-to-end BB scenario.

Since a successful BB attack works equally well in a GB scenario and a successful GB attack works in a WB scenario we restrict our empirical investigation to two machine learning based attacks, one for a GB, where $tok()$ is given and one for the BB scenario. We call our GB attack *InvBert Classify* and our BB attack *InvBert Seq2Seq*. Both models are detailed in Fig. 2 and described in the next section.

4. Experimental Design

In this Section, we describe two attack models, one for a GB and one for a BB scenario, introduced in Section 3. First, we introduce and discuss the datasets. Next, we explain both neural network structures and the general attack pipeline. The code and datasets are publicly available as a Github repository (see Sec. 8 and 9).

4.1 Data

As a data basis, we have chosen two text corpora that fulfil three conditions: First, the texts of the corpora should be similar to the protected works that are to be distributed in DTF. We restrict ourselves to English language prose texts and choose the corpora accordingly. Secondly, the corpora must be big enough to draw data sets of a size that allow models to be trained successfully. Furthermore, it is important to us that our results are reproducible, which is why we have chosen openly available data.

First, we scraped the *Archive of Our Own* (AO3)³, an openly available fanfiction repository, using a modified version of AO3Scraper⁴. During the preprocessing, we filtered out mature, extreme, and non-general audience content using the given tags. We split the AO3 data into the following three topics (based on the ten most common tags *Action*, *Drama* and *Fluff*⁵) to get different samples. Table 1 shows the exact size and number of samples of each subset.

Before training our models, the datasets were each split into non-overlapping training and evaluation datasets. Training is performed on the complete training data set (100%) or on a subset of this data (10%, 1%, 0.1%) to assess how the amount of training data affects the text reconstruction ability of the models.

As fanfiction mostly resembles contemporary literature, we gathered a fourth dataset from Project Gutenberg⁶, a non-commercial platform but with a focus on archiving and distributing historical literature, including western novels, poetry, short stories, and drama. Consequently, our Gutenberg train/eval set contains a mix of different genres. This data set shows a slightly higher Type-Token-Ratio indicating a higher

3. <https://archiveofourown.org>

4. <https://github.com/radiolarian/AO3Scraper>

5. "Feel good" fan fiction designed to be happy, and nothing else, according to https://en.wikipedia.org/wiki/Fan_fiction

6. <https://www.gutenberg.org/>

Name	Filesize	Tokens	Unique	Ratio
Action	372.97 MB	72,086,159	152,847	0.002120
Drama	304.51 MB	59,133,691	136,759	0.002313
Fluff	327.80 MB	64,002,888	144,492	0.002258
Gutenberg	257.94 MB	48,807,783	173,716	0.003559

Table 1: Size and number of contained training samples of the collected data sets, number of unique tokens (types) and Type-Token-Ratio.

lexical variation in contrast to the AO3 datasets (see Table 1). Gutenberg’s content is sorted by *bookshelves*, we have selected prose genres in Modern English (Classics, Fiction, Adventure etc.) not removing any metadata.

4.2 Models & Pipelines

In Sec. 3 we argued that machine learning models are promising candidates for inversion attacks. We propose two models, one for a GB and one for a BB scenario:

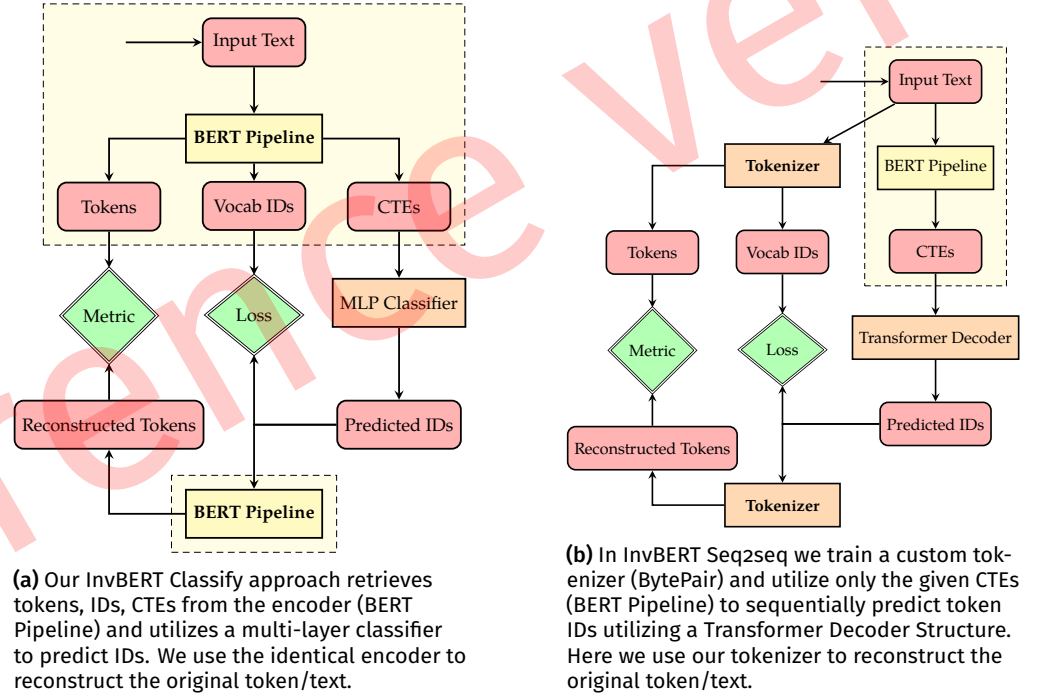


Figure 2: Flowchart for each approach. Givens are enclosed in a dotted yellow area and attack-specific modules to be estimated are filled with orange. Data objects are highlighted in red, while green represent the evaluation/objective function.

InvBERT Classify (GB): Here, we have access to the CTEs and the tokenizer $tok()$. As the tokenizer is a look-up table, which can be queried from both directions, the inverse $tok^{-1}()$ to $tok()$ is also provided, effectively simplifying the problem of finding an approximation of the inverse $tok^{-1}(enc^{-1}())$ of the whole pipeline to just $enc^{-1}()$. We train a multi-layer perceptron to predict the vocabulary IDs given CTEs. As we use the given tokenizer, CTEs and IDs have a one-to-one mapping, and our attack boils down to a high-dimensional token classification task.

InvBERT Seq2Seq (BB): Here, we only have access to the CTEs. Without the tokenizer,

we lose the one-to-one mapping and cannot infer the token CTE ratio. Thus, we
have to train a custom tokenizer and optimize a transformer decoder structure to
predict our sequence of custom input IDs. The decoder utilizes complete sentence
CTEs as generator memory and predicts each token ID sequentially.

We use the *Hugging Face API*⁷ to construct a batch-enabled BERT Pipeline capable of
encoding plain text into CTEs and decoding (sub-)token IDs into words. All param-
eters inside the pipeline are disabled for gradient optimization. Our models and the
training/evaluation routine are based on *PyTorch modules*⁸. We utilize AdamW as an
optimizer and the basic cross-entropy loss. Our model implementations have $\sim 24M$
(InvBert Classify) and $\sim 93M$ (InvBERT Seq2Seq) trainable parameters. We train on a
single Tesla V100-PCIe-32GB GPU and do not perform any hyperparameter optimiza-
tion. Further, we use in each type of attack the identical hyperparameter settings to
ensure the highest possible comparability.⁹ A training epoch for a model takes up to 8
hours depending on the dataset and type of attack.

4.3 Evaluation Metrics

We evaluate the 3-gram, 4-gram, and sentence precision in addition to the BLEU metric
(Papineni et al. 2002). The objective of our model is to reconstruct the given input as
closely as possible. BLEU defines our lower bound in terms of precision, as it is based on
n-gram precision allowing inaccurate sentences with matching sub-sequences. Since the
BLEU metric might be too imprecise to quantify if a reconstruction captures the content
of a sentence and style of the author, we preferred to use complete sentence accuracy
in our quantitative evaluation. There, we only count perfectly correct reconstructions,
resulting in a significantly higher bound in contrast to BLEU. While the BLEU metric
can give us an indication of how closely the reconstruction candidate resembles the
original text, we consider correct reconstructions to be a clear sign that possible copyright
violations are imminent when publishing.

5. Empirical Results

In this section we will first present our qualitative results, before showing some examples
of different reconstruction results.

5.1 Quantitative Evaluation

We quantitatively evaluated the trained models in-domain by calculating their sentence
accuracy over all samples of their corresponding test set. Additionally, we conducted
the reconstruction across all other evaluation datasets to measure the out-of-domain
performance. A condensed representation of our in-domain results is presented in Fig. 3,
while the full results are included in appendix 7.

The InvBERT Classify model achieves a very high in-domain as well as out-domain
sentence reconstruction accuracy when trained on 100% and 10% of the training data-

7. <https://huggingface.co>

8. <https://pytorch.org>

9. The parameters used for the experiments can be found in the configuration files of the repository

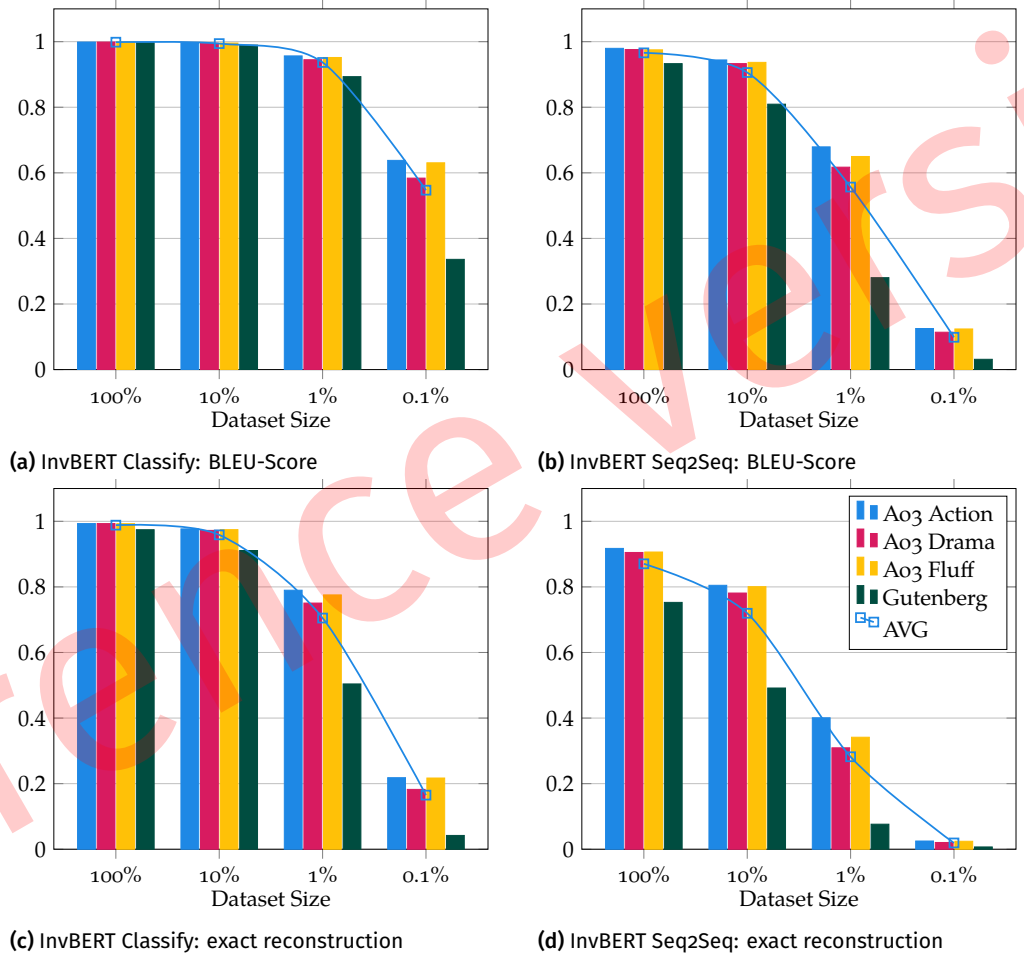


Figure 3: Both reconstruction approaches compared by their in-domain BLEU score (a), (b) and exact sentence reconstruction accuracy (c), (d) on the evaluation data sets.

set. Thus, we can reconstruct around $\approx 97\%$ of the original content without errors. 257
 Even when just utilizing 1% of the training datasets, our model scores $\approx 65\%$ sentence 258
 reconstruction accuracy. This likely still is enough to violate copyright laws since the 259
 remaining 35% of sentences get very close to the originals. In comparison, the sentences 260
 generated with a model trained on only 0.1% of the data no longer resemble the original 261
 input data. 262

The consistently high BLEU-scores achieved by both models, even with smaller data 263
 sets (BLEU-scores $> 60\%$ even if only 1% of the data is used), indicate that the text 264
 reconstructions are very close to the original text and that perhaps just individual tokens 265
 could not be reconstructed exactly. 266

We observe that the performance on the samples of the AO3 dataset is very consistent. 267
 The performance considerably drops on the Gutenberg corpus. We assume that the 268
 more heterogeneous content in combination with input shuffling during training yields 269
 a more challenging data set than our AO3 crawl. In particular, the smaller the train 270
 subsets, the smaller the number of samples of a certain genre inside our Gutenberg 271
 corpus. Additionally, the Gutenberg corpus contains noise like metadata and unique 272
 tokens in the form of title pages and table of contents which we did not clean. The 273
 differences are negligible when using 100% or 10% of the training data set, but become 274
 clear on 1% or 0.1% train data usage, where the accuracy differs around 20%. 275

The InvBERT Seq2Seq2 model reaches slightly worse results while also being much more 276
 sensitive to the training data size and the type of dataset. This is to be expected since 277
 this approach utilizes a more complex network architecture that sequentially predicts 278
 the reconstruction parts. We attribute the differences to the more complex task and the 279
 higher number of trainable parameters. 280

5.2 Qualitative Evaluation 281

To put our previously made assumption about their reconstruction quality to the test, 282
 we applied our models to 15 quotes from the Harry Potter book series.¹⁰ The calculated 283
 metrics in Table 2 show that the performance on these real-world examples are consistent 284
 with the quantitative results on our test data.¹¹ 285

InvBERT Classify completely reconstructs the samples when trained on 100% or 10% of 286
 the training dataset. Only when using 1% or 0.1% of the train data, the model predict 287
 false but semantically similar content. Contrary, InvBERT Seq2Seq starts to produce 288
 substantial errors in its reconstruction while using 10% of the train data, and with less 289
 data, the predictions do not resemble a reasonable reconstruction attempt neither on 290
 the syntactic nor semantic level. 291

5.3 Discussion 292

Our exemplary manual evaluation corroborates the results from our quantitative ex- 293
 periments. Both attacks can, if enough data is available, successfully reconstruct the 294
 original content. In conclusion, according to our assessment, all scenarios (WB, GB 295

10. Retrieved from <https://mashable.com/article/best-harry-potter-quotes>.

11. Reconstructions of the 15 quotes by all 32 models trained on the different data sets of different sizes can be found in the repository provided, in the respective logfiles of the models.

SRC:	if you want to know what a man's like, take a good look at how he treats his inferiors, not his equals.	i'll just go down and have some pudding and wait for it all to turn up ... it always does in the end.
-------------	---	---

InvBERT Classify

100%	<i>exact reconstruction</i>	<i>exact reconstruction</i>
10%	<i>exact reconstruction</i>	<i>exact reconstruction</i>
1%	if you want to know what a man's like, take a good look at how he treats his subordinates , not his equals.	<i>exact reconstruction</i>
0.1%	if you want to know what a man's like, take a good look at how he treat his enemies , not his friends .	i'll just go down and have some dinner and wait for it all to come up ... it always does in the end.

InvBERT Seq2seq

100%	<i>exact reconstruction</i>	<i>exact reconstruction</i>
10%	if you want to know what a man's like, take a good look at how he treats his inferior tors , not his equals.	<i>exact reconstruction</i>
1%	if you want to know what a man's like, take a good look at his partners , not his partners .	i'll just go down and wait for some chocolate and wait for it all to turn up in the end ... it always does in the end.
0.1%	if you have a little to get a little, but you're a little look at him, not like he're a little look.	i'll go and get up up the rest of the rest, it just just just have been going to get up.

Table 2: Example of Harry Potter quotes Rowling 2006 and their predictions. Differences are highlighted: **red** as error and **yellow** as false, but semantically acceptable. 'exact reconstruction' represents identical reproduction.

and BB) cannot be considered safe. Even in the “safest” BB scenario without a given tokenizer, reconstruction is feasible.

Collecting training data has proven to be very easy, as there are many corpora available digitally that are sufficiently similar to modern English-language texts. The word order information that BERT can extract from this data is apparently sufficient to reconstruct texts from CTEs derived from texts that are not allowed to be published.

Thus, copyright violations are imminent when publishing CTEs as DTFs.

6. Conclusion and Future Work

To conclude, we first summarize our contributions and findings, before outlining open research questions.

6.1 Summary and Conclusion:

Derived Text Formats (DTFs) are an important topic in Digital Humanities (DH). There, the proposed DTFs rely on deleting important information from the text, e.g., by using term-document matrices or paragraph-wise randomising of word order. In contrast, Contextualized Token Embeddings (CTEs), as produced by modern language models, are superior in retaining syntactic and semantic information of the original documents. However, the use of CTEs for large-scale publishing of copyright protected works as DTFs is hindered by the risk that the original texts can be reconstructed.

In this paper we first identify and describe typical scenarios in DH when analyzing text

using CTEs is helpful to different degrees. Next, we list potential attacks to recover the original texts. We theoretically and empirically investigate what attack can be applied in which scenario.

Our findings suggest, that if a certain number of training instances (known mappings of sequences of CTEs produced by the encoder to the original sentences) are given or can be obtained it is not save to publish CTEs. Even the safest BB scenario that we covered in this paper is not resistant against reconstruction attacks. Consequently, all GB and WB scenarios are even more vulnerable.

6.2 Future Work:

While researchers in DH have to judge the usefulness of CTEs as DTFs, finding a copyright protected way of publishing content is also relevant for the field of Natural Language Processing (NLP) in general. There, CTEs have only been investigated in regarding privacy risks, but not copyright protection. After all, the problem of reproducibility of scientific results from restricted corpora is not limited to the DHs.

However we encourage to establish a novel research niche, the focus of this paper is to define the task of reconstructing text from CTEs of literary works. Accordingly, we only covered the most obvious lines of attack, there are more scenarios that require additional investigation.

Another potential scenario that has not been discussed in this paper is the publication of CTEs without any (means to generate) training data. Although this scenario seems conceivable, there are practical reasons that virtually rule it out: First, to be of any value for DH researchers, the bibliographic meta data (author, title, ...) of the literary work has to be published along with the CTEs. In addition, the rich information encoded in CTEs (e.g., compared to a bag-of-words representation) is more likely to be useful when used in conjunction with more detailed information such as sentence boundaries. Second, ensuring that no training data can be obtained from a released sequence of CTEs seems only feasible in very special cases. If (parts of) the literary works in the corpus can be obtained in a digitized format through other means, it might be possible to align them with the sequence of CTEs and generate a training set. How sentences can be aligned remains the key research challenge in such a scenario, but as soon as an alignment can be established it becomes an invertible BB scenario.

Also, there is the question of finding a compromise scenario where the complete sequence of CTEs is not published or noise is added, as it has been done with DTFs. Examples are shuffling the sequence, random deletion of a portion of the CTEs, or representation of certain CTEs by linguistic features. What benefits CTEs provide in such scenarios is also a question for future research.

While we covered the most obvious lines of attack in this paper, there are more scenarios that require additional investigations: Potential combinations of different DTFs or meta-data might allow new lines of attack, for instance, if n-grams plus CTEs are published for the same text. Also, a mapping between the used embedding and a different embedding, based on the incorporated linguistic information they share, might be possible.

CTEs generated by more modern language models than BERT are also of interest for

future research. These models still keep growing in size and capabilities and so do the complexity of the CTEs they generate. It is to be investigated whether texts represented by these embeddings can still be reconstructed using approaches like ours, but we assume that it is rather a matter of scaling the reconstruction model accordingly, than rendering our general approach infeasible.

Opposite to the attack perspective, an open research question is, if there are novel types of DTFs, beyond CTEs, which are more expressive and more safe.

Another related issue that we did not discuss, is the suitability of quantitative metrics for measuring copyright violations. Ultimately, it is a legal consideration, if a reconstruction accuracy, e.g., above a certain BLEU-score, violates copyright laws. This is beyond the scope of this paper.

Ultimately, publishers and libraries need to decide if they release DTFs of their inventory. However, based on our findings we advise against it, since it is likely that training samples might be obtained. Still, we believe that more research is needed to find compromise solutions that balance usefulness while ensuring safety from reconstruction. What contribution CTEs can provide is still an open question.

For researchers, this is an exciting challenge, since it requires both, theoretical studies regarding computational complexity, but also empirical experiments with real-word corpora in real-world settings.

7. Supplementary Material

8. Data Availability

The AO₃ corpus cannot be made available, for the same copyright reasons discussed in this paper. However, it can be recrawled to replicate our experiments. The Gutenberg corpus is freely downloadable and usable: <https://anonymous.4open.science/r/invbert-BF31>.

9. Software Availability

The code to replicate our findings is available on GitHub, once the paper is accepted (during review as an anonymous repository): <https://anonymous.4open.science/r/invbert-BF31>.

10. Author Contributions

Kai Kugler: Conceptualization, Data curation, Validation, Writing

Simon Münker: Methodology, Software, Formal Analysis, Visualization, Writing

Johannes Höhmann: Methodology, Visualization, Writing

References 392

- Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 393-395
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prithvi Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. 400-405
- Carlini, Nicholas, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel (2021). "Extracting Training Data from Large Language Models". In: *USENIX Security Symposium*. 406-409
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 410-414
- Krishna, Kalpesh, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer (2020). "Thieves on Sesame Street! Model Extraction of BERT-based APIs". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. 415-418
- Mahloujifar, Saeed, Huseyin A. Inan, Melissa Chase, Esha Ghosh, and Marcello Hasegawa (2021). "Membership Inference on Word Embedding and Beyond". In: *ArXiv abs/2106.12384*. 419-424
- Melis, Luca, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov (2019). "Exploiting unintended feature leakage in collaborative learning". In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 422-423
- Mikolov, Tomáš, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 424-427
- Pan, Xudong, Mi Zhang, Shouling Ji, and Min Yang (2020). "Privacy risks of general-purpose language models". In: *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE. 428-430
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 431-433

- Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 434–436.
- Raffel, Colin, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *ArXiv abs/1910.10683*. 437–439.
- Rigaki, Maria and Sebastian Garcia (2020). “A Survey of Privacy Attacks in Machine Learning”. In: *arXiv: 2007.07646 [cs.CR]*. 440–441.
- Rowling, J. K. (1998). *Harry Potter and the Sorcerer’s Stone*. 442.
- (2006). *Harry Potter and the Half-Blood Prince*. 443.
- Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke (2020a). “Abgeleitete Textformate: Prinzip und Beispiele”. In: *RuZ-Recht und Zugang* 1.2. 444–446.
- (2020b). “Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen”. In: *Zeitschrift für digitale Geisteswissenschaften*. 447–448.
- Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov (2017). “Membership inference attacks against machine learning models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 449–451.
- Song, Congzheng and Ananth Raghunathan (2020). “Information leakage in embedding models”. In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. 452–454.
- Song, Congzheng and Vitaly Shmatikov (2019). “Auditing Data Provenance in Text-Generation Models”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. Ed. by Ankur Teredesai, Vipin Kumar, Ying Li, Römer Rosales, Evimaria Terzi, and George Karypis. 455–459.
- Thomas, Aleena, David Ifeoluwa Adelani, Ali Davody, Aditya Mogadala, and Dietrich Klakow (2020). “Investigating the Impact of Pre-Trained Word Embeddings on Memorization in Neural Networks”. In: *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings*. Brno, Czech Republic: Springer-Verlag, 273–281. ISBN: 978-3-030-58322-4. 460–464.
- Vulic, Ivan, Edoardo Maria Ponti, Robert Litschko, Goran Glavas, and Anna Korhonen (2020). “Probing Pretrained Language Models for Lexical Semantics”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics, 7222–7240. [10.18653/v1/2020.emnlp-main.586](#). 465–470.
- Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019a). “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett. 471–475.
- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019b). “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural 476–479.

- Language Understanding". In: *7th International Conference on Learning Representations*, 480
ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. 481
- Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, 482
Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva 483
Santos, Philip E Bourne, et al. (2016). "The FAIR Guiding Principles for scientific 484
data management and stewardship". In: *Scientific data* 3. 485

Dataset	Precision				Precision				Precision			
	BLEU	3-gram	4-gram	Sent	BLEU	3-gram	4-gram	Sent	BLEU	3-gram	4-gram	Sent
	action _{100%: 5,903,010 lines}				action _{10%: 590,818 lines}				action _{1%: 59,330 lines}			
Action	0.9989	0.9986	0.9982	0.9931	0.9961	0.9954	0.9936	0.9769	0.9566	0.9479	0.9298	0.7899
Drama	0.9982	0.9979	0.9971	0.9889	0.9945	0.9934	0.9910	0.9668	0.9516	0.9418	0.9222	0.7702
Fluff	0.9982	0.9978	0.9970	0.9889	0.9943	0.9932	0.9908	0.9668	0.9541	0.9448	0.9262	0.7776
Gutenberg	0.9758	0.9709	0.9615	0.8693	0.9585	0.9502	0.9345	0.7842	0.8509	0.8229	0.7732	0.4430
	drama _{100%: 4,854,969 lines}				drama _{10%: 484,128 lines}				drama _{1%: 48,569 lines}			
Action	0.9981	0.9977	0.9977	0.9884	0.9933	0.9920	0.9891	0.9601	0.9334	0.9201	0.8938	0.7041
Drama	0.9989	0.9986	0.9986	0.9931	0.9954	0.9944	0.9924	0.9723	0.9449	0.9339	0.9115	0.7507
Fluff	0.9981	0.9977	0.9977	0.9884	0.9936	0.9922	0.9895	0.9619	0.9407	0.9288	0.9054	0.7289
Gutenberg	0.9763	0.9716	0.9716	0.8725	0.9563	0.9476	0.9310	0.7710	0.8273	0.7956	0.7395	0.3986
	fluff _{100%: 5,251,248 lines}				fluff _{10%: 524,322 lines}				fluff _{1%: 52,696 lines}			
Action	0.9972	0.9966	0.9954	0.9827	0.9912	0.9894	0.9856	0.9480	0.9280	0.9137	0.8853	0.6876
Drama	0.9973	0.9967	0.9956	0.9831	0.9914	0.9896	0.9859	0.9491	0.9328	0.9193	0.8928	0.7061
Fluff	0.9988	0.9986	0.9981	0.9928	0.9958	0.9949	0.9930	0.9746	0.9518	0.9421	0.9224	0.7756
Gutenberg	0.9726	0.9671	0.9564	0.8453	0.9495	0.9393	0.9201	0.7305	0.8126	0.7784	0.7186	0.3722
	gutenberg _{100%: 2,728,188 lines}				gutenberg _{10%: 273,263 lines}				gutenberg _{1%: 27,240 lines}			
Action	0.9833	0.9798	0.9728	0.9007	0.9651	0.9579	0.9438	0.8104	0.8460	0.8163	0.7623	0.4446
Drama	0.9846	0.9814	0.9750	0.9089	0.9676	0.9608	0.9477	0.8223	0.8561	0.8280	0.7773	0.4686
Fluff	0.9806	0.9766	0.9687	0.8878	0.9616	0.9536	0.9383	0.7947	0.8451	0.8152	0.7618	0.4433
Gutenberg	0.9972	0.9966	0.9955	0.9744	0.9894	0.9873	0.9830	0.9110	0.8933	0.8727	0.8346	0.5041
	gutenberg _{0.1%: 2,755 lines}											
Action									0.3440	0.2685	0.1707	0.0496
Drama									0.3551	0.2793	0.1799	0.0525
Fluff									0.3510	0.2759	0.1765	0.0503
Gutenberg									0.3362	0.2627	0.1693	0.0421

Table 3: InVbERT Linear trained on every data sizes and evaluated across all eval datasets.

Dataset	Precision				Precision				Precision			
	BLEU	3-gram	4-gram	Sent	BLEU	3-gram	4-gram	Sent	BLEU	3-gram	4-gram	Sent
	action _{100%: 5,903,010 lines}				action _{10%: 590,818 lines}				action _{1%: 59,330 lines}			
Action	0.9797	0.9762	0.9692	0.9174	0.9443	0.9376	0.9190	0.8049	0.6792	0.6360	0.5522	0.4009
Drama	0.9675	0.9625	0.9521	0.8722	0.9290	0.9222	0.9000	0.7611	0.6679	0.6263	0.5418	0.3496
Fluff	0.9632	0.9579	0.9468	0.8625	0.9252	0.9186	0.8961	0.7547	0.6656	0.6241	0.5397	0.3468
Gutenberg	0.8299	0.8048	0.7664	0.5252	0.7284	0.6919	0.6358	0.3930	0.4009	0.3515	0.2631	0.1430
	drama _{100%: 4,854,969 lines}				drama _{10%: 484,128 lines}				drama _{1%: 48,569 lines}			
Action	0.9581	0.9508	0.9376	0.8356	0.9084	0.8934	0.8656	0.7058	0.5964	0.5513	0.4581	0.2789
Drama	0.9760	0.9719	0.9636	0.9050	0.9331	0.9219	0.8999	0.7813	0.6173	0.5689	0.4769	0.3093
Fluff	0.9589	0.9521	0.9396	0.8470	0.9122	0.8979	0.8715	0.7216	0.6033	0.5579	0.4654	0.2931
Gutenberg	0.8189	0.7928	0.7523	0.5084	0.6998	0.6610	0.6018	0.3696	0.3492	0.3036	0.2166	0.1206
	fluff _{100%: 5,251,248 lines}				fluff _{10%: 524,322 lines}				fluff _{1%: 52,696 lines}			
Action	0.9501	0.9416	0.9262	0.8117	0.9050	0.8894	0.6104	0.6926	0.6079	0.5661	0.4757	0.2856
Drama	0.9551	0.9475	0.9333	0.8315	0.9112	0.8894	0.6104	0.7156	0.6185	0.5744	0.4851	0.3029
Fluff	0.9751	0.9709	0.9626	0.9064	0.9369	0.9265	0.6104	0.8007	0.6499	0.6013	0.5147	0.3416
Gutenberg	0.8015	0.7731	0.7289	0.4706	0.6502	0.6104	0.5499	0.3499	0.3497	0.3077	0.2215	0.1223
	gutenberg _{100%: 2,728,188 lines}				gutenberg _{10%: 273,263 lines}				gutenberg _{1%: 27,240 lines}			
Action	0.9028	0.8904	0.8619	0.6513	0.8051	0.7848	0.7313	0.4522	0.3474	0.2736	0.1852	0.0915
Drama	0.9124	0.9019	0.8761	0.6813	0.8164	0.7976	0.7461	0.4766	0.3546	0.2809	0.1916	0.0975
Fluff	0.8964	0.8838	0.8545	0.6419	0.7993	0.7784	0.7246	0.4494	0.3450	0.2718	0.1836	0.0929
Gutenberg	0.9330	0.9241	0.9071	0.7528	0.8094	0.7839	0.7419	0.4917	0.2803	0.2175	0.1406	0.0764
	gutenberg _{0.1%: 2,755 lines}				gutenberg _{0.1%: 2,755 lines}				gutenberg _{0.1%: 2,755 lines}			
Action												
Drama												
Fluff												
Gutenberg												

Table 4: InVBERT Seq2Seq trained on every data size and evaluated across all eval datasets.

What do characters do?

The embodied agency of fictional characters

Andrew Piper¹ 

1. Languages Literatures and Cultures, McGill University, Montreal, Canada.

Citation

Andrew Piper (2023). "What do characters do? The embodied agency of fictional characters". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-21

Date received 2023-02-17

Keywords

literary characters, fiction, narratology, nlp, theory of mind, embodied cognition

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. This paper uses machine learning to provide the first ever large-scale estimates of the distribution of actions of literary characters. Understanding these distributions is an important undertaking because they can help us better understand how different genres, cultures and time periods simulate personhood, i.e. what it means to be a person in the world. Prior research has emphasized fiction's capacity to promote "social cognition" (Theory of Mind) on the part of readers, where the complex cognitive states of literary characters are thought to facilitate deeper reasoning about human motivations and behavior. The data and models used here add an important dimension to this theory by highlighting how the actions that have increasingly distinguished fictional characters from their non-fictional counterparts over time entail forms of embodiment rather than explicit invocations of cognitive or emotional behavior. These results suggest that embodiment has emerged as a central value in the fictional representation of personhood.

1. Introduction

When contemporary writers tell stories, what do their characters *do*? What does the distribution of actions across characters look like and how has this changed from the past? And what can this knowledge about the behavior of fictional characters tell us about the meaning and function of fictional storytelling?

Understanding the actions of fictional characters is important because it can give us access to how personhood – what it means to be an agent in the world – is simulated across time, cultures, and genres. This process of fictional characterization provides insights into the values associated with human or non-human agency (Eder et al. 2010; Frow 2014; Jannidis 2004; Phelan 1989).

Traditions in both empirical and theoretical research have strongly focused on the concept of social cognition when it comes to the role of fictional characters. As Kidd et al. (2016) have written, "We propose that reading fiction can be an exercise in advanced ToM [Theory of Mind] that prompts readers to represent and engage with characters' nuanced mental states." As Palmer (2004) has stated even more emphatically, "narrative fiction is, in essence, the presentation of fictional mental functioning" (5). Similarly, as Anderson has written about the genre of the novel in particular, "The novel has a special capacity and license to convey the phenomenology of the thinking life, and it has demonstrated a special interest in forms of thinking since its inception" (Anderson et al. 2019, 131). The fictional simulation of mental worlds, so this line of thinking suggests,

provides the opportunity to develop more sophisticated forms of social cognition on the part of readers (i.e., Theory of Mind). As Zunshine writes, “We like reading fiction because it lets us try on different mental states and seems to provide intimate access to the thoughts, intentions, and feelings of other people in our social environment” (Zunshine 2006, 25). According to these theories, characters stand at the centre of fiction and minds at the centre of character.

A core challenge for this theoretical framework is the incorporation of knowledge about how such mentalizing actually takes place within fiction. If novels have a special capacity to convey the phenomenology of the thinking life, how is this manifested within the language of novels more broadly? Computational models of text analysis can be useful to provide more detailed information about the linguistic representation of characters’ actions, and by extension the mental life of characters. This work thus represents a continuation of prior work aimed at understanding the large-scale representation of literary characterization (Bamman et al. 2014; Cheng 2020; Heuser and Le-Khac 2012; Piper 2018; Underwood 2019), with a particular focus on character agency.

In order to estimate the distribution of character actions, this paper utilizes the predictive annotations provided by bookNLP (Bamman 2021) to identify actions associated with characters and then classify those actions according to higher-level categories provided by the WordNet hypernym taxonomy (see Table 1 for a full list). BookNLP is a particularly valuable resource for this task because it has been trained on literary data (Bamman et al. 2019). This workflow is then applied to two datasets: the CONLIT dataset, which consists of a collection of 2,754 works of English prose published since 2001 drawn from twelve different genres (Piper 2022), and the Hathi1M dataset (Bagga and Piper 2022), which consists of a collection of 1,671,370 randomly sampled pages of English prose published between 1800 and 2000.

As this paper will show, the actions that distinguish fictional characters from their non-fictional counterparts largely encompass forms of embodiment, such as touching, smiling, shrugging, moving, sensing, etc., rather than explicitly emphasizing cognitive or emotional actions, such as thinking, wondering, or reflecting. Fictional agency distinguishes itself as embodied agency, a fact that has only grown stronger over time. This is not to suggest that fiction is not invested in the representation of inner mental states any more than non-fictional narratives are. But it does suggest that there is a sensori-motor preference surrounding fictional agency that future work on character and social cognition will want to consider more fully. What value does the translation of virtual agency through the human body have for readers?

2. Data and Methods

All texts in the two primary datasets mentioned above are first processed through the large model of bookNLP (Bamman 2021). BookNLP is a natural language processing pipeline that includes part-of-speech tagging, dependency parsing, entity recognition, character name clustering, and super-sense tagging. The entity tagging model in particular has been trained on literary data (Bamman et al. 2019). This pipeline thus allows one to extract an ID for every detected character, the grammatical position of the character, the verb tokens associated with the character, and finally a “super-sense” tag of the verb

type. A schematic representation of the full process is illustrated in Figure 1.

This paper will be relying on the super-sense tags in particular to estimate more general categories of character actions. Super-sense tags in bookNLP are generated using a predictive model trained on SemCor, which is based on the taxonomies provided by WordNet’s hypernym trees. Instead of relying on individual keywords for analysis, the super-sense tagging aggregates individual words into more general behavioral categories, but does so using predictive models rather than dictionaries to account for the problem of polysemy. For example, the model accurately classifies the verb “found” in different senses, where Example 1 represents a cognitive event while Example 2 represents a “perceptual” event according to the super-sense taxonomy. Table 1 provides a list of all verb types with their most-frequently associated words from the contemporary data. According to the bookNLP documentation, the overall accuracy of super-sense tagging is estimated to be 76%. In the results section we describe a more fine-grained manual validation exercise, which suggests even higher-level accuracy for the key categories of interest here.

Capturing Polysemy with bookNLP

1. “I **found** the work in the small outpatient clinic difficult, as I was certain that many things were getting lost in translation.” [Cognition]
2. “He then **found** himself in a **group** around a television journalist who had just published his memoirs.” [Perception]

type	top tokens	type	top tokens
body	smile, laugh, wear, sleep, feel	emotion	want, like, feel, love, hope
change	start, begin, get, make, die	motion	go, come, walk, turn, leave
cognition	know, think, remember	perception	see, look, hear, have, feel
communication	say, ask, tell, call, mean	possession	have, get, find, give, lose
competition	fight, play, shoot, win, fire	social	do, try, make, let, work
consumption	need, use, eat, have, drink	stative	be, keep, wait, live
contact	stand, sit, put, pull, open	weather	light, burn, blow
creation	make, do, imagine, write		

Table 1: Top tokens for contemporary fiction for each verb type according to the bookNLP super-sense tags.

3. Results

Table 4 provides an overview of the distribution of fictional character actions in the CONLIT data according to the super-sense schema described above. As we can see in the left column, the most frequent actions undertaken by characters are acts of *communication* followed by *motion* and *cognition*. However, when we look at the relative frequencies of these actions between fictional and non-fictional narratives measured using a G^2 log-likelihood ratio statistic (Dunning 1993) (right column), the strongest positive predictors of fictional character behavior are all embodied forms of action (contact, body, perception, motion). Communication, far from being dominant in fiction despite its overall frequency, is actually weakly indicative of non-fiction relative to fictional discourse. Indeed, the only verb types that are statistically distinctive of fictional discourse

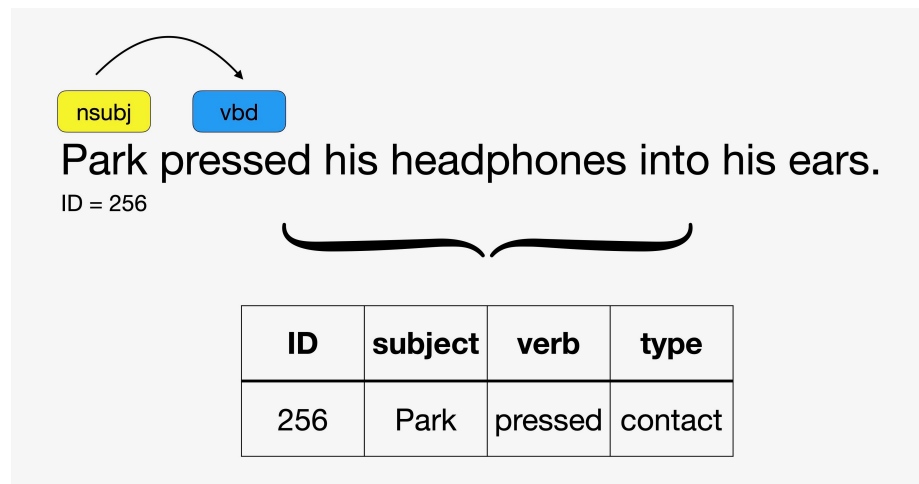


Figure 1: The process of character behavior identification. First the character is identified as either a named-entity or co-referencing pronoun, given a meta-ID and then the verb token(s) associated with that character are identified. For every character identified this way we store the verb token(s) and the associated super-sense tag.

are the ones that indicate embodied actions (though as we will see embodiment is not limited to just these types). Even if we rely only on the overall frequencies, the combined frequency of the distinguishing embodied character actions (contact, body, perception, motion) are just about two-times more frequent than cognitive and emotion types combined.

Frequency		Log-Likelihood Ratio	
Type	Count	Type	G^2
communication	3344071	contact	116743
motion	2056877	body	47340
cognition	1965067	perception	39404
contact	1443898	motion	33655
perception	1363000	weather	67
social	945900	consumption	-73
possession	896984	emotion	-273
emotion	796297	cognition	-476
stative	717342	stative	-1457
change	619910	communication	-5226
body	599845	possession	-21834
consumption	263173	change	-26977
creation	223784	competition	-30030
competition	89309	creation	-46578
weather	1754	social	-69318

Table 2: Overall counts of actions undertaken by fictional characters in the CONLIT data set (left column) along with the relative frequency as measured using Dunning's log-likelihood ratio (right column) comparing fictional and non-fictional characters. Positive / negative values indicate actions positively / negatively associated with fiction.

Table 4 gives us some indication of the extent to which the distinguishing qualities of

character actions in fiction revolve around behavior associated with different forms of embodiment. Fictional characters spend considerably more time standing, sitting, turning, walking, and smiling than their non-fictional counterparts. On the other hand, they appear to engage in relatively similar levels of explicit mentalizing (knowing, thinking, wanting, hoping, liking).

In order to better understand these relationships more broadly across our data, we can aggregate our verb types into two larger classes, one for “embodied” actions and one for “cognitive” and then calculate the fraction of all actions comprised by these types. To do so, I combine the types for motion+contact+body for the former and cognition+emotion for the latter. I thus frame “embodiment” for the purposes of this paper as a form of corporeal movement and “cognition” as the combination of thinking and emotional feeling. I return to this issue in the discussion section to review limitations and possible alternatives to this approach. The three forms of movement captured here are by no means exhaustive of embodied agency, but they can give us some insights into the nature of the distributions of these kinds of actions across time and genres.

To test the validity of these categories, we manually annotate 750 tokens randomly drawn from the CONLIT data. A set of three student readers were asked whether a character was a) “physically moving” with any body part, b) “thinking about something,” or c) none of the above. A true positive occurs if the types motion, body, or contact are predicted for a) and the types cognition or emotion for b). Importantly, for the purposes of this exercise we consider these as mutually exclusive, a point to which I will return in the discussion section. We might think of this as a means of identifying a “primary” understanding of the action for which there may be secondary features (e.g. a movement that indicates a mental state). The validation thus captures the extent to which the super-sense categories align with reader judgments about the behavior of characters.

As we can see in Table 3, both categories exhibit very high precision with respect to our combined super-sense types. Thus we can be quite confident that when a character action is labeled as “cognition” that the character is indeed mentalizing. Similarly, we can assume that the three designated types used here almost always capture someone’s bodily movement.

Type	Precision	Recall	F1
Cognition	0.98	0.94	0.96
Embodiment	0.98	0.85	0.91

Table 3: Accuracy of bookNLP super-sense annotation for our two categories “cognition” and “embodiment.” Note that we aggregate the three types body, contact and motion to capture embodiment and cognition and emotion to capture the cognition category.

Nevertheless, the lower recall for “embodiment” highlights the way embodied movements can be spread across more of our super-sense categories than just the three used here. Table 4 shows examples of the way bookNLP classifies verbs where characters are engaging in bodily movement under other headings. Observing all errors in our validation, we do not see any single category over-represented, suggesting that bodily movement is not systematically aligned with any single other category than the three used here.

Sentence	bookNLP Label
"I handed him a cheque."	possession
"Michelle took another sip of coffee."	consumption
"Russ nodded ."	communication
"I burst into the hall."	change

Table 4: Examples of motions labeled by bookNLP according to super-sense tags other than the primary three used here (motion, contact, body). Word in bold is the annotated token.

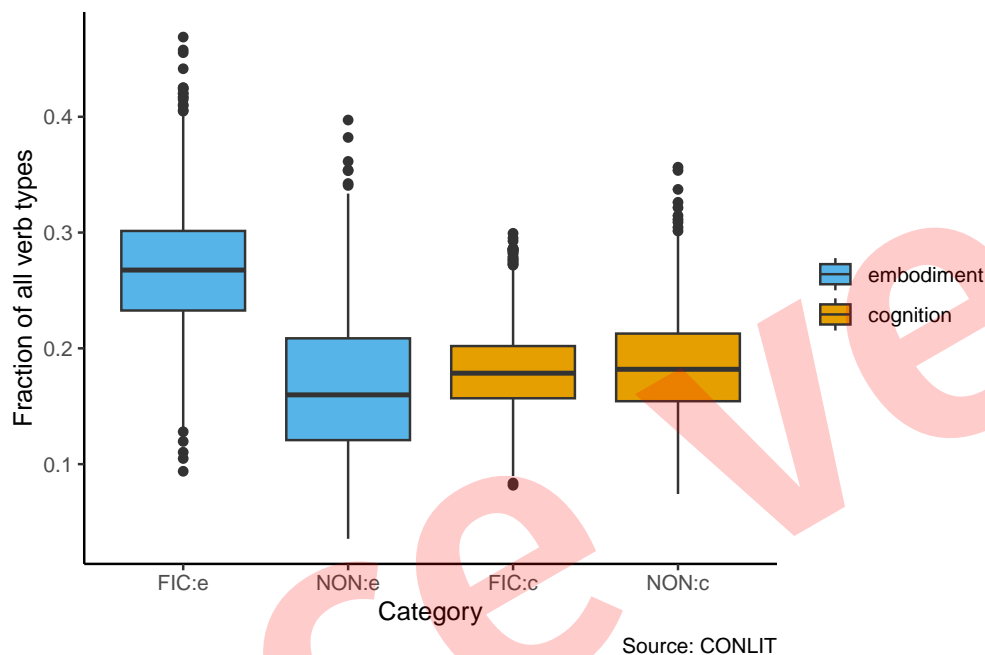


Figure 2: Relative rates of embodied actions compared to cognitive for fiction and non-fiction.

Figure 2 illustrates the effect size associated with embodied actions compared to cognitive types in the CONLIT data. As we can see, the effect size for the embodied class is very large ($d=1.82$) while that for the cognitive class is negligible ($d=-0.17$). Figure 3 illustrates the inter-genre differences for the embodied class of verbs, highlighting the relative consistency across fictional genres, with the exception of Romance.

Using the Hathi1M data, we can also observe these actions' behavior across historical time. As we can see in Figure 4, actions associated with our embodied verb types have risen considerably over time within fictional prose, while those associated with cognitive and emotional actions have remained largely stable. It is important to note that unlike the CONLIT data, non-fiction in Hathi1M is not exclusively narrative non-fiction but consists of considerable amounts of non-narrative non-fiction, thus the comparison with non-fiction are less reliable than for CONLIT.

If we break down our historical data back into the individual types (Figure 5), we see how verbs of "motion" have experienced the largest overall raw increase, but that the relative change across all three classes between the beginning of the nineteenth century and the end of the twentieth is similar across classes (ranging from a 46% increase for motion verbs to a 58% increase for body verbs).

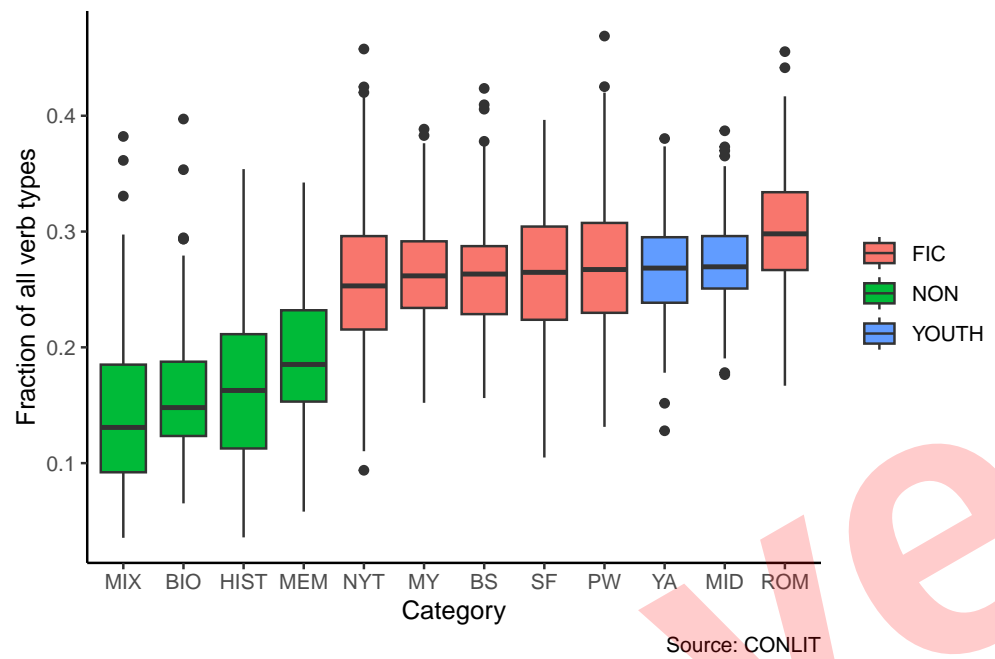


Figure 3: Rates of embodied actions across all genres in the CONLIT data.

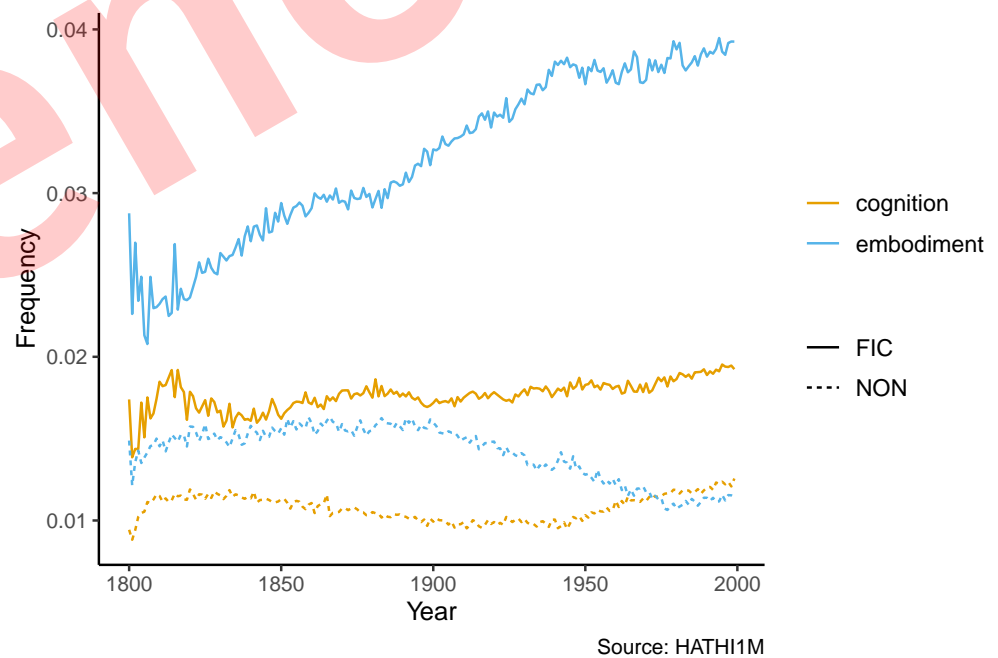


Figure 4: Rate of combined embodiment and cognition verbs for fiction (solid lines) and non-fiction (dotted lines) for the past two centuries.

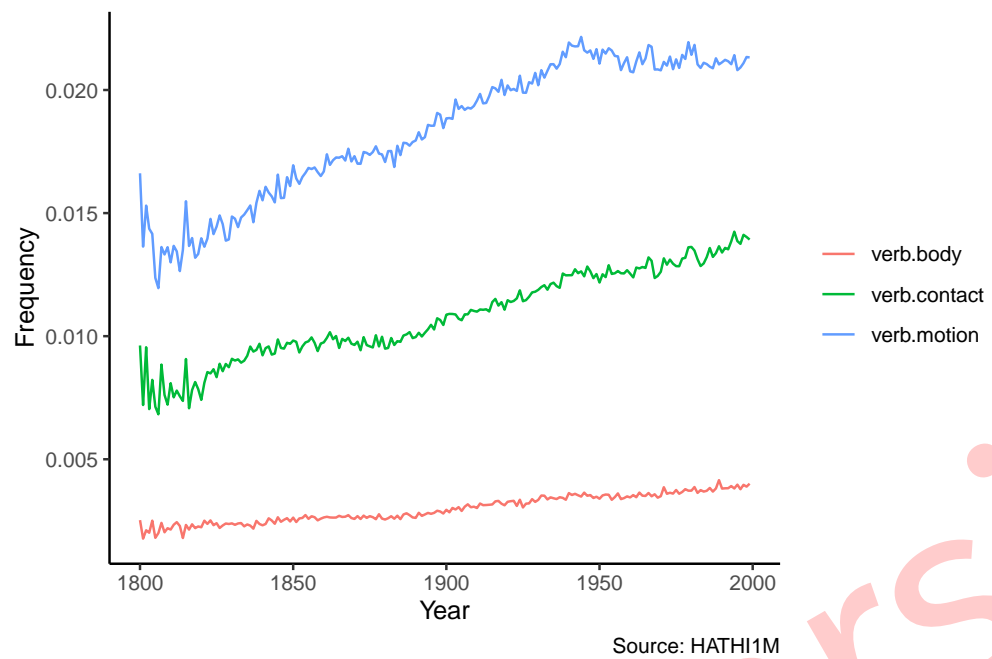


Figure 5: Frequency of the three embodied types in fiction over the past two centuries.

4. Discussion

155

The data and models presented here indicate that the primary way in which fictional narratives distinguish themselves is through a set of embodied actions associated with characters. The distinctive quality of personal agency in fiction is manifested through the agent's body. This is consistent across genres and is also illustrated in the historical data as we see the strongest growth among actions belonging to the embodied types. It appears that fictional narration, at least in an English-speaking context, has settled on a particular framework of representing agency rooted in the body.

These findings help corroborate prior work on the rising interest in embodiment through the practice of characterization (Heuser and Le-Khac 2012; Underwood 2019). They also raise a host of questions and challenges for future research. One of the main challenges posed by this research is the alignment between super-sense tags and the concepts of "embodiment" and "cognition." As we saw with the validation exercise above, embodied behavior is captured across a variety of verb types. The three used here provide very high precision when estimating the rates of embodied behavior but miss some of its more widespread uses. Nevertheless, given the very large effect sizes and the absence of what appear to be systematic errors, we can be confident about the distinctiveness of fiction's investment in embodied agency.

On the other hand, the "embodied" actions that are captured by the super-sense tags used here (motion, contact, body) are not necessarily opposed to the act of "cognition" or mentalizing more generally. To return to the opening theoretical framework of "Theory of Mind," it is safe to assume that when characters are smiling, for instance, they are conveying something about their internal cognitive state for readers to interpret. As Palmer (2004) argues, "Mental and physical sides of action and behaviour coexist and interpenetrate to the point where they are difficult to disentangle" (120).

Character Mentalizing

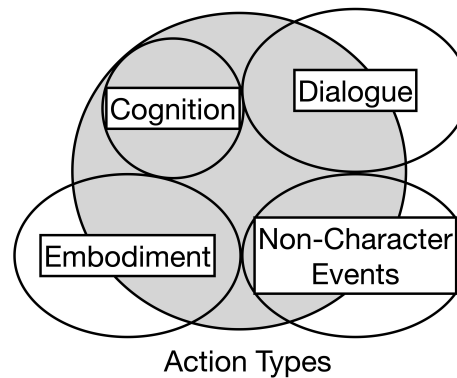


Figure 6: Frequency of the three embodied types in fiction over the past two centuries.

Indeed, potentially *all* actions may contribute to a reader's mentalizing about a character's mental state, but with differing degrees of intensity. These could include events not directly undertaken by characters (like floods or fires), but that they directly experience, along with the particular nature of the contents of dialogue, which also represent mental states beyond the straightforward action of "communication."

Future models that wish to account for the "phenomenology of the thinking life" within fictional narratives will therefore require new annotation schemes that could potentially incorporate the super-sense tagging scheme used here (see Figure 6 for a schema). Such annotation schemes could approach texts at the passage level and assess the intensity or depth of mentalizing represented. The verb types along with other linguistic features that best predict the assessment of mental depth would then help us better understand what features trigger the social cognition that is assumed to be a hallmark of fictional narrative as well as identify those literary spaces that are invested in other kinds of representational work.

Overall, the data and models used here provide further evidence that there has emerged a larger consensus around the process of characterization that foregrounds embodied agency at the heart of simulating fictional persons. Whereas prior work has focused on the prevalence of body parts, this work allows us to gain insights around the behavior of characters and its relationship to embodiment. Far from portraying characters as "thinking black boxes," which readers learn to decode, it seems more appropriate to see fictional narrative in its contemporary form as a cultural technique of modeling "embodied cognition," i.e. what it means for an agent to be embedded in an environment (Caracciolo and Kukkonen 2021). As Thelen et al. (2001) write, "From this point of view, cognition depends on the kinds of experiences that come from having a body with particular perceptual and motor capabilities that are inseparably linked and that together form the matrix within which reasoning, memory, emotion, language, and all other aspects of mental life are meshed" (1). Seen in this way, fiction's value is the way it helps us see thought as something that is produced through one's interaction with an environment and not as an abstract process of reasoning in a vacuum. Understanding this relationship more precisely and how culturally specific it is is a promising area for future work.

5. Data Availability 211

Data can be found here: <https://figshare.com/s/6ac2def06de96502f551> 212

6. Software Availability 213

Software can be found here: [see_above](#) 214

7. Acknowledgements 215

This research was funded by the Social Sciences and Humanities Research Council of Canada (895-2013-1011). 216
217

8. Author Contributions 218

Andrew Piper: Conceptualization, Analysis, Writing 219




References 220



- Anderson, Amanda, Rita Felski, and Toril Moi (2019). *Character: Three Inquiries in Literary Studies*. University of Chicago Press. 221
222
- Bagga, Sunyam and Andrew Piper (2022). "HATHI 1M: Introducing a Million Page Historical Prose Dataset in English from the Hathi Trust". In: *Journal of Open Humanities Data* 8. 223
224
225
- Bamman, David (2021). *BookNLP. A natural language processing pipeline for books*. <https://github.com/booknlp/booknlp>. Accessed: 2022-01-30. 226
227
- Bamman, David, Sejal Popat, and Sheng Shen (2019). "An annotated dataset of literary entities". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2138–2144. 228
229
230
231
- Bamman, David, Ted Underwood, and Noah A Smith (2014). "A bayesian mixed effects model of literary character". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 370–379. 232
233
234
- Caracciolo, Marco and Karin Kukkonen (2021). *With bodies: Narrative theory and embodied cognition*. Ohio State University Press. 235
236
- Cheng, Jonathan (2020). "Fleshing out models of gender in English-language novels (1850–2000)". In: *Journal of Cultural Analytics* 5.1, 11652. 237
238
- Dunning, Ted E (1993). "Accurate methods for the statistics of surprise and coincidence". In: *Computational linguistics* 19.1, 61–74. 239
240
- Eder, Jens, Fotis Jannidis, and Ralf Schneider (2010). *Characters in fictional worlds*. de Gruyter. 241
242
- Frow, John (2014). *Character and person*. Oxford University Press. 243
- Heuser, Ryan and Long Le-Khac (2012). *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*. <https://litlab.stanford.edu/LiteraryLabPamphlet4.pdf>. Accessed: 2023-06-08. 244
245
246

- Jannidis, Fotis (2004). *Figur und Person: Beitrag zu einer historischen Narratologie*. Walter de Gruyter. 247
248
- Kidd, David, Martino Ongis, and Emanuele Castano (2016). "On literary fiction and its effects on theory of mind". In: *Scientific Study of Literature* 6.1, 42–58. 249
250
- Palmer, Alan (2004). *Fictional minds*. U of Nebraska Press. 251
- Phelan, James (1989). *Reading people, reading plots: Character, progression, and the interpretation of narrative*. University of Chicago Press. 252
253
- Piper, Andrew (2018). *Enumerations: Data and Literary Study*. University of Chicago Press. 254
- (2022). "The CONLIT Dataset of Contemporary Literature". In: *Journal of Open Humanities Data* 8. 255
256
- Thelen, Esther, Gregor Schöner, Christian Scheier, and Linda B Smith (2001). "The dynamics of embodiment: A field theory of infant perseverative reaching". In: *Behavioral and brain sciences* 24.1, 1–34. 257
258
259
- Underwood, Ted (2019). *Distant horizons: digital evidence and literary change*. University of Chicago Press. 260
261
- Zunshine, Lisa (2006). *Why we read fiction: Theory of mind and the novel*. Ohio State University Press. 262
263

Need a Good Book about Privacy?

Evaluating Dictionary-Based Corpus Query for Detecting the Topic of Privacy in Literary Texts

Erik Ketzan¹ 
Jennifer Edmond¹ 
Carl Vogel² 

1. Centre for Digital Humanities, Trinity College Dublin , Dublin, Ireland.
2. School of Computer Science and Statistics, Trinity College Dublin , Dublin, Ireland.

Citation

Erik Ketzan, Jennifer Edmond, and Carl Vogel (2023). "Need a Good Book about Privacy? Evaluating Dictionary-Based Corpus Query for Detecting the Topic of Privacy in Literary Texts". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-21

Date received 2023-01-31

Keywords

computational literary studies, privacy, corpus query

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. This paper evaluates the usefulness of querying Vasalou et al.'s Privacy Dictionary (2011), a dictionary of 600+ words and phrases, in 131 canonical English-language novels from the long 19th century. We evaluate the word frequencies compared with a classification of the novels based on scholarly attention to the topic of privacy in each particular text. We report evidence of low- to low/medium strength of correlation between 3 of the 8 categories of the Privacy Dictionary and this classification. As a final step, by identifying the novels in our corpus which score highest in relative word frequency in these 3 categories, we suggest novels which have not yet received scholarly study on the topic of privacy but which may be promising for such studies.

"Nothing comes up oftener to-day than the question of the rights of privacy." Henry James (1900, 63).

1. Introduction

This paper evaluates the usefulness of querying a pre-existing dictionary of words relating to the topic of privacy in a large literary corpus, with the goal of uncovering texts which may benefit from further scholarship on the topic of privacy in literary texts. The aim of this paper is thus neither distant reading (Moretti 2000), nor macroscopic literary inquiry (Underwood 2017), nor the tracing of a crisply defined textual feature, such as one might see in e.g. corpus stylistics (Wynne 2006). Such macro- and micro-DH approaches have been analogized to a telescope and microscope (Eve 2019), but here, rather, we employ a digital method — dictionary-based corpus query — as an exploratory spotlight, to shine a light on candidate texts which may be promising for future research on a particular topic. This paper is an preparatory step in an ongoing project on how literary texts may inform the history of discourse relating to Artificial Intelligence (AI), including privacy, identity formation, and anonymity (Edmond et al. 2023).

How well does a dictionary of words relating to the topic of privacy classify texts in which privacy has been interpreted as a notable feature of the literary work? Despite decades of lexicon i.e. dictionary-based query of texts for computational linguistic, stylistic, and other computational literary studies, the methodology for such an evaluation is

unestablished, and the problem is compounded when the topic selected is vague (as “privacy” is) rather than crisp (i.e. it would be far easier to evaluate whether a dictionary of animal names, for instance, correlates positively with novels in which animals feature prominently).

Here, we first report the results of querying Vasalou et al.’s Privacy Dictionary (Vasalou et al. 2011), a dictionary of 616 words and phrases for automated content analysis on privacy-related texts, divided into 8 categories, in 131 canonical English-language novels from the long 19th century. We suggest some minor adjustments to Vasalou et al.’s Privacy Dictionary, and introduce our methods for evaluation by creating a classification of literary texts on the topic of privacy, namely, a binary classification of novels in our corpus based on one factor: whether we can identify a scholarly article or monograph which discusses the topic of privacy at length in that novel. We report evidence of low- to low/medium strength of correlation between 3 of the 8 categories of the Privacy Dictionary — which Vasalou et al. dubbed “Intimacy,” “PrivateSecret,” and “NormsRequisites” — and this classification.

As a final step, by identifying the novels in our corpus which score highest in relative word frequency in these 3 categories, we uncover novels about which scholars have not yet commented extensively on the topic of privacy in, but which may be fruitful for extensive research on “Privacy in Novel X”. From our corpus, we suggest that scholars should investigate the topic of privacy in, inter alia, Maria Edgeworth’s *Castle Rackrent* (<1800), Stephen Crane’s *The Red Badge of Courage* (1895), and George Meredith’s *The Egoist* (1879). And our method could be applied to corpora of “the great unread” (Cohen 1999, Moretti 2000) to uncover further candidates for the study of privacy in literature. While our method has many limitations due to the application of dictionary-based corpus query to a vague and manifold topic — privacy — at worst, our method can serve as a “tool for thought” (following Rheingold 2000), and at best, uncover texts which can contribute to studies on literature’s ability to inform our current robust debates on privacy, as well as ongoing debates on the “contested spaces” of the public and private spheres in the long nineteenth century (Clark 1996).

2. Privacy

Conceptual/theoretical studies of privacy are voluminous and diverse, which poses a challenge for our research, but the prevalence of discourse around privacy in our current times suggests that the investigation is worthwhile. One categorization of the vast literature on privacy is proposed by Tavani (Tavani 2007), who groups theories of privacy into four categories: 1) nonintrusion (“being let alone”); 2) seclusion (“one’s being secluded from others”; 3) control (“one has privacy if and only if one has control over information about oneself”) and; 4) limitation (“one has privacy when information about oneself is limited or restricted in certain contexts”). While a consistent, uniform theory of privacy has proven elusive (Vasalou et al. 2011, 2095), discourse on privacy has gained new importance with the rise of the so-called “surveillance capitalism,” in which large Internet companies “challeng[e] social norms associated with privacy” (Zuboff 2015, 85). This growing discourse surrounding online privacy has been met with a slew of NLP work on e.g. privacy violation detection (Silva et al. 2020), privacy

language detection in medical data (Alawad et al. 2020), and privacy leaks in social networks (Canfora et al. 2018).

The long nineteenth century is particularly fertile for the study of privacy, as privacy as a concept underwent important evolutions in society, both in its literature, which increasingly explored privacy as a theme, and in law, with the foundation of the recognition of privacy as a legal right separate from copyright or defamation. Our current conceptualisation of privacy in the Anglophone world was born around the early nineteenth century, as Erica Longfellow writes: “the definition of privacy that arouses the most debate for us, ‘the state or condition of being alone, undisturbed, or free from public attention, as a matter of choice or right,’ did not come into use until 1814. That new definition signaled a change in the paradigm of public and private” (Longfellow 2006, 315). Per Koehler, “The invasion of privacy [...] as a theme [...] assumed an especially important role in the Victorian novel” (Koehler 2016, 64). Koehler interprets this as resulting from shifting class distinctions in nineteenth century Britain: as “[e]xternal manifestations of status became less fixed [...] the newly empowered middle classes fashioned the ethos of respectability. [...] The obsession with respectability, in turn, generated an increased longing for privacy,” and that “[i]n response to dazzling economic, social, and economic changes, from the Romantic period onward privacy came to be enshrined as a positive value and claimed as an important individual right” (Koehler 2016, 65). Meanwhile, privacy as a right in common law in the United States and Britain percolated throughout the 19th century, through such cases as *Prince Albert v. Strange* (which distinguished “a breach of trust [or] confidence” from traditional concepts of property, 1849, quoted in Warren and Brandeis). “The injection of private experience into public space [...] became a crucial aspect of American life as the century wore on”, per Ackerman (1997, 2), and Warren and Brandeis’ landmark law journal article “The Right to Privacy,” was published in 1890, “widely recognized by scholars and judges, past and present, as the seminal force in the development of a ‘right to privacy’ in American law” (Bratman 2001, 624). The contours of privacy were naturally manifold in British, American, and myriad local contexts, but these many vibrant literary and legal discourses around privacy underscore the long nineteenth century as a foundational period in the development of the privacy concept.

3. Experiment: Corpus Query of Privacy Dictionary

Here, we explore the results of querying Vasalou et al.’s Privacy Dictionary (Vasalou et al. 2011), a dictionary of 616 words and phrases for automated content analysis on privacy-related texts, divided into 8 categories, in our bespoke corpus of 131 canonical English-language novels from the long 19th century. While no selection of corpus-as-canon can be innocent (Mark and McGurl 2015, 5), we selected all novels from the 19th and early 20th centuries originally written in English in Clarence Green’s Corpus of the Canon of Western Literature (Green 2017), which results in 131 texts.

Lexicon- or dictionary-based query in Natural Language Processing (NLP) is “purely descriptive and is to count words according to large lists of words of a specific category” (Schmidt et al. 2021). Querying dictionaries of words as a model of thematics dates to the earliest days of natural language processing, and many of the limitations of querying

Category	words / phrases	Vasalou et al.'s description	Sample
<i>NegativePrivacy</i>	143	"words that relate back to privacy concerns and risks as well as judgments about the source and type of violation"	afraid, being watched, deceitful
<i>NormsRequisites</i>	107	"the norms, beliefs, and expectations in relation to achieving privacy"	consent, control of, discreet
<i>OutcomeState</i>	38	"words that describe the static behavioral states and the outcomes that are served through privacy"	anonymous, liberty, security
<i>PrivateSecret</i>	58	"descriptors or words that express the 'content' of privacy," "what aspects people regard as being private"	secret, confidential, sensitive information
<i>Intimacy</i>	117	"words that portray and measure different facets of small-group privacy," "words that refer to the psychological requisites in opening up to another person as well as the emotional closeness that develops between people"	confide, friendship, gave my support
<i>Law</i>	43	"words employed to describe legal definitions of privacy"	illegal, policy, statute
<i>Restriction</i>	150	"the closed, restrictive, and regulatory behaviors employed in maintaining privacy," "the behaviors that people take to protect their privacy"	controls, hidden, lies to
<i>OpenVisible</i>	58	"words that represent the dialectic openness of privacy"	disclosed, posted, reveals

Table 1: Categories of Privacy Dictionary by Vasalou et al. 2011

a large literary corpus selected only for “canonicity” in the English language with a Privacy Dictionary created in the 21st century are obvious: finiteness and subjectivity of term selection, semantic change in lexis over time, uneven chronological distribution of lexis, and differences in e.g. British and American English (although in the latter case, the Privacy Dictionary often includes both orthographies). Yet dictionary-based query of historical and literary texts remains an established method in DH, either as a first step in more sophisticated methods (e.g. Blanke et al. 2020) or as the experiment itself (Hogenraad 2018).

Vasalou et al.'s Privacy Dictionary, intended as “linguistic resource for automated content analysis on privacy-related texts,” is divided into 8 categories (Table 1).

Vasalou et al. created their dictionary based initially on two datasets: one-on-one interviews based on a number of privacy-related topics and scraping 859 blog posts from a popular blogging platform, Blogger, in which the bloggers discussed issues relating to privacy violations (Vasalou et al. 2011, 2098). After “construct[ing] theoretically sound categories of semantically similar words”, as well as checking the semantic relatedness of words in each category using word vectors provided by the LSA website¹ and comparing the frequencies and a number of statistics on a corpus of written descriptions of privacy violations written by university staff and students, and a control corpus unrelated to privacy. Vasalou et al. envisioned the application of their Privacy Dictionary to “privacy perceptions as they are expressed in online settings,” as well as social science, for instance “comparing technology users’ language and the language employed by academics and policymakers” (Vasalou et al. 2011, 2102, 2104). A number of subsequent studies have applied the Privacy Dictionary to research in social media (Islam et al. 2014), privacy-aware software systems (Casillo et al. 2022), privacy policies in B2B and B2C e-commerce (Vakeel et al. 2017), and hotel guest reviews (D’Acunto et al. 2021). The only humanistic application of the Privacy Dictionary that we are aware of is a study of science fiction film reception, in which Milne et al. (Milne et al. 2021) performed a basic relative frequency query of the 8 categories of the Dictionary in online film reviews,

1. <http://wordvec.colorado.edu>

reporting “a dramatic increase of the [privacy] terms used in the media after the movie release than before” (756). The Privacy Dictionary has not previously been applied to explicitly literary texts, but given the rigor with which it was created, and the variety of research questions to which it has been applied, we select it as the best reference dictionary available for our task.

For our experiments, we made some relatively minor modifications to Vasalou et al.’s wordlist. First, we apply a more consistent approach to lemmas (including verb conjugation, tense, and plurals): e.g. Vasalou et al. include *block*, *blocked*, *blocking*, but not *blocks*. Vasalou et al. include *be watched* and *being watched*, but not *been watched*, *am watched*, etc. As literary texts can be written in different tenses, most often past and present, we made 120 such modifications to the word list. Next, we performed manual inspection of results to check for different word senses that may skew results. For a list of hundreds of words in 131 texts, this was no small task, and settled on the pragmatic methodological step of limiting inspection to top 10 results in each Privacy Dictionary category query. After this inspection, we ultimately excluded only one word from the Dictionary completely: *judge*, as in our literary texts it resulted in more false positives than true positives due to word sense; e.g. *judge* in literary texts more often appears as a noun (a judicial official), than a verb, which would denote a privacy concern. In all following steps, we apply this modified Privacy Dictionary which incorporates our changes.

3.1 Privacy Dictionary Category: Intimacy

Vasalou et al. describe their category of “Intimacy” as “words that portray and measure different facets of small-group privacy. It includes words that refer to the psychological requisites in opening up to another person as well as the emotional closeness that develops between people” (Vasalou et al. 2011, 2100). In our corpus, the category of intimacy words is highly skewed by only 8 words/lemmas/n-grams, which account for 94.9% of the results: *friend/-s*, *family*, *conversation*, *trust*, *confidence*, *their own*, *group/-s*, *friendship*. *Friend/-s* and *family* alone account for 62.9% of results, so when visualizing the results for the Intimacy sub-dictionary, it is largely a query of these words (Figures 1 and 2).

This observed distribution — with the total frequency mostly due to a small percentage of very frequent words — follows Zipf’s Law (see Brezina 2018, 44). These visualizations underscore how, despite the 600+ words and phrases in the modified Privacy Dictionary, when applied to literary texts, its model results in a very small number of highly frequent words, a concern for its application to our task. Having set out to locate literary texts in which “privacy” is a topic, the first sub-model is largely an extremely crude query of *friend/-s* and *family*.

How can we evaluate this query — and the queries of other sub-categories of the Privacy Dictionary — as evidence that the literary text is concerned with the topic of privacy? Without some gold standard to evaluate these queries against, the scholar can all too easily conjure the kinds of “leaps” from data to interpretation that Stanley Fish criticized in digital stylistics (Fish 1980, 89-90). As tentative observations, we could mention that the beloved classic on the themes of family and love, *Pride and Prejudice*, scores highest in these intimacy words, and the very title of Thackeray’s *Vanity Fair*

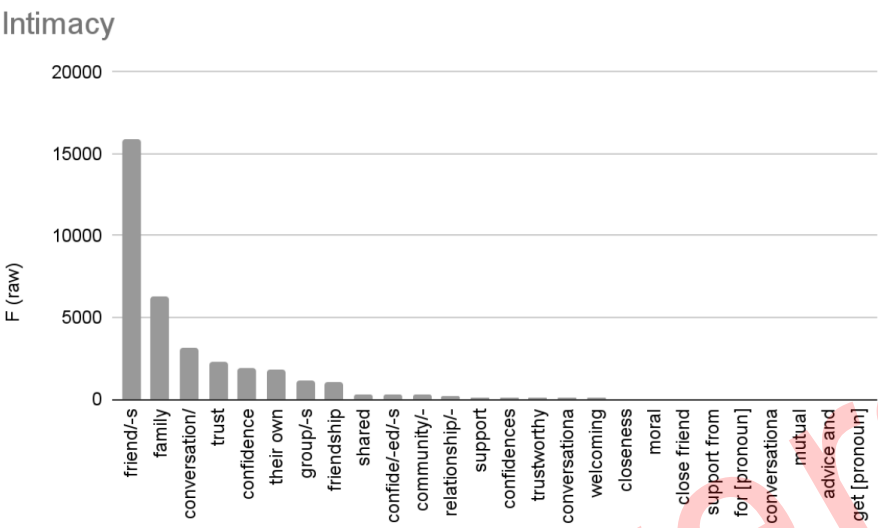


Figure 1: Raw frequency of Intimacy category of modified Privacy Dictionary in 131 canonical English novels.

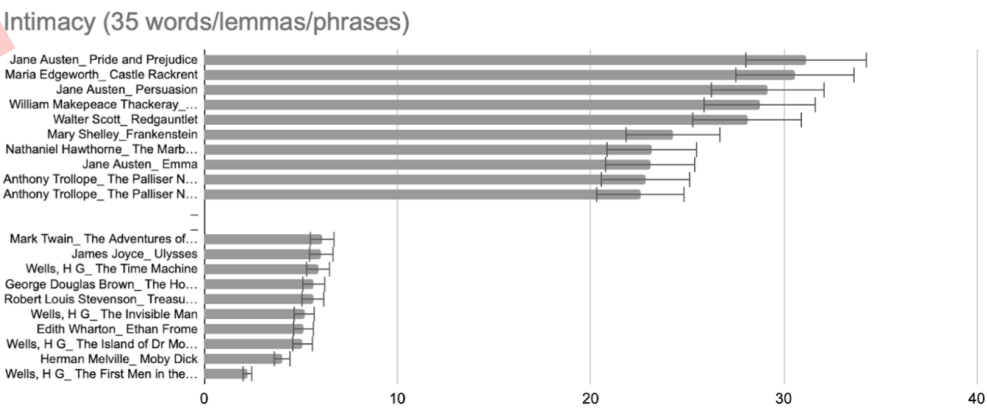


Figure 2: Texts with highest and lowest relative frequency of words in Intimacy category of modified Privacy Dictionary, per 10k word tokens.

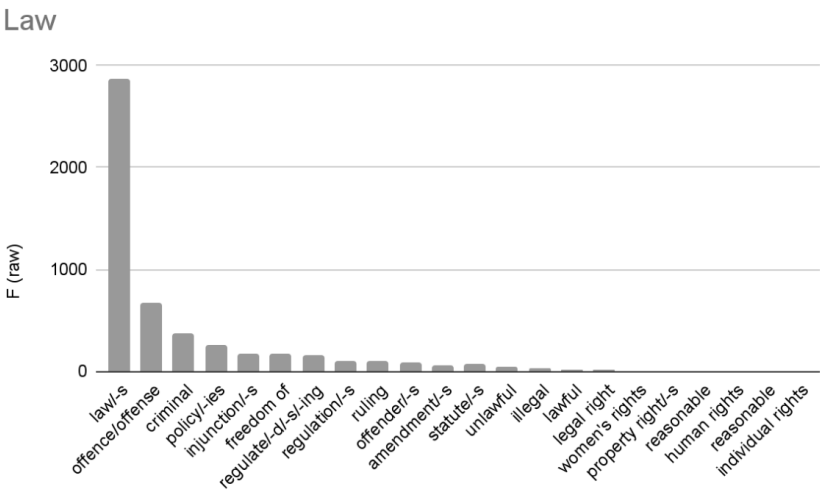


Figure 3: Raw frequency of Law category of modified Privacy Dictionary in 131 canonical English novels.

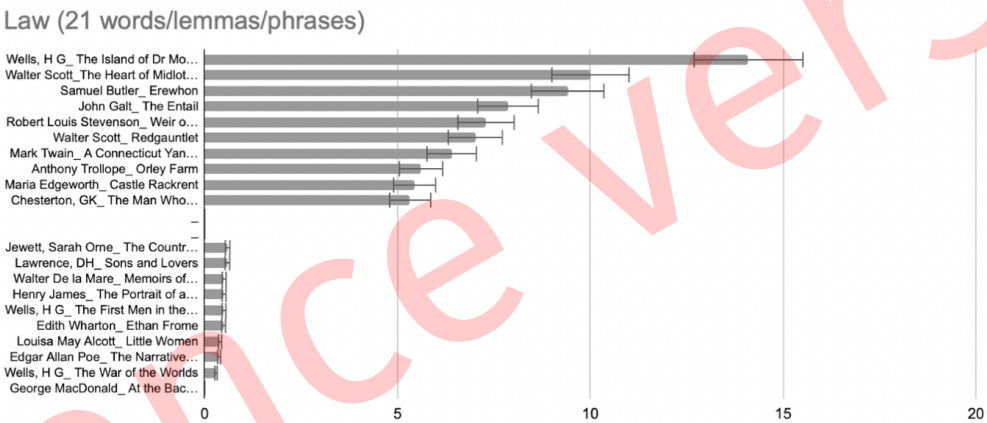


Figure 4: Highest and lowest relative frequency of words in Law category of modified Privacy Dictionary, per 10k word tokens.

(scoring third) has become a metaphor for societal interaction. Amongst the lowest scoring in this model of intimacy, meanwhile, is Melville’s *Moby-Dick*, the narration of Captain Ahab’s lonely quest of revenge, amidst sailors cramped on a whaling ship with precious little privacy. These general observations, however, require some comparison. While comparison corpora would often fill this methodological step, there is no obvious comparison corpus in which the topic of “privacy” is high. We create one below, but first, we present some more of the dictionary queries in our experiment.

3.2 Privacy Dictionary Category: Law

Vasalou et al. describe this category as “words employed to describe legal definitions of privacy” (Vasalou et al. 2011, 2100). In our literary corpus, results are again extremely skewed by a few highly frequent words: 67.5% of results are due to only two words out of 21, *law/-s* and *offense/offence*, while 93.3% of results are due to only 9 words and their plurals: *law*, *offense/offence*, *criminal*, *policy*, *injunction*, *freedom of*, *regulate* (verb lemma), *ruling*, *regulation* (Figures 3 and 4).

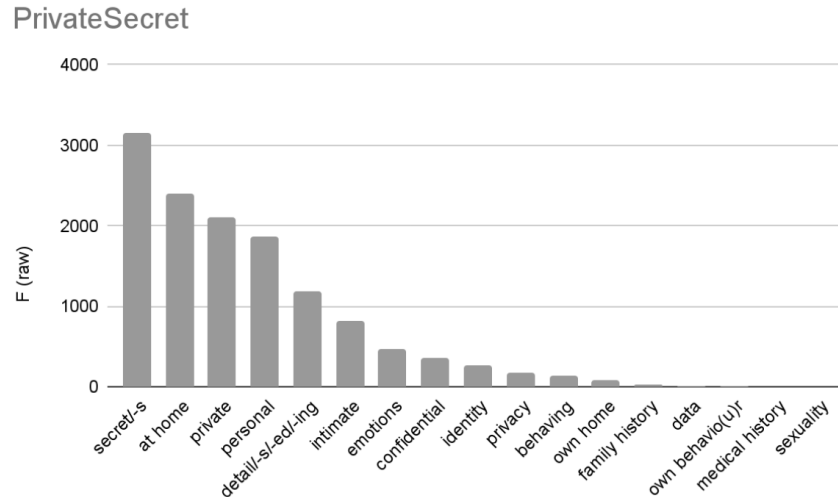


Figure 5: Raw frequency of PrivateSecret category of modified Privacy Dictionary in 131 canonical English novels.

With this query mostly based on the frequency of *law/laws* and *offense/offence*, it is likely that this is little indication of privacy as a topic, but rather, law and legal matters which figure strongly in the fabula of the novels. And indeed, law is central to the plot in Scott's *The Heart of Midlothian* (Ward 1997), John Galt's *The Entail* (1823) is a multi-generational drama involving a legal inheritance (the "entail"), while Stevenson's *Weir of Hermiston* features a conflict between high-born Scot Archie Weir and his cruel father, a judge. *The Island of Dr. Moreau* scores highest due to arguably false positives, as *Law* is an oft-repeated proper noun in the text (frequency: 61), meaning the system of rules created by Dr. Moreau that his creature-men must follow, in such dramatic passages as, "Evil is he who breaks the Law," chanted the Sayer of the Law." So far, the Privacy Dictionary is proving a crude method to query the topic of privacy in literary texts, especially with the sub-dictionary of Law.

3.3 Privacy Dictionary Category: PrivateSecret

Vasalou et al. describe this category as words which "express the 'content' of privacy. This category can be used to understand precisely what aspects people regard as being private" (2100). We expand their list of 58 words and phrases, which includes *alone*, *secure**, *prevent**, and *protect**, to include more word forms (e.g. not only *details*, but *detail.**). This modified query is presented in Figures 5 and 6.

Skew is a concern, as only 5 words and lemmas account for 82.5% of results: *secret/-s*, *at home*, *private*, *personal*, *detail**. While this remains a fairly crude model of a topic, the query is based on a wider range and distribution of words. In a category of words centered on the concept of secrets, it makes sense that Conrad's *The Secret Agent*, a "spy novel" which contains *secret* in its title, scores highest, while two texts by Wilkie Collins, proto mystery novels, follow.

Many more observations could be made about each of these queries, discussing the "secrets" in more canonical texts or hypothesizing why "friends" and "family" are high in others. But, rather than interpreting the output of the word queries at greater length,

PrivateSecret (58 words/lemmas/phrases)

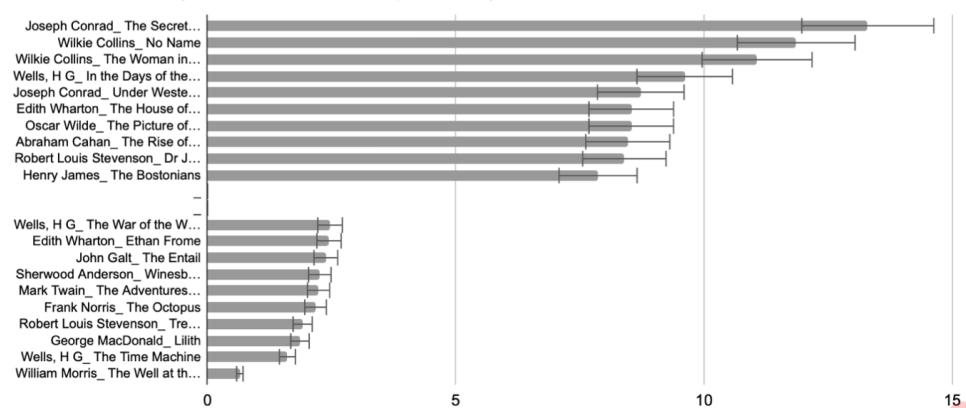


Figure 6. Texts with highest and lowest relative frequency of words in PrivateSecret category of modified Privacy Dictionary, per 10k word tokens.

Figure 6: Texts with highest and lowest relative frequency of words in PrivateSecret category of modified Privacy Dictionary, per 10k word tokens.

we add two steps of statistical evaluation which, we suggest, are more useful for our task. Visualizations of the rest of the sub-dictionary word frequency results are available at the GitHub page linked in Data Availability below.

4. Evaluation of Query with Scholarly Publications as Classification

Evaluation of a method is often ultimately a classification task. Here, we wish to see whether the Privacy Dictionary reveals literary texts which scholars have discussed at length with regard to the topic of privacy. But what gold standard could be used to classify texts which do and do not relate to privacy? Is a numerical score, such as 1-10 for “low and high privacy as a topic” even possible? While ever-shifting definitions present an essential challenge of comparing privacy discussions across time and texts, this extant analog scholarship on privacy in literary texts provides an opportunity to compare with the results of our corpus query.

We suggest a method based on existing scholarly publications. We created one binary metric for our 131 novels: whether a scholarly article or book explores “Novel X and Privacy” at substantial length.² For instance, querying academic publication databases such as JSTOR and Google Scholar, we found that there is an article titled “The British Postal Service, Privacy, and Jane Austen’s *Emma*” (Wheeler 1998). After reading this article to confirm a substantial discussion of privacy in the text, we assigned a score of 1 (rather than 0) to the specific text, Austen’s *Emma*. This often required some digging; for instance, in “Sexuality, Shame, and Privacy in the English Novel” by Ruth Bernard Yeazell, around four pages of discussion are devoted to issues of privacy in Elizabeth Gaskell’s *North and South*; we thus include this as a score of 1 for *North and South*.

The main benefit of this method is that it provides some kind of classification to compare

2. The methodology for this step includes scholarly commentary on “privacy” and Text X, the “private sphere,” and “surveillance” and “secrets” insofar as aspects of privacy are also discussed.

the Privacy Dictionary queries against. The limitations of this method are many. While we experimented with speeding up the process through automatic queries of titles and abstracts, such as through JSTOR's Constellate API,³ the ability of this API to query titles and abstracts only led us to complete this step manually, which was a considerable task. Technical considerations and scholarly labor aside, these articles and books are heterogeneous in discipline, methodology, and way in which they explore various aspects of privacy ("an elusive and multidimensional concept whose meaning is culturally and historically contingent," per Bennett 2008, xi) in texts from vastly different time periods and genres, united only in their canonicity. Most of the studies we found are typical of contemporary literary criticism: they investigate some aspect of privacy in the text, with the goal of adding knowledge to author-/genre-/time period-specific studies, e.g. "Fierce Privacy in *The Wings of the Dove*" (Lescinski 1990) or "Feminism and the Public Sphere in Anne Brontë's *The Tenant of Wildfell Hall*" (Carnell 1998). A number of other studies come from American law journals, as privacy is, among other things, a legal concept that has generated considerable jurisprudence and commentary. For our purposes, we include law journal articles that have substantial discussions of privacy as it relates to the literary text itself. For instance, in "The Huck Finn Syndrome in History and Theory: The Origins of Family Privacy", Macias (2010) introduces the competing interests in the law of family privacy through an example in Twain's *Adventures of Huckleberry Finn*; while Macias's aim is an analysis of American lawmaking and court cases, he devoted substantial discussion to literary scholars' interpretations of Twain's text. For our purposes, we thus consider Macias's article as evidence that the topic of privacy has been interpreted in Twain's text. Finally, there are scholarly commentaries which consider the issue of privacy as it impacts the privacy of the author, e.g. the reading and scholarship of a deceased author's private letters and unauthorized texts (e.g. Schloss 2000) — we have not included such studies in our classification table, but rather, academic texts which consider privacy within the diegetic world of the text.

The result is a classification table with scores for each author and text in our corpus of 131 canon novels, a sample of which is in Table 2 and the full table available at the GitHub page linked in Data Availability below.

As a first experiment in comparing the Privacy Dictionary word queries in our corpus of canonical novels with this new classification based on scholarly attention, we first looked at the word frequencies per category in novels with scholarly attention to the topic of privacy (1) and without (0) using log likelihood, a well-established metric in corpus linguistics (Brezina 2018). These results, calculated by Rayson's Log Likelihood Calculator (Rayson 2003),⁴ are in Table 3.

These results show that 3 of the 8 bags of words in the modified Privacy Dictionary showed no statistically significant difference between the words in novels we marked 1 (scholarly attention to privacy in that text) and 0 (the absence of). If there is no significant difference in the word queries, then it can be inferred that word frequency cannot correlate with either 1 or 0 in these text samples. We thus conclude that these categories have no value for our task, again, which is to use the Privacy Dictionary to identify literary texts in which privacy might be a substantial topic. Also, 2 of the 8

3. <https://constellate.org>

4. <https://ucrel.lancs.ac.uk/llwizard.html>

Text	Privacy Scholarship	Citation
Charles Dickens, <i>David Copperfield</i>	1	Bulman, Jessica. "Publishing Privacy: Intellectual Property, Self-Expression, and the Victorian Novel." <i>Hastings Communications and Entertainment Law Journal</i> 26, no. 1 (2003): 73-118.
Charles Dickens, <i>Hard Times</i>	0	
Charlotte Bronte, <i>Jane Eyre</i>	1	Spacks, Patricia Meyer. "The Privacy of the Novel." <i>NOVEL: A Forum on Fiction</i> 31, no. 3 (1998): 304-316.
Chesterton, GK, <i>The Man Who Was Thursday</i>	0	
David Lindsay, <i>A Voyage to Arcturus</i>	0	
Edgar Allan Poe, <i>The Narrative of Arthur Gordon Pym</i>	0	
Edith Wharton, <i>Ethan Frome</i>	0	
Elizabeth Gaskell, <i>North and South</i>	1	Yeazell, Ruth Bernard. "Sexuality, Shame, and Privacy in the English Novel." <i>Social Research</i> 68, no. 1 (2001): 119-144.
EM Forster, <i>Howard's End</i>	0	

Table 2: A sample of our classification of literary texts in our corpus by: (1) we have identified scholarly literature which discusses the topic of privacy in this specific text at length, or (0) we have not identified such scholarship.

Privacy Dictionary Category	(1) novels with scholarly commentary on privacy			(0) novels with no scholarly commentary on privacy			LL	Significance
	Raw frequency	Relative F (per 100k word tokens)	Word tokens in corpus	Raw frequency	Relative F (per 100k word tokens)	Word tokens in corpus		
<i>Intimacy</i>	17,807	169.7	10,496,128	17,172	135.5	12,668,752	440.04	p < 0.0001; very highly significant
<i>Law</i>	2,289	21.8	10,496,128	2,911	23.0	12,668,752	-3.51	p < 0.05; not significant
<i>OpenVisible</i>	12,045	114.8	10,496,128	14,568	115.0	12,668,752	-0.03	p < 0.05; not significant
<i>NormsRequisites</i>	7,259	69.2	10,496,128	6,995	55.2	12,668,752	180.50	p < 0.0001; very highly significant
<i>NegativePrivacy</i>	17,642	168.1	10,496,128	21,926	173.1	12,668,752	-8.38	p < 0.01; moderately significant
<i>Restriction</i>	13,418	127.8	10,496,128	16,071	126.9	12,668,752	0.43	p < 0.05; not significant
<i>OutcomeState</i>	13,811	131.6	10,496,128	17,061	134.7	12,668,752	-4.11	p < 0.01; moderately significant
<i>PrivateSecret</i>	6,491	61.8	10,496,128	6,587	52.0	12,668,752	98.19	p < 0.0001; very highly significant

Table 3: Log likelihood comparison between modified Privacy Dictionary word frequencies in novels with and without scholarly attention to privacy.

	Median relative frequency		Rank sum		rho	p
	0	1	0	1		
<i>Intimacy</i>	12.31	15.44	4380	4266	0.316	0.0002
<i>Law</i>	1.71	1.7	5153	3493	-0.002	0.9814
<i>NegativePrivacy</i>	17.21	16.95	5324	3322	-0.072	0.4113
<i>NormsRequisites</i>	4.89	6.55	4424	4222	0.298	0.0006
<i>OpenVisible</i>	11.03	11.2	5033	3613	0.047	0.5916
<i>OutcomeState</i>	12.61	13.15	5057	3589	0.037	0.6713
<i>PrivateSecret</i>	4.73	5.88	4318	4328	0.341	0.000066
<i>Restriction</i>	12.33	12.28	4981	3665	0.069	0.4357

Table 4: Descriptive statistics and rank correlation coefficients (Spearman's rho) with p values. The text category corresponds to 0 (no evident scholarly attention in relation to privacy; n=77) and 1 (evidence of scholarly attention in relation to privacy; n=53).

categories show a significant correlation, but a negative one. This means that the *fewer* 287 of the presumed privacy words are present, the more likely it is that a novel is marked 288 as 1, and has a scholarly study article that explores privacy at length in that novel. We 289 conclude that these categories — NegativePrivacy and OutcomeState — also cannot aid 290 our task, because our aim is to find a bag of words positively relating to privacy, not the 291 inverse. This leaves us with 3 categories of the modified Privacy Dictionary which are, 292 so far, promising for our task: Intimacy, NormsRequisites, and PrivateSecret. 293

To evaluate correlation between our sub-dictionary queries and our binary classification, 294 we use a non-parametric rank biserial correlation, noting the approximation of the 295 Cureton test (Cureton 1956; see also Glass 1966) with Spearman's rank correlation 296 statistic rho (which, in turn, is equivalent to Pearson correlation of ranks of scores instead 297 of the scores themselves). When the score ranking yields no ties, as in the data here, 298 the Spearman approximation is exact. Thus, we assess Spearman correlations between 299 relative frequencies and the binary classifications of works in which the frequencies are 300 noted. As a non-parametric test, there are no particular requirements beyond sample 301 size (we take 20 observations as a heuristic minimum; cf. Berk 1978), and here 53 302 items are classified in category 1 and 77 items classified in category 0. To illustrate 303 the distributions of relative frequencies for each of the word lists in relation to these 304 categories, we provide tables of median relative frequency and also rank-sums for the 305 two classification categories (rank sums being the values obtained from considering the 306 relative frequencies in rank order, with the least relative frequency assigned the least 307 rank, and summing the ranks that fall into each of the binary classification categories; 308 note that it is "easier" for the rank sum of 77 items to exceed the rank sum of 53 items). 309 The result is that three categories of the modified Privacy Dictionary positively correlate 310 at a statistically significant level with scholarly attention to privacy in individual texts 311 (Table 4). 312

Table 4 reports Spearman's rho for the rank correlations between the relative frequencies 313 of observations of dictionary items and the binary text categorization described above. 314 Adopting the convention $\alpha = 0.05$, and adjusting alpha by dividing by the number 315 of tests ($\alpha' = 0.00625$), the correlation is significant for the lists associated with 316 Intimacy, NormsRequisites and PrivateSecret. As rho can range from -1 to 0 to 1, the 317 rho values for these categories, from 0.298 to 0.341, can be interpreted as a low to 318

Text	Relative F	Scholarship on Privacy in Text
Jane Austen_ Persuasion	48.72	1
Jane Austen_ Pride and Prejudice	47.50	1
William Makepeace Thackeray_ Vanity Fair	45.07	1
Maria Edgeworth_ Castle Rackrent	44.09	0
Walter Scott_ Redgauntlet	42.86	0
Jane Austen_ Mansfield Park	41.19	1
Jane Austen_ Emma	39.74	1
Walter Scott_ Waverley	39.32	1
Wilkie Collins_ No Name	38.24	0
Anthony Trollope_ The Palliser Novels 3	37.98	1
Anthony Trollope_ The Palliser Novels 4	37.94	1
Anthony Trollope_ The Palliser Novels 2	37.46	1
Mary Shelley_ Frankenstein	37.04	1
Wilkie Collins_ The Woman in White	36.65	1
William Makepeace Thackeray_ The History of Henry Esmond	36.12	0
Henry James_ The Ambassadors	35.24	1
Anthony Trollope_ The Palliser Novels 1	35.06	1
James Hogg_ The Private Memoirs and Confessions of a Justif	34.89	1
Nathaniel Hawthorne_ The Marble Faun	34.78	1

Table 5: Combined relative F (per 10,000 word tokens) of Intimacy, NormsRequisites, and PrivateSecret categories in modified Privacy Dictionary.

low/medium strength of correlation.

These experiments now allow us to turn to our task: identifying literary texts in which the topic of privacy may be a notable concern and has gone unnoticed by scholars. Having identified 3 sub-dictionaries of the modified Privacy Dictionary as most promising for the task, this allows us to potentially identify novels which rank highest in terms of “privacy potential” (as a simple sum of the relative word frequencies in these 3 sub-dictionaries). See Table 5.

Many of the highest results are simply less well-known novels by authors whom scholars have already investigated on the topic of privacy, such as Walter Scott and Anthony Trollope. To identify both texts and authors whom scholars have yet to investigate the topic of privacy in, see Table 6.

5. Conclusion and Future Work

We report that our statistical method shows evidence of low- to low/medium strength of correlation between 3 of the categories of the Privacy Dictionary with observed scholarship on the topic of privacy in specific literary texts, and may assist our aim of identifying candidate texts which may be promising for research on the topic of privacy, including Maria Edgeworth’s *Castle Rackrent* (1800), Stephen Crane’s *The Red Badge of*

Text	Relative F	Scholarship on privacy in text
Maria Edgeworth, <i>Castle Rackrent</i>	44.09	0
Stephen Crane, <i>The Red Badge of Courage</i>	33.59	0
George Meredith, <i>The Egoist</i>	33.50	0
John Galt, <i>The Entail</i>	33.22	0
George Gissing, <i>New Grub Street</i>	30.96	0
Samuel Butler, <i>Erewhon</i>	29.62	0
Samuel Butler, <i>The Way of All Flesh</i>	29.44	0
Norman Douglas, <i>South Wind</i>	28.63	0
Thomas de Quincey, <i>Confessions of an English Opium Eater</i>	28.54	0
Joseph Conrad, <i>The Secret Agent</i>	28.20	0

Table 6: Combined relative F (per 10,000 word tokens) of Intimacy, NormsRequisites, and PrivateSecret categories in modified Privacy Dictionary.

Courage (1895), and George Meredith's *The Egoist* (1879). Recalling that our project began with the desire to trace discourses relating to AI in literary texts, it is interesting that one of our methods' suggested texts is Samuel Butler's 1892 novel *Erewhon*, a satirical utopian fiction that has been considered a seminal text in the conceptual history of artificial intelligence (Brownsword 2017), which merits future investigation.

On a legal note, Vasalou et al.'s Privacy Dictionary is provided to researchers who request it under a Creative Commons No Derivatives license, which is how we obtained it, but an interesting and open legal question is that, as we are based at a university in the European Union and now beneficiaries of the text and data mining laws ushered in by the 2019 Directive on Copyright and Related Rights in the Digital Single Market (popularly known as the "Text and Data Mining Directive"), we and other institution-based EU researchers may no longer be bound by the No Derivatives restriction of the Creative Commons license when text and data mining CC-licensed material for scientific research. Article 3 of the Text and Data Mining Directive provides that "research organisations and cultural heritage institutions [may] in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access," and that "Any contractual provision contrary to the exceptions provided for in Articles 3 [...] shall be unenforceable." This could be interpreted that the No Derivatives restriction of the license (or any other CC restrictions) may not be enforceable in this research context, a position that the Creative Commons organization has commented on tangentially, in relation to Article 4 of the Text and Data Mining Directive, noting that "Because there are many different methods for conducting text and data mining, however, there may be some types of mining activities that will implicate the licensed rights" (Lazarova et al. 2021). Arguably, beneficiaries of the Text and Data Mining Directive would not be restricted by the No Derivatives restriction when performing text and data mining on material to which they have lawful access, such as a resource like the Privacy Dictionary, but it remains an open question as to whether the creation and publication of a derivative dictionary outside of the scientific publication would go past the act of "text and data mining." In our case, in the spirit of scholarly courtesy, we

requested Vasalou’s permission to modify the dictionary for our experiments, which she graciously agreed to.

Future work can expand our experiments through larger corpora. As noted in the Introduction, now that we have obtained results for a dictionary method that correlates with scholarly attention to privacy in specific texts, we could apply our method to “the great unread” of the long nineteenth century and identify non-canonical candidate texts which may reveal heretofore unknown commentary or even contributions to the rich conceptual history of the evolution of privacy.

One of many limitations to our classification method is that papers on “Novel X and Privacy” may simply be more likely in authors with a very wide scholarly reception. As “privacy” as a primary topic for literary scholarly discourse is far from a very common or obvious one, it may be possible that the authors/texts with the greatest amount of scholarly commentary in general, such as Austen, might contain a paper on “Austen and privacy,” simply because Austen studies has produced so many “Austen and Topic X” papers. This could be investigated by additionally classifying our canonical texts as “more canonical” (e.g. Austen, Dickens, Conrad, Thackeray) and “less canonical” (e.g. Abraham Cahan, Miles Franklin, Sarah Orne Jewett).

While there has been sophisticated computational work on interpreting the semantic coherence of automatically-generated topic models (e.g. Lau et al. 2014), we did not explore such methods in this paper, as the Privacy Dictionary was not automatically generated, but created by researchers following considerable theoretical and empirical methods. However, there may be promising paths in evaluating the Privacy Dictionary and other bespoke, human-made dictionaries through the methods applied to topic model evaluation.

Finally, now that we have created a classification of “scholarly attention to privacy in Text X” and “not” (Data Availability), future work could explore the distinctions between these sub-corpora through more sophisticated methods than dictionary query, e.g. stylometric signals and machine learning, which could replace our method.

6. Data Availability

A .csv file of the canonical English novels with the citations to scholarship we used in our classification, as well as .png files of visualizations of our dictionary queries may be found at https://github.com/erikannotations/JCLS_Privacy

7. Acknowledgements

The authors would like to thank the editors and insightful comments of two anonymous peer reviewers.

8. Author Contributions

Erik Ketzan: Conceptualization, Writing, Statistical Analysis

Jennifer Edmond: Writing	402
Carl Vogel: Statistical Analysis	403


References 404


- Alawad, Mohammed, Hong-Jun Yoon, Shang Gao, Brent Mumphrey, Xiao-Cheng Wu, Eric B Durbin, Jong Cheol Jeong, Isaac Hands, David Rust, Linda Coyle, et al. (2020). "Privacy-preserving deep learning NLP models for cancer registries". In: *IEEE transactions on emerging topics in computing* 9.3, 1219–1230. 405
- Berk, Ronald A (1978). "Empirical evaluation of formulae for correction of item-total point-biserial correlations". In: *Educational and Psychological Measurement* 38.3, 647–652. 406
- Blanke, Tobias, Michael Bryant, and Mark Hedges (2020). "Understanding memories of the holocaust—A new approach to neural networks in the digital humanities". In: *Digital Scholarship in the Humanities* 35.1, 17–33. 407
- Bratman, Ben (2001). "Brandeis and Warren's The Right to Privacy and the Birth of the Right to Privacy". In: *Tenn. L. Rev.* 69, 623. 408
- Brezina, Vaclav (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press. 409
- Brownsword, Roger (2017). "From Erehwon to AlphaGo: for the sake of human dignity, should we destroy the machines?" In: *Law, Innovation and Technology* 9.1, 117–153. 410
- Canfora, Gerardo, Andrea Di Sorbo, Enrico Emanuele, Sara Forootani, and Corrado A Visaggio (2018). "A NLP-based solution to prevent from privacy leaks in social network posts". In: *Proceedings of the 13th International Conference on Availability, Reliability and Security*, 1–6. 411
- Casillo, Francesco, Vincenzo Deufemia, and Carmine Gravino (2022). "Detecting privacy requirements from User Stories with NLP transfer learning models". In: *Information and Software Technology* 146, 106853. 412
- Clark, Anna (1996). "Contested Space: The Public and Private Spheres in Nineteenth-Century Britain". In: *Journal of British Studies* 35.2, 269–276. 413
- Cureton, Edward E (1956). "Rank-biserial correlation". In: *Psychometrika* 21.3, 287–290. 414
- D'Acunto, David, Serena Volo, and Raffaele Filieri (2021). "'Most Americans like their privacy.' Exploring privacy concerns through US guests' reviews". In: *International Journal of Contemporary Hospitality Management*. 415
- Edmond, Jennifer, Vera Yakupova, and Erik Ketzan (2023). "A Tyrannical Social Something: Drawing Lessons for 21st-Century 'Privacy-Protecting' Technology from Long 19th-Century Literature". In: 416
- Eve, Martin Paul (2019). *Close reading with computers: textual scholarship, computational formalism, and David Mitchell's Cloud Atlas*. Stanford University Press. 417
- Fish, Stanley (1980). *Is there a text in this class?: The authority of interpretive communities*. Harvard University Press. 418
- Glass, Gene V (1966). "Note on rank biserial correlation". In: *Educational and Psychological Measurement* 26.3, 623–631. 419
- Green, Clarence (2017). "Introducing the Corpus of the Canon of Western Literature: A corpus for culturomics and stylistics". In: *Language and Literature* 26.4, 282–299. 420

- Hogenraad, Robert (2018). "Smoke and mirrors: Tracing ambiguity in texts". In: *Digital Scholarship in the Humanities* 33.2, 297–315. 445 446
- Islam, Aylin Caliskan, Jonathan Walsh, and Rachel Greenstadt (2014). "Privacy detective: Detecting private information and collective privacy behavior in a large social network". In: *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, 35–46. 447 448 449
- Koehler, Karin (2016). *Thomas Hardy and Victorian Communication: Letters, Telegrams and Postal Systems*. Springer. 450 451
- Lau, Jey Han, David Newman, and Timothy Baldwin (2014). "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539. 452 453 454 455
- Longfellow, Erica (2006). "Public, Private, and the Household in Early Seventeenth-Century England". In: *Journal of British Studies* 45.2, 313–334. 456 457
- Mark, Mark Algee-Hewitt and Mark McGurl (2015). *Between canon and corpus: six perspectives on 20th-century novels*. Universitätsbibliothek Johann Christian Senckenberg. 458 459
- Milne, George R, Begum Kaplan, Kristen L Walker, and Larry Zacharias (2021). "Connecting with the future: The role of science fiction movies in helping consumers understand privacy-technology trade-offs". In: *Journal of Consumer Affairs* 55.3, 737–762. 460 461 462 463
- Moretti, Franco (2000). "Conjectures on world literature". In: *New left review* 1, 54. 464
- Rayson, Paul Edward (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Lancaster University (United Kingdom). 465 466
- Rheingold, Howard (2000). *Tools for thought: The history and future of mind-expanding technology*. MIT press. 467 468
- Schmidt, Thomas, Johanna Dangel, and Christian Wolff (2021). "Senttext: A tool for lexicon-based sentiment analysis in digital humanities". In. 469 470
- Silva, Paulo, Carolina Gonçalves, Carolina Godinho, Nuno Antunes, and Marilia Curado (2020). "Using NLP and machine learning to detect data privacy violations". In: *IEEE INFOCOM 2020-IEEE conference on computer communications workshops (INFOCOM WKSHPS)*. IEEE, 972–977. 471 472 473 474
- Tavani, Herman T (2007). "Philosophical theories of privacy: Implications for an adequate online privacy policy". In: *Metaphilosophy* 38.1, 1–22. 475 476
- Underwood, Ted (2017). "A Genealogy of Distant Reading." In: *DHQ: Digital Humanities Quarterly* 11.2. 477 478
- Vakeel, Khadija Ali, Saini Das, Godwin J Udo, and Kallol Bagchi (2017). "Do security and privacy policies in B2B and B2C e-commerce differ? A comparative study using content analysis". In: *Behaviour & Information Technology* 36.4, 390–403. 479 480 481
- Vasalou, Asimina, Alastair J Gill, Fadhila Mazanderani, Chrysanthi Papoutsis, and Adam Joinson (2011). "Privacy dictionary: A new resource for the automated content analysis of privacy". In: *Journal of the American Society for Information Science and Technology* 62.11, 2095–2105. 482 483 484 485
- Wynne, Martin (2006). "Stylistics: corpus approaches". In. 486
- Zuboff, Shoshana (2015). "Big other: surveillance capitalism and the prospects of an information civilization". In: *Journal of information technology* 30.1, 75–89. 487 488

The Authorship of Stephen King's Books Written Under the Pseudonym "Richard Bachman": A Stylometric Analysis

Dorothy Modrall Sperling¹ 

Mike Kestemont² 

Vincent Neyt² 

1. DEPARTMENT, Karel de Grote University of Applied Sciences and Arts, Antwerp, Belgium.

2. DEPARTMENT, University of Antwerp, Antwerp, Belgium.

Citation

Dorothy Modrall Sperling, Mike Kestemont, and Vincent Neyt (2023). "The Authorship of Stephen King's Books Written Under the Pseudonym "Richard Bachman". A Stylometric Analysis". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-03

Date received 2023-01-31

Keywords

Stephen King, stylometry, pop culture, authorship attribution, contemporary English-language fiction

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. Between 1977 and 1984, Stephen King published five novels under the pseudonym "Richard Bachman". Reviewers noted similarities between King's and Bachman's writing styles when *Thinner* (1984) was published, ultimately leading to King's unmasking. We investigate, using the Juola protocol, whether computational techniques can correctly identify King as the author of the Bachman books out of a selection of contemporary candidate authors – Dean Koontz, Peter Straub, and Thomas Harris. We also perform a post-hoc analysis of the use of pop-culture references and brand names in Bachman, King, Koontz, Straub, and Harris novels, based on comments in reviews of Bachman and King novels. The references extracted from the Bachman books occurred significantly more often in King's texts than in the others', showing that attentive readers could have "heard King's voice" in the Bachman books through what a reviewer denigratingly called King's "compulsion to list brand-name products and his affinity for pop-cult teenage junk". These results contribute to the vexed issue of explainability, which is a recurrent challenge in author identification for literary texts.

1. Introduction

In February 1980, Stephen King published a lengthy essay called "On Becoming a Brand Name" in *Adelina*, a men's magazine. The essay records King's reaction to being referred to as a "Brand Name Author" in the modern American horror genre. He did not wish to oppose himself to being given that label; he was proud of his accomplishments at that point as a regular in the *New York Times* bestseller list and receiver of advances of well over one million dollars for a new novel. Instead, he wrote the essay to tell the story of how he had arrived at being labelled as such, emphasizing especially that he didn't exclusively write horror. In fact, before starting on what was to become his first published novel, *Carrie* (1974), he had already completed four other novels—none of them horror. King described in detail when he wrote those four pre-*Carrie* books, and how he tried, unsuccessfully, to get them published. He does not mention their titles, but refers to them as "Book #1", "#2", and so on. It is understandable that an author would rather remain vague about unpublished material. But in Stephen King's case, he had no choice but to be vague because two of those four books *had* by then been

published under a secret pseudonym: Richard Bachman. 16

Up to the present-day, King has published no less than seven novels under the pen name Richard Bachman: *Rage* (1977), *The Long Walk* (1979), *Roadwork* (1981), *The Running Man* (1982), *Thinner* (1984), *The Regulators* (1995), and *Blaze* (2007). The uncovering of Bachman's identity sparked shortly after *Thinner* came out. Both readers and reviewers noticed how similar the text was to King novels in style, theme, and narrative drive. In early 1985, King was forced to admit to *The Bangor Daily News* that he was Bachman (Smith 1985). He later acknowledged that it was inevitable that Bachman's true identity would come to light at some point; he had been getting letters from the publication of the very first Bachman book, which he attributed to readers recognizing his "voice" (King 1985, v–vii). 17 18 19 20 21 22 23 24 25 26

In this article, we will use techniques from computational author identification to determine – albeit, only retrospectively – whether a stylometric analysis of the Bachman books could indeed have discovered Stephen King's distinctive voice in the texts.¹ In stylometry, attribution and verification are commonly distinguished. Authorship attribution is the process of attributing an unknown document to an author within a set of candidate authors (Koppel et al. 2011). Authorship attribution is distinct from authorship verification in that authorship verification involves comparing an unknown text to a corpus of texts by a known author, where the aim is to determine whether the unknown document is also by that author. In authorship verification, no other candidate authors are necessarily involved, although they might serve as a point of comparison (Potha and Stamatatos 2014). 27 28 29 30 31 32 33 34 35 36 37

This paper is structured as follows: below, we will first survey the seminal critical response to the Bachman novels. The next section describes the materials used for this study, i.e., the (control) authors and the books selected for analysis. We go on to introduce the experimental setup that we chose for the analysis of the Bachman books and motivate our choice for the Juola protocol (Juola 2015). After discussing the results of our analysis, we move on to the issue of explainability, which is rarely discussed in the context of verification. We describe an experimental method to analyze counts of brand names and pop-culture references in the corpus, one of the "Bachmanesque" features that were explicitly mentioned by early critics. 38 39 40 41 42 43 44 45 46

2. Early Criticism of Bachman Novels 47

Rage, *The Long Walk*, *Roadwork*, and *The Running Man* appeared as paperback originals with very small print runs. They were hardly reviewed at all. We were only able to find two reviews of *Roadwork* (Slotek 1981; Strachan 1981), and one of *The Running Man* (Frank 1982)—none of them compared Bachman to King. Sam Frank praised Bachman's "vivid gut-level prose," his "straight-ahead storyline," "multi-dimensional characters," and "stark, dank, volcanic descriptions" (Frank 1982, 6). Jim Slotek's review of *Roadwork* 48 49 50 51 52 53

1. Van Cranenburgh and Ketzan 2021 also apply computational stylometry to the work of Stephen King to quantify the literariness of 73 King novels and novellas. In their conclusion they suggest that "an exploration of King/Bachman would merit a dedicated, mixed-method study," to determine "whether the Bachman novels (some separated by decades) can convincingly be argued to share distinctive features" and if Bachman is "a signal, or merely noise" (Van Cranenburgh and Ketzan 2021, 196). We propose our paper as a first step in such an exploration.

stated: "Bachman's frenetic, staccato writing style and his habit of throwing in one pop-culture backdrop after another (The Rolling Stones, Merv Griffin, the ever-present television) grab the reader's eye like headlines" (Slotek 1981, 10). That this reviewer noticed enough pop-culture references in the novel to give them such attention in a short review is interesting, since the use of brand names and names of musicians, actors, songs, movies, and television programs was a "sin" that critics commonly attributed to King. For instance, one negative review of *The Talisman* (1984), which King wrote in collaboration with Peter Straub, posited that the novel "inherited the worst traits of both its parents. From King it has acquired, among other things, his compulsion to list brand-name products, his affinity for pop-cult teenage junk and his penchant for the endless repetition of cryptic italicized phrases" (Lehmann-Haupt 1984, C-15). Reviewer Roger Grooms drew the same conclusion: "Combined novel has King's flavor. [...] When the frost is on the 'punkin', the ubiquitous Stephen King pops up with yet another tale of gruesome goings-on, generally taking place right in the middle of our pop culture'" (Grooms 1984, E5).

Around the same time as the publication of *The Talisman* (fall of 1984), Richard Bachman's *Thinner* arrived in bookstores. The book received more critical attention than the previous four Bachman books because it was published in hardback and review copies were widely distributed. While some reviews made no mention of King (e.g.: O'Neil 1984; Levin 1984; Williams 1984), and one reviewer remarked that the plot of *Thinner* contained "several gaps that a writer like King doesn't fall through" (Denger 1985, 6D), there were also many reviewers, booksellers, and early readers who heard King's voice in *Thinner*. One review included the tentative remark that "Bachman's style is remindful of Stephen King" (Anonymous 1984, 4G), but others were less cautious. *Locus Magazine* wrote: "King does not acknowledge it, but this horror novel about dieting sure sounds like him" (as quoted in Ganley 1985, 5). W. Paul Ganley elaborated on this quote in his own review: "this novel is pure King in style, syntax, character development, dialog, plot structure, humor, gross outs, and even in technical mistakes" (ibid.). Mark Graham announced that "this thriller raises an authorship question": "If King didn't write the end of this little narrative, his doppelganger did" (Graham 1984, 26-N).

It is striking that so many readers – independently from one another – seem to have picked up on similarities between Bachman's and King's writing style: out of curiosity alone, this observation sufficiently motivates the question whether computational methods would have been able to pick on these resemblances too. Using an established method from authorship verification, we shall assess below whether that is indeed the case. Perhaps equally striking, however, is that the reviewers' comments that we have assembled remain rather vague as to why *precisely* they connected Bachman to King; in fact, we could only find a single textual feature mentioned which was concrete enough to be counted using standard digital text analysis techniques: the use of brand names and other references to popular culture. Therefore, we will also explore whether King's supposed trademark use of this device can be used to single him out as the most-likely author of the Bachman books. As such, this paper contributes to the issue of post hoc explainability in author identification for literary texts.

3. Related Work

97

With the emergence of computing technology in Humanities scholarship, quantitative authorship studies initially focused primarily on two approaches: (1) unsupervised methods (such as exploratory visualization techniques, such as dendrograms and PCA scatterplots), which are still very popular in computational literary studies; or (2) casted the problem of author identification as a standard text classification problem. The latter approach casts attribution as a machine learning task, where exactly one label, from a set of (mutually exclusive) labels, has to be assigned to a previously unseen document. In such a setup, a text classifier can be trained on a set of reference documents for which the authorship can be established beyond reasonable doubt. This particular setup is often referred to as "closed-set attribution", because the set of candidate authors is well delineated and fixed. Excellent performance has been reported in this area, although there still exist important limitations regarding text length, text variety (genre), as well as the number of author classes to be learned.

In many authorship attribution experiments, a text's authorship is determined by calculating the similarity of that text to texts by a set of candidate authors, i.e. a form of "lazy learning". In Burrows's "Delta" method, for example, a text's authorship is predicted based on the frequencies of the frequently-occurring 150 words in the entire set of reference texts (Burrows 2002). "Delta" is a measure that quantifies the similarity between a target text and texts written by potential authors. It represents the standardized difference between the observed frequencies of the 150 most frequently occurring words in the target text and the expected frequencies based on reference texts (Evert et al. 2017). Similarity-based authorship attribution, combined with machine learning methods, have been found to perform well in identifying the true authors of novels written by pseudonymous authors. For example, Jaques Savoy used Burrows's Delta, along with Labbé's distance, nearest shrunken centroids (NSC), naïve Bayes, k-nearest neighbors, and character n-grams to identify Domenico Starnone as the probable author of pseudonymous Italian novelist Elena Ferrante's novels (2018). Eder, Tuzzi, and Cortelazzo corroborate Savoy's prediction that Starnone wrote Ferrante's novels (Eder 2018, Tuzzi and Cortelazzo 2018).

In the real world, however, there are many practical scenarios that do not fit the attribution scenario ideally, mainly because the set of potential candidate authors might be prohibitively large (the "needle in a haystack problem") and, consequently, it might be difficult (or even impossible) to construct a training data set that can be guaranteed to include the true author of an anonymous document. In such cases, it is problematic that text classifiers will always attribute an anonymous document, no matter what, to one the available authors, even if none of the available authorial labels in reality applies. Koppel and colleagues have published seminal papers in this area, highlighting that open-set attribution is a much more difficult, but also much more realistic approach to author identification "in the wild" (Koppel and Y. Winter 2014a; Koppel et al. 2007; Koppel et al. 2009). Especially open-set attribution, or authorship verification, has emerged as an established formulation of the problem. Here, algorithms still work with a limited set of candidate authors in the foreground, but specifically take into account the possibility that the correct author might not be included, effectively introducing a "back-off option" where the system returns "None of the above".

Apart from a series of applied case studies in literary studies, the authorship track in the annual shared task at the PAN workshop has played a major role in benchmarking existing approaches in this domain.² Below, we will especially draw inspiration from the "imposter approach" that was seminally introduced by Koppel and colleagues. Variations of this approach have ranked particularly high in recent editions of the shared task on authorship attribution at PAN. Importantly, the imposter approach is dependent on a pool of "imposter authors" or "distractors", to which anonymous documents and candidate authors (e.g. in a foreground corpus) can be compared. Researchers have observed that a larger and more diverse imposter pool is invariably beneficial to the performance of the imposter approach. Unsurprisingly, the most successful applications of the method have been applied for text varieties that were abundant and easy to collect online, such as blog posts. For many text varieties, however, it is much more difficult to collect large imposter sets, such as contemporary fiction, because of intellectual rights or digitization backlog.

This is a practical disadvantage of the method that is hard to circumvent in practice. Interestingly, this limitation is bypassed in the so-called Juola protocol that nevertheless still shares characteristics with the imposters approach. Apart from this practical advantage, Juola's seminal case study (perhaps the most mediatized in the history of the field) bears important resemblances to the Bachman case. After initial speculation in the (social or traditional) media, the high-profile author was twice relatively quick to self-identify as the author behind the pseudonymously published novels. In both cases, moreover, the authors have no reason to consciously alter their writing styles and lacked a clear incentive to publish in an alternative "mode". This was not the case, for instance, in the well-known French controversy surrounding Romain Gary, where the author actively resisted self-attribution, even after having been called out (Tirvengadam 1996). In the case of Rowling, however, the initial speculation was not based on stylistic similarities, which was the case for Bachman.

Patrick Juola was actively involved in the verification of the authorship of *The Cuckoo's Calling*, published under the pen name "Robert Galbraith" (Juola 2013a). This successful research initiative later led him to publish a so-called "protocol" (Juola 2015) that included methodological guidelines as to how such cases could be reliably tackled in the future. Juola compared Galbraith's novel to books by (the small number of) contemporary British female crime novelists (ibid.), that served as distractors. He ultimately operated in an open-set context, because there was no guarantee that the correct author was included among the candidates. As with our Bachman case, sampling a larger pool of imposters was infeasible, because of the intellectual rights that lie on contemporary literature and also often challenging to obtain in a digital format. Thus, the Juola protocol does not implement any sampling of imposters, although it does engage in a stochastic component in the form of feature sampling: a large and diverse feature set is engineered for each of the documents involved and the similarity is measured across these sets to produce a ranking of candidate authors that stable across different feature sets. This approach is reminiscent to the iterative bootstrapping of features in the imposter approach.

Surveying the state of the art in author identification is challenging, because case studies

2. <https://pan.webis.de/> See the overview papers listed there (e.g. Bevendorff et al. 2021)

and benchmark task differ enormously, for instance across languages, historical periods, dataset sizes, documents lengths, text varieties involved or the number of candidate authors. Recently, neural models (e.g. Boenninghoff et al. 2019), in particular large foundation models in the form of embedders such as BERT, have yielded promising results. Currently, transformer models appear to be among the best performing author identification method when (1) one is dealing for a language variety for which pretrained language models are available (2) there is a substantial amount of text data available per author. A important survey of modern authorship attribution methods found that performance heavily depends on the number of available words per author in a dataset (Tyo et al. 2022). Tyo et. al found that experiments with datasets containing less than 100,000 words per author, traditional n-gram based models achieved a higher accuracy than BERT-based models (76.50% and 66.71%, respectively). They note that, in general, in experiments with fewer words per author, traditional word- and character-level n-grams outperform more sophisticated deep learning techniques. Their observation is supported across the literature — for example, Alkatori et al. found that using word- and character-level n-grams yield higher accuracy than transformer models in an authorship attribution task using a dataset of Guardian articles, with an average of 41 thousand words per author in their corpus. (Altakrori et al. 2021). Thus, while neural methods are promising, they often come with requirements that cannot always be met.

While BERT-based models achieve state-of-the-art accuracy on authorship attribution tasks with extensive word datasets per author, applying such models in the humanities presents challenges due to the technical complexity of deep learning, which may be unfamiliar to researchers in this field. Moreover, model inspection and feature analysis are even more challenging with deep models like BERT compared to traditional n-gram approaches. n-gram models offer simplicity and transparency, making them more suitable for authorship attribution tasks in the humanities where interpretability is crucial. Therefore, in this paper, we opt for a simple similarity-based authorship attribution method that relies on transparent word- and char-level n-gram document representations.

4. Materials

Our corpus of distractor novels is made up of texts by three horror-thriller writers: Dean Koontz, Peter Straub, and Thomas Harris. These writers were chosen because, like King, all three are American, male authors that published popular novels in the 1970s, 80s, 90s, and 2000s in the same genre. The corpus includes 20 novels by King, 5 by Harris, 12 by Straub, and 17 by Koontz. It consists of all novels by Harris and Straub and a selection of books by Koontz and King up until 2007, which is when *Blaze* was published. The books were obtained in EPUB format and converted to UTF-8-encoded plain text files. Table 1 includes a selection of lexical statistics for each of the texts. The number of unique tokens, i.e., types, may be affected by text length: type-token ratios tend to be lower in longer texts because words are more likely to reoccur in longer texts (Richards 1987). Therefore, type-token ratios (TTR) were extracted from the first 10,000 tokens of each book.

Table 1: Basic statistics about the books used in this corpus. Included are the total number of tokens (word count), unique tokens (word types), and the type-token ratio (TTR) in the first 10,000 tokens.

Author	Title	Date of Publication	Word Count	Number of Word Types	TTR (First 10,000 Tokens)
Bachman	1966	<i>The Long Walk</i>	87,333	7,928	0.197
	1968	<i>Roadwork</i>	93,047	8,786	0.209
	1970	<i>Rage</i>	55,909	6,203	0.209
	1973	<i>Blaze</i>	82,444	7,720	0.187
	1981	<i>The Running Man</i>	67,769	8,579	0.252
	1984	<i>Thinner</i>	99,272	8,582	0.221
	1995	<i>The Regulators</i>	120,909	9,456	0.217
Harris	1975	<i>Black Sunday</i>	96,485	9,194	0.247
	1981	<i>Red Dragon</i>	105,648	9,136	0.214
	1988	<i>The Silence of the Lambs</i>	99,299	8,878	0.222
	1999	<i>Hannibal</i>	126,831	11,588	0.240
	2006	<i>Hannibal Rising</i>	67,575	7,671	0.231
King	1974	<i>Carrie</i>	62,275	7,509	0.247
	1975	<i>'Salem's Lot</i>	156,566	12,117	0.251
	1977	<i>The Shining</i>	165,734	11,574	0.218
	1978	<i>The Stand</i>	479,256	20,363	0.217
	1979	<i>The Dead Zone</i>	156,648	11,689	0.246
	1994	<i>Insomnia</i>	251,490	13,842	0.215
	1995	<i>Rose Madder</i>	180,040	11,171	0.205
	1996	<i>The Green Mile</i>	135,954	8,754	0.212
	1996	<i>Desperation</i>	199,619	11,581	0.202
	1997	<i>Wizard and Glass</i>	265,321	13,470	0.209
	1998	<i>Bag of Bones</i>	215,488	13,045	0.214
	1999	<i>The Girl Who Loved Tom Gordon</i>	63,368	6,102	0.201
	2001	<i>Dreamcatcher</i>	214,223	13,051	0.195
	2002	<i>From a Buick 8</i>	128,836	9,159	0.195
	2003	<i>Wolves of the Calla</i>	251,658	13,232	0.199
	2004	<i>Song of Susannah</i>	132,659	10,018	0.198
	2004	<i>The Dark Tower</i>	283,647	14,861	0.214
	2005	<i>Colorado Kid</i>	35,265	4,260	0.201
	2006	<i>Cell</i>	126,858	9,235	0.208
	2006	<i>Lisey's Story</i>	192,276	11,719	0.220
Koontz	1968	<i>Star Quest</i>	35,233	5,057	0.256
	1970	<i>Beastchild</i>	48,544	5,900	0.224
	1972	<i>Warlock</i>	57,259	6,825	0.240
	1974	<i>After the Last Race</i>	84,411	7,895	0.221
	1976	<i>Night Chills</i>	94,934	8,477	0.241
	1977	<i>The Vision</i>	66,100	6,508	0.215
	1980	<i>Whispers</i>	177,987	11,509	0.246
	1981	<i>The Eyes of Darkness</i>	89,325	8,299	0.245
	1983	<i>Phantoms</i>	138,378	10,878	0.225
	1986	<i>Strangers</i>	264,107	15,553	0.270
	1988	<i>Lightning</i>	140,257	10,676	0.234
	1990	<i>The Bad Place</i>	147,579	11,311	0.248
	1992	<i>Hideaway</i>	130,796	11,235	0.253
	1994	<i>Winter Moon</i>	118,019	10,376	0.250
	1998	<i>Fear Nothing</i>	130,790	11,553	0.249
	2000	<i>From the Corner of His Eye</i>	217,821	15,649	0.267
	2002	<i>By the Light of the Moon</i>	127,314	11,920	0.302
	2004	<i>The Taking</i>	86,104	9,857	0.280
	2006	<i>The Husband</i>	86,316	8,797	0.231
	2007	<i>The Good Guy</i>	84,424	8,305	0.218
Straub	1975	<i>Julia</i>	85,342	7,853	0.211
	1977	<i>If You Could See Me Now</i>	110,624	8,394	0.236
	1979	<i>Ghost Story</i>	187,951	11,283	0.217
	1980	<i>Shadowland</i>	159,291	10,621	0.219
	1982	<i>Floating Dragon</i>	225,917	13,022	0.252
	1988	<i>Koko</i>	210,205	12,503	0.238
	1990	<i>Mystery</i>	181,264	10,243	0.221
	1993	<i>The Throat</i>	254,781	12,635	0.218
	1995	<i>Hellfire Club</i>	191,789	11,760	0.232
	1999	<i>Mr. X</i>	186,987	13,010	0.244
	2003	<i>Lost Boy, Lost Girl</i>	89,152	8,044	0.244
	2004	<i>In The Night Room</i>	103,235	9,070	0.232

5. Task Operationalization

228

In this section, we investigate whether it is possible to identify Stephen King, only post hoc of course, as the author of the Bachman books. Our approach reproduces some of the key characteristics of Juola's authorship verification protocol, who recommends building various feature sets (word lengths, most frequent words, character 4-grams, and word bigrams) to calculate similarities between the target and distractor texts (Juola 2013b). The author who wrote the text most similar to a target text is predicted to be the author of the target text.

As in Juola's protocol, we convert calculated similarities to ranks. Juola takes similarities between target texts and known-author texts and ranks each author by their comparative similarity to the target text. Likewise, we take the cosine distances between book segments, i.e., sequences of consecutive tokens drawn from a book. We calculate the cosine distance between a Bachman segment and a segment by one of the candidate authors in our corpus, and rank the candidate authors by cosine distance. Consider a Bachman segment that has the cosine distances 0.43 for a Harris segment, 0.30 for a King segment, 0.70 for a Koontz segment, and 0.67 for a Straub segment. For this Bachman segment, King occupies rank 1 because the King segment had the smallest cosine distance to the Bachman segment, Harris has rank 2 because his segment was second closest, and so on.

Our prediction algorithm is based on Koppel and Winter's many-candidates method of authorship attribution (Koppel and Y. Winter 2014b). It relies on several varied ("bootstrapped") feature sets used to make predictions. Using diverse feature sets to represent texts reduces false similarities between target texts and known-author texts that cannot be reproduced with different feature sets. Half of the features from this varied feature set are randomly subsampled to calculate the similarity between a target text and a known-author text, repeated in k iterations. Finally, candidate authors are each assigned a score representing the proportion of iterations in which the candidate's segment was most similar to the target segment.

First, we produce a corpus of segments. To create this corpus, we tokenize and remove punctuation from all books in our corpus. Each book is then split into segments of 1,000, 5,000, and 10,000 consecutive tokens. Trailing segments of less than each of the aforementioned lengths are not included in the dataset.

Second, we produce a large feature set from the segments. We generate the feature set by vectorizing segments using combinations of different vectorizer settings. Half of the vectorizers use tf-idf weighting, and half of the vectorizers do not use tf-idf weighting. We use char and word analyzers with ranges of 2 to 4, and 1 to 3, respectively; i.e., word vectorizers create features of unigrams, bigrams, and trigrams, and char vectorizers create features of bigrams, trigrams, and 4-grams. Each vectorizer caps the number of features/columns it extracts at 10,000 to limit the number of n-grams that only appear in one book.

In total, the combination of 2 tf-idf settings (true and false), 2 analyzers (word and char), and n-gram ranges of 3 (for both word and char n-grams) creates 12 distinct vectorizer settings to generate 12 feature spaces. The 12 feature spaces are concatenated to create

one combined feature space with a maximum of 120,000 columns. Each new feature	271
space was scaled with min-max scaling before being appended to the final feature space.	272
Third, we calculate the cosine distance between each Bachman segment vector and a	273
random segment vector by each distractor author.	274
Below is our algorithm in pseudo-code:	275
For each segment length (1,000, 5,000, and 10,000 tokens) s :	276
1. Split all books in the corpus into segments of s tokens	277
2. Initialize an empty list of feature spaces l	278
3. For each collection of vectorizer settings v :	279
(a) Instantiate a vectorizer i with collection of settings v	280
(b) Create a feature space f by vectorizing the corpus of segments using vectorizer	281
i	282
(c) Append the feature space f in the list of feature spaces l	283
4. Horizontally concatenate the list of feature spaces l into a 2-dimensional feature	284
space array a	285
5. For each row r_B representing a Bachman segment in the feature space array a ,	286
repeat 1,000 times:	287
(a) For each candidate author in the corpus (King, Koontz, Straub, and Harris):	288
i. Randomly select a row r_C representing a segment by this candidate	289
author in the feature space a	290
ii. Randomly sample 10,000 distinct features from the feature space a	291
iii. Calculate the cosine distance between r_B and r_C using these 10,000	292
randomly-chosen features	293
6. Convert cosine distances into ranks (1 = segment has lowest cosine distance to	294
Bachman segment, 4 = segment has highest cosine distance to Bachman segment).	295
This algorithm created 1,000 cosine distances between each Bachman segment and a	296
segment by a candidate author – 4,000 cosine distances in total per Bachman segment.	297
We describe the results of this analysis below.	298

6. Authorship Attribution Results

299

Table 2: Bachman book titles, ranks, and the proportions of 5,000-token segments from each book that were predicted to be written by each author in the corpus. Tables for data collected from 1,000- and 10,000-token segments can be viewed in Appendix A.

Title	Rank	Harris	King	Koontz	Straub
<i>The Long Walk</i>	1	0.018	0.661	0.077	0.244
	2	0.101	0.238	0.214	0.446
	3	0.311	0.080	0.386	0.223
	4	0.569	0.021	0.323	0.087
<i>Roadwork</i>	1	0.079	0.491	0.108	0.322
	2	0.181	0.290	0.183	0.346
	3	0.322	0.156	0.303	0.219
	4	0.418	0.063	0.406	0.114
<i>Rage</i>	1	0.043	0.526	0.062	0.370
	2	0.141	0.324	0.158	0.377
	3	0.342	0.117	0.362	0.178
	4	0.474	0.033	0.418	0.075
<i>Blaze</i>	1	0.055	0.616	0.085	0.244
	2	0.184	0.244	0.196	0.376
	3	0.354	0.102	0.305	0.239
	4	0.407	0.038	0.414	0.140
<i>The Running Man</i>	1	0.086	0.456	0.165	0.293
	2	0.193	0.267	0.220	0.321
	3	0.309	0.172	0.281	0.238
	4	0.413	0.105	0.334	0.148
<i>Thinner</i>	1	0.032	0.604	0.104	0.260
	2	0.117	0.262	0.229	0.392
	3	0.296	0.103	0.363	0.239
	4	0.555	0.032	0.304	0.109
<i>The Regulators</i>	1	0.022	0.706	0.085	0.187
	2	0.115	0.203	0.250	0.432
	3	0.296	0.069	0.373	0.262
	4	0.567	0.021	0.293	0.118

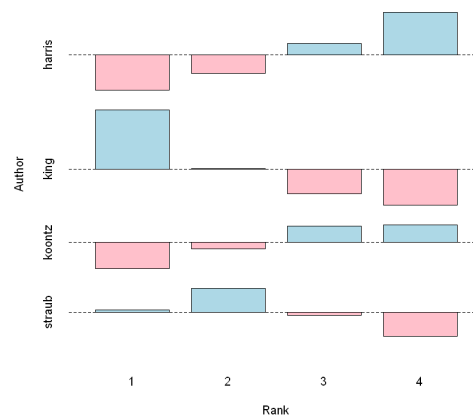


Figure 1: Association plot showing the direction and significance of correlation between the number of times a candidate author received a similarity ranking (1, 2, 3, or 4) to Bachman segments and candidate author in 5,000-token segments. See Appendix B for plots of rankings in 1,000- and 10,000- token segments.

In order to test whether authors received certain rankings significantly more or less often than if rankings were sampled from a random distribution, we performed a chi-squared test and visualized Pearson residuals using an association plot. The association plots show how ranking counts vary in a meaningful way across authors. They indicate that King was considerably more likely to be predicted as the author of a Bachman segment (ranking = 1). The other candidate authors were significantly less likely to be predicted as the author of a Bachman segment. In plots created using 5,000 and 10,000-token segments, segments written by King were significantly less likely to be 2nd, 3rd, or 4th most similar to Bachman segments (see Appendix B). However, in the plot created using 1,000-token segments, segments written by King were significantly more likely to be both the 1st and 2nd most similar to Bachman segments.

For every Bachman book, Stephen King was predicted as the author of a segment in a much greater proportion of iterations than all other candidate authors. *The Regulators* had the highest proportion of iterations in which King was the predicted author, at 50.8%, 70.6%, and 81.8% of iterations with 1,000, 5,000, and 10,000-token segments. By contrast, *The Running Man* had the lowest proportion of iterations in which King was the predicted author for 1,000, 5,000 and 10,000-token segments, with 36.8%, 45.6%, and 47.5%, respectively. King nevertheless still received the highest proportion of iterations compared to other authors.

7. Brand Name and Pop-Culture References Analysis

King's work has been described as combining the tradition of American naturalism with the classic supernatural horror genre (Bradley 1998, 96). As he himself has vehemently stated, King was in no way the first writer to take horror out of its classic gothic settings and transport it into small-town America. He has claimed that Richard Matheson, Robert Bloch, Jack Finney, and the TV show *The Twilight Zone* created the genre of modern American horror, which lies at the root of his poetics: "Those things formed my idea of

what a horror story should do: the monster shouldn't be in a graveyard in decadent old Europe, but in the house down the street" (Underwood and Miller 1989, 93). The craft, in King's opinion, is to "create any kind of environment that the reader can identify with totally" (Thomases and Tebbel 1981, 95) and then to "inject [it] with the fantasy element" (Underwood and Miller 1989, 113).

Inserting brand names and references to pop-culture is a tool in creating the familiarity necessary for the optimal effect of the horrific. In the essay "Dean Koontz and Stephen King: Style, *Invasion*, and an Aesthetics of Horror", Michael R. Collings posits that "brand-name descriptions, carefully established realism of setting and character, common images, and themes may themselves become, not trademarks of a single author (as we frequently assume when we talk of King's brand names), but characteristics of dark fantasy itself, part of the realism of presentation that C. S. Lewis argued was essential to fantasy at any level" (Collings 1998, 76). Including brand names in their work is a practice King and Koontz share, Collings believes, because it is inherent in the genre, "although far more so in King than in Koontz" (ibid.). Thus, in our opinion, it would be more interesting to approach the use of brand names and pop-culture references as a genre convention than as an idiosyncrasy of King's style, and to test whether King does indeed use brand names and references to pop-culture significantly more often than other authors in the genre—so much so that attentive readers could have deduced that the brand names used by Bachman "sounded" like King.

As a first step, we aimed to quantify the references to brand names and pop-culture of all five authors. We manually compiled lists of such references in books from the authors in our corpus: four of the five Bachman books published before King was uncovered as the author (*Rage*, *The Long Walk*, *Roadwork*, and *Thinner*),³ and three novels by Koontz, Straub, King, and Harris. We extracted references from three texts by each to avoid the risk of choosing one novel that may not accurately represent the author's overall writing style. This allowed us to establish an average use of references for all authors. Where possible, we selected novels that were published during the same period as the early Bachman books: between 1977 and 1984.⁴ These were the novels that fans of the genre of modern horror would have read not long before *Thinner* came out.

The concepts "popular culture" and "brand name" are difficult to define and delineate, so we have opted to cast a wide net: we include not only names of contemporary celebrities (e.g., "Kitty Carlisle", "Sting"), brand names of commercial products (e.g., "Ford", "Shell"), TV shows ("I Love Lucy", "Star Trek"), and movies ("Wizard of Oz", "Psycho"), but also the names of newspapers, hotels, airlines, sports teams, banks, musicians, writers, painters, literary characters, and book titles. Each reference was counted only once, no matter how many times it occurs in the text. In King's *Cujo*, for instance, a Ford Pinto plays an important role, but it only counts as one of the 235 references found in the novel. The result is presented in table Table 3, and the complete lists of extracted references for each author can be viewed in Appendix C.

3. Since *The Running Man* is set in a then-distant future, it hardly contains such references, so we disregarded it from this experiment.

4. In the case of Thomas Harris, who only published one novel in our chosen time period, we used *Black Sunday* (1975), *Red Dragon* (1981) and *The Silence of the Lambs* (1988)

Table 3: Unique references to brand names and pop-culture in 3 books by King, Koontz, Straub and Harris, and in 4 books by Bachman.

Author	Title	References	Word Count	Refs per 100,000 words
King	<i>Firestarter</i>	202	153,219	132
	<i>Cujo</i>	235	119,497	196
	<i>Pet Sematary</i>	215	144,961	148
		average:		168
Koontz	<i>The Eyes of Darkness</i>	60	89,325	68
	<i>Phantoms</i>	93	135,058	68
	<i>Darkfall</i>	63	102,550	62
		average:		66
Straub	<i>Ghost Story</i>	212	182,732	116
	<i>Shadowland</i>	115	159,291	72
	<i>Floating Dragon</i>	137	225,917	60
		average:		82
Harris	<i>Black Sunday</i>	114	96,485	118
	<i>Red Dragon</i>	87	105,648	82
	<i>The Silence of the Lambs</i>	142	99,299	144
		average:		114
Bachman	<i>Rage</i>	141	55,909	252
	<i>The Long Walk</i>	55	87,561	62
	<i>Roadwork</i>	225	93,272	242
	<i>Thinner</i>	218	99,272	220
		average:		190

To enable comparison, the number of references in each text was normalized to a standard rate per 100,000 words. We then calculated the average use of references for the total of all novels per author. As indicated in the table, Bachman included an average of **190** unique references per 100,000 words, followed by King with **168**, Harris with **114**, Straub with **82**, and Koontz with **66**.

In the second phase of our analysis, we sought to determine whether there was a significant overlap between the brand names and pop-culture references used by King in his Bachman books and those used in the novels published under his own name. This overlap could potentially reveal King's "voice" in the Bachman books through his selection of references.⁵ To test this, we examined how many of the 528 references found in Bachman's works also appeared in the texts of the other authors. All Bachman, King, Koontz, Straub, and Harris books were analyzed using a software library (*SpaCy*) capable of automatically tagging named entities, including those consisting of multiple tokens (e.g., "I Love Lucy"). Our algorithm is as follows:

1. For all books (Bachman, King, Koontz, and Harris books), repeat 100 times:
 - (a) Initialize a total cultural reference count of 0
 - (b) Randomly select a 10,000-token segment in the book
5. Having such a distinctive stylistic trait as an author also creates the risk that an imposter could insert similar references into a story to falsely present it as a King novel. Nonetheless, within the scope of this test case, the ease with which this trait can be imitated does not pose a problem.

- i. For each manually-collected pop-culture reference 383
 - A. Count the number of SpaCy-extracted named entities whose text 384
 - matches the manually collected pop-culture reference and add it to 385
 - the total cultural reference count 386
- (c) Store the total cultural reference count in a list of cultural reference counts 387
 - for each book. 388

Our algorithm produces 100 pop-culture reference counts per book in the corpus – 64,000 389
 counts in total. Differences in the central tendencies of these pop-culture reference counts 390
 by author are compared in a pairwise fashion with a Wilcoxon rank-sums test. Wilcoxon 391
 rank-sums tests compare pop-culture reference counts in Bachman versus King books, 392
 Bachman versus Koontz books, Bachman versus Harris books, and Bachman versus 393
 Straub books. In addition, Wilcoxon rank-sums tests compare pop-culture reference 394
 counts in King versus Koontz books, King versus Harris books, and King versus Straub 395
 books. 396

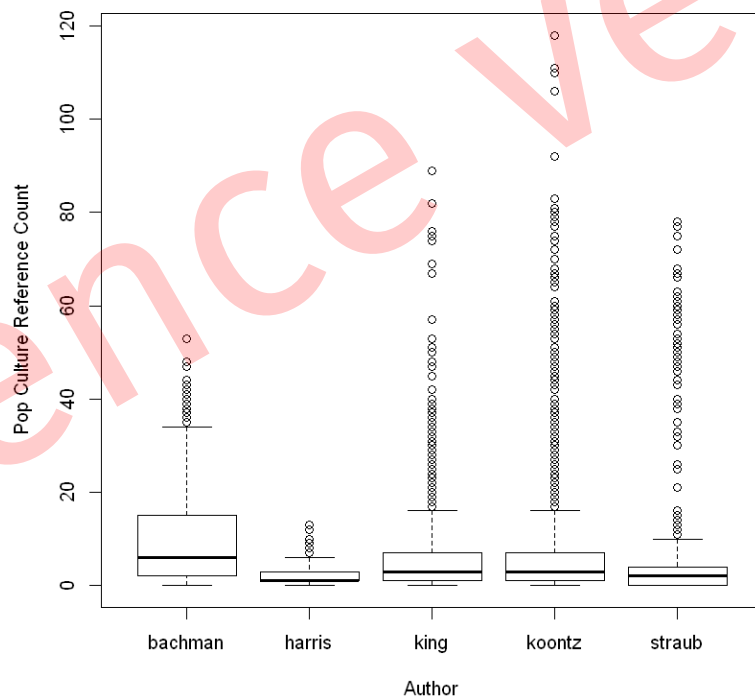


Figure 2: Boxplot showing the absolute frequencies of pop-culture references in 100 randomly-selected 10,000-token segments from each Bachman, Harris, King, Koontz, and Straub book.

The cultural references extracted from Bachman books were, as could be expected, 397
 significantly more common in Bachman segments than in segments by any other author 398
 in the corpus. A one-tailed two-sample Wilcoxon rank sums test indicated that the 399
 median pop-culture reference count was significantly higher in Bachman segments than 400
 in King segments ($W = 879455, p < .001$) (see Appendix D 6). Median pop-culture 401
 reference counts in Bachman segments were also found to be significantly higher than 402
 in segments by Harris, ($W = 272510, p < .001$), Koontz ($W = 894245, p < .001$), and 403

Straub ($W = 615746, p < .001$) . That is to be expected, however, since the reference list was compiled on the basis of these very texts.

Of the distractor authors, King segments contained the most references extracted from Bachman books. One-tailed two-sample Wilcoxon rank-sums tests were applied to each pair of distractor authors – King and Straub, King and Harris, King and Koontz (see Appendix D 7). Median pop-culture reference counts in King segments were found to be significantly higher than in segments by Straub ($W = 1506216, p < .001$), Koontz ($W = 2096930, p = 0.004$), and Harris ($W = 672109, p < .001$). Appendix E 8 contains a breakdown of pop-culture reference counts by book title.

8. Discussion

The results of our analysis suggest that computational methods can correctly identify King as the real author of the Bachman books. However, the chosen segment length matters – larger segments lengths seem to produce more extreme, and probably less trustworthy proportions for any given rank and author. The proportions of iterations in which a Bachman segment was predicted to be written by King increased with segment length. Likewise, authors that were more consistently ranked third or fourth, like Dean Koontz and Thomas Harris, had higher proportions of iterations with rank 3 or 4 in longer segment lengths. This trend is consistent with the observation that larger text sizes (5,000 tokens and over) tend to increase the probability that a text's authorship will be correctly attributed (Eder 2015). The extreme proportions in 10,000-token segments are likely the result of skew by a smaller sample size.

While *Thinner* led to readers outing King as the true author of the Bachman books, *Thinner* did not have the highest proportion of iterations predicting King as the author of its segments, *The Long Walk* did, a text written so early in his career that it could be classified under juvenilia. It was *Thinner* that lead to King's unmasking, not because it was more "King-like" than the previous Bachman books, but because it had a larger readership and was much more widely reviewed. *The Regulators*, the only novel conceived specifically for the alter ego "Richard Bachman"—an alter ego that had surely taken form in King's mind by the mid-nineties—scored second-lowest (scoring only fractionally better than *The Running Man*), which in our opinion is mainly due to the different text-types which take up a substantial portion of the novel: long letters and sections from a screenplay; but it could also be because of King's conscious effort, as he stated afterwards, to find "a good voice and a valid point of view that were a little different from my own" (King 1996, ix), something he did not do with the previous novels that were only published as Bachman but not written as him.

Interestingly, Peter Straub consistently has the highest value for rank 2. King and Straub, who were friends and collaborators, have commented on each other's styles over the years. Straub praised the style of *The Shining* as the "reverse" of a literary style which "made a virtue of colloquialism and transparency", an "unprecedentedly direct style" (Straub 1984, 10). Because of this directness and transparency, it has been claimed that King in fact has "no style" (Bradley 1998, 116). The language, King believes, should not pose an interference between the reader and the story, it should be accessible to all, allowing the reader to "get through the barrier of print and into the story without

too much effort", and for that to happen, the "writer's voice" should be "low enough" (Dewes 1981, 63). Straub's style is more classic; referred to as "the good prose" by King "almost always structurally correct", "not flashy", and "unobtrusively strong" (King 1982b, 30). King has noted stylistic differences between Straub's early novels: he called *Julia* (1975) an "English ghost story", its diction also being "English—cool, rational, almost disconnected from any kind of emotional base" (King 1982a, 285); *If You Could See Me Now* (1977) has a "Chandleresque first-person narrative" (King 1982b, 31); and *Ghost Story* (1979)—Straub's third novel in a row to feature a ghost—is written in a "Jamesian diction" (ibid.). The slightly meandering style in Straub's early works might in part account for Straub emerging from the experiment as the second-most-likely author of the Bachman books. Of relevance here also is that Straub has stated that King's early novels *Salem's Lot* and *The Shining* heavily influenced him as a writer: "[King's] aims and ambitions were very close to my own [...] [*The Shining*] was like a roadmap of where to go: [King] armored my ambition" (Straub 1984, 9–10). The differences between their styles did not propose a challenge when they collaborated on *The Talisman*. As Straub told an interviewer in 1985: "Our styles seemed to melt together. The book has its own sound; it doesn't sound like me and it doesn't sound like Steve. [...] There were times when I deliberately imitated Steve's style and there were times when he deliberately, playfully, imitated mine" (D. E. Winter 1985, 64). Straub's rank 2 in our experiment is an indication that a comparative stylistic analysis of the works of King and Straub, including their two collaborations, would be a fruitful avenue of further research.

The results of the pop-culture experiments reveal that King's use of brand names and pop-culture references was part of his style from the first stages of his career, when he wrote the early Bachman books. *Rage*, in particular, which he began at the age of nineteen, contains a high amount of references: 141 in a 55,909-word text. *Rage* is not a horror novel, but a suspense novel set in contemporary USA, as is *Roadwork*, which has an equally high amount of references: 225 in a 93,272-word text. Converted to averages per 100,000 words, these two books each contain over three times as many references as any of the novels by Koontz and Straub. *Rage* and *Roadwork*, both non-horror books, show that King's use of brand names and popular culture as a technique to create a world that is familiar to the reader transcended (and in a sense predated) his views on the poetics of the modern horror genre; the real world is there even when the fantasy element is not.

The manual counts of references indicate that all four modern horror practitioners used the technique in their novels. This tentatively confirms the claim that it is inherent to the genre. However, King employed it to a much greater extent than Straub and Koontz, which irked some literary critics in the nineteen eighties. A larger-scale study of this kind, with a corpus consisting of more works by more authors, would provide a more accurate picture of the genre's dependence on realism and familiarity in setting, for instance in comparison with other genres in popular fiction, such as romances, mystery novels, and thrillers. There is a challenge, however, in the degree to which the extraction of the references can be automated. Reliable named entity recognition is a vital first step in automating this kind of work. However, a filtering process is necessary to remove or flag irrelevant entries such as character and place names for it to be feasible for all remaining entries to be manually vetted, since the named entities in a novel easily run

into the thousands.

493

9. Conclusion

494

Style is elusive and idiosyncratic. While some reviewers were confident King was the real author of the Bachman books, they remained vague about the similar stylistic features they had discovered in the texts of both authors. In our paper, we showed that an authorship attribution algorithm is able to predict King to be the author of Bachman texts significantly more often than the distractor authors. But because it predicts authorship by randomly sampling large feature sets in thousands of iterations, the algorithm is a black box, which, comparable to the reviewers, only outputs scores and does not supply insights into the similarity in voice.

495

496

497

498

499

500

501

502

Our research might be extended by applying machine learning and deep learning authorship attribution techniques. In this paper, we use a similarity-based authorship attribution method instead of machine learning or deep learning methods. Because pre-trained BERT models have demonstrated state-of-the-art accuracy in authorship attribution experiments where each candidate author in the corpus has a large number of words, we expect BERT-based models to achieve better results with our corpus. It remains very much the question, however, whether such gains in performance would eventually scale to lesser-resourced (e.g. historic) literary cultures.

503

504

505

506

507

508

509

510

This paper also explored how pop-culture references may be important for identifying King as the real author of the Bachman books. Brand names and pop-culture references extracted from Bachman books were found to be significantly more common in King books than in books by Koontz, Straub, or Harris. Moreover, *Rage*, *Roadwork* and *Thinner* contained much more references than any of the contemporary novels by Koontz, Straub, and Harris, which again pointed towards King being the author (*Cujo* being similarly packed with pop-culture, for instance). These findings indicate a promising direction for further exploration of this technique as an inherent feature in modern horror fiction, but also shed some light on the intuition of many readers at the time that the Bachman books had actually been written by the author who, as he suggested in his essay in *Adelina*, had become a brand name himself.

511

512

513

514

515

516

517

518

519

520

521

A.

522

Table 4: Table containing Bachman book titles, ranks, and the proportions of 1,000-token segments from each book that were predicted to be written by each author in the corpus.

Title	Rank	Harris	King	Koontz	Straub
<i>The Long Walk</i>	1	0.077	0.484	0.156	0.283
	2	0.164	0.282	0.241	0.313
	3	0.289	0.156	0.311	0.244
	4	0.471	0.077	0.292	0.160
<i>Roadwork</i>	1	0.125	0.400	0.171	0.304
	2	0.196	0.297	0.228	0.279
	3	0.284	0.194	0.288	0.235
	4	0.395	0.109	0.314	0.182
<i>Rage</i>	1	0.092	0.433	0.130	0.345
	2	0.174	0.308	0.221	0.298
	3	0.288	0.175	0.320	0.216
	4	0.446	0.084	0.329	0.141
<i>Blaze</i>	1	0.117	0.452	0.161	0.269
	2	0.202	0.282	0.226	0.289
	3	0.296	0.171	0.285	0.248
	4	0.384	0.095	0.328	0.193
<i>The Running Man</i>	1	0.132	0.368	0.211	0.289
	2	0.202	0.288	0.237	0.274
	3	0.284	0.208	0.268	0.241
	4	0.382	0.137	0.285	0.196
<i>Thinner</i>	1	0.087	0.452	0.177	0.283
	2	0.167	0.289	0.249	0.295
	3	0.282	0.172	0.299	0.247
	4	0.464	0.087	0.274	0.175
<i>The Regulators</i>	1	0.088	0.495	0.169	0.249
	2	0.176	0.271	0.252	0.301
	3	0.289	0.154	0.299	0.257
	4	0.447	0.081	0.280	0.192

Table 5: Table containing Bachman book titles, ranks, and the proportions of 10,000-token segments from each book that were predicted to be written by each author in the corpus.

Title	Rank	Harris	King	Koontz	Straub
<i>The Long Walk</i>	1	0.010	0.733	0.052	0.205
	2	0.077	0.207	0.199	0.518
	3	0.326	0.052	0.408	0.213
	4	0.587	0.008	0.341	0.064
<i>Roadwork</i>	1	0.073	0.519	0.092	0.317
	2	0.180	0.275	0.168	0.377
	3	0.352	0.150	0.284	0.214
	4	0.395	0.056	0.456	0.092
<i>Rage</i>	1	0.036	0.561	0.045	0.358
	2	0.131	0.317	0.134	0.418
	3	0.367	0.098	0.367	0.168
	4	0.466	0.023	0.455	0.056
<i>Blaze</i>	1	0.033	0.702	0.056	0.209
	2	0.175	0.207	0.174	0.443
	3	0.402	0.068	0.303	0.226
	4	0.389	0.022	0.467	0.121
<i>The Running Man</i>	1	0.079	0.475	0.152	0.294
	2	0.203	0.247	0.209	0.341
	3	0.343	0.165	0.266	0.226
	4	0.376	0.113	0.374	0.138
<i>Thinner</i>	1	0.024	0.667	0.086	0.223
	2	0.102	0.234	0.223	0.441
	3	0.306	0.079	0.373	0.242
	4	0.568	0.020	0.318	0.093
<i>The Regulators</i>	1	0.008	0.818	0.049	0.125
	2	0.087	0.141	0.255	0.517
	3	0.300	0.034	0.396	0.270
	4	0.605	0.007	0.300	0.087

B. 523

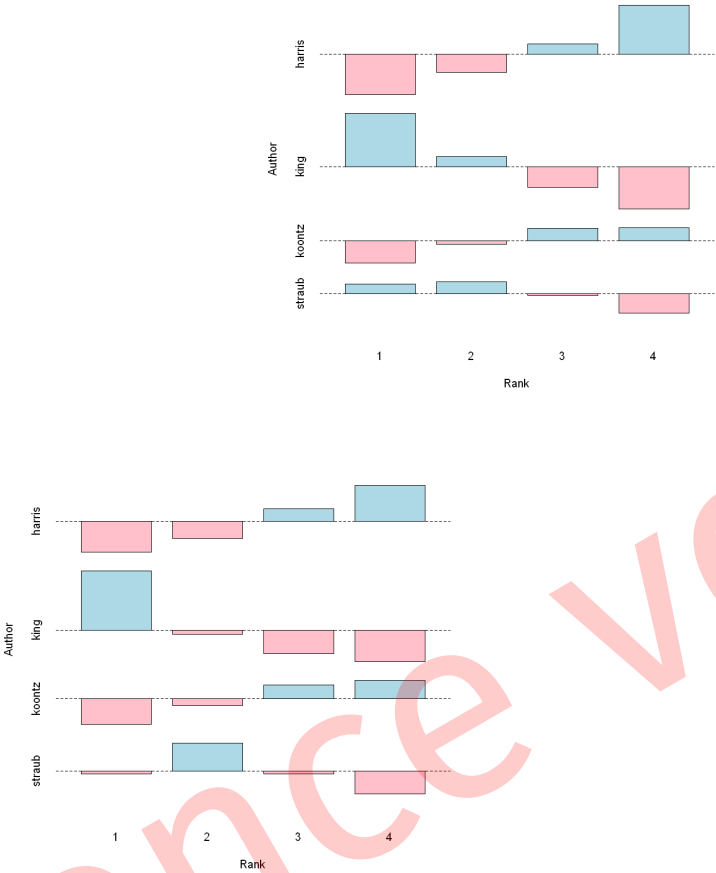


Figure 3: Association plots showing the direction and significance of correlation between the number of times a candidate author received a similarity ranking (1, 2, 3, or 4) to Bachman segments and candidate author in 1,000- (left) and 10,000- (right) token segments.

C. 524

C.1 Bachman 525

List of pop-culture references (brands, celebrities, products, fictional characters, movies, 526
TV shows, etc.) extracted from *Rage* (1979), *The Long Walk* (1979), *Roadwork* (1981), and 527
Thinner (1984). 528

A&S Tires, A. Gordon Pym, AP wire, Abdul Allhazred, Adreizi Brothers, Agatha Christie, 529
Ahab, Albert Einstein, Alfie, Alligators All Around, American Express, Amoco, Amos ’n 530
Andy, Amway, Anacin, Anaïs, And Justice for All, Andy Devine, Annie Oakley, Apple, 531
Arco, Arlene Dahl, Art Linkletter, Aureomycin, Avis, BO, BMW, Bach, Bally, Band-Aid, 532
Banjo Rag, Banker’s Life Insurance, Barbie, Bausch & Lomb, Be-Bop, Be-bop-a-lula, she’s 533
my baby, Beach Boys, Beatles, Beechcraft, Ben Alexander, Bermuda, Bertrand Russell, 534
Beverly Hill-billies, Big Mac, Bill Cullen, Black Jack gum, Blackglama, Bob Hope, Bobby 535
Sherman, Bombardier Skidoo, Bonneville, Brain from Planet Arous, Brian Wilson, Briggs 536

& Stratton, Broderick Crawford, Bruce Springsteen, Bud, Budweiser, Buick, Burger King,	537
Buttercup, Cadillac, Caesar's Palace, Calvin Klein, Camel, Campbell, Canada Mints,	538
Captain Midnight, Captain Queeg, Chancellor-Brinkley, Chargers, Charles Manson,	539
Charmin, Chatty Cathy, Cheez-Doodles, Chesterfield, Chevrolet, Chevy, Chevy Impala,	540
Chevy Nova, Chipwich, Chivas, Chris-Craft, Chrysler Imperial, Chryster, Chuck Berry,	541
Cimarron, Clint Eastwood, Coca-Cola, Coke, Colt, Colt Woodsman, Con-Tact, Cracker	542
Jack, Curly, D-Con, Dairy Freez, Dallas Cowgirls, Dan Fortune, Dannon, Darvocet, Dar-	543
von, Datsun, David Cassidy, David Janssen, Delco, Delta 88, Denny's, Detroit Redwings,	544
Detroit Tigers, Dial M for Murder, Dialing for Dollars, Diamond International, Dick	545
Cavett, Dingo, Dior, Dirty Harry, Disney, Disney World, Dodge, Dodge Custom Cab,	546
Dolby, Don Rickles, Donald Westlake, Dorito, Dorothy Sayers, Dos Passos, Dotto, Dr.	547
Caligari, Dr. Scholl, Dragnet, Dungeons and Dragons, Duz, Dylan, Eames, Eberhard	548
Faber, Econoline, Edith Head, Electrolux, Ellery Queen, Elton John, Elvis, Empirin,	549
Ernest Hemingway, Exorcist, Exxon, F Troop, Facing the Lions, Family Feud, Fantasia,	550
Farmer Brown, Fat Sammy's, Father Brown, Ferrari, Field and Stream, Firebird, Flair	551
Fineliners, Flatt and Scruggs, Fontainebleau, Ford, Ford Pinto, Formica, Forrest Tucker,	552
Francis Gary Powers, Frederick's of Hollywood, Frisbee, GI Joe, GM, Garfield, Garry	553
Moore, Gary Davis, Ghidra, Gilligan's Island, Gisele MacKenzie, Glade Pine Fresh,	554
Godzilla, Good-bye Yellow Brick Road, Goodwill, Gravy Train, Great Northern, Great	555
Western, Green Door, Greenbriar Boys, Greyhound, Griff, Gucci, Gulf, Guy Lombardo,	556
Guy Madison, H. Rider Haggard, HBO, Hal March, Hamburger Helper, Hammond	557
Innes, Henry Glassman, Henry James, Henry Youngman, Herman Wouk, Hertz, Hesse,	558
Hexlite, Hey, Mr. Sun, Highway Patrol, Hogan's Heroes, Holiday Inn, Home Box Office,	559
Honda, Horchow, Hot Stuff, Howard Cosell, Howdy Doody, Hula Hoops, Humphrey	560
Bogart, I Dream of Jeannie, I Love Lucy, Incredible Shrinking Man, J. C. Penney's, J.	561
Press, J. W. Dant, J.C. Whitney, J.J. Cale, Jack Barry, Jack Benny Program, Jack Narz, Jack	562
and Jill, James Bond, James Cain, Jaws, JayCee, Jell-O, Jerry Jeff Walker, Jimmy Cagney,	563
Jimmy Stewart, Jingles, Jock Mahoney, Joe Friday, John Agar, John Carradine, John	564
Chancellor, John Cheever, John Cougar Mellancamp, John D. MacDonald, John Travolta,	565
Joker's Wild, Jordache, Judy Blume, Julia Child, KLH, Kalishnikov, Karen Carpenter,	566
Kent, Kewpie, King Kong, Kitty Carlisle, Kleenex, Kluge, Kodachrome, Krazy Glue,	567
Kurt Vonnegut, LTD, Lacoste, Larry, Lawrence Belch, Lee Strasberg, Lenny Bruce, Let	568
It Be, Let it Bleed, Let's Make a Deal, Levi's, Lionel Richie, Lipton, Lobstermen, Long	569
Island Dragway, Loony Tunes, Lorne Greene, Louis Lamour, Mace, Malboros, Mam-	570
moth Mart, Manhattan Transfer, Mantovani, Marcus Welby, Marlboro, Marlboro Light,	571
Marty Milner, Maurice Sendak, Mauser, Max Von Sydow, Maxwell House, McDonald,	572
McDonald's, Mercedes, Mercury, Merv Griffin, Mets, Michael Jackson, Michelangelo,	573
Mick Jagger, Mickey Mouse, Midnight Rambler, Milky Way, Miller, Miller Lite, Miracle	574
Chopper, Moe, Molotov, Monkey Man, Monkey Trial, Monocle, Monopoly, Monte Hall,	575
Morns, Mothra, Motorola, Mr. Bojangles, Mr. Hyde Jekyll, Mr. Rogers, Mr. T., Muen-	576
ster cheese, Mustangs, Myron Floren, NFL, National Enquirer, National Geographic,	577
National Lampoon, New Paltz, New York Times, Nike, Niques, Nite Owl, Nivea, Nor-	578
man Bates, Norman Vincent Peale, Num-Zit, O. Henry, OK Corral, Olds, Oldsmobile	579
Ninety-Eight, Olivetti, Olivia Newton John, Omar the Tentmaker, Oshkosh, Outdoor	580
Life, Pall Mall, Panthers, Pat Benatar, Paul Harvey, Paul Stuart, Pavlov, Peanuts comic	581
strips, Peggy Sue, Penthouse, Pepsi, Pepsi-Cola, Pepto Bismol, Pequod, Perrier, Perry	582
Mason, Peter Rabbit, Philco, Phillies Cheroot, Phillips, Pig Pen, Piper Cub, Plymouth,	583

Polaroid, Ponderosa golf, Pontiac, Porsche, Psycho, Q-tip, Quoddy, RCA, REA express, 584
 Radio Shack, Ramada Inn, Range Rider, Raquel Welch, Ravi Shankar, Reader's Digest, 585
 Red Sox, Richard Petty, Richard Stark, Richard Widmark, Rin Tin Tin, Ring-Dings, Rinso, 586
 Ripley's Believe It Or Not, Ritz, Ritz crackers, Robert Redford, Rodan, Roloids, Rolex, 587
 Rollerdrome, Rolling Stones, Rolls-Royce, Ronald McDonald, SH Green Stamps, SOI, 588
 Saab, Sadie Hawkins, Saf-T-Glass, Salvation Army, Samsonite, Sara Lee cheesecake, 589
 Saran Wrap, Saville Row, Schooner, Schwinn, Scripto, Sergeant Preston, Sears, Seven 590
 Flags Over Georgia, Seven-Up, Shell, Sheraton, Sherman tank, Shop and Save, Shop 591
 'n Save, Shop 'n Save, Slaughterhouse Five, Slurpies, Slurpy, Smith Wesson, Sony, 592
 Soo Line, Soupy Sales, Souther Comfort, Southern Pacific, Space Command, Spencer 593
 Tracy, Spider John Koemer, Star Trek, Starsky and Hutch, Sterno, Stetsons, Sting, Stop 594
 and Shop, Stranger in Paradise, Subaru, Sunoco, Sylvester Stallone, T. S. Eliot, TRS-80, 595
 Tarr Brothers, Technicolor, Tensor, Tensor study lamps, Texaco, The \$64,000 Question, 596
 The Ballad of John and Yoko, The Day the Earth Stood Still, The Gift of the Magi, The 597
 Gong Show, The Guardian, The Manchester Guardian, The New Price is Right, The 598
 New York Review of Books, The Postman Always Rings Twice, This Is Your Life, This 599
 Savage Rapture, Thomas Carlyle, Thousand Island dressing, Three Stooges, Three's 600
 Company, Thunderbird, Thus Spake Zarathustra, Tic Tac Dough, Time Magazine, Timex, 601
 Tinkertoys, To Tell the Truth, Tolkien, Tom Paxton, Tom Rush, Tom Wicker, Tony Curtis, 602
 Toyota, Toys Are Joys, Trans Am, Trifles, Trix, True Argosy, Tukkan the Terrible, Twenty 603
 One, Twinkies, Twinky, Universal Pictures, VW, Vantage 100, Victor Canning, Vince 604
 Lombardi, Volvo, Von Ronk, WGAN-TV, Waldenbooks, Waldorf-Astoria, Walkman, 605
 Wall Street Journal, Walt Disney, Walter Cronkite, Warner Anderson, Warner Brothers, 606
 Washex, We Gotta Get It On Again, Weatherby, Wet-Nap, What's My Line, Where are 607
 They Now, Whoppers, Wild Bill Hickok, Wilt Chamberlain, Winslow Homer, Wizard of 608
 Oz, Woody Woodpecker, Wranglers, Wyatt Earp, Yankees, Yorick, You can't always get 609
 what you want, Your Hit Parade, Your Show of Shows, Zenith, Zippo, lazy Susan 610

C.2 Harris

611

References extracted from *Black Sunday* (1975), *Red Dragon* (1981), and *The Silence of the* 612
Lambs (1988): 613

280ZX, A Nurse to Marry, AK-47, Abdel Awad, Ain't Misbehavin, Air Force C-141, Aldo 614
 Ray, Alka-Seltzer, American Aermotor, American Express, Amex, Antoine's, Avon, Baby 615
 Ruth candy bar, Baeder Chemical, Band-Aid, Barry Manilow, Bartlett's Familiar, Base- 616
 ball Joe, Batard-Montrachet, Beaujolais, Beaver Cleaver, Beechcraft, Begin the Beguine, 617
 Bell Atlantic, Beretta, Betty Skelton, Big Mac, Black Mountain Rag, Blondie, Blooming- 618
 dale, Boeing, Bolex Super Eight camera, Bonwit Teller, Britches, Bronco, Buffalo Bill, 619
 Bufferin, Buick, Bulldog .44 Special, C. S. Forrester, Camaros, Canoe after-shave, Canoe 620
 beer, Captain Video, Cardinals, Cash for Your Trash, Celotex, Cessna, Charles James, 621
 Chateaubriand, Checker, Chemical Mace, Cher Bono, Chevrolet, Cinzano, Citroën, 622
 Clorox, Coke, Cole Porter, Colt, Coromandel screen, Corsair, DC motor, Danny Kaye, 623
 Datafax, Deborah Harry, Decca, Delta, Demerol, Disney, Doc Watson, Dos Equis beer, 624
 Duccio, Duke Keomuka, Edith Piaf, e.e. cummings, El Diario-La Prensa, Elvis, Emily 625
 Dickinson, Erythromycin, Evelyn Waugh, Evyan, Ezio Pinza, F-4 Phantom, Fats Waller, 626
 Federal Express, Ferragamo, Flaying of Marsyas, Fleet, Flicker, Ford, Fotomat, Fox lock, 627
 Franklin Mint locomotives, GM, GOODYEAR DOUBLE EAGLE, Galatoire's, Garfinkel, 628

Georgia Power Company, Glaser Safety Slugs, Glenn Gould, Greyhound, Grumman	629
Gulfstream, Géricault, H. Allen Smith, Howard Hughes, Huckins, Huey, J. Edgar Hoover,	630
Jack Daniel's, Jane Austen, Jell-O, Jimmy Hoffa, Johnny Carson, Joy of Cooking, Kaiser,	631
Katyusha rocket, Kevlar, Kewpie doll, Kiss, Kleenex, Kodak D-76 developer, Kool-Aid,	632
Kotex, L. L. Bean, Land Rover, Lean Cuisine, Lee Harvey Oswald, Levis, Lewis Carroll,	633
LifeSaver, Lincoln Versailles, Listerine, Litton Policefax, Llama automatic pistol, Lomotil,	634
Lord & Taylor, Lucite, Lutece, Lycra, Lysol, L'Air du Temps, Mace, Madonna, Mag Na	635
Port, Magic Marker, Magnum, Man Mountain Dean, Marcus Aurelius, Mark Five gas	636
mask, Mary Janes, Max Shulman, Melvin Purvis, Mets, Miami Dolphins, Miami Herald,	637
Mike Hailwood, Moe, Monteleone Hotel, Montrachet, Morocco Mole, Mounds candy	638
bar, Movietone News, Mr. Hide, NBS Sports Spectacular, NFL, National Football Con-	639
ference, National Geographic, Nero, New York Times, Nikon, Nomex, Norman Vincent	640
Peale, Novocaine, Odor-Eaters, Orkin, Over the Sea to Skye, Packard, Pan Am, ParkRite,	641
Perelman, Perrier, Personality Plus, Peter Jennings, Phantom F-4, Phone-Mate, Picasso,	642
Pinto, Pittsburgh Steelers, Plimsoll, Plutos, Plymouth, Polaroid CU-5, Port-O-San, Prince	643
Andrew, Quonset, R. L. Polk and Company, Rand McNally, Reebok, Remington 870,	644
Remy Martin, Reynolds 5130, Rice Stadium, Rinso white, Rinso bright, Ritalin, Rolodex,	645
Romeos, Rose Marie Reid, Rubik's Cube, Rybovich, Saks, Sam Browne belt, Sancerre,	646
Sapporo beer, Satellite Monroe, Schmeisser, Sears Best, Secret Squirrel, Sedan de Ville,	647
Segovia, Servco Supreme, Shea Stadium, Sheetrock, Sikorsky, Sinderella, Skycrane,	648
Smith Wesson, Smithsonian's National Museum of Natural History, Smokey the Bear,	649
Southeastern Bell, Southern Bell, Sperry-Rand, Sports Illustrated, Startron, Stevie Won-	650
der, Sting-Eez, Studebaker, Styrofoam, Super Bowl, Superdome, Tanqueray, Tater Tots,	651
Teflon, Telex, The Knifemakers Guild, The Look of Love, The Young and the Restless,	652
Thorazine, Threave, Times, Titian, Toto, Trans-Am, Trumpy, Tulane Stadium, Tulane's	653
Green Wave, Twinkie, UNICEF, United Coal, VanSleek Farfoon, Vanderbilt, Velcro, Vicks	654
VapoRub, Visa, Vogue, Voice Privacy system, Volkswagen, Vonnegut, W. W. Greener,	655
WPIK-TV, Walkman, Washington Post, Washington Redskins, Weight Watchers, Western	656
Union, Whiskey River, Wile E. Coyote, Windex, Winston, Wolf's Ears, World Cup soccer,	657
Wratten, Xerox, Yoo-Hoo, ZPG-1, Zamfir, Master of the Pan Flute, Zodiac, the Bargain	658
Center, the Chicago Tribune, the China Mail, the Court of Two Sisters, the Daily News,	659
the Fairmont Hotel, the Florida State League, the Goldberg Variations, the Harmon Tro-	660
phy, the Holiday Inn, the International Herald-Tribune, the Intra Bank, the Los Angeles	661
Times, the Marriott Hotel, the Monteleone Hotel, the National Broadcasting System, the	662
National Gallery, the Navy's Blue Angels, the New Orleans Saints, the New York City	663
Aquarium, the New York Post, the Oilers, the Ramlet el Baida, the Reader's Digest, the	664
Royal Orleans Hotel, the Sugar Bowl Classic, the Super Bowl, the Tigers of Louisiana	665
State University, the Times, the Top of the Mart, the World Series, the Yellow Pages	666

C.3 King 667

References extracted from *Firestarter* (1980), *Cujo* (1981), and *Pet Sematary* (1983): 668

A P, AFC Championship, AMC Matador, Ace bandage, Adidas, Adolph's Meat Ten-	669
derizer, Agway Market, Albany Airlines, Alfalfa, Ali MacGraw, Alka-Seltzer, All My	670
Children, Allagash, Allegheny, Alpoburgers, American Casket Company, American	671
Express, Amex, Amoco, Amway, Anacin, Andrea Doria, Andy Warhol, Antonioni, Arco	672
gas, Ariel Sharon, Armstrong ceiling, Arte Johnson, As the World Turns, Astroturf, Atari,	673

Atlanta Braves, Atlantic, Audrey Rose, Auld Lang Syne, Avis, Avon, B.J. and the Bear,	674
Bally, Bang caps, BankAmericard, Barbie, Bass, Bat-Cycle, Bearcat scanner, Beatrix Pot-	675
ter, Becton-Dickson syringe, Bell helmet, Ben-Gay, Bentley, Bermuda onion, Bermudas,	676
Bespin Warrior, Bette Davis, Beulah Land, Big Mac, Big Red Machine, Bijou, Bill Blass,	677
Bird's Eye orange juice, Biscayne, Biz, Black Beauty, Black Label, Bloomingdale, Blue	678
Cross-Blue Shield, Blue Horse tablet, Bluto, Bob Hope, Bob Seger, Bob Stanley, Bomba the	679
Jungle Boy, Bone-Phone, Boris Karloff, Boston Post, Botany 500, Boy Scouts, Braniff Air-	680
lines, Brillo, Brookings-Smith Mortuary, Brooks Brothers, Buddy Hackett, Bugs Bunny,	681
Buick, Burger King, Busch, Butterball, Cadillac, Caesar, Caldor, Camaro, Camel, Camera	682
Store, Campbell, Canada Dry, Candy Man, Carlos Castenada, Casco Bank and Trust,	683
Charles Dickens, Checker, Cheerios, Chester, Chesterfield Kings, Chevette, Chevrolet,	684
Chevy, Chicago Tribune, Chrysler, Chuggy-Chuggy-Choo-Choo, Cisco, Citgo, Claymore,	685
Clearasil, Clint Eastwood, Clio, Coca-Cola, Cocoa Bears, Coke, Colt, Con-Tact paper,	686
Count Chocula, Crawly-Gator, Crayola, Cream, Credence, Cremora, Cuisinart, D. Duck,	687
Dairy Queen, Dale Carnegie, Danskin leotard, Dark Victory, Darth Vader, Darvon, Dave	688
Garroway, Dave and Frank Blair, Decoster Egg Farms, Dee Dee Ramone, Del Monte,	689
Delco battery, Delta, Diamond matches, Dilly Bar, Diners Club, Dingos, Dinty Moore,	690
Disney, Disney World, Doctor Doolittle, Dodge, Douglas MacArthur, Dow Chemical,	691
Downy fabric softener, Dr. Cyclops, Dr. Denton suit, Dr. Seuss, Duke Wayne, Dukes of	692
Hazzard, Dumbo, Dwight Frye, Eastern Airlines, Eastern Bank, Eeyore, Egg McMuffin,	693
Ellsworth American, Elmer Fudd, Elmer's Glue, Elvis Presley, Encyclopedia Britannica,	694
Erica Jong, Ernie, Esso, Ever-Lock, Exxon, Family Fun Lanes, Fay Wray's, Festus, Flexible	695
Flyer, Ford, Forest Lawn, Franco Harris, Frankenberry, Frankenstein, GEN. PATTON,	696
Gaines Meal, Gelusil, Gene Autry, Gene Simmons, General Hospital, George Carlin,	697
George Romero, George and Gracie, Georgia Charger whiskey, Gerrypack, Gilbey's gin,	698
Gillette Foamy, Gilligan's Island, Girl Scouts, Goofy, Gordon R. Dickson, Gravy Train,	699
Greedo, Greyhound, Grover, Gucci, Gunsmoke, Hamburger Helper, Han Solo, Hefty	700
bag, Heinz, Herbert Tareyton, Hershey, Hertz, Hillerich Bradsby, Hitachi TV, HoJos,	701
Holiday Inn, Home Box Office, Honda Civic, Hoosier cabinet, Horseman, pass by, Hot	702
One Hundred, Hotpoint, Howard Johnson's Motor Lodge, Hush Puppies, Hush Puppy,	703
IBM, Igor, Immelmann, Indian motorcycle, Injun Joe, International Harvester, Isodil, It's	704
a Good Life, J & B whisky, J. B. Rhine, J. C. Whitney & Co., J. Fred Muggs, J. J. Cale, J.	705
Walter Thompson, J.R., JERRY FALWELL, Jack Daniel's, Jacob Marley, Jaguar, James	706
Bond, Jaundaflo, Jeep, Jefferson Airplane, Jerome Bixby, Jerry Garcia, Jim Beam, Jim	707
Morrison, Joan Baez, Jockey shorts, Joe DiMaggio, Joe Green, John Hurt, Johnny Carson,	708
Johns Hopkins, Johnson's No More Tears shampoo, Johnson's Wax, Karl Malden, Keds,	709
Keebler, Kellogg's, Kelvinator, Ken, Kermit, King Kong, Kleenex, Kodachrome, Kodak,	710
Kool-Aid, L. L. Bean, La-Z-Boy, Lark cigarette, Latex, Laugh-In, Lawnboy, Lawrence	711
Welk, Lee Riders, Lengyll, Lester, Lestoil, Lincoln Continental, Little Black Sambo, Little	712
Golden Books, Little House on the Prairie, Lord Buxton, Lou McNally, Loudon Wain-	713
wright, Love Me Tender, Love Story, Love of Life, Lovecraft, Lucky Charms, Luger, Luke	714
Skywalker, Löwenbräu, MIKE WALLACE, Magic Kingdom, Magic Mountain, Magnum,	715
Mammoth Mart, Marek stove, Marshal Dillon, MasterCard, Matchbox, Matt Dillon,	716
Maurice Sendak, Max Factor, McCheese, McDonald, Menachem Begin, Mercedes, Mer-	717
cer Mayer, Michael Jackson, Micheloeb, Mickey Mantle, Mickey Mouse, Milky Way bar,	718
Mille Bourne, Miller, Mixmaster, Mondavi, Monopoly, Mr. Coffee, Murray Leinster,	719
Myer, NBC, Nabisco, Napoli's, Narnia, Necromancer, New Franklin Laundry, New	720

York Mets, Nicklaus, Nipper, the RCA dog, Norman Rockwell, Northeast Bank, Noël	721
Coward, Old MacDonald, Olympia, Omar the Tentmaker, Orasin, Orson Welles, Orville,	722
Oscar, Oscar the Grouch, Oz the Great and Terrible, PATTI SMITH, PAUL HARVEY, Pall	723
Mall, Pampers, Panasonic, Pancho, Parcheesi, Pearl Kineo, Pentel pen, Penthouse Forum,	724
Pepsi-Cola, Peter Pan, Phil Donahue, Phone-Mate, Piggly Wiggly, Piggy, Pilot Razor	725
Point, Pinto, Planet Mongo, Planet Quark, PlaySkool, Pledge, Plymouth, Polaroid, Ponch	726
and John from CHiPS, Pooh, Popeye, Popov, Popsicle, Porsche, Puffer and Sons, Pyrex,	727
Queen Victoria, Quell, Quonset, Raleigh, Ralston-Purina, Razberry Zingers, Reader's Di-	728
gest, Red Man, Red Rose bag, Red Sox, Redball Flyer, Revell airplane, Richard Dreyfuss,	729
Richard Scarry, Ritz-Carlton, Roadrunner, Robbie the Robot, Robert A. Heinlein, Robert	730
Gordon, Robert Parker, Rockaway Beach, Rolling Stone, Rolls-Royce, Rolodex, Roman	731
Meal bread, Ronald McDonald, Rube Goldberg, Run Through the Jungle, SMERSH,	732
Sam Cunningham, Sammy's Pizza, Sara Lee, Saran Wrap, Saturday Night Live, Scarlett,	733
Schlitz, Schwinn, Scrabble, Scrooge, Sea World, Search for Tomorrow, Sears, Seeborg,	734
Seiko, Sesame Street, Shakey's, Shakin' Stevens, Shedd's Peanut Butter, Sherlock Holmes,	735
Shop 'n Save, Shuffle Off to Buffalo, Silex hotplate, Slim Jims, Smith & Wesson, Smith	736
Brothers' Wild Cherry, Smucker's, Snackin' Cakes, Snickers, Snoopy, Sonny Bono, Sony,	737
Space Invaders, Speedaway sled, Spic 'n Span, Spiderman, Spode china, Springsteen,	738
Star Blazers, Star Wars, Starsky & Hutch, Sterno, Steve Martin, Stonehenge, Studebaker,	739
Styrofoam, Sugaree, Sunkist, Sunoco, Superwoman, Sweet 'n Low, T. S. Eliot, TWA	740
flightbag, Tarzan, Texaco, The Bangor Daily News, The CBS Morning News, The Cat	741
in the Hat, The Chicago Tribune, The Creature from the Black Lagoon, The Crosswits,	742
The Deer Hunter, The Deering Ice Cream Parlor, The Doctors, The Doobie Brothers,	743
The Drac Pack, The Grateful Dead, The Headless Horseman, The Hundred Acre Wood,	744
The Jimmy Durante Hour, The Kingston Trio, The Little Rascals, The Man from Glad,	745
The Maytag Repairman, The Mellow Tiger, The Monkey's Paw, The Muppet Show, The	746
New England Patriots, The New York Times, The New York Yankees, The New Zoo	747
Revue, The PTL Club, The Pulitzer Prize, The Ramones, The Red & White grocery store,	748
The Reds, The Rolling Stones, The Rookies, The San Diego Padres, The Super Bowl,	749
The Taste Freeze, The Temptations, The Tigers, The Today show, The Toledo Blade,	750
The Tonight Show, The United Van Lines, The Washington Post, The Weapon Shops of	751
Ishtar, The Whitehall Hotel, The Wind and the Willows, The Young and the Restless,	752
Thermos, This Ole House, Thorazine, Thunderbird, Tide, Tiffany box, Tiger tank, Tiger,	753
Time magazine, Tipperary, Titanic, Toad, Tom Rush, Tom Watson, Tonka bulldozer,	754
Tonto, Toonerville Trolley, Top Job, Trace Optical, Trinitron, Tuborg, Tuinal, Tupperware,	755
Turtle Wax, Twinkie, U.S. of Archie, UN Plaza, Underalls, Underwood, United Airlines,	756
United Cerebral Palsy, Upjohn, Utica Club beer, Valium, Van Donen, Van Vogt, Vantage	757
cigarette, Vega, Vermont Maid Syrup, Victor Jory, Visa, W. W. JACOBS, WACZ, WCSH,	758
WOXO, Walter Mitty, Watson, Watson's Hardware, Weight Watchers, Wells Fargo truck,	759
Wendigo, Wheel of Fortune, Where the Wild Things Are, White Lightning, White Line	760
Fever, Wilbur, Wilkie Collins, Willy Loman, Willy Wonka's Great Glass Elevator, Winch-	761
ester, Winnebago, Winnie the Poe, Winston Churchill, Wonder Bread, Woolco, Wurlitzer	762
jukebox, Wyeth, Wyman, Xerox, Zayre, Zenith television, Zig-Zag paper, Zippo	763

C.4 Koontz

764

References extracted from *The Eyes of Darkness* (1981), *Phantoms* (1983), and *Darkfall* (1984): 765
766

7-Eleven, ABC, Abominable Snowman, Alan Alda, Alan Jackson, Albert Einstein, Al- 767
ice in Wonderland, Alien, Amelia Earhart, American Express, Andrew Wyeth, Ann 768
Landers, Bagley, Baloney, Barry Fitzgerald, Barry Manilow, Batman, Batmobile, Beat- 769
les, Beethoven, Bell JetRanger, Benny Goodman, Bermuda Triangle, Bernaise, Big Mac, 770
Biosan-4, Botticelli Madonna, box of Cheer, Bulova, Busby Berkeley, CBS, Cadillac Seville, 771
Cartier, Celica, Cessna, Charles Dickens, Charles Manson, Cheerios, Chevrolet, Chevy, 772
Chewbacca the Wookie, Chivas Regal, Clairol, Coke, Coleman gas, Coors, Copernicus, 773
Culligan, Dear Abby, Dennis the Menace, Dickens, Disneyland, Donatella, Dr Pepper, 774
Dr. Faustus, Dracula, E.T., Eleanor Rigby, Electronic Battleship, Elmore Leonard, Elvis, 775
Eroica, Explorer, Follett, Ford, Formica, Forsythe, Francis Bacon, Frank Sinatra, Franken- 776
stein, Frosty the Snowman, Fudge Fantasies, Garth Brooks, General Electric, George 777
Alexander, George Bernard Shaw, George Plimpton, Godzilla, Goodwill Industries, 778
gouda, Gore, Graveyard, Groucho Marx, Gucci, Hallmark, Hank Thomas, Harley, Heck- 779
ler & Koch, Honda, Hostess Twinkies, Howdy Doody, Irish Spring, Jack the Ripper, 780
Jacqueline Bisset, Jalape 241, James Bond, Jasper Johns, Jeep, Joel Bandiri Presents, Judge 781
Crater, K-Mart, Kleenex, Kraft Swiss cheese, Lalique, Land Rover, Lazarus, Levolor, 782
Lexus, Life-Savers, Listerine, Lovecraft, Lyndon Johnson, M-1 semiautomatic, Ma Bell, 783
MacLean, Magnum, Mario's Pizza, Marquis de Sade, Mary Celeste, McDonald's, Memo- 784
rex, Mennen's Skin Conditioner, Mercedes, Mercedes-Benz, Mickey Mouse, Millionaire's 785
Row, Mother Teresa, Mumm, Mussolini, NBC, Nash Rambler, New York Philharmonic, 786
Noah, Norman Rockwell, O.J. Simpson, Pepsi, Plexiglas, Pontiac Trans Am, Pop Tarts, 787
Pulitzer, Purina Cat Chow, Queen Anne, R. L. Stine, Rancho Circle, Raquel Welch, Rem- 788
ington, Remy Martin, Robert Redford, Rolex, Rolls-Royce, Rubik's Cube, Rudolph the 789
Red-Nosed Reindeer, San Francisco Chronicle, Saran Wrap, Scott Baio, Sears, Seiko, 790
Sharon Tate, Sheraton, Sherpa, Sidney Poitier, Silver Bells, Skylane RG, Smith & Wesson, 791
Spam, St. Francis of Assisi, Star Wars, Superman, The Book of Job, The Cessna Turbo, 792
The Mad Hatter, The New York Times, Thomas Mann, Time, Timex, Tiny Taylor, Tolstoy, 793
Tom Dooley, Tonka Toys, Toyota, Tums, Vaseline, Vince Foster, Walt Disney, Walter 794
Raleigh, Whiffle Ball, Wild Turkey, the Associated Press, the Bermuda Triangle, the New 795
York Times, the Plaza Hotel, the Saturday Evening Post, the Twilight Zone, the Wall 796
Street Journal 797

C.5 Straub

798

References extracted from *Ghost Story* (1979), *Shadowland* (1980), and *Floating Dragon* (1982): 799
800

A Praise of His Lady, ABC, Abraham Lincoln, Adidas, Adolf Eichmann, Agatha Christie, 801
Alan Alda, Alan Ladd, Albert de Salvo, Amanda Cross, American Express, Andre 802
Previn, Andrew Wyeth, Anne Bancroft, Anthony Powell, Aramis cologne, Archer Ho- 803
tel, Archie Goodwin, Aretha Franklin, Arnold Palmer, Art Carney, Art Deco, Arthur 804
Fonzarelli, Arthur Schlesinger, Audi, Audie Murphy, Audubon, BMW, Baldwin, Bambi, 805
Bass Weejuns, Beatle, Ben Jonson, Ben Sidran, Benny Goodman, Betamax, Big Mac, 806

Bill Perkins, Bill Terry, Bloomingdale, Blue Mountain beans, Bluto, Bo Diddley, Bobby	807
Hackett, Bokhara rug, Bose, Bowie knife, Brooks Brothers, Brothers Grimm, Bruno,	808
Bruno Hauptmann, Buck Rogers, Bud, Budweiser, Buick, Burberry, Burger King, Burl	809
Ives, Butch Cassidy, CBS, Cadillac, Camaro, Campanella, Campari, Carrie, Cary Grant,	810
Chaplin, Charles Addams, Charlie Antolini, Charlie Farrell, Charlie's Angels, Chaucer,	811
Chevrolet, Chevy, Chiquita Banana, Christopher Isherwood, Cinderella, Claire Bloom,	812
Clark Gable, Claude Rains, Coke, Constance Talmadge, Coors, Cornelius Agrippa,	813
Cortez, Corvette, County Fair, Cuisinart, Currier and Ives, Cynara, D. H. Lawrence,	814
Daimler, Daniel Boone, Dar Duryea, Death Star, Desi Arnaz, Diet Pepsi, Dingo, Disney,	815
Dodge, Dolly Parton, Dom Perignon, Don Juan, Dorothy Sayers, Dr Faustus, Dr Pep-	816
per, Dracula, E. B. White, Eames chair, Edgar Allen Poe, Edvard Munch, Eliphas Levi,	817
Elizabeth Jane Howard, Ella Fitzgerald, Elvis Presley, Emma Bovary, Ernest Dowson,	818
Ernest Hemingway, Ernie Kovacs, Eugene Palette, Exxon, F. SCOTT FITZGERALD, F.	819
W. Dixon, Falada, Flexible Flyers, Florence Nightingale, Fludd, Ford, Foreign Intrigue,	820
Four Roses, Frank Sinatra, Frankenstein, Fred Astaire, Freud, Frosty the Snowman,	821
Gary Cooper, Gatsby, Gene Shalit, General Pershing, George Shearing, Glenn Miller,	822
Golden Chicken, Goodwill, Grace Bumbry, Great Expectations, Gremlin, H. P. Lovecraft,	823
Halston dress, Hank Williams, Hansel, Harpo Marx, Harry Carey, Jr., Harry Truman,	824
Henry James, Herman Wouk, Holden Caulfield, Houdini, Huckleberry Finn, Humphrey	825
Bogart, Hush Puppies, I Love Lucy, IBM, Ichabod Crane, Ilie Nastase, Isobel Archer, Ivan	826
the Terrible, Ivy Compton-Burnett, J. D. Salinger, JOHN BARRYMORE, Jack Nicholson,	827
Jackie Gleason, Jaguar, James Bond, James Cagney, James Dean, James Fenimore Cooper,	828
James Stewart, Jameson, Jane Pauley, Janet Gaynor, Jann Wenner, Jean Harlow, Jean	829
Rhys, Jell-O, Jim Beam, Jim Reeves, Jimmy Durante, Jimmy Nervo, Joan Crawford, Joe	830
McCarthy, John D. MacDonald, John Dean, John Denver, John Ford, John Gilbert, John	831
Held, John Kennedy, John Scarne, John Updike, Johnnie Ray, Johnnie Walker, Johnny	832
Carson, Johnny Unitas, Johnny Walker Black, Josh Randall, Joyce Carol Oates, Katharine	833
Hepburn, Keds, Kitaj, Kiwanis, Kleenex, Knute Rockne, La Belle Dame Sans Merci, La	834
Grande Illusion, Labatt, Laura Ashley, Le Baron, Lewis and Clarke, Liane D'Eve, Lincoln,	835
Links Golf, Lion's Club, Little Red Riding Hood, Lonnie Donegan, Loretta Lynn, Louise	836
Brooks, Lully, MG, Madame Blavatsky, Madame Bovary, Mae West, Magic Fingers, Mai-	837
die Scott, Manson family, Margaret Drabble, Margaux, Marilyn Monroe, Mark Hopkins,	838
Marlboro, Mars bar, Mary Astor, Mary Miles Minter, Mary Pickford, Matchbox, Mather,	839
Mazda, McDonald's, Mello Yello, Melvyn Douglas, Mercedes, Miami Herald, Mickey	840
Mouse, Midnight Tango, Miss Marple, Monopoly, Monsieur Verdoux, Morgan, Mount	841
Rushmore, NBC, Nansen, Napa Valley Chardonnay, Nathaniel Hawthorne, National	842
Broadcasting Company, Nero Wolfe, Night of the Living Dead, Noble Sissle, Norma	843
Shearer, Norman Rockwell, Nosferatu, O. Henry, Old Parr, Oliver Hardy, Oreos, Orson	844
Welles, Ouija board, Packard car, Pampers, Pandora's Box, Pansy Osmond, Paul Hor-	845
nung, Paul Scott, Paul Stuart, Pepsi, Pepsi-Cola, Pernod, Perrault, Peter Lorre, Peter Pan,	846
Peugeot, Picasso, Piggly Wiggly, Pissarro, Pitch 'n Putt golf, Playboy, Plaza Hotel, Pocket	847
Books, Polka Dots and Moonbeams, Ponce de Leon, Princeton University Press, Pulitzer-	848
prize, Puvis de Chavannes, R. D. Jameson, R. P. Blackmur, Randolph Scott, Rasputin,	849
Raymond Chandler, Remington, Remy Martin, Renoir, Rex Bell, Rex Stout, Rex, the	850
Wonder Horse, Rialto theater, Richard Barthelmess, Richard Speck, Rip Van Winkle,	851
Robert Burns, Robert Ferguson, Robert Frost, Robert Redford, Robert Reed, Rolling	852
Stone, Rosa Forte, Rose Room, Rotary, Sam Shepard, Scott Fitzgerald, Sears, Sergeant	853

York, Shaker chair, Sheldon Leonard, Shiraz carpet, Sir Walter Scott, Smith and Wesson, 854
 Snickers, Snoopy, Snowy Breasted Pearl, Song for a Sucker Like You, Sony, Spencer Tracy, 855
 Spiderman, Starsky and Hutch, State Farm, Stephen Crane, Stetson, Steve McQueen, 856
 Steve Miller, Stilton, Stolichnaya, Studebaker, Stychen Tyme, Sunoco, Surfin' Bird, Sweet 857
 Sue, Ted Koppel, Teddy Knox, Texas Ranger, The Alamo, The Archer Hotel, The Brady 858
 Bunch, The Cheshire Cat, The Elks, The Everly Brothers, The Far Side of Paradise, The 859
 Garrick Club, The Hands of Dr. Orlac, The Hardy Boys, The Honeymooners, The House 860
 of the Seven Gables, The Invisible Man, The Jaycees, The John Birch Society, The Kansas 861
 City Times, The Key of Solomon, The Legend of Sleepy Hollow, The Little White Cloud 862
 That Cried, The Mad Hatter, The Making of a Surgeon, The Marble Faun, The Million 863
 Dollar Roundtable, The Narrative of A. Gordon Pym, The National Rifle Association, 864
 The Odyssey, The Phil Donahue Show, The Red Badge of Courage, The Rhetoric of Irony, 865
 The Scarlet Letter, The Statler Hilton, The Three Musketeers, The VFW, The Wizard of 866
 Oz, There's a Small Hotel, Thomas Mann, Tiffany lamp, Tom Brokaw, Tom Seaveris, 867
 Tommy Flanagan, Tony Archer, Tottel's Miscellany, Toyota, Treasure Island, Tupperware, 868
 Tweedledum and Tweedledee, UPS, Uri Geller, Valium, Van Helsing, Vandyke, Vanity 869
 Fair, Vanny Chard, Vergil, Village Pump restaurant, Vilma Banky, Vincent Price, Vir- 870
 ginia Woolf, Volvo, W. C. Fields, W. H. AUDEN, Wabash Cannonball, Waldenbooks, 871
 Waldorf-Astoria, Walter Cronkite, Wayne Booth, When The Red, Red Robin Goes Bob, 872
 Bob, Bobbin Along, William Bendix, William Powell, Willie Mays, Willie Nelson, Wilton, 873
 Wimbledon, Winnebago, Woodward and Bernstein, Wyatt Earp, Xerox, YPSL, Yoda, 874
 Young Brothers department store, Zoot Sims 875

D.

876

Table 6: Table reporting the results of Wilcoxon rank-sums one-sided tests comparing pop-culture reference counts extracted from randomly-sampled 10,000-token segments in Bachman, Straub, Koontz, King, and Harris books. Tests compared pop-culture reference counts in Bachman and Straub segments, Bachman and Koontz segments, Bachman and King segments, and Bachman and Harris segments.

Group	N	W	p
Bachman	700	615,746	< 0.001
Straub	1,200		
Bachman	700	894,245	< 0.001
Koontz	2,000		
Bachman	700	879,455	< 0.001
King	2,000		
Bachman	700	272,510	< 0.001
Harris	500		

Table 7: Table reporting the results of Wilcoxon rank-sums one-sided tests comparing pop-culture reference counts extracted from randomly-sampled 10,000-token segments in King, Straub, Koontz, and Harris books. Tests compared popular reference counts in King and Straub segments, King and Koontz segments, and King and Harris segments.

Group	N	W	p
King	2,000	1,506,216	< 0.001
Straub	1200		
King	2,000	2,096,930	0.004
Koontz	600		
King	2,000	672,109	< 0.001
Harris	500		

E.

Table 8: Table containing the mean, median, minimum, maximum, and standard deviations of pop-culture reference counts extracted from randomly sampled 10,000-token segments from the Bachman books and books in the distractor corpus.

Author	Book Title	Mean	Median	Minimum	Maximum	Standard Deviation
Bachman	<i>Blaze</i>	5.77	4.0	0	15	3.47
	<i>Rage</i>	21.72	18.5	7	48	10.27
	<i>Roadwork</i>	23.30	20.5	9	53	11.25
	<i>The Long Walk</i>	1.76	2.0	0	4	1.16
	<i>The Regulators</i>	3.39	3.0	0	18	3.00
	<i>The Running Man</i>	1.56	1.0	0	6	1.58
	<i>Thinner</i>	16.37	12.0	5	47	9.43
Harris	<i>Black Sunday</i>	1.22	1.0	0	4	1.19
	<i>Hannibal</i>	2.60	2.0	0	13	2.09
	<i>Hannibal Rising</i>	0.79	1.0	0	2	0.82
	<i>Red Dragon</i>	2.13	1.0	0	7	2.10
	<i>The Silence of the Lambs</i>	3.78	2.0	0	10	3.15
King	<i>Bag of Bones</i>	3.68	3.0	0	22	2.96
	<i>Carrie</i>	2.85	2.0	0	8	2.11
	<i>Cell</i>	2.42	2.0	0	8	2.08
	<i>Colorado Kid</i>	4.99	6.0	1	7	1.78
	<i>Desperation</i>	1.73	1.0	0	8	1.78
	<i>Dreamcatcher</i>	4.24	3.0	0	15	3.53
	<i>From a Buick 8</i>	25.61	25.5	8	45	7.58
	<i>Insomnia</i>	3.36	2.0	0	12	2.50

Continued on next page

Table 8 – Continued from previous page

Author	Book Title	Mean	Me- dian	Minimum	Maximum	Stan- dard Devia- tion
	<i>Lisey's Story</i>	3.54	3.0	0	14	3.52
	<i>Rose Madder</i>	2.79	2.0	0	12	2.92
	<i>'Salem's Lot</i>	6.76	4.0	0	33	8.04
	<i>Song of Susannah</i>	2.61	2.0	0	8	2.47
	<i>The Dark Tower</i>	2.22	2.0	0	9	2.14
	<i>The Dead Zone</i>	4.82	5.0	0	15	2.96
	<i>The Girl Who Loved Tom Gordon</i>	13.16	13.0	6	25	4.25
	<i>The Green Mile</i>	9.29	6.0	0	34	9.76
	<i>The Shining</i>	3.47	2.0	0	14	2.90
	<i>The Stand</i>	23.55	20.0	0	89	21.40
	<i>Wizard and Glass</i>	1.19	0.0	0	10	2.03
	<i>Wolves of the Calla</i>	1.31	0.0	0	10	1.95
Koontz	<i>After the Last Race</i>	2.80	2.0	0	13	2.86
	<i>Beastchild</i>	0.18	0.0	0	1	0.39
	<i>By the Light of the Moon</i>	59.96	59.0	35	118	17.83
	<i>Fear Nothing</i>	1.41	1.0	0	6	1.40
	<i>From the Corner of His Eye</i>	3.86	3.0	0	26	4.47
	<i>Hideaway</i>	5.84	6.0	1	20	3.52
	<i>Lightning</i>	8.96	6.0	2	46	8.49
	<i>Night Chills</i>	1.13	1.0	0	4	1.13
	<i>Phantoms</i>	1.53	1.0	0	11	2.48
	<i>Star Quest</i>	0.00	0.0	0	0	0.00
	<i>Strangers</i>	4.17	4.0	0	28	3.47
	<i>The Bad Place</i>	4.82	4.0	0	30	5.98
	<i>The Eyes of Dark- ness</i>	4.73	4.0	0	14	3.72
	<i>The Good Guy</i>	6.78	3.0	0	30	7.50
	<i>The Husband</i>	12.88	12.0	3	25	5.64
	<i>The Taking</i>	1.31	1.0	0	5	1.35
	<i>The Vision</i>	4.32	4.0	0	10	2.46
	<i>Warlock</i>	27.99	27.5	6	58	14.89
	<i>Whispers</i>	2.17	1.0	0	11	2.40
	<i>Winter Moon</i>	3.32	2.0	0	17	3.78
Straub	<i>Floating Dragon</i>	2.61	2.0	0	10	2.61

Continued on next page

Table 8 – Continued from previous page

Author	Book Title	Mean	Me- dian	Minimum	Maximum	Stan- dard Devia- tion
	<i>Ghost Story</i>	28.23	23.0	0	78	24.31
	<i>Hellfire Club</i>	2.19	1.0	0	14	3.16
	<i>If You Could See Me Now</i>	1.76	2.0	0	5	1.11
	<i>In The Night Room</i>	3.73	3.5	0	11	3.06
	<i>Julia</i>	0.45	0.0	0	3	1.04
	<i>Koko</i>	3.05	2.0	0	16	3.42
	<i>Lost Boy, Lost Girl</i>	1.55	1.0	0	4	1.30
	<i>Mr. X</i>	2.42	2.0	0	8	1.59
	<i>Mystery</i>	1.88	1.0	0	14	2.59
	<i>Shadowland</i>	1.96	1.0	0	9	2.35
	<i>The Throat</i>	2.99	2.0	0	15	3.41



Figure 4: Boxplot visualizing pop-culture reference counts by book title. Counts were extracted from randomly sampled 10,000-token segments from the Bachman books and books in the distractor corpus.

F. Data Availability

878

Due to copyright restrictions, the full texts and segments of the King, Straub, Harris, 879 and Koontz books used in our experiments cannot openly be shared. Data extracted 880

from full texts can be found here: <https://doi.org/10.5281/zenodo.7956049> 881

G. Software Availability 882

Software can be found here: https://anonymous.4open.science/r/king_bachman_authorship_verification-B16D 883
884

H. Acknowledgements 885

Funded in part by the FWO research project "Creating Suspense Across Versions: Genetic Narratology and Stephen King's IT" at the University of Antwerp. 886
887

I. Author Contributions 888

References 889





- Altakrori, M., J.C.K. Cheung, and B.C. Fung (Nov. 2021). "The topic confusion task: A novel evaluation scenario for authorship attribution". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4242–4256. 890
891
892
- Anonymous (Nov. 1984). "No Dieting after 'Thinner'". In: *The Kingsport Times-News*, 4G. 893
- Bevendorff, Janek, BERTa Chulvi, Gretel Liz De La Peña Sarracén, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, et al. (2021). "Overview of pan 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*. Springer, 419–431. 894
895
896
897
898
899
900
- Boenninghoff, Benedikt, Steffen Hessler, Dorothea Kolossa, and Robert M. Nickel (2019). "Explainable Authorship Verification in Social Media via Attention-based Similarity Learning". In: *IEEE International Conference on Big Data (IEEE Big Data 2019)*, Los Angeles, CA, USA, December 9–12, 2019. 901
902
903
904
- Bradley, Linda (1998). "The Sin Eater: Orality, Postliteracy, and the Early Stephen King". In: *Stephen King*. Ed. by Harold Bloom. Bloom's Modern Critical Views. Philadelphia: Chelsea House, 95–124. 905
906
907
- Burrows, John (Sept. 2002). "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship". In: *Literary and Linguistic Computing* 17.3, 267–287. ISSN: 0268-1145. 10.1093/llc/17.3.267. eprint: <https://academic.oup.com/dsh/article-pdf/17/3/267/2743069/170267.pdf>. <https://doi.org/10.1093/llc/17.3.267>. 908
909
910
911
- Collings, Michael R. (1998). "Dean Koontz and Stephen King: Style, *Invasion*, and an Aesthetics of Horror". In: *Discovering Dean Koontz: Essays on America's Bestselling Writer of Suspense*. Ed. by Bill Munster. Cabin John: Wildside Press. 912
913
914
- Denger, Laurie (Jan. 1985). "Bachman Novel has Thrills, Chills, Gaps". In: *The Dayton Daily News*, 6D. 915
916
- Dewes, Joyce Lynch (Mar. 1981). "Interview: Stephen King". In: *Mystery Magazine*. 917
- Eder, Maciej (2015). "Does size matter? Authorship attribution, small samples, big problem". In: *Digital Scholarship in the Humanities* 30.2, 167–182. 918
919

- Eder, Maciej (2018). "Elena Ferrante: a virtual author". In: *Drawing Elena Ferrante's Profile*, 31–46. 920
921
- Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt (June 2017). "Understanding and explaining Delta measures for authorship attribution". In: *Digital Scholarship in the Humanities* 32.suppl₂, ii4–ii16. ISSN: 2055-7671. 10.1093/llc/fqx023. eprint: <https://academic.oup.com/dsh/article-pdf/32/suppl\2/ii4/21298943/fqx023.pdf>. <https://doi.org/10.1093/llc/fqx023>. 922
923
924
925
926
927
- Frank, Sam (July 1982). "Running Man Beats the Odds". In: *The San Francisco Examiner*, 6. 928
- Ganley, W Paul (1985). "Thinner, by Richard Bachman". In: *Fantasy Mongers* 13. 929
- Graham, Mark (Dec. 1984). "Fit for a King: This Thriller Raises an Authorship Question". In: *The Rocky Mountain News*, 26–N. 930
931
- Grooms, Roger (Nov. 1984). "Combined Novel has King's Flavor". In: *Palladium-Item*, E5. 932
933
- Juola, Patrick (Aug. 2013a). "How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling". In: *Scientific American*. <https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoo-s-calling/>. 934
935
936
937
- (July 2013b). *Rowling and 'Galbraith': an Authorial Analysis*. <https://languagelog.ldc.upenn.edu/nll/?p=5315>. Language Log. UPenn. 938
939
- (Oct. 2015). "The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions". In: *Digital Scholarship in the Humanities* 30.suppl₁, i100–i113. ISSN: 2055-7671. 10.1093/llc/fqv040. eprint: <https://academic.oup.com/dsh/article-pdf/30/suppl\1/i100/1038325/fqv040.pdf>. <https://doi.org/10.1093/llc/fqv040>. 940
941
942
943
944
- King, Stephen (1982a). *Danse Macabre*. London: Futura. 945
- (1982b). "Peter Straub: An Informal Appreciation". In: *Program Book, World Fantasy Convention '82*. Ed. by Kennedy Poyser. New Haven, CT: World Fantasy Convention. 946
947
- (1985). *The Bachman Books*. New York: New American Library. 948
- (1996). "The Importance of Being Bachman". In: *The Bachman Books*. New York: Plume. 949
950
- Koppel, Moshe, Jonathan Schler, and Shlomo Argamon (2009). "Computational methods in authorship attribution". In: *Journal of the American Society for Information Science and Technology* 60.1, 9–26. <https://doi.org/10.1002/asi.20961>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20961>. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20961>. 951
952
953
954
955
- (2011). "Authorship attribution in the wild". In: *Language Resources and Evaluation* 45.1, 83–94. 956
957
- Koppel, Moshe, Jonathan Schler, and Elisheva Bonchek-Dokow (2007). "Measuring Differentiability: Unmasking Pseudonymous Authors". In: *Journal of Machine Learning Research* 8.45, 1261–1276. <http://jmlr.org/papers/v8/koppel07a.html>. 958
959
960
- Koppel, Moshe and Yaron Winter (2014a). "Determining if two documents are written by the same author". In: *Journal of the Association for Information Science and Technology* 65.1, 178–187. <https://doi.org/10.1002/asi.22954>. eprint: <https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22954>. <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.22954>. 961
962
963
964
965

- Koppel, Moshe and Yaron Winter (2014b). "Determining if two documents are written by the same author". In: *Journal of the Association for Information Science and Technology* 65.1, 178–187. 966
- Lehmann-Haupt (Nov. 1984). "An Ungainly Offspring". In: *The Sunday Herald-Times*, C–15. 969
- Levin, Bob (Dec. 1984). "Novel Cursed by Cliches, Thin Characterizations". In: *The Atlanta Constitution*, 9–J. 970
- O'Neil, Ann W (Dec. 1984). "A Horrifying Weight-loss Plan". In: *The Philadelphia Daily News*, 51. 971
- Potha, Nektaria and Efstathios Stamatatos (2014). "A profile-based method for authorship verification". In: *Hellenic Conference on Artificial Intelligence*. Springer, Cham, 313–326. 972
- Richards, Brian (1987). "Type/token ratios: What do they really tell us?" In: *Journal of child language* 14.2, 201–209. 973
- Slotek, Jim (June 1981). "Roadwork, by Richard Bachman". In: *The Ottawa Citizen*, 10. 974
- Smith, Joah H (Feb. 1985). "Pseudonym Kept Five King Novels a Mystery". In: *The Bangor Daily News*, 1. 975
- Strachan, Don (Mar. 1981). "Soft cover". In: *The Los Angeles Times*, 8. 976
- Straub, Peter (1984). "Meeting Stevie". In: *Fear Itself: The Horror Fiction of Stephen King*. Ed. by Tim Underwood and Chuck Miller. New York: New American Library. 977
- Thomases, Martha and John Robert Tebbel (Jan. 1981). "Interview with Stephen King". In: *High Times Magazine*. 978
- Tirvengadam, Vina (1996). "Linguistic Fingerprints and Literary Fraud". In: *Digital Studies/le Champ Numérique* 2.1. 979
- Tuzzi, Arjuna and Michele Alberto Cortelazzo (2018). "It takes many hands to draw Elena Ferrante's profile". In: *Drawing Elena Ferrante's Profile*, 9–30. 980
- Tyo, Jacob, Bhuwan Dhingra, and Zachary C. Lipton (2022). "On the state of the art in authorship attribution and authorship verification". In: *arXiv preprint arXiv:2209.06869*. 981
- Underwood, Tim and Chuck Miller, eds. (1989). *Bare Bones: Conversations on Terror with Stephen King*. New York: Warner Books. 982
- Van Cranenburgh, Andreas and Erik Ketzan (2021). "Stylometric Literariness Classification: the Case of Stephen King". In: *Proceedings of LaTeCH-CLfL 2021*. <https://aclanthology.org/2021.latechclfl-1.21.pdf>. 983
- Williams, Nick B (Nov. 1984). "Thinner". In: *The Los Angeles Times Book Review*, 11. 984
- Winter, Douglas E. (Feb. 1985). "Stephen King, Peter Straub, and the Quest for the Talisman". In: *Twilight Zone Magazine*. 985

Extracting Geographical References from Finnish Literature

Fully Automated Processing of Plain-Text Corpora

Harri Kiiskinen¹ 
Asko Nivala² 
Jasmine Westerlund² 
Juhana Saarelainen² 

1. Library, Tampere University , Tampere, Finland.
2. Department of Cultural History, University of Turku , Turku, Finland.

Citation

Harri Kiiskinen, Asko Nivala, Jasmine Westerlund, and Juhana Saarelainen (2023). "Extracting Geographical References from Finnish Literature. Fully Automated Processing of Plain-Text Corpora". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-21

Date received 2023-01-27

Keywords

named entity recognition, geographic information system, geoparsing, linked open data, literary geography, Finland

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. In *Atlas of Finnish Literature 1870-1940* project, we extract geographical information from a Finnish-language corpus of literary texts published between 1870 and 1940. The texts are transformed from plain texts to TEI/XML, and further processed with named entity recognition and linking tools. The results are presented in web-based environment. This article describes the technical structure of the analysis chain, the tools used and the metaprocesses used to manage the research dataset.

1. Introduction

1.1 Project

Literary geography asks where literature is located and why. The term was already in use in the early 1900s (for the definition and history of literary geography, see Piatti 2008, 20, 65–121). As a modern research paradigm, literary geography emerged in the late 1990s as part of the spatial turn. Franco Moretti's *Atlas of the European Novel* can be considered a classic of this approach (Moretti 1998). Literary geography can also use cultural geography methods and qualitative analysis, in which case it does not necessarily make use of maps (see e.g. Tally 2018). However, it often aims to map place names or other variables, in which case it can also be called literary cartography.

Place names and spatial information can be annotated and collected manually from texts. However, language technology allows spatial information to be extracted using computational methods. In the 2000s, literary geography has adopted the methods of digital humanities, combining the natural language processing (NLP) with geographic information systems (GIS). Gregory and Hardie have used a part of speech tagger (POS) to extract proper nouns from the *Lancaster Newsbooks Corpus* (1653–1654) and filtered the results to place names using a gazetteer. They have then imported the place names onto maps using GIS and used density maps to analyse the results (Gregory and Hardie 2011). In a similar project, works describing the English Lake District, were parsed for place names and the place names were references using a Gazetteer (Gregory et al. 2019). The Edinburgh Geoparser, that was also used in these projects, has still been developed further (Alex et al. 2019).

Named entities in literature have been studied mostly in English-language texts, but research has also extended to smaller language areas. For example, the ELTeC (the European Literary Text Collection) COST action has done significant work on the study of named entities in non-English European literature (Frontini et al. 2020). Transformers and BERT language models have significantly improved Named Entity Recognition (NER) tasks in non-English languages due to their ability to capture contextual information and learn representations from vast amounts of unlabeled text (Labusch et al. 2019). By employing self-attention mechanisms, transformers can effectively model long-range dependencies and capture intricate patterns in the data. BERT, in particular, introduced the concept of pretraining and fine-tuning, enabling the model to learn from massive amounts of text before being fine-tuned on specific NER tasks (Devlin et al. 2018). This approach has led to substantial advancements in NER performance also in Finnish language datasets (Luoma et al. 2020).

Our *Atlas of Finnish Literature 1870–1940* project applies similar methods for the first time to the study of Finnish literature. The objectives of the project are:

- to recognise place names mentioned in Finnish literature between 1870 and 1940
- to store geo-annotated texts and their metadata in a database
- to geocode and enrich the identified place names by linking them to linked open data (Wikidata, DBpedia etc.)
- to publish interactive maps showing the locations mentioned.

NER can be used to extract proper nouns – such as people, places and organisations – from unstructured text. As our project focuses on literary geography, we are mainly interested in political and geographical toponyms, as well as street and building names. After the recognition process, the locations are disambiguated, and geographic coordinates are retrieved by linking them to the linked open data. The location names can then be enriched using semantic web databases such as DBpedia and Wikidata. This allows, for example, the separation of cities from natural landscapes (e.g. rivers, lakes and mountains).

The aim of our project is to present a new interpretation of the geography imagined by Finnish-language literature in 1870–1945. The identification of named spatial entities offers a new method of exploring the territories to which fictional texts refer that has not been previously applied in Finnish scholarship. Based on the preliminary results, it seems that during the period under study, spatial references in literature were particularly related to nation-building and the definition of the territory of Finnish culture in relation to Sweden in the west, Russia in the east and the Sámi regions in the north. Historical novels seem to contain considerably more references to existing place names than other genres. Germany, Italy and France were the settings for many of the novels, and the migration from Finland to America is also reflected in the literature, but there are surprisingly few references to the British Isles. On the other hand, the polyphonic nature of literature should be emphasised: for example, labour literature had different aims from nationalist historical novels. However, we will present the literary-historical results in future publications once the mapping and historical analysis of the data has been completed. The purpose of this paper is to describe the construction of the compu-

tational infrastructure of the study: corpus design, NER, NEL and the production of annotated XML-TEI. 66
67

In this paper, we will describe the text corpus used and the methods applied and developed in the project. In the introduction, we describe the characteristics of Finnish literary history, our two corpora and the principles used to collect and verify the metadata. In Section 2, we then describe the automated process by which the plaintext is converted into XML-TEI format, segmented into discrete works and chapters, and how the named entities are identified and linked. In Section 3 we describe how the data processing has been implemented with the message-broker system and what kind of manual checks have been made on the results to control the accuracy of the results. 68
69
70
71
72
73
74
75

1.2 Literary Historical Context and Primary Sources 76

Finland is a bilingual country, that was part of Sweden from the Middle Ages to 1809. After the Finnish war 1808–1809, Finland was annexed to Russia and became an autonomous Grand Duchy. Finnish was the language of common people, even though the first printed books in Finnish were published already in the 1540s. Great majority of the Finnish intelligentsia were Swedish-speakers still at the end of the nineteenth century. The century was a time of modernisation for the Finnish language. Both Finnish and Swedish were defined as national languages in 1919. 77
78
79
80
81
82
83

At this stage, our project only covers texts written originally in Finnish. As said, Finland is a bilingual country, and Swedish-language fiction was also published during the period 1870–1940. In the future work stages of our project, we intend to study them as well. However, named entity recognition in Swedish requires the development of a separate NER and disambiguation system. For this reason, we limit our research at this stage to Finnish-language fictional works – including novels, dramas and poems. A comparison of Finnish- and Swedish-language literature could reveal interesting results concerning, for example, whether Finnish literature in Swedish describes more southern cities, coastal areas and archipelagos. On the other hand, Swedish-language material would also provide an opportunity to study a much earlier historical period when no significant Finnish-language literature was published. We will endeavour to carry out this study at a later stage. 84
85
86
87
88
89
90
91
92
93
94
95

Seitsemän veljestä (*Seven Brothers*, 1870) by Aleksis Kivi is usually considered as the first novel written in Finnish. The year 1870 is also the starting point of our research project. Additionally, it has been also suggested that it was only by circa 1880 that Finnish had developed into its modern form as a literary language, previous decades 1810–1880 being labelled as “early modern period” and the time before that as “old literary Finnish”. Therefore, our project focuses on first decades of Finnish taking its modern shape. Finnish literature underwent an enormous increase and progress during the period we study. Many new writers, including Minna Canth, Juhani Aho, Arvid Järnefelt and Santeri Alkio, rose to popularity. The Finnish nationalism and the Fennoman movement actively promoted Finnish language and literature. The labour movement and women’s right movement facilitated social discussion and had a great impact on literature as well. In 1917, Finland was declared independent from Russia and in 1918 Finland went through a traumatic civil war. However, the European influences were also important for the development of Finnish literature. Realism arrived from 96
97
98
99
100
101
102
103
104
105
106
107
108
109

Scandinavia in the 1880s. In the 1920s, the *Tulenkantajat* (the Flame Bearers) advocated cosmopolitanism and sought to build international connections. Leftist writers came strongly forth in the 1930s. Economic depression and political crises briefly affected the number of literary publications in the 1930s, but the numbers started to raise already at the end of the decade. The end point of our project is the end of the Second World War (1945). *Terminus ad quem* is chosen on two criteria. First, the Second World War marked a thematic break in Finnish literature. Second, in Finland literary works are released from copyright 80 years after the death of the author, meaning that much of the literature published since the 1940s is not in the public domain.

The main dataset of the project is the public domain *Projekti Lönnrot* corpus of digital books in Finnish from the nineteenth and early twentieth centuries. The corpus has been created by volunteers, who have corrected the spelling of the digitised books. Therefore the plain texts are mostly free of noise and other errors from Optical Character Recognition (OCR) scanning, i.e. misidentified letters. *Projekti Lönnrot* includes classic works of Finnish literature but also popular fiction and other more marginal genres. However, the corpus also contains translations from other languages and non-fiction books, which we have filtered out. This ready-to-use resource in public domain has provided an excellent starting point for our project.

The second corpus we use for the project is *Project Gutenberg*, which also contains transcribed Finnish fiction texts. Much of its work is included in *Projekti Lönnrot*, but there are also many supplementary texts: only texts that are not in the Lönnrot collection have been manually selected. We have written a parser to segment the Gutenberg texts into chapters. After this step, the data has been processed with the same pipeline as the Lönnrot corpus.

1.3 Metadata

Collecting and curating the metadata has constituted a major part of the first year of our project. Collecting metadata consisted of five phases that were interlocked with the process of collecting the full texts:

1. In the case of *Project Gutenberg*, searching and selecting the Finnish works included in the multilingual collection. By contrast, the works in *Projekti Lönnrot* are all published in Finnish, although some are translated from other languages.
2. Excluding the translations. *Projekti Lönnrot* and *Project Gutenberg* include many Finnish translations of works written elsewhere and/or in other languages; if the translator is Finnish, the work has been defined as Finnish literature as well.
3. Manually reviewing all the Finnish works to exclude non-fiction and to find the ones matching the time span of our project.
4. Solving the question of the identity of the writer. Should collectors and editors of folk poetry be regarded as writers, and to which extent? The use of pseudonyms also caused problems in the collection of metadata.
5. The division of edited collections into individual works. *Project Gutenberg* and *Projekti Lönnrot* include many volumes of collected works. Often the information about the original publishing date of each individual work was not easy or even

possible to find. Many poems, dramas and short stories were originally published
in newspapers or periodicals. In these cases, we have included all metadata that
could be found (original publishing date, original title) in addition to the metadata
of the edited collection.

2. Process chain

2.1 Parsing plain text to TEI

The source files from the *Projekti Lönnrot* are plain text files, very similar to what the
Project Gutenberg uses. The files are encoded by human volunteers. In the process of
encoding, the texts are systematised to a certain extent, but as perhaps can be expected,
the results are not fully systematical: *e.g.* book sections are usually indicated with a
separator of four empty lines, but there can also be five. Moreover, the unit of digitisation
is a physical volume which, for example, in case of collected works and anthologies
results in many individual works appearing in one *Projekti Lönnrot* item.

In addition, the resulting text files use various encoding systems, UTF-8, Win-1252 and
ISO-8859-1 among others. This is a problem for Finnish texts, since there are various
special characters (“ö” “ä” “å” and their capitalised versions) used that are encoded with
different code points in each coding system. These letters are very common, and also
significant: the meanings of the words “läski” and “laski”, for example, have nothing to
do with each other. For processing the texts, it was necessary to transform everything
to Unicode.

This was done by analysing each file with the Unix file utility.¹ The utility return the
guessed encoding for a file based on an analysis of coding points in the texts, and in
most cases, the use of this reported encoding as argument to Java’s file input/output
routines resulted in a correct read.

The actual conversion problem was connected with the aims of the whole project. The
most simple way to process the texts would be to treat them as plain texts, tokenise
them and run through relevant tools. This approach is often used when NLP tools
are developed and measured, and the results are usually in some kind of standardised
format, like CoNLL-U² or similar. The purpose of this project, however, is to use the
NLP tools only as a first stage in a production context for geoparsing the texts and
representing their geographical information on maps. We do not focus on these results
as such, but use them further in working with the texts on a different level. Therefore,
the CoNLL-U type format is not suitable for the project, but we need to integrate the
annotations produced by the NLP processes into XML documents that are more suited
to both visual presentation of the text and dissemination of the annotated documents.

In order to fill the main obligations of the project, we decided that the texts resulting
from the processing pipeline should be TEI-encoded XML files in order to facilitate their
further use. This suggested a possible approach where the texts were converted to TEI
as early as possible. These texts were then to be used as sources for further analyses.

1. file -b -mime-encoding

2. <https://universaldependencies.org/format.html>

- The conversion process was divided in several phases: 191
1. parse the text files 192
 2. convert to parse-based XML 193
 3. convert this XML to TEI 194

The parser for the texts was written using the EBNF³ notation. The actual parse process 195
was done with Clojure code, running the Instaparse library (Engelberg 2022). 196

The parse was then converted to an XML document with the same structure that was 197
defined in the parser definition. This step is not strictly necessary, but producing TEI 198
directly from the parse results is very complicated, requires a lot of manual coding and 199
is therefore error-prone. A better solution is to export the parse results as XML, and 200
then further process this XML with a suitable XSLT to TEI (See `parse-to-tei.xsl` in 201
the project repository). 202

As this is the step where the individual works in the source volumes are recognised (see 203
the parser definition, especially the element `work1`, in the file `parser.bnf` in the project 204
repository), this was the occasion to create individual ID's for these works. The IDs 205
were created following a simple scheme: "`lonnrot_<basename>_<serial>`". At first, we 206
have the string "`lonnrot`" or "`gutenberg`" to signify the data corpus, then the `basename` 207
of the file which is the same as the volume ID in the *Projekti Lönnrot* or Project Gutenberg 208
corpus, and as a third element, the `serial` number of the recognised `work1` element. This 209
ID is stored as the `xml:id` attribute of the TEI element. 210

As a result of this conversion process, we have a set of files in TEI/XML format that 211
corresponds very closely with the original *Projekti Lönnrot* files. 212

2.2 Splitting to works and adding metadata 213

The problem with the first conversion process is that the resulting files closely resemble 214
the structure of the original text files. They do not contain any metadata, which is 215
difficult to extract reliably from the original files. Moreover, due to the structure of 216
the parser, the structure of physical volumes that contain more than one "work" by an 217
author or multiple authors are preserved. This means, that for these volumes, there is 218
the main TEI document reflecting the actual volume, containing each work as separate, 219
subordinate TEI documents.⁴ 220

The separation between different works is marked with (at least) six empty lines in the 221
Projekti Lönnrot files. A problem rises at the beginning of the digitised volumes, because 222
the first work contained in the physical volume is often, but not always, separated from 223
the header data by six empty lines. So for example a book containing a single work 224
can have six empty lines between the book bibliographic data and the actual work, but 225
not always; in many cases, the actual text begins after only four or five lines, which is 226
otherwise used to separate different chapters within the works. 227

This problem is also present in volumes containing multiple works, in the form of a 228

3. Extended Backus-Naur form.

4. This is supported by and allowed for by the TEI specification.

missing separator between the book header and the first work. 229

In practice, the main `tei` element can be the whole work or not; if it has one subordinate `tei` element, this can mean that there is one work on the volume, but there can be also two works, the first just being without a separator from the volume header. And whatever the number of the works, each has its own set of metadata (see Section 1.3) that must be added to the actual work. 230 231 232 233 234

The external metadata offered a solution in the form of providing information about how many works a volume should contain. This information, contained with the ability to extract all `tei` elements from the documents and count them, made combining work metadata with right works possible. The extracted `tei` elements were also turned into independent documents at this stage. 235 236 237 238 239

Another process that was implemented at this point was the identification of individual tokens in the text. The parsing process created individual tokens in the result data using TEI's elements `w`, `pc` and `num`. Since the idea was to further process the texts with external algorithms, it was necessary to think in advance how to merge the results back to the TEI documents, and for this reason, giving unique IDs to individual tokens was deemed necessary. 240 241 242 243 244 245

A simple naming scheme, based on the document ID (see above), was created: `<document_id>_token<serial>` where the `document_id` was the same used in the `xml:id` of the TEI element, and the `<serial>` was calculated using the XSL accumulator element. 246 247 248

As a result of this process, we have each “work” in the dataset (novel, play, etc.) in its own TEI file, with a unique ID used both as the `xml:id` attribute of the TEI root element as well as the base name for the XML file. Each work has a basic set of metadata defined in the TEI document header, including the author, the title, and the year of publication, as well as statements regarding the production of these digital editions of the works. 249 250 251 252 253

Each work has its main internal divisions marked with the TEI `div` elements, and paragraphs are surrounded with the `p` tags. The texts are tokenised using the respective TEI elements. Spaces are not marked, but they are left in the XML text. Since the whitespace handling of XML can be occasionally tricky, and some tools may mess up spaces, these can also be reconstructed later: the information about the lack of spacing is stored using the `join=“left”` attribute in the cases where the token does not have any whitespace before it. Thus, each token now has an ID that is unique for the whole project corpus. 254 255 256 257 258 259 260 261

As a result of this process, at the time of writing, we have a dataset of 848 texts in Finnish language, published for the first time between 1870 and 1944. These texts include altogether 20,356,701 tokens (see Table 1). 262 263 264

	1870s	1880s	1890s	1900s	1910s	1920s	1930s	1940s
fiction	216925	1795264	2422107	3115430	4703275	3638502	1201012	595485
drama	18060	149437	246392	497496	377157	201590	36375	15654
poetry	29430	74873	111124	195146	315331	91529	36486	4041
misc	0	0	0	22683	108038	115939	21920	0
TOTAL	264415	2019574	2779623	3830755	5503801	4047560	1295793	615180

Table 1: Sum of tokens per genre and decade

OntoNotesNE type	TEI element
GPE	<placeName>
LOC	<geogName>
PERS	<persName>
ORG	<orgName>

Table 2: Correspondence between the OntoNotesNE types and TEI elements.

2.3 Named Entity Recognition and lemmatisation

One of the key goals of the project is to automate the processes of place name recognition and referencing.

Many of the projects with similar aims are using geotagging processes based on either recognising certain types of names or using a gazetteer. For example, the *Edinburgh Geoparser* uses the latter approach, and requires therefore a pre-defined set of place names to work with (Alex et al. 2019). An example of the former method is the retrieval of Parisian street names from French novels between 1800 and 1914, that was based on the typical structure of street and other place names in the French language, where for example “rue” forms a part of most street names when referred to in the text (Moncla et al. 2017, 2019).

Instead of trying to create our own heuristics for this, we chose a very different approach by using Natural Language Processing algorithms created by the TurkuNLP group⁵, especially the *Finnish NER* (Named Entity Recognition) tool (Luoma et al. 2020). The *Finnish NER* system analyses the source text using a natural language model, and recognises various types of named entities in the text, including geopolitical place names, natural place names, persons, buildings, organisations, etc.⁶

The system takes as its input either a plain text file, which it tokenises itself, or an already tokenised list of words and punctuation where each token is on its own line. The latter is the only option if the purpose is to somehow connect the results with the original data, for in this case, there is a strict correspondence between the lines in submitted source data and received results.

As the words, punctuation, and numbers in the source documents were already tokenised in the previous stage, it is a simple task to extract a part of the source document as lines where each line contains the token ID and the content, separated with a tab. From this data, the content column can be separated, given as argument to the *Finnish NER*, and the results can be merged back with the original extract containing the token ID’s. This list can then be used as source data for an XSLT merging selected data from the results with the original TEI files.

In practice, this means choosing selected entity types from the NER results, finding the token ID’s covered by each named entity, and then surrounding these elements in the TEI with the corresponding tag.

Initially, it seemed that the TEI elements denoting places would be appropriate to

5. <https://turkunlp.org/>

6. Full list of recognised entity types can be found at the system web pages, and further description of the entity types in the OntoNotes Manual (Weischedel et al. 2012, 21)

OntoNotes entity types	Description
PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc.
ORGANIZATION	Companies, agencies, institutions, etc.
GPE	Countries, cities, states
LOCATION	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Vehicles, weapons, foods, etc. (Not services)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK OF ART	Titles of books, songs, etc.
LAW	Named documents made into laws
LANGUAGE	Any named language

Table 3: The list of the OntoNotesNE types

describe OntoNotesNE types (see Table 2 for correspondences between OntoNotesNE types and TEI elements). However, this mapping between the OntoNotesNE types produced by the *Finnish NER*, and the entity types offered by TEI is not without problems, for the TEI categories do not fully correspond to the OntoNotesNE types. For example, OntoNotesNE uses the FAC(ility) tag to denote buildings, roads, and other man-made architectural structures. In the TEI documentation, buildings are annotated under the name tag:

1 I never fly from <name key="LHR" type="place">Heathrow Airport</name> to <name key="FR" type="place">France</name>

(TEI Consortium 2022, 13.1.1 Linking Names and Their Referents)

To be able to tag facilities, historical events, or nationalities in texts, we need to use the XML tag “name”, where the type of the entity is given in the “type” attribute. We are currently considering a solution to this problem that would be in line with TEI principles, but at this point we have decided to use the name tag because it is suitable for all tagged entities and still meets the TEI specification.

As a result of this process, we have TEI documents with place names marked with name tags. In addition to the actual tags marking the entity, each annotation was given a unique ID using the xml:id attribute so that they can be referred to from elsewhere, just like was done with the tokens in the earlier stage.

For the next phase, Named Entity Linking, we need to have the place names in a form that can be used as search string in geodatabases. The NER process used above uses a natural language model, and does not lemmatise the tokens in the process; consequently, the resulting place names are still in their inflected form. In the Finnish language, common and proper nouns decline in 14 cases. For example, the nominative case for Berlin in Finnish is “Berliini”, and when someone is in Berlin, the inflected word form is “Berliinissä” (inessive case). Whereas “Berliini” is the nominative form of the word, and can be found as such in many databases, “Berliinissä” usually does not return any results.

Therefore, the results also need to be lemmatised in order to recover their root forms. For this purpose, we use another system created by the TurkuNLP group, the *Turku*

Neural Parser Pipeline (Kanerva et al. 2018, 2021) (TNPP). This system accepts the input in a similar format as the NER system, and returns results in the CoNLL-U format.

Since the input data for both systems is the same, the lemmatisation process was merged with the NER process. Because the results of the NER process had to be merged with the TEI files in four separate runs, the merging of lemmas just added a fifth data merge with XSLT, and as a result, the documents were lemmatised at the same time.

This is a side effect of using the TNPP system. Moreover, this system is based on the natural language model, and each token is analysed in the context of a sentence, where its position in relation to what comes before and after is significant. The entity names cannot therefore be extracted as individual names and lemmatised; but this also yields the lemmatisation of the whole documents. This POS (part of speech) data may yet prove to be useful in the later stages of the project, because we can link place names to their original context, including information such as which adjectives or verbs refer to them.

After running this process for the data files, we have a set of TEI document files, where in addition to the results of Section 2.2, each token has also the lemma of its content word, and the four name element types shown in Table 2 are marked in the texts.

2.4 Named Entity Linking

After the Named Entity Recognition process, the various entity types recognised are annotated with the respective TEI elements in the documents. This does not yet take us much further on the path of geographical analysis of the texts; we might be able to create some statistics about the density of various entity references in the works by different authors or of various genres, but these annotations do not allow for any cartographic analyses of the documents.

The place names need to have some additional data, coming from outside of documents, in order to be usefully analysed in any wider context. In principle, there are two ways to approach the geographical linking of the place names in the documents.

In the first option, for each place name, some kind of geographic location is looked for. This could be as simple as a reference point, expressed in any coordinate system (practically geographical data in any coordinate system can be transformed to WGS84, which is the format assumed by TEI and supported by all libraries and applications for geographical analysis), which could be stored together with the place name in order to provide a geographical location. This location can then be used to create geographical presentations of the texts, either individually or in larger sets.

In the second option, for each place name, the corresponding *place* entity in some reference system is found. This place entity is what we generally mean when we talk about different places, like “Berlin” is a place entity, but “52°31’N, 13°23’E” is a string containing the latitude and longitude in degrees of a point that could be used to represent the location referred to place name “Berlin” in the text. If instead of storing this string, we store a reference to the entity “Berlin, the Capital of Germany”, we can use this entity to gather more information about the places referenced in the documents.

The second option also provides a way to control the results of the linking process.

In the first option, what happens behind the scenes is obscured by storing only the results. Why is this particular location stored, and not some other? Without any other information than the geographical coordinates, it is not possible to understand why the reference in our text to “Berlin” suddenly shows up as a point in Wisconsin, USA, on our geographic display of the results. The second option is somewhat more complicated, but not too much not to be preferred.

For further processing of the data, another set of XSL transformation templates was created. In practice, each Named Entity Linking based on already recognised entity names uses some kind of gazetteer of entities. For geographical data, there is a myriad of options to choose from. No comprehensive comparison of these gazetteers was done at this time, but with some preliminary tests it was easy to gather that for example *Getty Thesaurus of Geographic Names*⁷ has limited coverage of the local place names that appear in the data. GeoNames⁸ has better coverage, and also an API that allows for more structured searches of the data. DBpedia Spotlight uses DBpedia for disambiguation and linking of named entities (Mendes et al. 2011). However, Finnish version of DBpedia is currently not maintained.

In the first phase of this project, Wikidata was chosen as the source for geographical database, mainly because of existing knowledge about Wikidata’s data model and the use of SPARQL as a query language. Moreover, Wikidata has a good-quality search engine for searching the entities in the data based on their preferred and alternative labels and textual descriptions. This search also includes a ranking algorithm, which allows for the retrieval of the most probable result. In contrast, DBpedia Spotlight uses a ranking algorithm based on the Inverse Candidate Frequency (ICF) weight (Mendes et al. 2011, 3).

In the first stage of entity linking, each source document was processed for its place and location annotations. For each annotation, the place name it referred to was obtained in its basic form using the lemmatised data, along with the unique ID of the annotation (see above).

When figuring out the search string, another problem manifested. A typical form of place name that appears also in these texts is “Suomen Suuriruhtinaskunta” (“Grand Duchy of Finland”), which is composed of one or more genitive forms (“Suomen” is the genitive of “Suomi”) before the base word in nominative (“suuriruhtinaskunta”, “Grand Duchy”). The parts in genitive are not inflected, so the inessive form of the place is “Suomen Suuriruhtinaskunnassa”, etc. A lemmatiser, however, returns the lemmas of each part, so the lemmatised form of the name is “Suomi suuriruhtinaskunta”, which usually does not return any results from any geodatabase.

The POS (part of speech) data provided by the lemmatiser is valuable in this case. In the case the place annotation contains more than one word, if the words before the last were in genitive form, this form should be returned instead of the lemmatised nominative, which then should be used only for the last word in the multiword annotation. In this way, these type of multiword place annotations returned their referent in a form that actually is usable in searching the place data. This heuristic was coded as a function in

7. <https://www.getty.edu/research/tools/vocabularies/tgn/>

8. <http://www.geonames.org/>

the XSL transformations, and used in the data retrieval process.

For each name, the local cache was checked for corresponding content, and if no content was found, data was looked for in the Wikidata using the SPARQL endpoint. A query template was defined for an efficient search of the place and location name labels in Finnish, limiting the results to those results that were also descendants of appropriate geopolitical and natural place types, respectively. The template for the geopolitical place name query is shown below (With the query term “York”).

```
1 PREFIX bd: <http://www.bigdata.com/rdf#>
2 PREFIX mwapi: <https://www.mediawiki.org/ontology#API/>
3 PREFIX wd: <http://www.wikidata.org/entity/>
4 PREFIX wdt: <http://www.wikidata.org/prop/direct/>
5 PREFIX wikibase: <http://wikiba.se/ontology#>
6
7 SELECT distinct ?place ?placeLabel ?country ?countryLabel ?location ?
   geonamesID ?openstreetmaprelid ?description WHERE {
8   service wikibase:mwapi {
9     bd:serviceParam wikibase:endpoint "www.wikidata.org";
10      wikibase:api "EntitySearch" ;
11      mwapi:search "York" ;
12      mwapi:language "fi" .
13      ?place wikibase:apiOutputItem mwapi:item .
14      ?num wikibase:apiOrdinal true .
15   }
16   ?place wdt:P31/wdt:P279+ wd:Q56061 .
17   optional {
18     ?place wdt:P17 ?country .
19   }
20   optional {
21     ?place wdt:P625 ?location .
22   }
23   optional {
24     ?place wdt:P1566 ?geonamesID .
25   }
26   optional {
27     ?place wdt:P402 ?openstreetmaprelid .
28   }
29   service wikibase:label {
30     bd:serviceParam wikibase:language "fi" .
31   }
32   optional {
33     ?place schema:description ?description .
34     filter (lang(?description)="fi")
35   }
36 } order by ?num
37 LIMIT 1
```


When adding the returned place and location information to the local cache data, the place is given an ID that can be used to refer to it. Repeating this process for all place and location names gathered from the texts created a list of annotation ID's and place ID's that was used to modify the TEI files by adding ref-attributes pointing to the place ID's to each place and location annotation. The contents of the cache were stored as records in a MongoDB-database, so the list can be reconstructed and replaced later.

The TEI files were then updated with the retrieved reference data. The ref-attributes of each annotation element were updated with content that can later be used to link each annotation to the place record in the MongoDB database.

2.5 Resulting dataset

Once the text has gone through all the steps in the process, an annotated sentence looks like this:

```
1 <p><pc xml:id="lonnrot-0585-1-token3137" n="3137">-</pc><pc xml:id="
  lonnrot-0585-1-token3138" n="3138">-</pc> <w xml:id="lonnrot-0585-1-
  token3139" n="3139" lemma="pikkunen" pos="ADJ" msd="Case=Nom|Degree=
  Pos|Derivation=Lainen|Number=Sing|Style=Coll">Pikkunen</w> <w xml:id="
  lonnrot-0585-1-token3140" n="3140" lemma="käärö" pos="NOUN" msd="Case=
  Nom|Number=Sing">käärö</w> <w xml:id="lonnrot-0585-1-token3141" n="
  3141" lemma="olla" pos="AUX" msd="Mood=Ind|Number=Sing|Person=2|Tense=
  Past|VerbForm=Fin|Voice=Act">olit</w><pc xml:id="lonnrot-0585-1-
  token3142" n="3142">,</pc> <w xml:id="lonnrot-0585-1-token3143" n="
  3143" lemma="kun" pos="SCONJ" msd="_">kun</w> <w xml:id="lonnrot
  -0585-1-token3144" n="3144" lemma="ankara" pos="ADJ" msd="Case=Nom|
  Degree=Pos|Number=Sing">ankara</w> <name key="/PERSON/PERSON_
  Heblarouva?lemma=Heblarouva" type="PERSON" xml:id="lonnrot-0585-1-
  annotation-PERSON-101"><w xml:id="lonnrot-0585-1-token3145" n="3145"
  lemma="Hebla#rouva" pos="NOUN" msd="Case=Nom|Number=Sing">Hebla-rouva<
  /w></name><pc xml:id="lonnrot-0585-1-token3146" n="3146">,</pc> <w
  xml:id="lonnrot-0585-1-token3147" n="3147" lemma="iso#äiti" pos="NOUN"
  msd="Case=Nom|Number=Sing|Number[psor]=Sing|Person[psor]=2">isoäitisi
  </w><pc xml:id="lonnrot-0585-1-token3148" n="3148">,</pc> <w xml:id="
  lonnrot-0585-1-token3149" n="3149" lemma="viedä" pos="VERB" msd="Mood=
  Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act">vei</w> <w
  xml:id="lonnrot-0585-1-token3150" n="3150" lemma="sinä" pos="PRON"
  msd="Case=Acc|Number=Sing|Person=2|PronType=Prs">sinut</w><lb/> <w
  xml:id="lonnrot-0585-1-token3151" n="3151" lemma="täältä" pos="ADV"
  msd="_">täältä</w> <w xml:id="lonnrot-0585-1-token3152" n="3152" lemma
  ="koti" pos="NOUN" msd="Case=Ill|Number=Sing|Person[psor]=3">kotiinsa<
  /w> <name key="/GPE/GPE_Siuntio?lemma=Siuntio&wikidata_id=Q984931"
  type="GPE" xml:id="lonnrot-0585-1-annotation-GPE-27"><w xml:id="
  lonnrot-0585-1-token3153" n="3153" lemma="Siuntio" pos="PROPN" msd="
  Case=Ill|Number=Sing">Siuntioon</w></name><pc xml:id="lonnrot-0585-1-
  token3154" n="3154">.</pc><lb/><lb/></p>
```

The sample cannot be prettyfied, since the contents of the <p> element is a mixed-mode

xml, where the whitespace between different elements is semantically important. 501

The sentence contains a reference to the GPE entity “Siuntio”, which is linked to the 502
Wikidata record Q984931. The key tags contain the information needed by the web front 503
end to link to the entity. The sample shows how each element has been given an ID that 504
is unique to the whole project dataset. Also, the discovered annotations, marked with 505
<name> elements, have unique ID’s. 506

3. Metachain 507

3.1 Managing the workflows 508

When working with datasets containing more than a some tens of data units, managing 509
the workflows becomes a challenge. 510

A division to three different types of workflow management is proposed here: 511

1. Running different algorithms and analysis tools individually on each unit of data 512
in the dataset; 513
2. Writing scripts to either chain many stages of the analysis in order to run this 514
whole chain for each data unit, running each stage for numerous data units, or as 515
an end product, to run many stages for all data units; 516
3. Manage the running of each stage and data unit with some external tool, only 517
triggering the required phases for relevant data units. 518

The first type of workflow management is usually the one chosen when experimenting 519
with new tools and datasets. Data units are analysed individually using a graphical 520
user interface, or with shell commands. This is manageable for a small number of data 521
units, but icreasing the amount of data and with repeated applications of the tools, this 522
soon becomes very cumbersome. 523

The usual solution is to turn the management process into a script, and let the script 524
run the analyses; the script are run on an external services, with more efficient servers 525
or cloud instances. This adopts the second type of workflow management. 526

This process tends to end up in difficulties when used in digital humanities. The 527
script-based workflow management is uncomplicated to use only in the cases where 528
the source data is “born digital”, meaning, it is originally produced as digital data 529
and therefore corresponds to a well-defined, strict data scheme. We have in this paper 530
already encountered a situation where the data was not what it was expected to be: the 531
Projekti Lönnrot corpus texts use different encoding systems, and there is no systematic 532
information anywhere about which file uses which encoding. In fact, the situation is 533
even worse: there are files that do not fully follow *any* known encoding. Often this kind 534
of situation is caused by only a single character in the file, but this kind of error in file 535
reading usually causes the program processing the file to throw an exception, if this has 536
been anticipated, or otherwise, just crash. 537

This kind of crash has to be managed somehow. The faulty file has to be recorded 538
somewhere, and it has to be decided in advance, how to continue the process. Further, 539

there has to be a way to manage the metadata so that if only the faulty file is skipped, 540
there is a way to retrieve a list of the faulty files, and also, when these are fixed, not 541
necessarily to process the whole dataset again, but only the ones that are missing. This 542
kind of scripting becomes quickly very complicated, and the superficially rather simple 543
script-based approach starts to acquire features that actually are already present in 544
other kinds of systems, which brings us to the third kind of workflow management. 545

This approach to project workflows is more of a methodology rather than a straightfor- 546
ward solution, but the fundamental concept is straightforward: utilise a system that can 547
automate processing chains, while also keeping track of results and errors, to manage 548
the various phases and units of data within a project. 549

In our case, the management of the processing chains was created around *asynchronous* 550
message queues. Each individual stage in the processing chains is devised as an inde- 551
pendent client, that both asks relevant queues for work to do and sends either results 552
or metadata about completed works to another queue. In relevant terminology, each 553
tool functions as a consumer, retrieving work from designated queues and publishing 554
completed results or potential errors to the appropriate queues. 555

The queues are managed by a *RabbitMQ* message broker, which is a simple yet powerful 556
tool for this purpose. It has a satisfactory browser-based user interface, which facilitates 557
the monitoring of the progress of various queues, as well as the examination of the 558
contents of queues, especially in instances where errors have been reported. 559

The actual clients can reside anywhere where it is possible to access the message broker. 560
This is beneficial because it enables the utilisation of remote resources for data processing. 561
For example, in our case, the main data servers are located on the local university 562
premises, but some CPU-intensive processes are run on virtual machines at the national 563
computational resource provider. 564

3.2 Manual intervention and human contribution 565

Luoma et al. (2020) propose that their Finnish NER is capable of reaching around 90% 566
of precision and recall for results in texts drawn from most domains. However, they 567
have not assessed the performance of the NER tagger using fiction from the 1870s to the 568
1940s. It is possible to distinguish false positives by checking the statistics of the NER 569
results, although some cases require going back to the annotated text to evaluate the 570
context in which the annotation belongs. But the only way to ensure how many spatial 571
entities NER has failed to identify is to read the texts and check the annotations one 572
by one. The manual review of the results is time-consuming even if only a randomly 573
sampled portion of them is checked. The NER results are currently under review in 574
our project, but the preliminary review of the results indicate that they are of sufficient 575
quality. 576

Two of the texts were manually controlled by the project members. In addition to 577
checking the recognised entities, also the missing cases were calculated, allowing us to 578
compute valid precision and recall values, as well as the F1-values. (See Tables 4 and 5.) 579
580

The Finnish language of the late nineteenth century may pose more challenges for the 581

Text	Precision	Recall	F1-value
Santeri Ivalo: <i>Anna Fleming</i> (1898)	0.92	0.97	0.94
Algot Untola: <i>Kuolleista herännyt</i> (1916)	0.93	0.84	0.88

Table 4: Precision, Recall, and F1-values calculated for geopolitical place name recognition in manually controlled texts: Santeri Ivalo's historical novel *Anna Fleming* from 1898 and Algot Untola's novel *Kuolleista herännyt* from 1916.

Text	Precision	Recall	F1-value
Santeri Ivalo: <i>Anna Fleming</i> (1898)	0.97	0.89	0.93
<i>Anna Fleming</i> (with NEL added) (1898)	0.81	0.61	0.70
Algot Untola: <i>Kuolleista herännyt</i> (1916)	0.33	0.5	0.40

Table 5: Precision, Recall, and F1-values calculated for geographical place name recognition in manually controlled texts: Santeri Ivalo's historical novel *Anna Fleming* from 1898 and Algot Untola's novel *Kuolleista herännyt* from 1916. For *Anna Fleming*, we also include respective values for the combined NER & NEL process.

NER algorithm, as the Finnish literary language was still taking shape at that time. We will now look at some problematic cases, using as an example the historical novel *Anna Fleming*, published 1898 by Santeri Ivalo. The novel, rich in geographical references, is set in Sweden in the sixteenth and seventeenth centuries and, to a large extent, in the area now known as Finland.

The main characters in the novel travel to places, reminisce over their stay in different places, talk about travelling to different locations and are related to people who live in different parts of Europe. The NER algorithm has worked very well for this challenging text, but it makes certain systematic errors. There are some examples of classification error, where a political entity is tagged as a geographical place or vice versa. Most mistakes were found with the names of medieval manors like Kuitia, Liuksiala and Kankainen. For example, Kuitia (a manor house founded in the fifteenth century in Southwest Finland) is mentioned 11 times in the first chapter of the book. It is tagged as a geographical formation six times, as a geopolitical entity (which is the correct tag) two times and as a person once. On four occasions, the place has been completely unmarked.

The names of historical or geographical entities that no longer exist also sometimes cause problems for the tagger. For example, "Danzig" (modern Gdańsk), is in the second chapter of the book tagged as a person, and so is "Iharinkoski" (Ihari rapids) in the third chapter. On the other hand, tagging geographical places usually works fine: "Vantaankoski" ("Vantaa rapids"), "Suomenlahti" ("Gulf of Finland") and many others were all tagged correctly, even if the names are Swedish ones like "Sandö" or "Estrnäs". Cities, provinces, or countries were usually tagged correctly. For example, "Suomi" ("Finland") occurs in the text in inflected forms "Suomessa", "Suomen", "Suomea", "Suomesta" and "Suomeen" but it is however correctly tagged every time (for example, 50 times in the second chapter).

Similar inaccuracies are present also in the novel *Kuolleista herännyt* (*Risen from the Dead*, 1916) by Algot Untola, who is probably better known by his pseudonym Maiju Lassila. In the novel, an uncultured and illiterate dockworker Jönni Lumperi from Helsinki gains 2000 marks from a lottery and is encouraged by a wealthy businessman into an

endeavour to turn the sum into millions. Jönni chases this dream of enrichment around Southern Finland, first to Tampere and its rural surroundings, then to Hämeenlinna and Riihimäki. The journey ends back to Helsinki with all the lottery winnings and more lost. The novel contains many geopolitical place names (GPE) and some natural locations (LOC) both real and fictional. NER has recognised quite well many of these, but some inaccuracies also occur. Manual proofing of GPE and LOC tags by NER has revealed 60 clear errors (the novel contains altogether 213 GPE and LOC tags). Some of them are most likely due from outdated spelling of place names, such as “Marseljeesi” (“Marseille” in current Finnish spelling) or “Söörnäinen” (nowadays “Sörnäinen”, a neighbourhood in Helsinki). Others errors are fictional non-GPE locations and biblical place names such as “Kaana” (“Cana”) etc. Sometimes person names are recognised as non-GPE locations and nouns as proper names. There is also a surprisingly high number of cases when Helsinki and Tampere were not recognised in a declined form even though the spelling is modern, and both are large and well-known cities in Finland. The errors do not seem systematic, and both cities are more often correctly recognised than not.

In the whole, inaccuracies are quite rare and in most cases do not repeat. Yet, it is worthwhile also to point out two cases of inaccuracies with special nature:

1. The chapter 17 introduces a new character named “Hesa”. This person name (PERS) is tagged 13 times out of 21 as GPE location, most likely due to the frequent modern use of “Hesa” as the nickname for Helsinki. In three cases when “Hesa” is followed by the family name “Ruokka” it is recognised as a person name and only in one case without the family name. Also, in four cases “Hesa” is not tagged at all, one of them followed by the family name. The “Hesa” case exemplifies that NER is capable to recognise even nicknames for places, but also that from this fact can follow unpredictable inaccuracies. Had the main character been named “Hesa” hundreds of false positives would have occurred.
2. A more peculiar case to explore in more detail is the Estate of “Punturi”. This estate is referred to in many ways: “estate”, “house”, “farm” and “land of Punturi”. In this case, the ambiguity of natural language makes it very hard to even control if the name is tagged correctly. As the name of the estate is the same as its owners, even a close reading of the context does not always give a clear answer to what is referred, the proper name of a specific land area, or the fact that a person named Punturi is an owner of a land area left unnamed. From 16 cases with reference to the estate, 10 are recognised as GPE and 6 cases are not tagged at all. In some cases, it is very clear that reference is to the name of the land, and it should be tagged as GPE, but also many ambiguous cases occur. However, as previously stated, even manual proofreading and close reading cannot provide a binary classification, as natural language does not function in this manner. It should be noted that as our project investigates fictional literary, which is by nature meant to be open to multiple interpretations, these ambiguities are an essential part of data material and should not be ignored, excluded or forced into binary categories.

4. Conclusion

653

In this article, we have described the process of converting plain text fictional texts into TEI XML format. The pipeline is fully automated as far as possible. The compilation and harmonisation of the metadata of the texts and the curation of the corpus has been carried out by a literary scholar with domain knowledge of the Finnish literary history of the period. The text corpus is then passed through a pipeline, which translates it into TEI XML format, segments it into separate chapters, lemmatises the text, searches for named entities and disambiguates and geocodes the spatial entities by linking them to open linked data.

The results of the NEL process are then checked by human readers familiar with the literature and culture of the period to assess the success of the identification. This cannot be done for the whole corpus, but requires the examination of randomly selected samples of works from different periods and by different authors. Although the basic idea behind our project – to put texts on maps – is simple, the Finnish-language corpus presents challenges: named-entity recognition in Finnish has been difficult to implement with sufficient accuracy in the past. *Finnish NER's* ability to recall different names in text and the precision of the recognition are over 90% on contemporary Finnish. However, *Finnish NER* has been trained on modern textual material, so its accuracy is unlikely to be as good with nineteenth-century texts. Yet, the language model can be tailored to historical data using transfer learning (see Labusch et al. 2019). We can experiment with this at a later stage of the project once we have enough human-corrected data available.

Finally, we will import the texts to an online interface (TEI publisher) where they can be read while exploring the maps drawn by the works and comparing the spatial regions of the different texts. As our project will produce a clean and structured TEI XML corpus of the texts, further scholarly examination of the data using other digital humanities methods will be convenient. For example, we can apply topic modelling or network analysis, as the corpus is already pre-processed by tokenisation, part of speech tagging, lemmatisation, etc. This allows us to extract the semantic features of the texts and associate them with the geocoded toponyms.

Our future plans also include studying and comparing Swedish-language Finnish fiction with Finnish-language fiction. Since the software we are developing is modular, adding Swedish name recognition to the pipeline will not be difficult. Moreover, due to its modular architecture, our pipeline, which has been released under an open licence, is also suitable for processing data in other languages. Thus, our work could be adapted to research and web publishing use with other lower-resourced European languages with relevant literary traditions.

5. Data and Code availability

689


The original *Projekti Lönnrot* data files are available at the project website at <http://www.lonnrot.net/>.

Relevant pieces of code are available at <https://anonymous.4open.science/r/extraction-georef-finlit-5249/>. The repository does not include any scaffolding code to run

the tools, just the resource files needed for the various transformations of the data.	694
6. Acknowledgements	695
The 2-year project <i>Atlas of Finnish Literature 1870–1940</i> is funded by the Alfred Kordelin Foundation under the Major Cultural Projects Programme (2022–2024).	696 697
7. Author Contributions	698
Harri Kiiskinen: Conceptualization, Writing – original draft (main author), Data curation, Software	699 700
Asko Nivala: Writing – original draft, Project supervision	701
Jasmine Westerlund: Writing – original draft, Metadata curation, Data verification	702
Juhana Saarelainen: Writing – original draft, Data verification	703
References	704
Alex, Beatrice, Claire Grover, Richard Tobin, and Jon Oberlander (Feb. 2019). “Geoparsing historical and contemporary literary text set in the City of Edinburgh”. In: <i>Language Resources and Evaluation</i> 53.4, 651–675. ISSN: 1574-0218. 10.1007/s10579-019-09443-x . http://dx.doi.org/10.1007/s10579-019-09443-x .	705 706 707 708
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: <i>arXiv preprint arXiv:1810.04805</i> .	709 710 711
Engelberg, Mark (2022). <i>Instaparse</i> . Version 1.4.12. https://github.com/Engelberg/instaparse .	712 713
Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković (Oct. 2020). “Named Entity Recognition for Distant Reading in ELTeC”. In: <i>CLARIN Annual Conference 2020</i> . Virtual Event, France. https://hal.science/hal-03160438 .	714 715 716 717
Gregory, Ian N., Christopher Donaldson, Andrew Hardie, and Paul Rayson (2019). “Modeling space in historical texts”. In: <i>The Shape of Data in the Digital Humanities</i> . Ed. by Julia Flanders and Fotis Jannidis. Routledge. Chap. 5, 133–149. ISBN: 9781315552941. 10.4324/9781315552941 .	718 719 720 721
Gregory, Ian N. and Andrew Hardie (2011). “Visual GISTing: bringing together corpus linguistics and Geographical Information Systems”. In: <i>Literary and Linguistic Computing</i> 26.3, 297–314.	722 723 724
Kanerva, Jenna, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski (2018). “Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task”. In: <i>Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies</i> .	725 726 727 728
Kanerva, Jenna, Filip Ginter, and Tapio Salakoski (2021). “Universal Lemmatizer: A Sequence to Sequence Model for Lemmatizing Universal Dependencies Treebanks”. In: <i>Natural Language Engineering</i> 27.5, 545–574. 10.1017/S1351324920000224 .	729 730 731

- Labusch, Kai, Clemens Neudecker, and David Zellhöfer (2019). "BERT for Named Entity Recognition in Contemporary and Historical German". In: *BERT for Named Entity Recognition in Contemporary and Historical German*. 732-734.
- Luoma, Jouni, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo (May 2020). "A Broad-coverage Corpus for Finnish Named Entity Recognition". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC2020)*. Marseille, France: European Language Resources Association, 4615-4624. ISBN: 979-10-95546-34-4. <https://aclanthology.org/2020.lrec-1.567>.
- Mendes, Pablo N, Max Jakob, Andrés García-Silva, and Christian Bizer (2011). "DBpedia spotlight: shedding light on the web of documents Proceedings of the 7th international conference on semantic systems". In: *DBpedia spotlight: shedding light on the web of documents Proceedings of the 7th international conference on semantic systems*, 1-8.
- Moncla, Ludovic, Mauro Gaio, Thierry Joliveau, and Yves-François Le Lay (Nov. 2017). "Automated Geoparsing of Paris Street Names in 19th Century Novels". In: *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*. 10.1145/3149858.3149859.
- Moncla, Ludovic, Mauro Gaio, Thierry Joliveau, Yves-François Le Lay, Noémie Boeglin, and Pierre-Olivier Mazagol (Mar. 2019). "Mapping urban fingerprints of odonyms automatically extracted from French novels". In: *International Journal of Geographical Information Science* 33.12, 2477-2497. ISSN: 1362-3087. 10.1080/13658816.2019.1584804. <http://dx.doi.org/10.1080/13658816.2019.1584804>.
- Moretti, Franco (1998). *Atlas of the European Novel, 1800-1900*. London; New York: Verso.
- Piatti, Barbara (2008). *Die Geographie der Literatur: Schauplätze, Handlungsräume, Raumphantasien*. Göttingen: Wallstein.
- Tally, Robert (2018). *Topophrenia: Place, Narrative, and the Spatial Imagination*. Bloomington: Indiana University Press.
- TEI Consortium (2022). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html> (visited on 12/07/2022).
- Weischedel, Ralph, Sameer Pradhan, Lance Ramsdew, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D. ang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster, Abdel-aati Hawwary, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston (Sept. 28, 2012). *OntoNotes Release 5.0. with OntoNotes DB Tool co.999 beta*. <https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf> (visited on 12/19/2022).

A sentence-based stylistic history of the Hungarian Novel

Botond Szemes¹ 

1. Institute for Literary Studies, Research Centre for the Humanities, Budapest, Hungary.

Citation

Botond Szemes (2023). "A Sentence-based Stylistic History of the Hungarian Novel". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-05-04

Date received 2023-01-26

Keywords

stylometry, sentence structure, clause linkage, literary history, classification, epistemology

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. The paper presents a method for the automatic identification of different types of compound and complex sentences in Hungarian through the analysis of conjunctions and their positions. This method opens up new perspectives in stylometry: on the one hand, conjunctions as function words provide a large amount of data for statistical analyses, and on the other hand, they also carry (grammatical) meaning - about the relations between clauses (e.g. opposition, conditionality). By examining the relative frequency of each type, it is possible to reveal the most typical relations between clauses in a given text or corpus. In this way, the style of novels can be described at the level of the sentence, while also revealing the rhetorical-logical structure and epistemological attitude of the texts, which is not usually reflected in the reading process. This method also provides an opportunity to identify different stylistic traditions in literary history.

1. Introduction

This paper offers a stylistic analysis of the sentence structures of 150 canonical Hungarian novels published between 1832 and 2005. In doing so, it proposes a method for examining the frequency of different types of compound and complex sentences (in other terminology: clause linkage [Kabatek and Vincis 2010; Raible 2001], junction [Raible 1992] or clause complexes [Kugler 2020]), which can serve as a starting point for future studies on both literary history and the linguistic construction of literary texts. "Clause complexes profile multiple referential scenes and their relations, integrated into a single, complex structure. (...) [They] are not structures produced by creating and concatenating clauses, they are not derivable from their parts; rather, they can be interpreted in terms of construction types (schemas) and their instantiations." (Kugler 2020) This characteristic highlights that compound-complex sentences can be examined not only from their syntactic structure, but also from a semiotic, pragmatic and discursive perspective. (Visapää et al. 2014b, p. 2–5) Such a "functional-topological framework" allows us not only to rely on a strict grammatical-syntactic taxonomy, but also to take into account the semantic-pragmatic dimension of relations. This is particularly important, since the traditional grammatical classification does not seem to be applicable in all cases - most contemporary theories also question the very division of subordination and coordination. This is shown both by cross-linguistic research (i.e. some of the grammatical types have no equivalent in other languages - see Cristofaro 2014) and also within a language: see for example Wolfgang Raible's analysis of the following example: "'On account of her illness, Joan remained at home.' In this case, the relation

that is expressed is again very clear: causality. Nevertheless, it would make no sense to speak of ‘subordination’.” (Raible 2001)¹

This is why the paper mainly draws on the grammatical *meanings* developed by clause linkages when analysing the different types from a quantitative point of view. Furthermore, this ‘grammatical meaning’ also carries information about the *rhetorical* structure of the text: following Christian Matthiessen and Sandra A. Thompson, we can conclude that clause relations play an important role in the organization of discourses, and are in fact grammaticalized versions of the cohesive rhetorical relations between larger units in a text. (Matthiessen and Thompson 1988) “The cross-linguistic study of clause linkage markers and the observation that they tend to fall into clearly definable semantic-pragmatic sets has led linguists recently to characterize somewhat more fully than in the past the conceptual and rhetorical functions of many types of clause combining.” (Hopper and Closs Traugott 2003, p. 177) Considering all this, the hypothesis of the paper is that the relative frequency of clause relations in a given text (or group of texts) can help in determining which types of relations are the most characteristic of this text, which in turn sheds light on its underlying rhetorical and logical properties.² For example texts with a relatively high number of conditional relations clearly have a different epistemological attitude (i.e. they arrange elements of the outside reality differently) than texts that rely mainly on adversative coordinations. Similarly, the absence of a certain type of clause linkage can also reveal a great deal about a text. Such is the case of Miklós Mészöy’s early work from the corpus, where there is hardly any causative, inferential and explanatory relations between the clauses – stylistically this is how he is able to express the main philosophical insights of the French existentialist literature (first of all Albert Camus), and a chain of unmotivated actions (*action gratuite*). (See data in *Data availability* for details)

The research project that served as the basis for the study took a similar approach by calculating the mean and median sentence lengths of the same 150 canonical Hungarian novels (for details see *Corpus*). The figures based on these measures show some trends, which already have certain implications regarding the major historical changes in the style of the Hungarian novel. Figure 1 shows the mean and median sentence lengths (in terms of number of words) and the regression curve fitted to the data points, while Figure 2 shows the same for 23 novels by famous Hungarian writer Mór Jókai published between 1846 and 1894. The negative slope in the 19th century, which can even be observed in Mór Jókai’s oeuvre, can be attributed not only to stylistic changes in the Hungarian literary tradition but also the growing role of the press, the widespread use of new writing tools, and reforms in how reading and writing were taught in schools. (Cf. Szemes 2020) This linear trend can be observed in the literature of several European languages (e.g. in Spanish - see Calvo Tello 2023; and other languages: see Schöch 2022).

In Figure 1 the last third of the twentieth century is marked not so much by a definite trend as the co-existence of different types of prose: while extreme values are produced by authors associated with long sentences, some novels from this era employ distinctly

1. I will however use these terms throughout the paper for the sake of simplicity, but the focus will be on the individual relationship types.

2. The word ‘logic’ is used in a broad sense: it should not be understood as a term from formal logic but simply as a definite relationship between independent clauses. Therefore, it might be more accurate to use the expression ‘the diagrammatic character of sentences.’ Cf. Stjernfelt 2010

short sentences. Note, that the trendline is slightly overfitted due to the outliers (see the detection of the outliers in Appendix Figure 11, and the overall trendline without them in Figure 12), but without outliers there is still a group of novels in the second half of the 20th century whose sentence lengths show a steady – linear – increase (see Figure 13 in Appendix). At the same time, these outliers are not in the corpus because of selection bias – the novels are in the center of the Hungarian canon from the 1970-1980s, the period called “prosa turn” (Szirák 2013) with internationally recognized authors like the Nobel Prize winner Imre Kertész, Péter Nádas, who has been a contender for the prize for years, or the International Man Booker Prize winner László Krasznahorkai. The dispersion in the second half of the 20th century is illustrated in Figure 3, where the texts are arranged chronologically and divided into groups of 30 texts, with the distribution of works consisting of “long”, “medium-length” and “short” sentences shown in the five resulting time frames. These three categories were created separately for each time frame based on average sentence lengths with the help of the k-means algorithm ($k=3$), following the method outlined in the 14th Pamphlet of the Stanford Literary Lab. That pamphlet illustrates the frequency of analepses and prolepses (flashbacks and flashforwards) in movies; not by plotting the films along a single trend line but by dividing movies from each decade into three clusters based on whether they have “extreme”, “moderate”, or “conservative” values. (Kanatova et al. 2017) Figure 3 indicates that from the 1970s onwards, works with long and short sentences start diverging more conspicuously than before, which suggests an increasing divide between coexisting prose styles. (Figure 13 in the Appendix uses the same method and shows novels from the 20th century grouped into three parts based on chronology and clusters them in each group just into “high” and “low” categories without the outliers.) Furthermore this dispersion could be the reason why there is no statistically significant relationship between year of publication and sentence length in the whole data set according to the one-way analysis of variance (ANOVA) test ($p\text{-value} = 0.29$), just in the 19th century ($p\text{-value} = 6.82e-09$) and in the whole dataset divided into three parts (1832-1899, 1900-1949, 1950-2005; $p\text{-value} = 0.0004$)

Visualizations of mean sentence lengths can therefore capture literary and stylistic developments and help in distinguishing between short-sentence and long-sentence prose traditions. However, the similarities and differences of these traditions as well as the internal structure of the sentences should be examined more thoroughly, since ‘long sentence’ does not necessarily have the same meaning across different authors and periods. (Cf. Allison et al. 2013)

2. Methods

In Hungarian orthography, clause boundaries are always marked by punctuation (commas, semicolons, colons, dashes), so identifying them is easier than in many other languages. Thus, combinations of punctuation marks and conjunction words or relative pronouns is enough to identify different clause relations, which can be detected with the help of basic regular expressions. Other scholars search for more complex grammatical features, which can also be used on unedited texts (e.g., transcriptions of colloquial language) and other languages. A more sophisticated method would be the practice of identifying conjunction words between finite verbs, since a clause prototypically

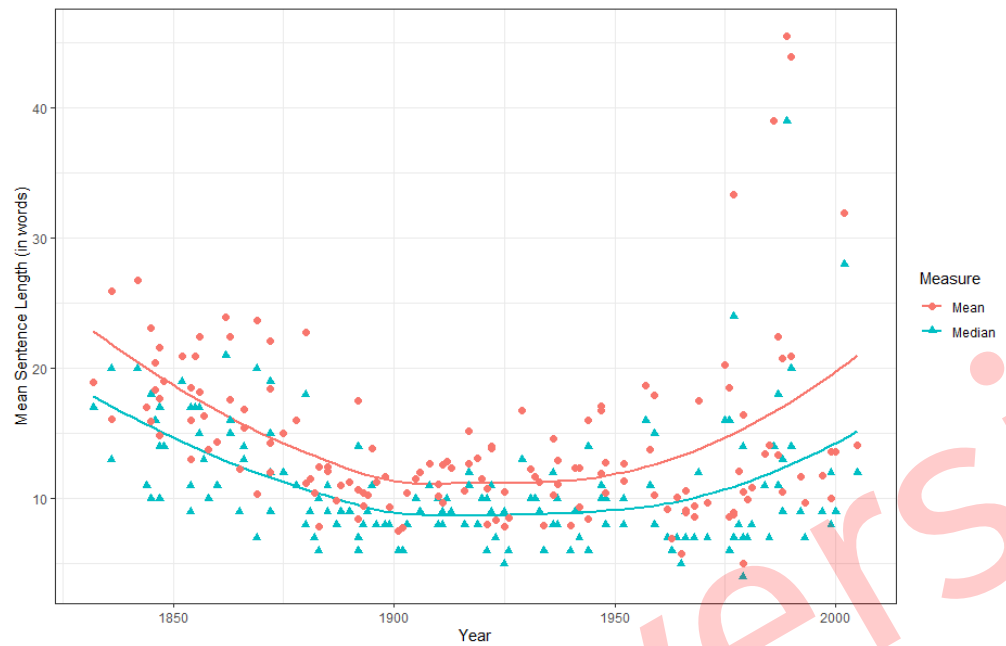


Figure 1: The mean and median of the sentence lengths of 150 Hungarian novels with loess trendline.

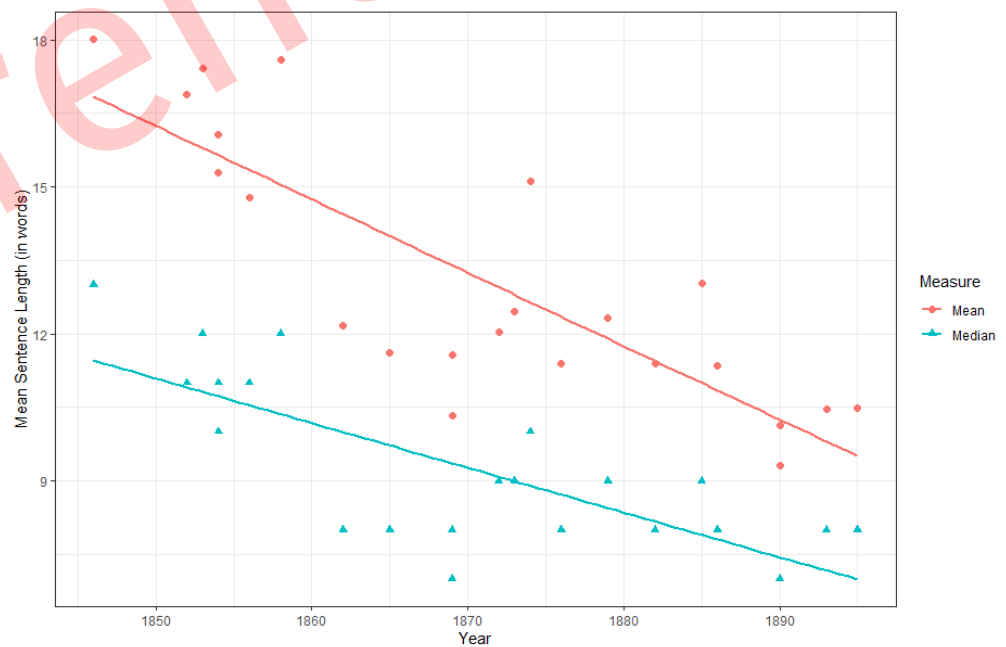


Figure 2: The mean and median of the sentence lengths of 23 novels by Mór Jókai with linear regression trendline. For mean $R^2 = 0.67$

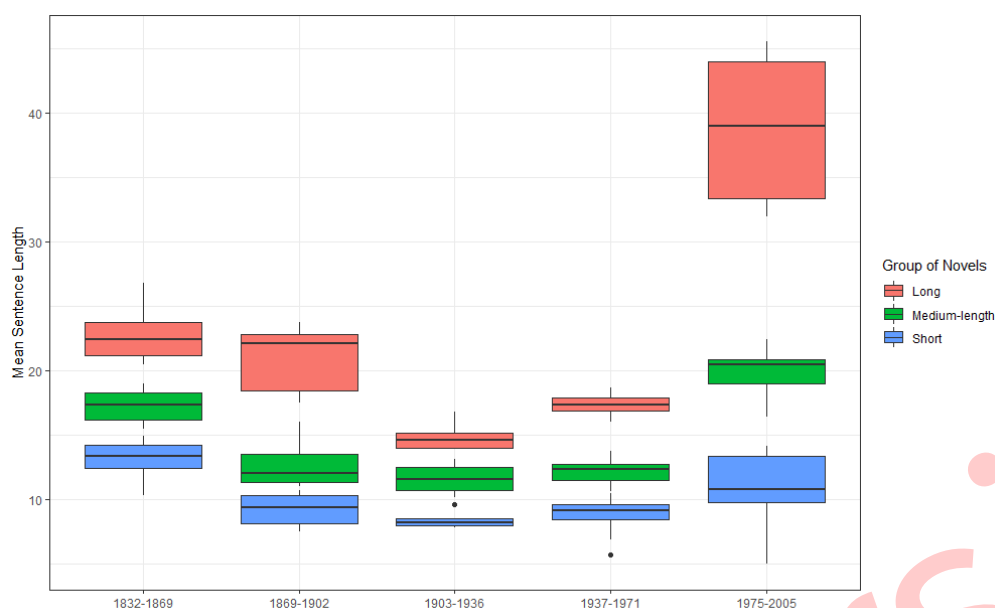


Figure 3: The distribution of novels with “long”, “medium-length”, and “short” sentences over five equivalent time frames, based on 150 novels.

consists of a finite verb as predicate and its elaboration. Although this might be a more general and nuanced approach, it is subject to three problems. Firstly, the *emagyar* NLP tool used for the present study (Váradi et al. 2018) cannot identify finite verbs with perfect accuracy, especially in texts from the first half of the 19th century. Secondly, not all conjunction words are placed between finite verbs; some of them can be found, for instance, at the beginning of a sentence. Thirdly, this approach does not take into account the cases in Hungarian where the clause is constructed not with a verbal but a nominal predicate (similar to the sentence: ‘that the house nice which tall.’ See Dömötör et al. 2020) Another option would be to rely on the outcome of a dependency parser. However, their accuracy is rather low (see Orosz et al. 2022) and, more importantly, they are based exclusively on syntactic categories (like “adverbial subordination”), while the research prefers to focus on the semantic and rhetoric dimensions of relations (see Introduction).

For regular expressions, the following features are of particular importance: (1) the position of the conjunction words/relative pronouns, i.e., their possible distance from the punctuation mark; (2) the prototypical meaning of the polysemantic conjunction words, and (3) whether the given conjunction word typically tends link clauses or phrases. For example the conjunction *azonban* (‘however’) can be placed anywhere in the clause thanks to the flexibility of the Hungarian word order (see: *Okos gyerek volt, az élet értelmét megfejteni azonban nem tudta*; ‘He was a clever child, he could not understand however the meaning of life’) – while *de* (‘but’) only establishes adversative relation between clauses right after the coma (it prorotipically develops such a relation between phrases anywhere else in the clause – e.g. ‘He is tall but strong.’)

We sought to answer these questions by manual analysis of 100-sentence samples from the corpus with the help of linguist annotators. If a conjunction word/relative pronoun created the same semantic type of clause relation at least 60 times out of 100 cases, it was classified under that type; otherwise, it was excluded from the study. For example, the

conjunction word 'that' [*hogy* in Hungarian] was not included because multiple uses occur with similar frequency. (This has a significant impact on the analysis: in a sample of 1,000 compound-complex sentences with 2,502 clause combinations from the corpus, 12.3% of them are of the *that-type* [sum = 321]). The relatively high margin of error in the categorization is due to the ultimate goal of the present research, i.e. not to detect predefined grammatical categories in texts but to analyze the rhetorical-diagrammatical properties of novels, which requires as much data as possible. Thus, relations between phrases with a similar grammatical function and meaning to that of clause relations were not excluded from search results in all cases, since this allows better identification of the predominant types of relations. (However, it is also worth looking at these separately, as in the case of the conjunction *and*, which coordinates clauses more often in speech-related texts, while coordinates phrases more often in formal-written texts - see Kytö and Smitterberg 2023.) The other extreme, that is, applying no limits and searching simply for conjunction words, is not efficient due to the problems caused by polysemantic words (e.g. *bár* means 'although' and 'bar' at the same time; it makes a difference whether the text expresses a concessive relationship or whether the characters are just thirsty), and by mixing semantically distinct grammatical structures, so the results would no longer reflect a clearly defined property of the novels. The application of this compromise between permissive and restrictive criteria is facilitated by Hungarian orthography, inasmuch as the relations between phrases are only marked with punctuation (and a pause in spoken language) if they have a grammatical meaning similar to that of clause relations. (e.g. *Hogy volt az embernek történelem előtti, azaz olyan korszaka, amelyről semmi, nemcsak írásban, de még szájról-szájra adott mondákban sem maradt fenn: az kétségen felül áll.* 'That there was a prehistoric age of man, that is, an age of which nothing has been preserved, not only in writing, but even in word of mouth: that is beyond doubt.' [Zsigmond Móricz, *Be Faithful Unto Death*])

During the research, conjunction words were classified into one of the following 12 categories (inspired by Seiler 1995 and Kortman 1997, adapted to Hungarian on the basis of Imrényi and Kugler 2018). After each category, the English translation of the most common conjunction word is given: 1. copulative ('and'), 2. adversative ('but'), 3. disjunctive ('or'), 4. inferential ('so', 'thus'), 5. explicatory (there is no strict equivalent in English grammar; 'namely', 'that is') 6. conditional ('if'), 7. concessive ('although'), 8. simile ('as', 'like'), 9. logical ('because'), 10. prototypical relative clause ('who', 'which'), 11. relative clause – space ('where'), 12. relative clause – time ('when').

A compound-complex sentence might fall into more than one category. For instance, the sentence 'I see no contradiction in your response, and thus, if I think about it, you have to be right' belongs to three categories, as it includes copulative, inferential, and conditional relations as well. Clause relations that are not marked by conjunction words or relative pronouns were disregarded in the study, for example: 'There are no conjunction words in this sentence; [so] it was left out of the results.' This draws attention to a crucial feature of the procedure: the research is not concerned with complex-compound sentences in general but only with those in which relations are elaborated grammatically. In the 1,000-sentence samples with 2,502 clause combinations, 35.3% of them belonged to the "not elaborated" category (sum = 883, while the number of linkages with conjunctions: 1,298, i.e. 51.9%) – so their exclusion is a strong limitation. But note they often include cases that do not primarily indicate a relation between clauses but a change of voice in

the narration (e.g. “‘It’s hot”, said the snowman’), or an unmarked subordination of the *that*-type (‘I don’t know [that] when he’s coming.’) What is more important that the elaboration by a conjunction makes the connection between the clauses more accessible, i.e. it does not create the relation but “profiles it, makes the nature of the relation accessible and thus guides the interpretative process”. (Imrényi and Kugler 2018) By contrast, in the case of unmarked relations (often called juxtapositions, syntactically and semantically the less integrated type of connection, see Raible 2001), multiple and subjective interpretations are possible, and the exact relationship between the clauses is less in the foreground for the reader. Moreover, conjunction words are particularly important for a quantitative analysis because, being function words, they are very common and thus provide enough statistical data to map the deep structures of a text (linking the method to previous researches in stylometry - cf. Chung and Pennebaker 2007; Kestemont 2014; Rybicki and Eder 2011), yet, unlike several other function words, they have (grammatical) meaning, which makes the results about their distribution and frequency easy to interpret. The history of the style of the Hungarian novel as outlined with quantitative methods is based on this twofold nature of conjunction words and relative pronouns.

3. Corpus

The 1830s were chosen as the starting point for the corpus, as this is when the rise of the Hungarian novel in the modern sense occurred. The earliest novel in the corpus is András Fáy’s 1832 novel *The House of Belteky*, “traditionally regarded as the first Hungarian domestic novel of manners.” (Czigány 1984) The number and type of novels from the first half of the 19th century available in digital archives – mainly in Hungarian Electronic Library (Magyar Elektronikus Könyvtár – MEK) – had a major impact on text selection criteria and the size of the corpus. Since archives mainly store canonical novels from this era, only works that can be considered canonical are included in the corpus even from later periods. In any case, the few hundred novels available in databases would not give an overview of Hungarian fiction as a whole, but they are able to accurately represent the current literary canon. Canonicity was defined in two ways; if either criteria was applicable to a given text, it was considered canonical. First, the ELTE DH Novel Corpus (Bajzát et al. 2021)³ was consulted. This corpus, which builds on MEK’s database, currently contains 400 novels dating back to the 1920s and its first version has been created in the framework of the ELTeC (European Literary Text Collection) international COST-Action project.⁴ This project evaluates canonicity on the basis of publication history: works are labelled as ‘high canonicity’ if they have had at least two new editions since 1979. Second, comprehensive studies of literary history were consulted, in particular in the case of novels published after the 1920s; works discussed in these studies were also considered canonical. It should also be noted that digitization and online availability can be seen as a form of canonicity: texts that are included in online databases such as MEK and especially the prestigious *Digital Literary Academy* (*Digitális Irodalmi Akadémia, DIA*) are in some sense part of the literary canon. Indeed, only four novels from this study’s final corpus are missing from these platforms.

3. <https://regenykorpusz.elte-dh.hu/> (accessed Jan 17, 2023)

4. <https://www.cost.eu/actions/CA16204/> (accessed Jan 17, 2023)

The corpus ends with Péter Nádas's 2005 novel *Parallel Stories*, excluding the last decades of contemporary Hungarian literature, as it would be difficult to find a criterion that would allow one to single out just a few canonical works from contemporary Hungarian literature.

The research, therefore, focuses on the history of the style of the Hungarian novel between 1832 and 2005. The corpus covering this period was created in two stages: first, 100 canonical novels by 58 authors were selected, with a minimum of 3 and a maximum of 8 novels from each decade, taking care not to over-represent any one period. To ensure proportional representation, a maximum of four novels per author was added to the collection, but preferably fewer; four novels were selected only if there was a significant time gap between them, which made it possible to examine whether the author's oeuvre followed a particular trend or whether the author had an artistic "fingerprint" unrelated to the period's trends. To examine this question more closely, a subcorpus consisting of 23 novels by Mór Jókai was created (see e.g. Figure 2). In the second stage, 40 new authors and 50 new novels were added, bringing the total number of texts to 150 and the number of authors to 98, thus making the corpus more balanced both chronologically and in terms of authors, and providing an opportunity to double-check the previous measurements (i.e. whether the trends and patterns identified earlier could also be observed in the extended corpus). However, this comes at the price of including 19 new works that are not mentioned in major studies on literary history and are listed in the *ELTE DH Novel Corpus* and in the *ELteC* as having 'low canonicity' – these novels are highlighted in the table of bibliographic data (see *Data availability*). In other words, for the sake of a more detailed historical analysis, canonicity was de-emphasized when expanding the corpus. But, at the same time, since the corpus does not omit any author who is discussed in literary histories and whose works were published during the same time period as the 19 'low canonicity' novels, and since the vast majority of texts in 10 equal-sized timegroup is labelled as "high canonicity", the collection arguably remains representative of the current literary canon even after its expansion to 150 items.

On average, the corpus contains a work for every 1.15 years. To examine the distribution over time, the 173 years were divided into 10 equal parts so that the number of novels in each unit (17 years) could be counted and compared: 15 novels are in in five groups, 14 in one, 13 in two, and 16 in two. Therefore, the difference between the periods with the highest and lowest values is only 3 novels. The period represented by the fewest works is the advent of the Hungarian novel (the period before the Hungarian Revolution and War of Independence of 1848–49): far fewer works are available digitally from this era than from the second half of the century, which can be explained by the simple fact that fewer novels were written then. On average, 1.53 novels per author are included in the corpus: 64 authors are represented by a single text, 19 by two, 12 by three and 3 by four. Only 11 of the 98 authors are women, a disproportion that reflects imbalances of the Hungarian literary canon and institutional practices in the history of Hungarian literature.

For further bibliographic details of the novels see *Data availability*.

4. Results 1

After identification, we can calculate the relative frequencies of each type of clause relationship: either relative to the length, i.e. the number of words in a novel; or relative to each other, i.e. proportionally. These values can be plotted in three ways: (1) focusing on historical changes of the relative frequencies along the timeline; (2) comparing the extent to which novels employ a certain type of relationship; and (3) concentrating on the proportions of the types within a single novel. In what follows, results for (1) and (3) will be reported.

Figure 4 shows the changes for prototypical relative clauses. The downward trend is caused both by the outliers of the 1840-50s and the fact that until the 1870s there is hardly any text in the lower regions of the graph. According to ANOVA there is a strong correlation between the frequency of relative clauses and the years of publication ($p\text{-value} = 3.4\text{e-}09$), which means that the differences indicated by the trend can be considered statistically significant. This trend is present even without the outliers (for outlier detection see Appendix Figure 11, for the trend both with loess trendline and segmented linear regression see Appendix Figure 14/a.) Figure 5 thus suggests a stylistic feature of the first half of the 19th century: authors from this era tend to describe the characters appearing in the sentences in detail by using separate clauses (e.g., “The project, which was supported by the Foundation, is now finished.”)

Relative clauses are most frequently used in a rhythmic, rhetorical style of prose that was very influential in this period in Hungarian literature. This style is characterized by what might be called a ‘periodic sentences’; a compound sentences in which one part (in most cases from the main clause) is elaborated by several relative clauses outlining different scenarios. (Herczeg 1981; for English context see: Carter and McRae 2005, p. 421) The following sentence (with the repetition of the subject) is taken from the novel *The Village Notary* from baron József Eötvös: *De ti, kiket nem bántottam soha életemben, s kik nyomorulttá tettetek, kik miatt nőm s gyermekim koldusbotra jutottak, kik belőlem gonosztevéőt csináltatok, kik kiűztetek az erdő vadjai közé, kik miatt e világon s az örökkévalóságban elkárhoztam...*; ‘you, who were the cause of my ruin! – you, who have caused my wife and children to beg their bread! – you, who made me a robber, who hunted me, who compelled me to herd with the beasts of the forest!’ (Eötvös 1850) Another version of this construction is when different characters are elaborated by parallel subordinations in the same level of the sentence: *...az apai szív örömében dagadozva áldá a pogány író t s barátját, ki neki e könyvet ajándékozta, s mindazokat, kik elkészítésében részt vettek* ‘And old Esaias blessed the pagan author who wrote the book, and the college-chum who made him a present of it, and even the very printer who had produced it’ (Eötvös 1850)⁵ Periodic sentences create a highly rhetorical language and make the text eloquent, but they can also be used for satirical-comical effect involving the accumulation of subordinations (especially in the passages criticising the Hungarian public conditions of the time.) This style of writing, despite its reliance on long sentences, is easy to understand due to its parallelisms, and has a strong rhetoric effect. In the Hungarian literary tradition it is significant until the last third of the 19th century – longer, than in Western European cultures (e.g. the decline its usage in England begins with William Wordsworth’s and Samuel Taylor

5. The English translation divides the original long sentences into several parts, so in selecting the examples I have focused on the English version to give a sense of the typical sentence structure of the original.

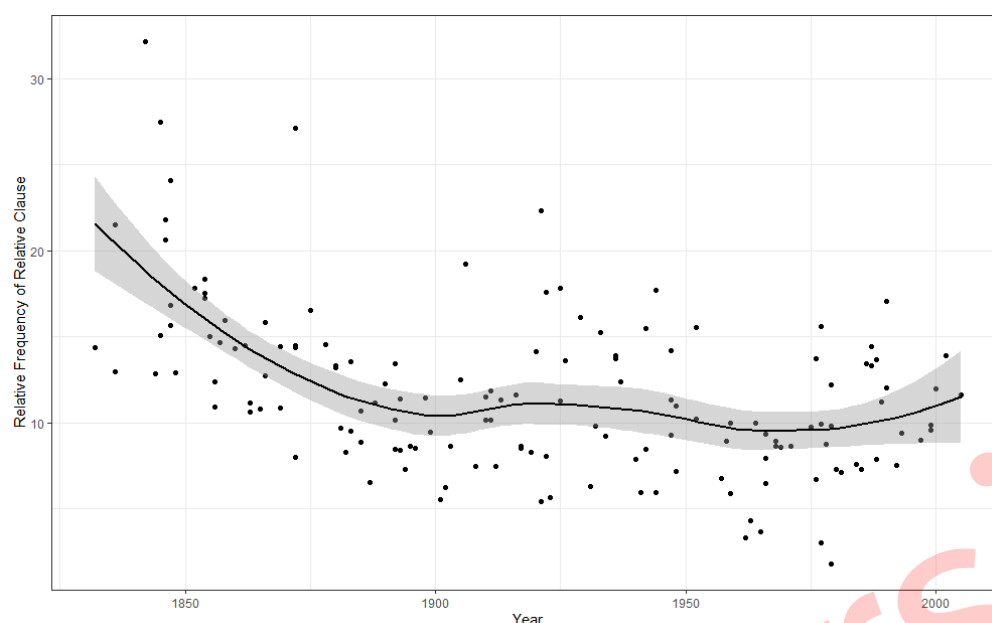


Figure 4: The changes of the prototypical relative clauses over time based on 150 novels with loess trendline, CI = 95%.

Coleridge's *Lyrical Ballads*, 1798 – Carter and McRae 2005, p. 420-22). It is because in Hungary, literacy education and the poetic tradition followed the ancient rhetorical model until the "Implementation of the Public Education Act" in 1868, which introduced compulsory education, and put the teaching of writing and reading on a new basis. Likewise, it is only in the second half of the century that a journalistic culture begins to emerge in which authors of texts based on shorter sentences and coordinations become successful (Mór Jókai played a leading role in this process – cf. Figure 2).

According to the data, another trend can be observed in the mid-19th century, one that also employs complex sentences, where the relation between the clauses tends to be coordination rather than subordination. These clauses might appear in the text as separate sentences — connecting them with conjunctions reinforces the logical and/or causal relationship between them. Linking clauses this way creates a more loosely edited, irregular prose than using relative clauses in a periodic sentence, but it allows authors to depict a dynamic sequence of events and to elaborate on motivations and situations. This long-sentence style reappears in Hungarian fiction in the 1970s and in fact becomes even more characteristic of a group of writers (especially for the main figures of the "prosa turn"), as can be seen in the graph showing the changes in the frequency of inferential clauses (Figure 5, ANOVA p-value is at the significance threshold: 0.04.) The trend without outliers is shown in Appendix Figure 14/b.

This figure and these results point to a similarity between certain novelists from different periods (and might even suggest cyclicity in the history of style). However, the differences between the two eras should also be noted. In the 19th century, inferential conjunctions usually introduce a sequence of events that logically follow each other and provide detail on a character's motivations. Making the successive nature of the events explicit is important because it clarifies the relation between complex structures that are otherwise difficult to understand. Yet, as early as the 19th century, authors start using inference ironically and parodically; they do so by employing inferential coordinations

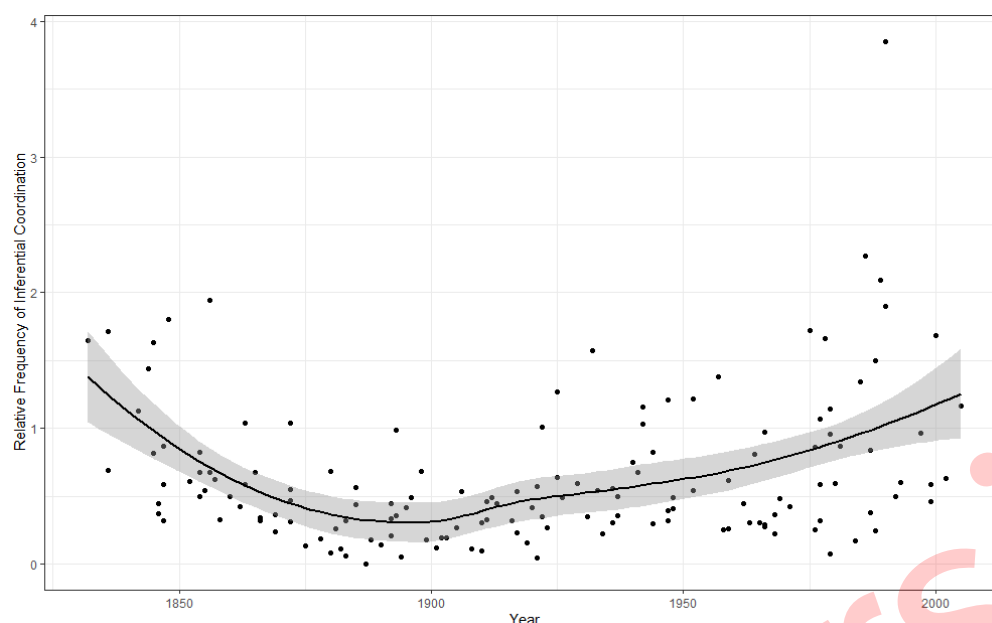


Figure 5: Changes in the frequency of inferential coordination over time based on 150 novels with loess trendline, CI = 95%.

or causal subordinations to make surprising connections between clauses that do not follow logically from each other. This can be observed in the following quotations from Ignác Nagy: *Az aratást tudniillik ma végezők be a dús alföld egyik gazdag birtokosánál a szegény emberek, mi természetesb tehát, mint, hogy a vendégszerető főtáblabíró úr nemes szomszédait ünnepélyes lakomára hívá meg, mert hiszen illő, hogy az urak is kifáradjanak valamiben, miután a parasztok már derekasan megizzadtak.* ‘The harvest was to be completed today by the poor people of the rich landowner, so what could be more natural than for the hospitable lord to invite his neighbours to a feast, for it is appropriate that the lords should also be tired after the peasants have sweated their hearts out.’ (Mosquitoes, my translation); or: *Az árkokban rothadó víz és szemét igen ocsmány bűzt terjeszt és ez rendkívül hasznos, mert a faluról bejövő uraságokat arra figyelmezteti, hogy mielőbb siessenek ismét vissza egészséges falusi levegőjökre;* ‘The water and garbage rotting in the ditches spreads a very foul stench, and this is very useful, because it warns men coming in from the village to hurry back to their healthy air.’ (Hungarian Secrets, my translation)

Focusing on the other end of the timeline, it should be noted that twentieth-century novels with long sentences exploit this subversive or ironic potential of inferences. These works develop detailed, intricate relationships in complex sentences but draw attention to the artificiality or even absurdity of these relationships. The artificiality of the sentence structure may be due to the fact that writers of the era recognized that the language of narrative fiction was only one among many discourses and could only represent a world that had always already been interpreted a certain way. (Kulcsár Szabó 1994) Reacting to this recognition, novelists employ a high number of not only inferential coordinate clauses but also explicatory coordinate clauses – a stylistic trait not shared by nineteenth-century novels (see data in *Data availability*). The accumulation of explanations and inferences clarifies the logical structure of the sentence and the represented world, but at the same time, the sheer number of inferences and explanations also draws attention to the artificiality or inaccuracy of these relationships (since they can always

be redescribed ever more accurately). This can be observed, for instance, Imre Kertész's, 362
 László Krasznahorkai's or Péter Nádas's novels in the corpus - which is closely related 363
 to the content of the texts. The works of Kertész for example describe events as the 364
 inevitable consequences of the functioning of the social order, and their aim is to explain 365
 this functioning as precisely as possible – whether it is the history of the Holocaust, 366
 Hungary of the 1950s or the Kádár era. The explanations show not only the natural 367
 consequence and cause of things, but also their uncanny absurdity: *Már egészen más 368*
dolog azonban – tették rögtön hozzá – az Arbeitslager, vagyis munkatábor: ott az élet könnyű, 369
a viszonyok és az élelmezés, járt híre, hasonlíthatatlan, s ez természetes is, hisz ott a cél is más 370
elvégre. 'An "Arbeitslager" or "work camp," on the other hand, it was immediately added, 371
 was something quite different: life there was easy, the conditions and food, the rumors 372
 went, bore no comparison, which is natural enough as the aim, after all, is also different.' 373
 (Kertész 2006) The same technique is applied in the *Kaddish for an Unborn Child* by Imre 374
 Kertész, where the English translation (similar to the quote above) uses disjunctive 375
 conjunctions and the term "more precisely" where the original employs explanatory 376
 relationships leading "to the point of absurdity": *ez a kérdés te vagy, pontosabban én vagyok, 377*
de általad kérdésessé téve, még pontosabban (és ezzel nagyjából doktor Obláth is egyetértett): 378
az én létezésem a te léted lehetőségeként szemlélve, vagyis én mint gyilkos, ha a pontosságot a 379
végletekig, a képtelenig akarjuk fokozni, és némi önkínzással ez meg is engedhető, hiszen, hál' 380
isten, késő, mindig is késő lesz már... 'and you are that question; or to be more precise, I 381
 am, but an I rendered questionable by you; or to be even more precise (and Dr. Obláth, 382
 too, broadly agrees with this): my existence viewed as the potentiality of your being, 383
or in other words, me as a murderer, if one wishes to take precision to the extreme, to 384
 the point of absurdity, and albeit at the cost of a certain amount of self-torment, since, 385
 thank God, it's too late now...' (Kertész 2004) The same is true of László Krasznahorkai's 386
 early prose, where the characteristics of the narrated world are also unfolded "from 387
 within", from the characters point of view, which makes the otherwise absurd events 388
 reasonable and relatable. This internal perspective is an important difference from 19th 389
 century novels, which develop inferences and logical connections in a similar way and 390
 in similar numbers – but almost always from an external perspective. Thus the ironic 391
 effect is more clearly encoded in those narratives, in so far as they maintain an external 392
 perspective that can look at the existing conditions from the outside; in the case of 393
 Kertész and Krasznahorkai (and some of their contemporaries), by contrast, the irony 394
 of the inferences is only created by the reader, since the character voices do not reveal 395
 the bizarre nature of the operations described. 396

The trends discussed so far mainly concern long-sentence prose, with outliers at both 397
 ends of the timeline. In contrast, the relative frequency of similes reaches its peak in 398
 the first half of the 20th century (see Figure 6). Based on the ANOVA test, a correlation 399
 between similes and year of publication can only be found between three major periods 400
 (1832-1909, 1910-1949, 1950-2005; p-value = 7.59e-07; without these yeargroups p-value 401
 = 0.84). The trend is also present without outliers – see Appendix Figure 14/c both with 402
 loess trendline and segmented linear regression. An analysis of the corpus suggests 403
 that novels from this period develop explicit relations between clauses less frequently; 404
 instead, writers tend to describe situations by placing individual scenes or images side 405
 by side. An exception to this, however, is the use of similes, whose aim is precisely to 406
 link distant areas together, sometimes resulting highly poetic formulations (e.g. *Délfelé* 407

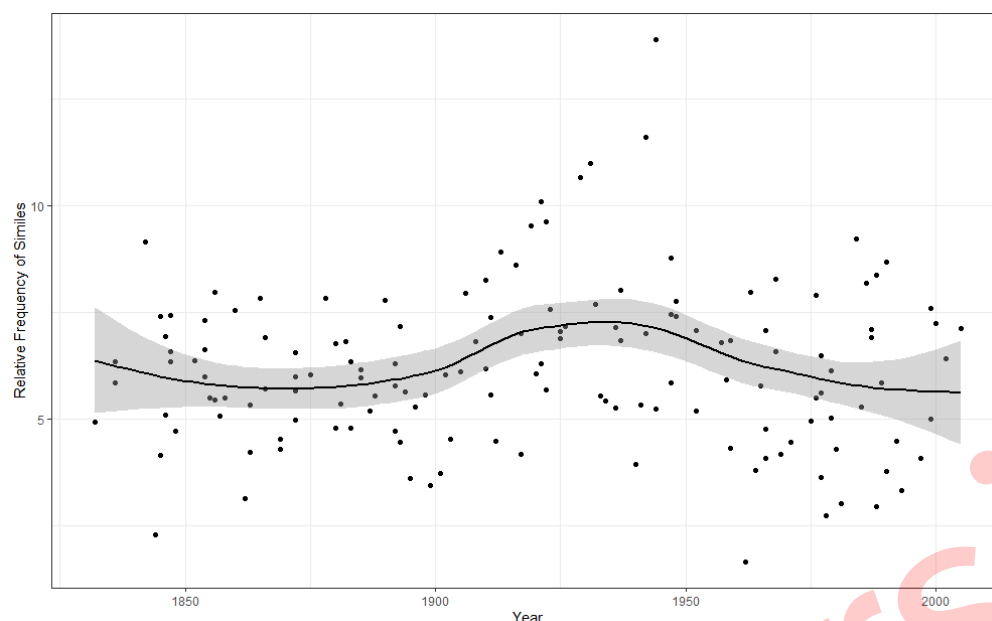


Figure 6: Changes of similes over time based on 150 novels with loess trendline, CI = 95%.

gyenge szél indul, forró, mint a gazdátlan büntudat. 'A light breeze comes up from the south, hot as uncontrollable remorse.' – Miklós Mészöly, *Saulus* - my translation). The large number of these kind of relations may be related to the increasing emphasis from the 20th century on the representation of the inner lives of the characters and the ever more frequent use of the technique of free indirect speech. As cognitive means of interpreting the world (cf. Lakoff and Johnson 1980) and as expressions of subjective perception ("he felt as if he were enclosed in a glass enclosure"), similes are particularly suitable for incorporating these inner worlds into the narrative.

Moreover this graph identifies a new and chronologically distinct paradigm. While the use of relative clauses was prevalent in the first half of the 19th century and the use of explanatory and inferential relations abundant in the late 20th century (and, to some extent, in the first half of the 19th century), the first half of the 20th century is characterized by similes. These stylistic paradigms of sentence structure also hint at historical differences in the epistemological approach to the world. Relative clauses provide a more detailed description of a character in the main clause by representing them in a new scene, which fixes the meaning of the sentence. By contrast, constant clarification and explanation open up the sentence to new interpretations; this assumes a less fixed or semantically less anchored access to the world but offers a more subjective, internal point of view. Similes, the third paradigm, can provide new information by comparing two ideas, or simply pointing to their common features; in this case, the focus is not on characterization or elaboration but on creating an analogy ("You are beautiful like a rose"), articulating differences ("Your dad is not as strong as mine"), or incorporating the inner world of a character into the third person narration.

5. Results 2

431

Another way to visualize the results is to focus on the internal structure of the novels. 432
 In this case, one does not calculate relative frequencies in a text (relative to the number 433
 of words) but the percentage of the relation types in proportion to each other. Here 434
 the copulative coordinations are left out, partly because they provide the least relevant 435
 information about the style of a text and partly because it is common for them to not 436
 be elaborated by conjunction words (e.g. juxtapositions like “I went to the restaurant, 437
 danced at the club, slept in the hotel.”) Thus, the results are only approximate and 438
 do not reflect the structures of the novels in their entirety – the figures only visualize 439
 the proportion of certain types of clause relations. For the sake of clarity, adversative 440
 and disjunctive coordinations; inferential and explicatory coordinations, and, finally, 441
 temporal and locative subordinations are put into joint groups in Figure 7, since the 442
 grammatical function of these relations is quite similar. A more detailed figure could 443
 easily be made, but the resulting graph would be somewhat crowded. Similarly, one 444
 could also visualize the proportion of types within joint categories (e.g., the proportion 445
 of temporal and locative subordinations). 446

There are multiple reasons these graphs deserve just as much attention as the figures 447
 showing historical tendencies in the change of the frequencies relative to the number of 448
 words. First, while the previous figures examined the types separately, these graphs 449
 show the proportion of them at once. Since some types (e.g., concessive, explicatory, and 450
 inferential relations) are almost always less frequent than others (e.g., relative clauses), 451
 what the graphs reveal is not necessarily the most frequent type in a text but the extent 452
 of the difference between types. Secondly, the figures offer a better representation of 453
 the characteristics of the novels that either had high values in several categories or that 454
 did not have high values in any of them but whose internal structure is interesting for 455
 some reason (see for example Iván Mándy’s novel *A Trafik* [*The Tabacconist*], where 456
 similes dominate a text that is otherwise poor in elaborated connections). Thirdly, these 457
 diagrams allow researchers to compare texts in a different way: since every type of 458
 relation is shown in the same graph, all types can be taken into account in the comparison, 459
 and the difference between the proportions can be seen at a glance. 460

In this figure, one can easily see which works tend to use similar sentence structure. 461
 Moreover, the similarity between novels written by the same author (e.g., József Eötvös) 462
 is unmistakable. This raises the question of whether the quantitative analysis of clause 463
 relations can help not just in exploring stylistic trends and interpreting individual texts, 464
 but also in authorship attribution. (This hypothesis is supported by Grieve, who explains 465
 the difference between authors in terms of registers rather than ideolects. See Grieve 466
 2023) In other words: to what extent can texts belonging to one author be distinguished 467
 from other authors on the basis of our data? Authorship researchers have proven on 468
 several occasions that there is a so-called “authorial fingerprint” beneath the thematic 469
 level of texts, which refers to the distribution of the most frequently – and therefore 470
 unconsciously – used words, mainly function words without specific meaning; and 471
 this distribution is approximately constant across texts of different genres written by a 472
 given author throughout his or her career. (Baayen 2001; Burrows 2002; Rybicki and 473

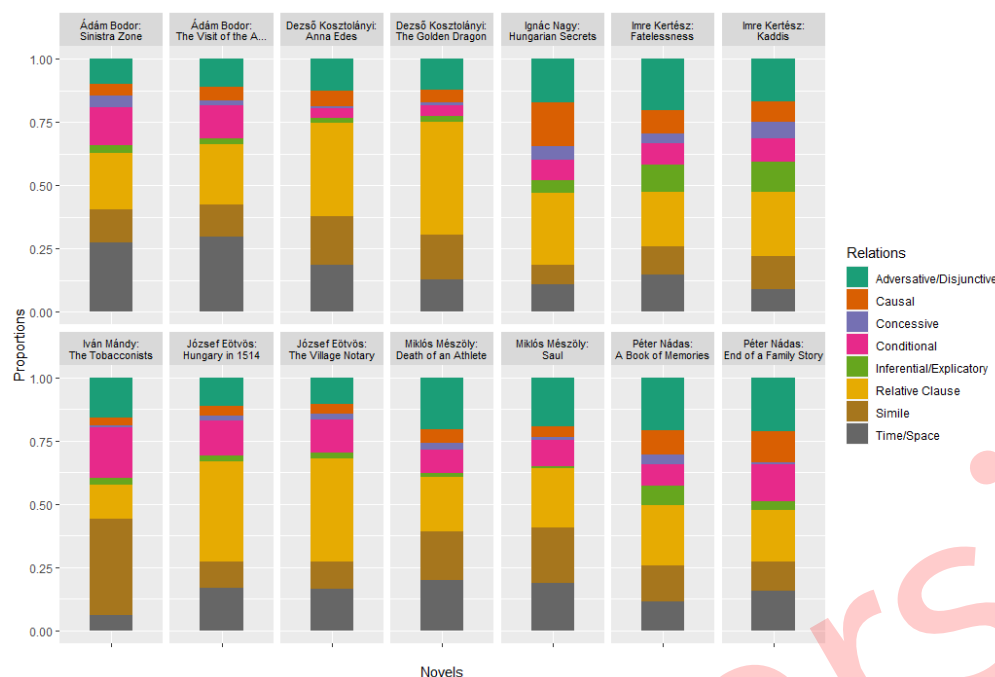


Figure 7: The internal structure of certain novels based on the categories under scrutiny.

Eder 2011)⁶ This means that based on the relative frequency of the most frequent words one author is distinguishable from another (irrespective of their social and aesthetic backgrounds). The question is whether this rule holds true even when considering only conjunction words and those relative pronouns that form a connection between clauses.

To test this question, both the frequencies relative to the length and relative to each other were taken as a starting point: a novel is thus associated with several measures (12 for relative frequencies and 8 for proportions, respectively) which can be used to place the texts in a multidimensional space (i.e. a multidimensional coordinate system). Such a multidimensional space, however, cannot be represented or imagined – but it helps us to describe the similarity and dissimilarity of texts in two ways. The first possibility is not to plot the datapoints in this space, but simply calculate the distance between their positions according to metrics that work the same in two dimensions, then to group the texts based to their proximity (in this case, using Ward’s method), and finally to visualize these groups in the form of a dendrogram.

When grouping novels in this way, the performance of several distance metrics was compared. Cosine distance proved to be the most effective in terms of the two types of data (relative frequency and proportion), while Manhattan distance performed best when considering the aggregated data set. (However, there seems to be a consensus among scholars that cosine distance is the most reliable metric for authorship attribution. Cf. Evert et al. 2017) The latter phenomenon is illustrated by the dendrogram of Figure 8, which shows 33 novels from the corpus, at least two of which have the same author (the same authorship is color-coded); this allows one to analyze the accuracy with which texts by the same author are placed on the same branch in the plot. In many

6. Cf.: “We assume that in a language, there is a subset of (traceable) linguistic features dependent on an individual idiolect rather than shared by writers of the same epoch, genre, gender, etc. In a word, we believe that some features of a written text can betray the person who wrote it, despite his/ her aesthetic, social, or historical conditions.” Eder 2002

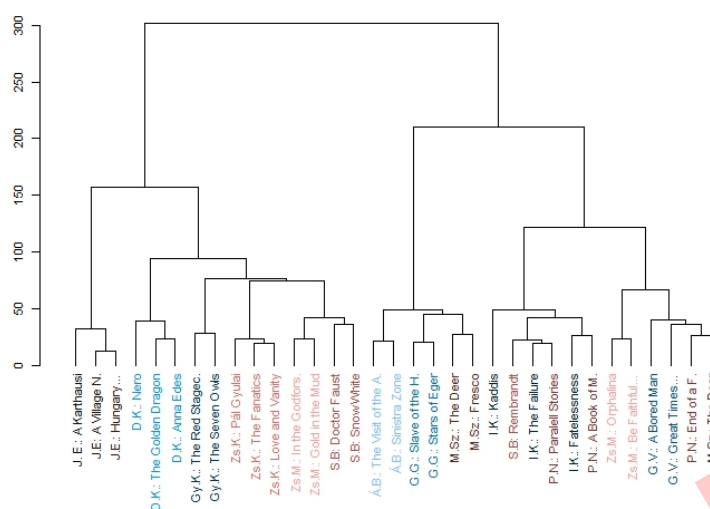


Figure 8: A clustering of 33 novels using Ward's method based on the relative proportion and relative frequency of relation types – Manhattan distance. Authors and titles are abbreviated - for details see *Data availability*

cases, works by the same author are grouped together correctly, but there are also some misclassifications (Adjusted Rand Index, ARI = 0.54).

The second possibility is to reduce the dimensions of the multidimensional space while preserving the spatial relationships of the data points as much as possible. The results of such a reduction based on principal component analysis (PCA) for relative frequencies are shown in Figure 9. Here, the difference and similarity between texts is a function of the position and the percentage assigned to the axes (PC1 and PC2) – a value that shows the extent to which the distance on the axes plays a role in distinguishing data points. Thus, texts that are similar according to the selected criterion are positioned close to each other; while works with the same authorship are shown in the same color. In addition, each type of relationship is also marked as a loading in the figure, the direction of which shows how these types influence the location of the texts as data points.

The separation of novels by the same authors can be described as rather successful (i.e. we can support the hypothesis), even if not perfect: dendrogram grouping does not work without errors, whereas in the PCA diagram these novels are in most cases in a similar position but cannot be clearly distinguished from other texts or groups of texts. This is due to the very characteristics under investigation. Namely, the use of conjunction words and relative pronouns belongs to a *semi-conscious* level of the text. While the distribution of other function words (such as articles) is completely independent of the author's intentions – which is why their frequency can bear the "fingerprints" of the author's style – the frequency of the different sentence structures is not entirely independent of the author's individual considerations and aesthetic design. These words operate somewhere between the conscious and the unconscious levels, inasmuch as their use (unlike the use of content words) is not controlled in the creative process, but conscious authorial choices can influence their frequency. This also reflects that the question of

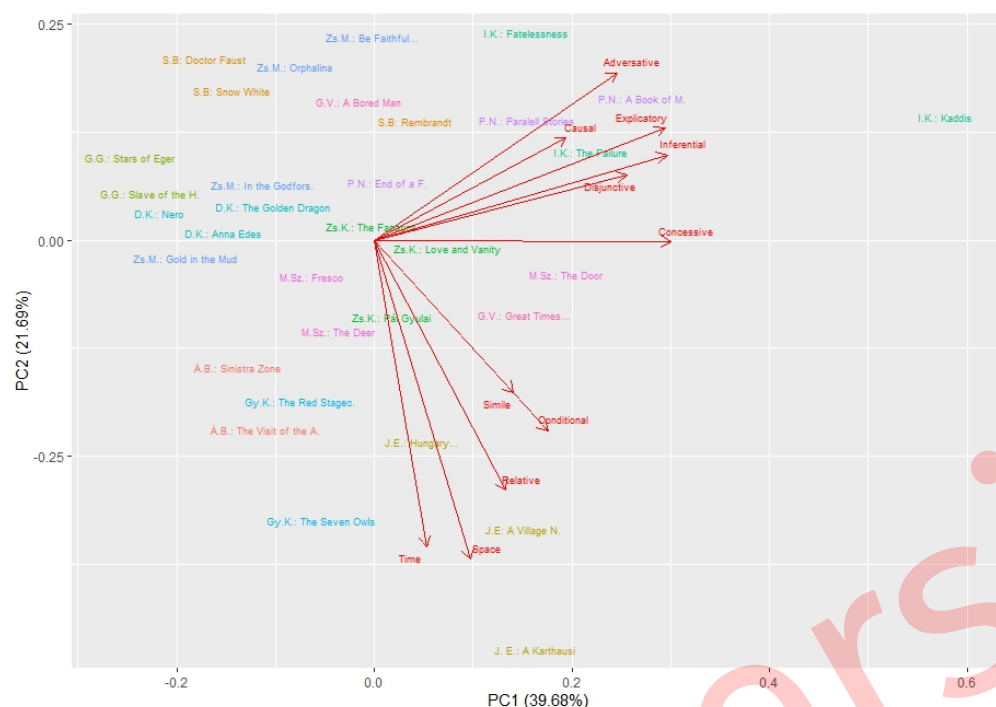


Figure 9: Principal component analysis by relative frequency of relation types, 33 novels.

which elements are under the author's control can be imagined along a continuum rather than along a conscious-unconscious dichotomy. Furthermore, this is why the experiment can be considered successful but without perfection, in that it is possible to group several texts by the same author, but at the same time it is not possible to group texts that are written in very different registers. Consequently, the analysis of the frequency of clause relations is above all not suited to answer questions of authorship attribution (or at least not by itself); its real use lies in identifying different stylistic traditions in the history of prose.

Figure 9 clearly shows that there are two distinct groups of relation types which could be characteristic for a text, that is, two directions can be distinguished with the help of the loadings: clause linkages traditionally fall into the category of coordinations (adversative, disjunctive, inferential, explicatory and concessive) and causal subordinations (that elaborate similar logical relation than some of the coordinations, mainly inferential and explicatory) operate in one direction; whereas comparative, conditional, relative, temporal, and locative subordinations operate in the other. The same holds true when the investigation is carried out on all the 150 novels (Figure 10). These results suggest that three traditions of complex and compound sentences exist in the history of canonical Hungarian fiction - which can only partly be explained by the diachronic changes presented already, since texts from different periods may be part of the same tradition. The first tradition tends to develop logical relations between coordinate clauses, similarly to how logical value is attributed mainly to conjunctions in formal logic. The second employs a style that tends to give more information on the actors (whether human or non-human) or the circumstances of the depicted scenes; in these, it is the number of relative, conditional clauses and similes that are high. The third tradition includes novels that favor no type of clause relations; they prefer short sentences and mark relations between clauses less frequently. Needless to say, these styles should be

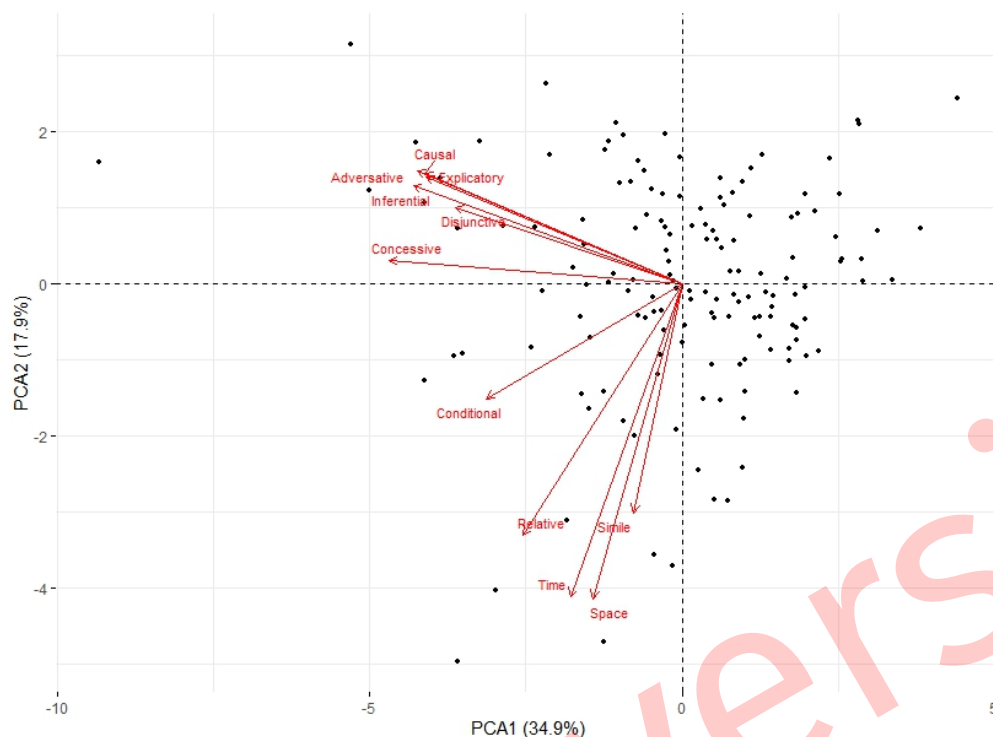


Figure 10: Principal component analysis by relative frequency of relation types, 150 novels.

seen as archetypes, and are rarely realized in pure form. Moreover, as we have seen earlier, these traditions are also subject to change over time.

6. Conclusion

An examination of the distribution of clause relations offers a better understanding of not only the linguistic structure of texts but also their diagrammatic, topological, and logical properties. Thus, the automatic identification of clause relations and the measurement of their frequency provides more than just a stylistic analysis: it can contribute to describing trends in literary history, interpreting individual novels, and distinguishing different traditions of prose style. The present study identified three traditions: a tradition that makes heavy use of subordinations, provides detailed descriptions of the elements in the main clause, and, thus, fixes the meaning of the sentence (mainly in the 19th century); a tradition that establishes logical relations between clauses, which keeps opening up the sentence to new interpretations (mainly in the second half of the 20th century); and a short-sentence tradition that relies chiefly on simple sentences, clause relations that are not elaborated, and similes (mainly in the first half of the 20th century). These conclusions are supported by various visualizations of the results; each type of visualization presents the values in a different layout to emphasize different aspects of the texts. Future directions for research include analyzing individual types in greater detail and breaking them down into subcategories, and complementing the research carried out so far by examining the relationships developed between sentences.

7. Appendix

568

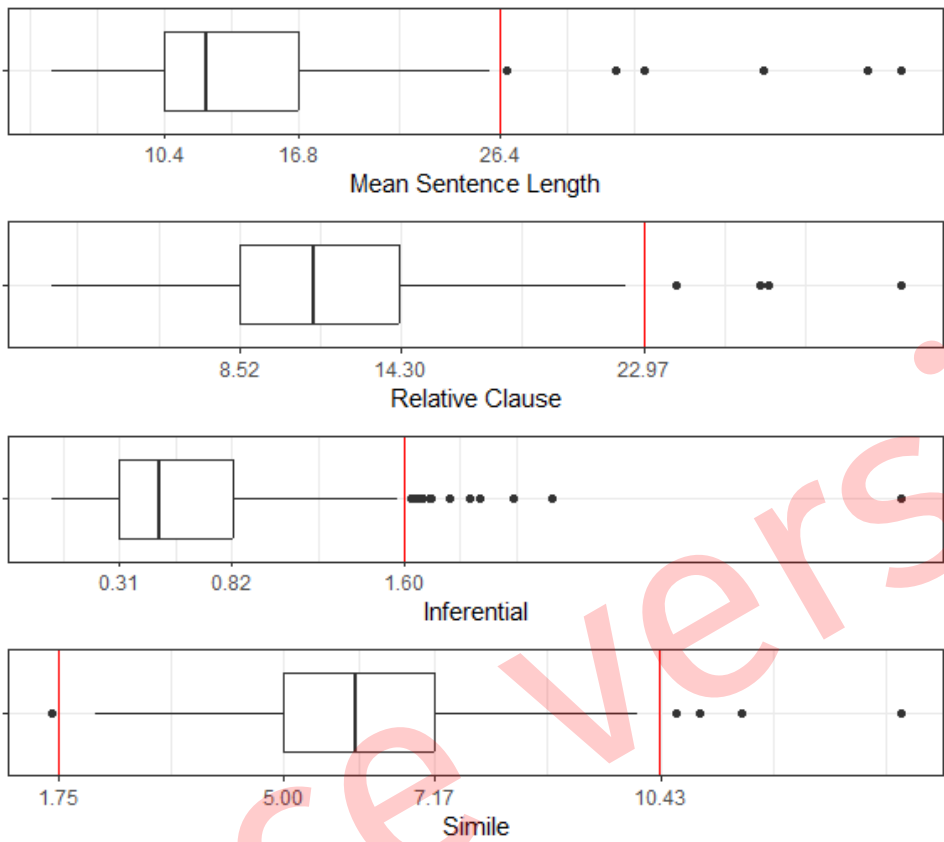


Figure 11: Outlier detection for mean sentence length, prototypical relative clause, inferential coordination and similes based on $Q3 + 1.5 * IQR$, and $Q1 - 1.5 * IQR$. The threshold values can be used to identify which novels are considered outliers. For details see *Data availability*

8. Data Availability

569

Data and codes can be found here: https://anonymous.4open.science/r/sentence_structure-46B0/

9. Acknowledgements

572

The author would like to thank Péter Tamás and Ben Nagy for their help with the translation, and the members of the Computational Stylistics Group in Krakow and the Department of Digital Humanities in ELTE University, Budapest for their methodological and theoretical contributions.

10. Author Contributions

577

Botond Szemes: Conceptualization, Analysis, Writing

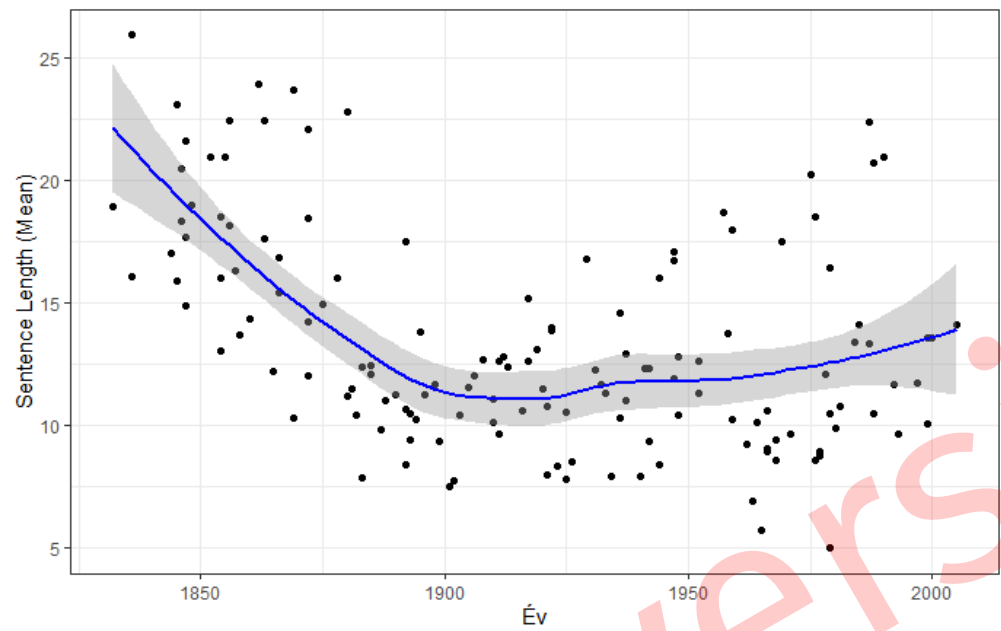


Figure 12: Trend in the changes of mean sentence length without the outliers (1832-2005)

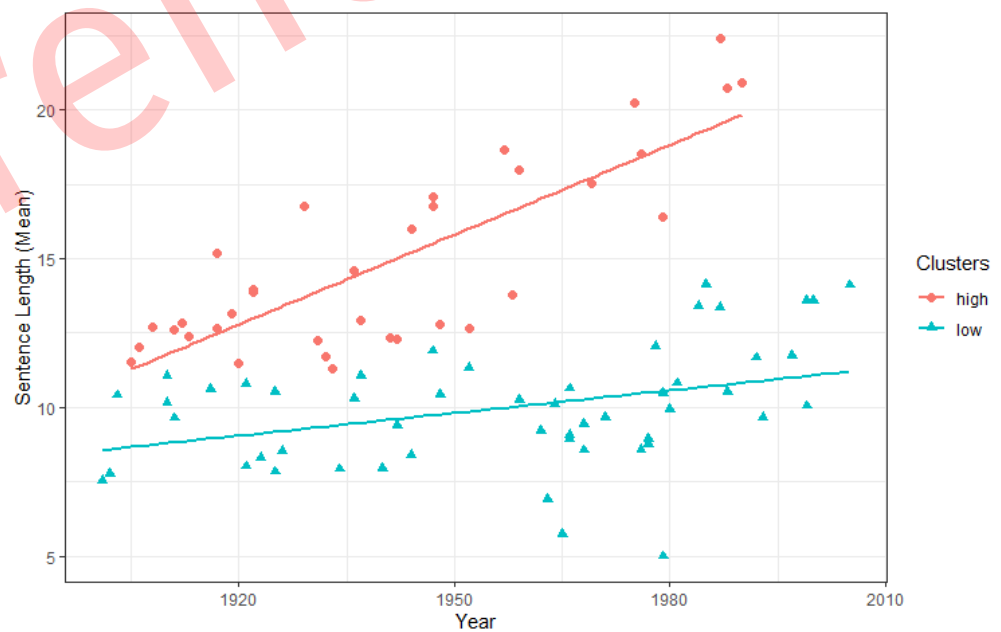
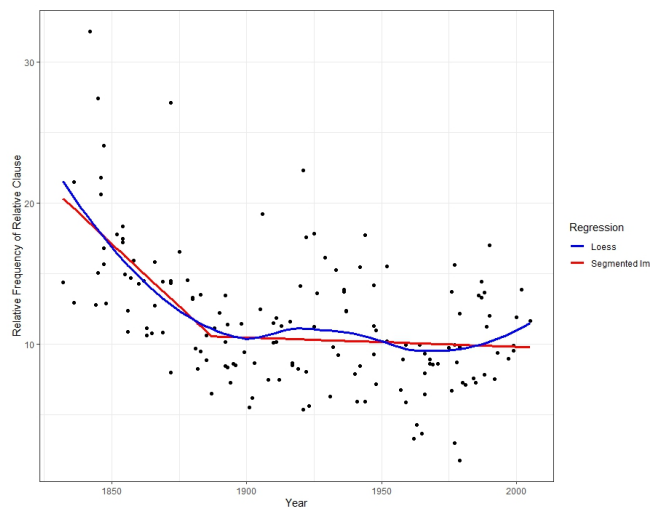
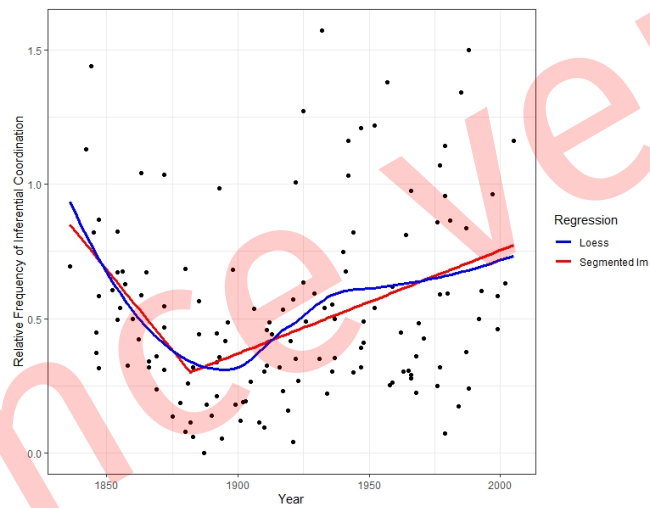


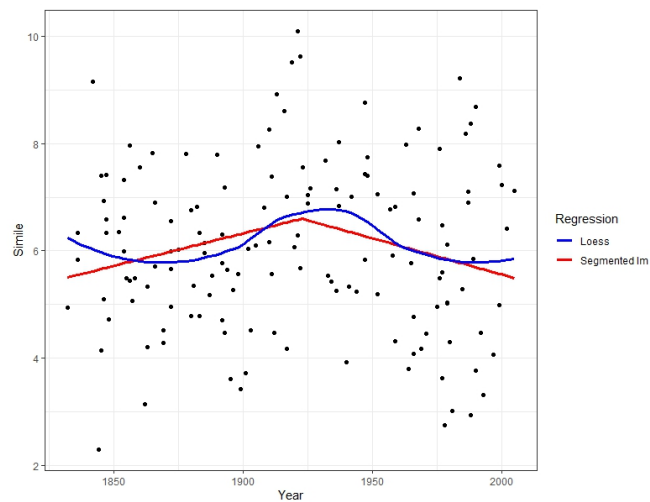
Figure 13: Trend in the changes of mean sentence length without the outliers (1900-2005), after clustering the data as described for Figure 3.



(a) Prototypical relative clause; for segmented lm $R^2 = 0.32$.



(b) Inferential coordination, for segmented lm $R^2 = 0.14$



(c) Simile, for segmented lm $R^2 = 0.02$.

Figure 14: Trends in a given clause relation without outliers based on loess trendline and segmented linear regression. Segmentation was done automatically using the *segmented* R package.

References

579

- Allison, S., M. Gemma, R. Heuser, F. Moretti, A. Tevel, and I. Yamboliev (2013). "Style at the Scale of the Sentence." In: *Stanford Literary Lab Pamphlets* 5. <https://litlab.stanford.edu/LiteraryLabPamphlet5.pdf>.
- Baayen, H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.
- Bajzát, T., B. Szemes, and E. Szlávich (2021). "Az ELTE DH Regénykorpusz és lehetőségei [The possibilities of the Novel Corpus from ELTE DH]". In: *Online térben az online térért. Workshop 30: országos online konferencia. 2021. április 6-9, Eötvös Loránd Tudományegyetem [Hungarian] Online spaces for online spacers. The 30th Workshop conference*. Ed. by J. Tick, Kokas K., and Holl A. Budapest: HUNGARNET Egyesület.
- Burrows, J. (2002). "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." In: *Literary and Linguistic Computing* 17. <https://doi.org/10.1093/llc/17.3.267>.
- Calvo Tello, J. (2023). *The Novel in the Spanish Silver Age*, Bielefeld. Bielefeld: transcript. <https://doi.org/10.1515/9783839459256>.
- Carter, R. and J. McRae (2005). *The Routledge History of Literature in English: Britain and Ireland*. London / New York: Routledge.
- Chung, C. and J. Pennebaker (2007). "The psychological functions of function words". In.
- Cristofaro, S. (2014). "Is there really a syntactic category of subordination?" In: *Contexts of Subordination. Cognitive, typological and discourse perspectives*. Ed. by L. Visapää, J. Kalliokoski, and H. Sorva. Amsterdam / Philadelphia: John Benjamins Publishing Company, 73–93. <https://doi.org/10.1075/pbns.249.03cri>.
- Czigány, L. (1984). *The Oxford history of Hungarian literature : from the earliest times to the present*. Oxford: Clarendon Press. <https://mek.oszk.hu/02000/02042/html/index.html>.
- Dömötör, A., Z. Gy. Yang, and A. Novák (2020). "Much Ado About Nothing Identification of Zero Copulas in Hungarian Using an NMT Model". In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Ed. by N. Calzolari. Marseille, France: European Language Resources Association. <https://www.aclweb.org/anthology/2020.lrec-1.591/>.
- Eder, M. (2002). "Style-Markers in Authorship Attribution: A Cross-Language Study of the Authorial Fingerprint." In: *Studies in Polish Linguistic* 7.1.
- Eötvös, J. (1850). *The Village Notary*. Trans. by O. Wenckstern. London: Longman, Brown, Green and Longmans.
- Evert, S., T. Proisl, F. Jannidis, I. Reger, S. Pielström, C. Schöch, and T. Vitt (2017). "Understanding and Explaining Delta Measures for Authorship Attribution". In: *Digital Scholarship in the Humanities* 32.2. <https://doi.org/10.1093/llc/fqx023..>
- Grieve, J. (2023). "Register variation explains stylometric authorship analysis." In: *Corpus Linguistics and Linguistic Theory*. <https://doi.org/10.1515/cllt-2022-0040>.
- Herczeg, Gy. (1981). *A XIX. századi magyar próza stílusformái [Stylistic Forms of 19th Century Hungarian Prose]*. Budapest: Akadémiai Kiadó.
- Hopper, P. and E. Closs Traugott (2003). *Grammaticalization*. Cambridge: Cambridge UP. <https://doi.org/10.1017/CB09781139165525>.
- Imrényi, A. and N. Kugler (2018). "Mondattan [Hungarian] Grammar of Sentences". In: *Nyelvtan*. Ed. by G. Tolcsvai Nagy. Budapest, Hungary: Osiris.

- Kabatek J.; Obrist, P. and V. Vincis (2010). "Clause linkage techniques as a symptom of discourse traditions: Methodological issues and evidence from Romance languages". In: *Syntactic Variation and Genre*. Ed. by H. Dorgeloh and A. Wanner. Berlin and New York: Mouton de Gruyter, 247–276. <https://doi.org/10.1515/9783110226485.2.2>
- Kanatova, M., A. Milyakina, T. Pilipovec, A. Shelya, O. Sobchuk, and P. Tinitis (2017). "Broken Time, Continued Evolution: Anachronies in Contemporary Films." In: *Stanford Literary Lab Pamphlets* 14. <https://litlab.stanford.edu/projects/broken-time/>.
- Kertész, I. (2004). *Kaddis for an Unborn Child*. Trans. by T. Wilkinson. New York: Random House.
- (2006). *Fatelessness*. London: Vintage.
- Kestemont, M. (2014). "Function Words in Authorship Attribution. From Black Magic to Theory?" In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Ed. by A. Feldman, A. Kazantseva, and S. Szpakowicz. Gothenburg: Association for Computational Linguistics. <https://doi.org/10.3115/v1/W14-0908>.
- Kortman, B. (1997). *Adverbial Subordination. A typology and history of adverbial subordinators based on european languages*. (*Empirical Approaches to Language Typology*, 18.) Berlin and New York: Mouton de Gruyter. <https://doi.org/10.1515/9783110812428>.
- Kugler, N. (2020). "Contextualizing Clauses." In: *Studia Linguistica Hungarica* 32, 76–90.
- Kulcsár Szabó, E. (1994). *A magyar irodalom története 1945–1991 [History of Hungarian Novel 1945–1991]*. Budapest: Argumentum.
- Kytö, M. and E. Smitterberg (2023). "Clausal and phrasal coordination in recent American English." In: *Corpus Linguistics and Linguistic Theory* 19.1, 23–46. <https://doi.org/10.1515/cllt-2022-0035>.
- Lakoff, G. and M. Johnson (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.
- Matthiessen, C. and S. A. Thompson (1988). "The Structure of Discourse and 'Subordination'". In: *Clause Combining, in Grammar and Discourse*. Ed. by J. Haiman and S. A. Thompson. Amsterdam / Philadelphia: John Benjamins Publishing House, 275–331. <https://doi.org/10.1075/tsl.18.12mat>.
- Orosz, Gy., Zs. Szántó, P. Berkecz, G. Szabó, and R. Farkas (2022). "HuSpaCy: an industrial-strength Hungarian natural language processing toolkit". In: *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged.
- Raible, W. (1992). *Junktion. Eine Dimension der Sprache und ihre Realisierungsformen zwischen Aggregation und Integration*. Heidelberg: Winter.
- (2001). "Linking clauses". In: *Language Typology and Language Universals*. Ed. by A. Burkhart, H. Steger, and H. E. Wiegand. Berlin / New York: de Gruyter, 590–617.
- Rybicki, J. and M. Eder (2011). "Deeper Delta across genres and languages: do we really need the most frequent words?" In: *Literary and Linguistic Computing* 26.3. <https://doi.org/10.1093/lc/fqr031>.
- Schöch, C. (2022). "Sentence length across ELTeC collections and Gutenberg Fiction". In: *Distant Reading Closing Conference, April 21-22, 2022*. <https://christofs.github.io/krakow22/>.
- Seiler, H. (1995). "Review of Raible 1992". In: *Vox Romanica* 54, 12–21.
- Stjernfelt, Frederik (2010). "The Extension of Peircean Diagram Category: Charting the Implications of a Diagrammatical Revolution in Semiotic". In: *Studies in Diagram-*

- matology and Diagram Praxis*. Ed. by Olga Pombo and Alexander Gerner. London: College, 57–73.
- Szemes, B. (2020). “Mondathosszúság és irodalomtörténet [Sentence Length and Literary History].” In: *Literatura* 46.3, 335–367.
- Szirák, P. (2013). “Im Sog des Schrifttextes. Der literalistic turn in der ungarischen Nachmoderne ab 1960/1970”. In: *Geschichte der ungarischen Literatur*. Ed. by E. Kulcsár Szabó. Berlin: De Gruyter, 502–548. <https://doi.org/10.1515/9783110241105.502>.
- Váradi, T., E. Simon, B. Sass, I. Mittelholcz, A. Novák, and B. Indig (2018). “E-magyar – A Digital Language Processing System”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by N. Calzolari. Miyazaki: European Language Resources Association.
- Visapää, L., J. Kalliokoski, and H. Sorva, eds. (2014a). *Contexts of Subordination. Cognitive, typological and discourse perspectives*. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- (2014b). “Introduction”. In: *Contexts of Subordination. Cognitive, typological and discourse perspectives*. Ed. by L. Visapää, J. Kalliokoski, and H. Sorva. Amsterdam / Philadelphia: John Benjamins Publishing Company, 1–17. <https://doi.org/10.1075/pbns.249>.

Why the Daisy sisters are different

A stylometric study on the oeuvre of Swedish author Henning Mankell and the Dutch translations of his work

Martje Wijers¹ 

1. Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, The Netherlands.

Citation

Martje Wijers (2023). "Why the Daisy sisters are different. A stylometric study on the oeuvre of Swedish author Henning Mankell and the Dutch translations of his work". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-21

Date received 2023-02-17

Keywords

stylometry, Cluster analysis, PCA, delta, zeta, Mankell, translation

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. In this paper, 32 books by the Swedish writer Henning Mankell were investigated using stylometric methods, to find out whether his style varies in different genres, if his style changed measurably over time, or if his books differ from each other stylistically for other reasons. The results show that the time of publication can play a role, but that other factors, such as dominant verb tense used and narrative perspective, as well as register, are more important in determining whether and how the style of novels differs. This study also gives more insight into frequently used methods in stylometry, such as cluster analysis and PCA, that give little information about the stylistic features that differ between texts. For this purpose, the original Swedish texts were also compared to the Dutch translations of the same texts to determine how translation and language influence the results of stylometric analyses.

1. Introduction

In a conversation at the university of Tulsa in 2011, Swedish author Henning Mankell told his colleague Michael Ondaatje:

I'm like the farmer, who knows, that the land shouldn't be used for the same crops many years in a row. I try to cultivate the land in my head in the same way...[] That's why I switch between styles and between novels, essays and theater. One of the decisive things for me is, when I have an idea for a story, to decide what kind of story it is. Is it a theater play? Is it a film script? A novel? A crime novel? (Jacobsen 2012, 31) ¹

Although Henning Mankell is most known for his detective series Wallander, he indeed wrote a variety of genres during his 42 years long career as a writer. He wrote literary novels, crime novels, non-fiction, theatre plays, film scripts and children's literature.

In this paper the whole oeuvre of Mankell is scrutinized using stylometric analyses to see if his style changed measurably over time, or if some books deviate stylistically from his other works for other reasons. In this study, style is used in the definition by Herrmann et al. (2015, 44): "Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively." In the current study style is analysed by quantitative features, as measured by word frequency patterns.

1. My translation

The original works by Mankell are also compared to their Dutch translations. The goal of this comparison is to investigate to what extent language and translation in general can influence the results of stylometric analyses. Apart from more insight into the styles in Mankell's oeuvre, this study will yield interesting observations about the selected methods, and it can give new insights about frequently used methods in computational literary studies, such as cluster analyses and principal components analyses. These methods are generally based on the Most Frequent Words (MFW) of a text, although little is known about which type of words are decisive and what factors should be taken into account in this type of analysis.

The current paper is inspired by the computational research project 'The Riddle of Literary Quality' (2012-2019). In this project, Karina van Dalen-Oskam and her colleagues at the Huygens Institute for the History of the Netherlands in collaboration with the Fryske Akademy and the Institute for Logic, Language and Computation at the University of Amsterdam investigated readers' perceptions of what (good) literature is and to what extent these perceptions can be linked to formal patterns in novels Dalen-Oskam 2021, 15–16² Five of the many novels that van Dalen-Oskam investigated were written by Mankell and particularly one of them, the literary novel *Daisy sisters*, stood out in several ways compared to the other novels written by translated male authors in the project. She looked into the novels by Mankell in more detail and the main conclusion was as follows:

It seems, therefore, that although Mankell published books in two different genres, Suspense and Literary novel, his style as reflected in his use of words, perhaps is not very different. The fact that *De Daisy Sisters* was an outlier does not disprove this because the original is much older (1982) and it is known that an author's writing style may develop over time just like languages and the conventions that apply to different genres [...]. Further research into Mankell's complete oeuvre would be needed to confirm this. Dalen-Oskam 2021, 76

So, Mankell's style did not differ very much between genres when looking at word use compared in a corpus including books by other writers, but the *Daisy sisters* deviated clearly from the other books, possibly because it was written much earlier. By looking at the broader oeuvre of Mankell, some of the questions that remained after the Riddle of Literary Quality was finished can be answered.

The corpus compiled for this study consists of 32 Swedish books written by Mankell in four genres: crime-fiction (N=15), literary novels (N=11), children's books (N=4) and non-fiction (N=2). For comparison purposes, ten books by the following best-selling Swedish writers were added to the corpus: Johannes Anyuru, Majgull Axelsson, Marianne Fredriksson, Lars Kepler, John Ajvide Lindqvist, Camilla Läckberg, and Håkan Nesser. Six of the books by other Swedish writers are literary novels and four are crime novels. The translation corpus contains 42 translations of all the above-mentioned works by Henning Mankell and other Swedish writers into Dutch.

2. An updated English version of this book will be published in English June 2023 under the title *The Riddle of Literary Quality: A Computational Approach*, Amsterdam University Press.

2. Multi-faceted Henning Mankell

Arvas and Nestingen (2011, 1) state that Mankell is the top selling Swedish crime-fiction author who, according to them “has sold 25 million copies, even outperforming Harry Potter in the German-language market.” Mankell was certainly one of Sweden’s most popular and well-read crime-fiction writers, although Berglund Berglund (2013, 10) puts these numbers in perspective. He shows that Henning Mankell was not necessarily the number one best-selling author in Sweden in the period 2004-2010, but that he indeed was among the top-sellers. He was in fact in fourth position after Camilla Läckberg, Stieg Larsson and Liza Marklund on the top 40 best-selling crime-fiction authors in Sweden (Berglund 2013, 81). Interestingly, compared to the even more popular authors Stieg Larsson, Camilla Läckberg and Liza Marklund, the books by Henning Mankell were borrowed much more frequently at libraries (Berglund 2013, 100–101).

Henning Mankell is an interesting author to investigate for multiple reasons. He modernized the already existing Swedish police novel that included criticism on modern society (started by Maj Sjöwall & Per Wahlöö) and he was the first Swedish author of crime novels to be published in many languages abroad with wide circulation. Therefore, Mankell played an important role in the rise of the Nordic noir genre (Berglund 2013, 114).

Mankell belongs to a group of authors that were already established writers of fiction before they started to write crime (in the late 90s) when there was a boom in crime-fiction in Sweden. His debut in the crime genre was in 1991 with *Mördare utan ansikte* (*Faceless killers*), but his debut as a writer of fiction was much earlier: in 1973 with *Bergsprängaren* (*the Rock Blaster*).

The fact that he has a broad oeuvre covering four genres over a time span of 42 years (1973-2015) also makes Henning Mankell useful for a computational study. Furthermore, his novels are widely translated into other languages. Almost his entire oeuvre is translated into Dutch. There is a limited number of Dutch translators from Swedish which makes it possible to compare translations by different translators. As mentioned earlier, these translations were sometimes published much later in Dutch than the original. It is important to bear in mind that a writer’s style can change over time, and so do ideas about translation (Can and Patton 2004; David L. Hoover 2020; Ríos-Toledo et al. 2022).

2.1 Mankell and the Riddle of Literary Quality

In the Riddle of Literary Quality, van Dalen-Oskam and her colleagues investigated if literary quality is measurable using stylometric methods. They selected 401 contemporary novels in Dutch published between 2007-2012 based on sales numbers and library borrowings in the three years prior to the survey (2009-2012) (Dalen-Oskam 2021, 44). The works included both novels originally written in Dutch as well as translated novels. These novels were rated for their literary quality on a scale from one (not literary at all) to seven (very literary) by almost 14000 readers in ‘Het Nationale Lezersonderzoek’ (the National Reader Survey) in 2013 (Dalen-Oskam 2021, 40–43). The ratings were then linked to the formal aspects of the books, such as vocabulary and sentence

length or contextual information, such as whether the author is male or female (van Dalen-Oskam).

One of the findings in The Riddle of Literary Quality was that many readers seemed to be somewhat more critical towards translated literary fiction compared to literary novels originally written in Dutch. In other genres, such as crime novels, the bias was just the opposite: on average, translated works received higher scores on literary quality than original Dutch crime novels (Dalen-Oskam 2021, 104–105).

However, there was a clear difference between books translated from English and books translated from other languages. Translated books from other languages than English scored higher on literary quality than works translated from English and books originally written in Dutch in the category literary novels as well as the category crime novels (Dalen-Oskam 2021, 105). Dalen-Oskam (2021, 112) suggests that readers are more critical toward translations from languages they know than from languages they are unfamiliar or much less familiar with.

In total there were 249 translated books in the survey. English was by far the language in which most of these books were written, namely 180. After English, the second most recurring original language was, somewhat surprisingly, Swedish (Dalen-Oskam 2021, 102). One Swedish author is represented with five books in the corpus: Henning Mankell. Three of these books are in the category crime. Remarkably, these three books end up relatively high in the ranking of literary quality among literary novels (Dalen-Oskam 2021, 178). The two literary novels, on the other hand, ended up among the lowest scoring literary novels, although the scores were still somewhat higher than his crime novels (Dalen-Oskam 2021, 193). The literary novel *Daisy sisters* turned out to have different frequency patterns of MFWs compared to other translated novels written by male authors. However, the frequency patterns of this book were remarkably close to the frequency patterns of one of the highest scoring translations: *Norwegian wood* by Haruki Murakami (Dalen-Oskam 2021, 190). Dalen-Oskam (2021, 189) wonders whether this could have something to do with the fact that both works were translated into Dutch much later than they were published in the original languages Swedish and Japanese (both in the eighties).

However, she did not have enough data in her corpus to investigate this assumption further. The corpus in the study I report on in this contribution, consisting of 32 books written by Mankell during his entire career, can confirm or reject this hypothesis. The following section reports on the results of the studies, and looks at genre differences, possible change over time and other factors that influence the clustering of texts.

3. Genre and style differences

When a book gets translated the genre classification chosen by the publisher could, at least theoretically be different in the source language. However, in the translations of the books by Henning Mankell into Dutch this is not the case. Squires (2007, 71–72) states that genre is a necessary part of book publishing. It is implemented in the whole publishing process, from cover design to advertising and what literary prizes the book qualifies for. The genre also determines on what shelf the book ends up in the bookstore

or library. Because of this, Squires (2007, 71–72) concludes that genre classification is not so much a literary boundary, but rather a marketing tool. Although this might be true to some extent, multiple studies in computational literary studies have shown that it is possible to distinguish genres based on style, measured by high frequency words (e.g. Dalen-Oskam 2021; Jautze 2014; Jautze et al. 2013; Jockers 2013).

Jockers (2013, 68–70) showed that genre and style are closely linked. Jockers and his colleagues looked at various subgenres in nineteenth century English novels. They divided the text into samples of 1000 words and performed an unsupervised clustering using the most frequent words (MFW). The high-frequency words turned out to not only be highly successful in distinguishing samples from the same author and novel, but also placed text samples that belonged to the same genre closely together. Jockers concluded that (sub)genres have a stylistic fingerprint that can be detected by looking at high-frequency words.

Jautze (2014) investigated whether the MFWs can distinguish chick lit from literary novels. She performed a stylometric analysis using the R package *stylo* (Eder et al. 2016) and found that chick lit was stylistically different from high literature. High literature turned out to have a more descriptive style, whereas chick lit seemed to be more informal.

In an earlier study, Jautze et al. (2013) compared high literature and chick lit syntactically and found that novels that are classified as high literature contain more complex sentences than chick lit. High literature was also found to be richer in prepositional phrases than chick lit.

To my knowledge, there are no studies that compare the style of high literature and crime novels, the genre that Henning Mankell is most known for. The genre is sometimes even referred to as literary crime novel, indicating that it has a higher literary quality than regular crime novels or thrillers. One might expect that it is harder to distinguish between high literature and ‘literary’ crime novels, especially if they are written by the same author.

To find out if an analysis of the MFWs can make this distinction, I performed a stylometric analysis on the Mankell corpus using the R package *Stylo* (Eder et al. 2016). The *Stylo* package automatically compiles a list of MFWs in the entire corpus and can check which words occur relatively frequently in the various texts, based on the Delta procedure for authorship attribution (Burrows 2002). Burrow’s delta looks at texts as a collection of data or ‘bag of words’ and disregards the context of sentences. The frequency of each word in the corpus is counted and the separate texts are compared to each other based on their frequency lists (MFWs). For this comparison, the relative, normalized z-scores are used, so differences in text length or the high impact of a small number of high-frequency words on the total outcome are ruled out (Eder et al. 2016). The distances between texts can, for instance, be visualized in a dendrogram representing the results of a cluster analysis, grouping texts that are similar to each other.

A cluster analysis was first performed on the Swedish corpus, to see if there are clear stylistic differences between various genres. The analysis is based on the 1000 most frequent words in the books. The results are visualized in Figure 1. As illustrated in figure 1, most books are neatly clustered by genre, where L stands for literary novel; C

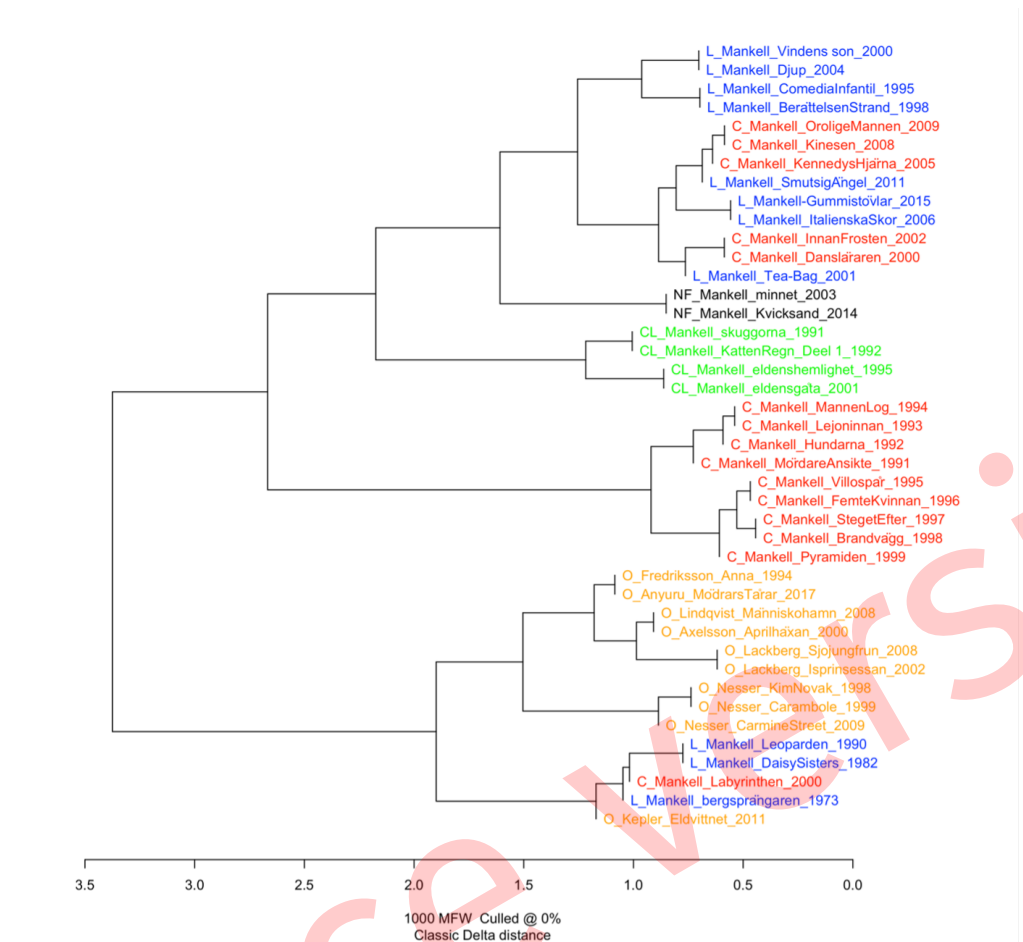


Figure 1: Cluster analysis of the Swedish books in the corpus based on the 1000 most frequent words (culling o, classic delta)

for Crime novel; NF for Non-Fiction and CL for Children's literature, although some 188
crime novels appear among literary novels or vice versa. This seems to be the case for 189
crime novels written from the year 2000 onwards. 190

The earlier crime novels: from the 1991 *Mördare utan ansikte* (*Faceless Killers*) until 191
Pyramiden (*The Pyramid*) from 1999, all belonging to the Wallander series, are in a 192
separate cluster. This cluster has two subclusters: one for the books written in the first 193
half of the 1990s (1991-1994), and one for the Wallander books published in the second 194
half of the 1990s (1995-1999). Remarkably, Mankell's last Wallander book *Den oroliga* 195
mannen (*The troubled man*), that was published in 2009, falls outside of this cluster. This 196
could again be explained by the fact that this last book of the series was written ten 197
years after the previous Wallander book, and that Mankell's style changed over time. 198
The crime novel *Innan frosten* (*Before the frost*) from 2002, which is written from the 199
perspective of detective Wallander's daughter, does not belong to the Wallander cluster 200
either. However, this book is closer in time to the other Wallander books, indicating that 201
there are other factors that weigh in. 202

The books by other authors than Mankell are also clearly different from the books by 203
Mankell. The crime novel *Carambole*, from the Van Veeteren series by Håkan Nesser, for 204
instance, is very comparable genre-wise to Mankells Wallander series. However, author 205
seems to be a stronger factor in the clustering of the text than genre, because *Carambole* 206

ends up in a separate cluster and clusters with other literary novels by Nesser. Genre, 207
in its turn, plays a more important role than time overall. If we for instance look at the 208
two non-fiction books by Mankell, they clearly form a cluster, even though they were 209
published eleven years apart. 210

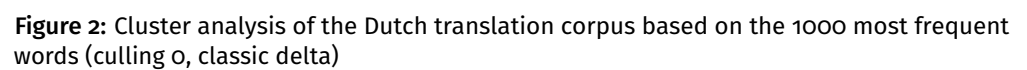
However, something remarkable is going on with the three oldest books by Mankell 211
in the corpus. These three literary novels: Mankell's debut, *Bergsprängaren* (*The Rock* 212
Blaster) from 1973, *Daisy Sisters* from 1982 and *Leopardens öga* (*The Eye of the Leopard*) from 213
1990 appear closer to other authors and even cluster with the 2011 crime novel *Eldvittnet* 214
(*The Fire Witness*) by Lars Kepler. The same is true for the crime novel *Labyrinthen* (*The* 215
Labyrinth) from a much later period (2000), that clusters with Mankell's early literary 216
novels. *Labyrinthen* is different from other works, because it was originally written as a 217
film script and later turned into a novel. This might have influenced the style of this 218
particular novel. 219

The fact that the three early Mankell novels are stylistically different from his later works 220
seems to indicate that Mankell's writing style and word choice indeed has changed over 221
time and confirms the findings by Van Dalen-Oskam that *Daisy Sisters* is different from 222
other novels by Mankell. However, the texts and their MFWs have to be investigated in 223
more detail to see how his style has changed and to ensure there are no other factors at 224
play. 225

The Dutch translation corpus was analyzed using the same procedure as shown for the 226
Swedish corpus to see if the texts appear in different clusters when they are translated. 227
The Dutch corpus consists of the same books by Mankell and by the ten books by 228
the aforementioned Swedish authors (Johannes Anyuru, Majgull Axelsson, Marianne 229
Fredriksson, Lars Kepler, John Ajvide Lindqvist, Camilla Läckberg, and Håkan Nesser). 230
This comparison can give important information about what type of MFWs influence 231
the clustering of texts in stylometric analyses. The results are shown in Figure 2. The 232
different titles were all labeled by genre first (L for literary novel; C for Crime novel; NF 233
for Non-Fiction, CL for Children's literature and O for different author than Mankell). 234
The second tag is the translator's initials, followed by the author's last name and two 235
years, the first one indicates the year the original novel was published, the second one 236
stands for the year the translation was published. 237

Overall, the results are similar to the results of the cluster analysis of the Swedish corpus. 238
However, the genre differences seem to be slightly bigger in the translated works. Unlike 239
the results in the Swedish corpus, all the non-Wallander crime novels end up in one 240
cluster together. Two novels stand out in particular: the literary novel *Tea bag* from 2001, 241
that appears close to Mankell's later crime novels and *Labyrinthen*, which just like in the 242
Swedish novels clusters with the three early literary novels by Mankell. 243

Another noticeable difference between the Swedish and the Dutch cluster analysis, is 244
that unlike in the Swedish originals, the early literary novels in the translations are more 245
similar to other works by Mankell than to the other Swedish authors, although Lars 246
Kepler's *Eldvittnet* shows up in this cluster again. 247



4. Network analysis of Mankell's oeuvre

248

As pointed out by Eder 2015, there are a couple of problems with cluster analysis pertaining to the distance, linkage and number of features (MFWs) used for analysis. Outcomes can differ depending on the MFWs used and there is no real consensus about what the optimal number of MFWs is. These problems can partially be overcome by using a bootstrap consensus tree, because it repeats measurements for multiple numbers of MFWs, and looks for the most robust groupings across different measurements.

However, Eder 2015, 55–56 notes there is still some arbitrariness involved in the production of a consensus tree, such as how many times the analysis should be repeated, for how many words in total are considered and the underlying algorithm used for linkage. A bigger caveat for the current study, however, is that a consensus tree only looks for the closest ranking text, which means it mainly looks at the strongest similarities. In most cases, this is the authorial fingerprint.

In this paper, the central question is rather why some works within one oeuvre deviate from the majority of works and what other factor beside the author are decisive for clustering of texts. These weaker patterns might better be detected by producing a network analysis as proposed by Eder 2015. In a network analysis, not only the closest text in rank is taken into account, but also the second and third closest neighbours. These links are visualized in a network in which close similarities are shown with thicker lines and weaker links with thinner lines.

I performed a bootstrap consensus tree in Stylo and used the CSV output to create a network analysis in the open-source tool GEPHI Gephi using the ForceAtlas2 algorithm. I ran a Modularity Analysis (resolution 0.6) in GEPHI which detects communities in the network, helping to distinguish closely related topological subgroups of nodes from each other and to make clusters more visible in the network. Finally, I applied eigenvector centrality, to measure the influence of nodes in the network. Ranking the function size of nodes indicates the centrality of a work for the cluster it is in.

The results of the network visualizations are shown in Figures 3 and 4. A short description of the clusters is given in the titles in red for ease of interpretation.

In general, the results shown earlier in the cluster analyses are confirmed by the consensus networks. Works cluster mainly by author and genre, although there is some overlap between crime novels and literary novels. There is a separate cluster for Mankell's Wallander series and the older literary novels are in a separate cluster.

However, there are some remarkable differences between the Swedish consensus network and the translated Dutch one. In the Swedish network the older Mankell novels cluster with two literary novels and a crime novel by Nesser as well as a crime novel by Kepler. In the Dutch network, they are only grouped together with the crime novel by Kepler only. The consensus network of the original Swedish corpus distinguishes six clusters, whereas the consensus network of the translated corpus contains eight. Unlike the cluster analysis, where the texts were clustered more clearly by genre in the translated corpus, the network looks somewhat more messy in the translated corpus with smaller and less clearly defined clusters.

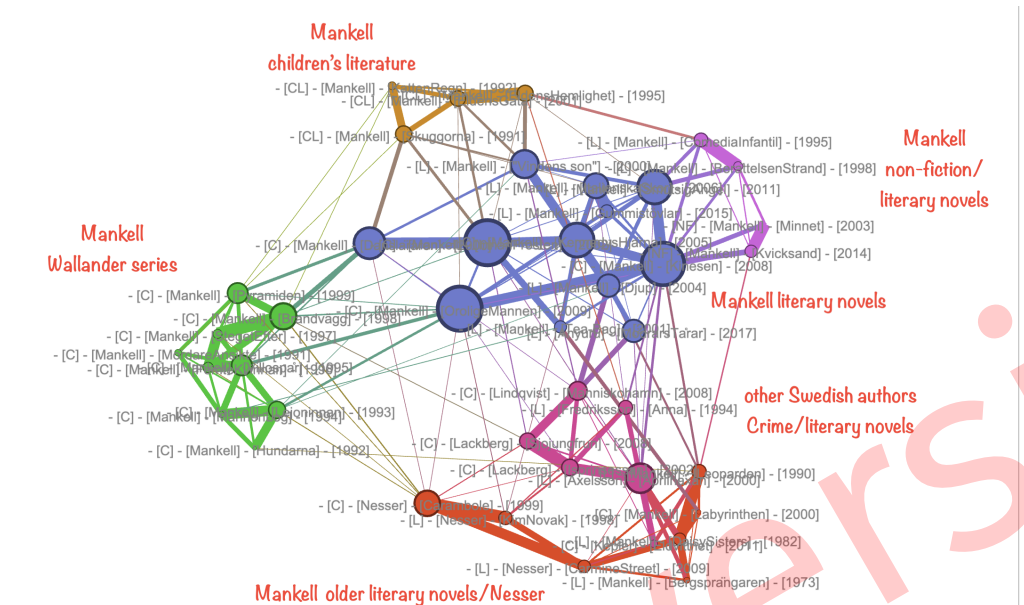


Figure 3: Consensus network of the Swedish corpus: classic Delta distance, 100–1000 MFWs, modularity 0.6

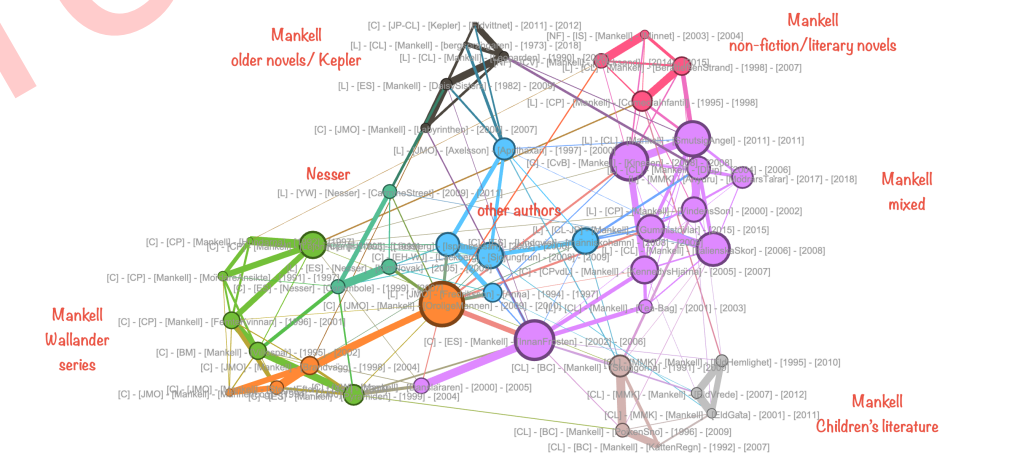


Figure 4: Consensus network of the translated Dutch corpus: classic Delta distance, 100–1000 MFWs, modularity 0.6

Most importantly for the current study, the same four novels that showed up as outliers in the cluster analyses (*Bergsprängaren* (*The Rock Blaster*) from 1973, *Daisy Sisters* from 1982, *Leopardens öga* (*The Eye of the Leopard*) from 1990 *Labyrinthen* (*The Labyrinth*)), again form a separate cluster. In the following section, the MFWs associated with these works are analysed to see why these particular novels stand out from the rest of Mankell's novels.

5. A closer look into the MFWs

To look at more dimensions in the data, a Principal Components Analysis (PCA) was performed on the Swedish corpus. Like a cluster analysis, a PCA also analyzes the MFWs in the dataset, but they are visualized in a scatterplot instead of a dendrogram. In a PCA multiple features are combined in an artificial variable, the so-called principal component that explains the largest proportion of the variance in the data (Jockers 2013, 65–67). On the x-axis, the first principal component is shown. The first principal component is often related to the author (David L. Hoover 2020) The y-axis shows the second principal component. The second principal component is less obvious to interpret, it could be explained by variables like chronology or genre (David L. Hoover 2020). These two principal components are unrelated.

I performed a classic PCA on the Swedish data in Stylo, with the Classic Delta and the correlation option, analyzing the 1000 MFWs. The results of the PCA of the Swedish corpus are presented in Figure 5 below. The x-axis, showing the first principal component, explaining 12.2% of the variance in the data, can clearly be linked to author and is in line with the findings in the cluster analysis. The same four books that were mentioned earlier are deviant from Mankell's other works and more similar to the other Swedish writers in the corpus. The books in question are the crime novel *Labyrinthen* from 2000, the two oldest books in the corpus, namely *Bergsprängaren* (*The Rock Blaster*) from 1973 and *Daisy Sisters* from 1982 and *Leopardens öga* (*The Eye of the Leopard*) from 1990. Unlike in the cluster analysis, we can now see that Lars Kepler's *Eldvittnet* is further away on the x-axis and probably clustered with these books because of the variance in the data that is represented on the y-axis.

Figure 5 shows that author and genre are still the most important factors in distinguishing between texts. However, there are a few books by Mankell that clearly behave differently and that end up closer to books by other authors. What makes these four stand out from the rest of Mankell's works?

If we perform the same PCA again, but with the option 'loadings' in Stylo, showing which words occur significantly more frequently in the texts they are close to in the graph, we might get a first impression about an important difference between the four atypical books and the rest of Mankell's work. In Figure 6, the results of the PCA with the option loadings are shown. This analysis was performed on the 100 MFW, because a figure with 1000 words would become illegible. This also means that the distribution of the novels on the graph is somewhat different. For instance, Kepler's novel *Eldvittnet* is now closer to Mankell's *Labyrinthen*, whereas Nesser's *Maskarna på Carmine street* (2009) appears close to Mankell's older novels. Importantly, Mankell's four diverging novels still stand apart from his other novels. In Figure 6, they are shown a bit below the upper

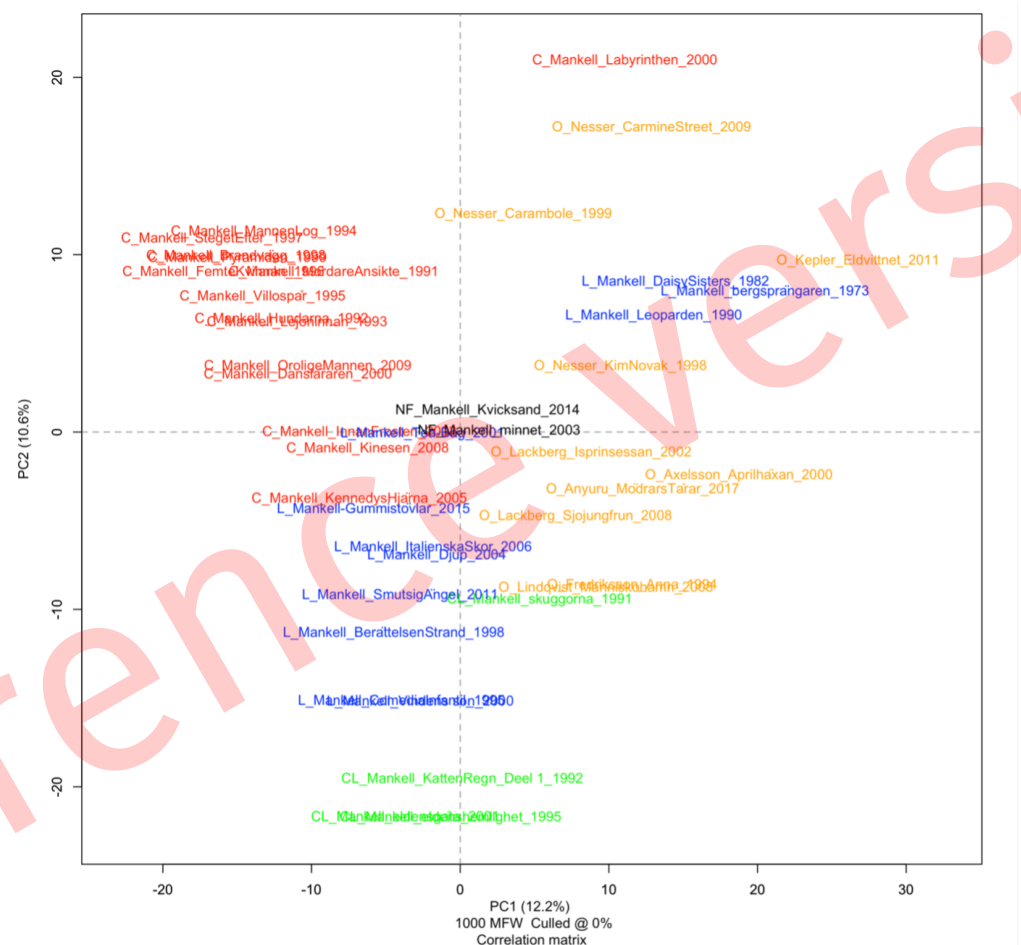


Figure 5: Principal component analysis of the Swedish corpus (1000 MFW, Classic Delta correlation, culling o)

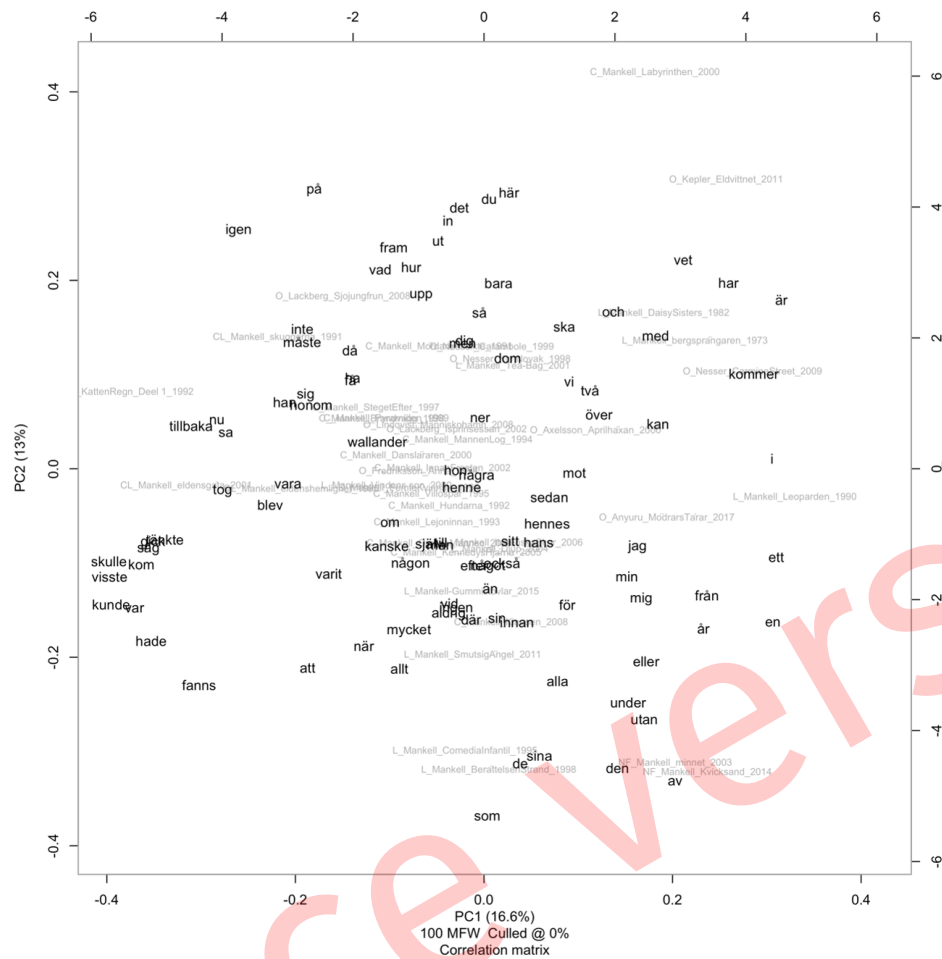


Figure 6: Principal Components Analysis showing the 100 MFW in the Swedish corpus

right corner. The words that are associated with these novels, and less with the other books, are: *kommer* ‘come’, *vet* ‘know’, *har* ‘have’, *är* ‘is/are’, *och* ‘and’ and *med* ‘with’. The first four are verbs in the present tense, whereas the verbs associated with other works are all in past tense or past participles.

This indicates that rather than the chronology, the tense primarily used in the narrative, established by verb tense, might be a decisive factor in why the four mentioned books are different from other Mankell books. On closer inspection, these books as well as *Eldvittnet* by Lars Kepler are primarily written in the present tense, whereas the other works by Mankell are primarily written in the past tense. Of course, this may be related to a chronological development: over time a writer can also change their preference for which tense to narrate a story in.

The same procedure was followed for the translated Dutch corpus. The results are shown in Figure 7 and 8. In Figure 7 the PCA for the translated Dutch corpus is shown, which is in many ways comparable to the results of the Swedish PCA. One remarkable outcome is that Mankell’s Children’s books are quite different on the x-axis, where this was not the case at all in the Swedish results. Another remarkable finding is that some novels by other writers in the corpus, namely Håkan Nesser, Camilla Läckberg and Marianne Fredriksson, end up very close to the literary novels by Mankell and in between books by Mankell in different genres.

Figure 7: PCA of the translated Dutch corpus based on the 1000 MFWs

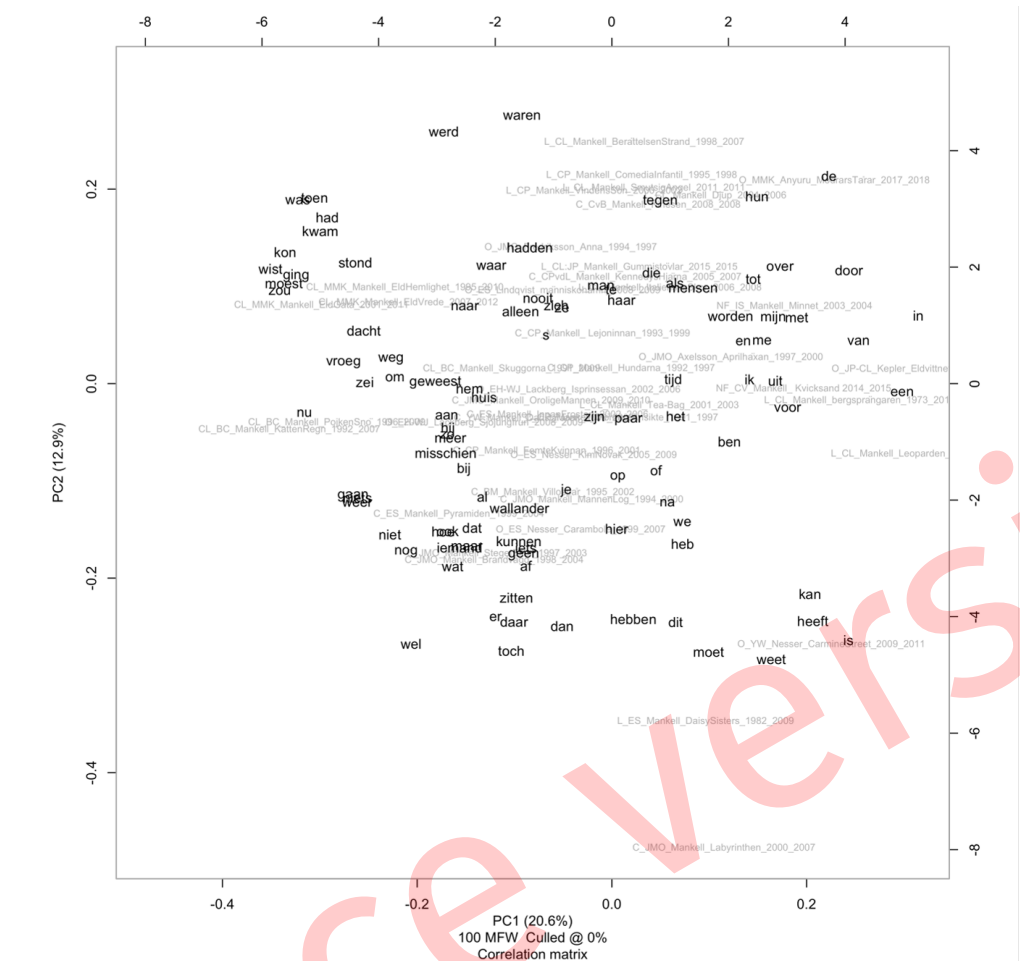


Figure 8: PCA (loadings) of the translated Dutch corpus based on the 100 MFWs

Otherwise, the same four books (*Leopardens öga*, *Bergsprängaren*, *Labyrinthen*, and *Daisy Sisters*) diverge in the translation corpus. The PCA with the loadings function (Figure 8) clearly shows that this is likely caused by the narrative tense again. Words that occur more frequently in these books are: *moet* ‘has to’, *kan* ‘can’, *is* ‘is’, *heeft* ‘have’ and *weet* ‘know’ whereas past tense verbs occur more frequently in other works. An important difference between Swedish and Dutch is that Swedish only has tense marking on verbs whereas Dutch has tense and person marking. This also means that Swedish verbs probably tend to end up higher in the list of MFWs because there are fewer possible forms compared to Dutch where the same verb is spread out over more possible forms.

5.1 Verb tense and perspective

In the remaining part of this section, I will elaborate on the results of the study so far, to get more insight into the stylistic methods used and the studied texts. It is important to look beyond the analyses of Delta distances to see what is behind the measurements and which words are decisive in the clustering of texts.

The results so far, show that verb tense is an important factor for the outcome in stylistic methods based on the MFWs, because there are many verbs (with tense marking) among the MFWs. In similar lines, narrative perspective might also play an important role, because pronouns are very frequent words. In order to get a good indication of

the predominant narrative perspective in the books in the current corpus, I applied Van Rossum's I-index to the data (Van Rossum et al. 2020). Van Rossum et al. (2020) applied both a machine learning and a narratology-based approach in which they computed the ratio of pronouns. Both methods turned out successful in determining the narrative perspective of texts, although the second approach was slightly more robust and yielded a perfect 1.00 score. This perfect score was possible, because Van Rossum et al. (2020) cleaned the data from dialogue. The narrative perspective was already known, so the predictions could be tested for their accuracy. For now, this is not possible in the Mankell corpus, but the ratio of pronouns can still give a good indication of a book's narrative perspective.

Van Rossum's I-index (Van Rossum et al. 2020) is focused on the first person narrative perspective but can be applied to other perspectives as well. I computed the I-index and the he-index, she-index and (singular) you-index (du-index) for both the Swedish originals and the Dutch translations. For the he-index (han-index), for instance, I did this by adding the relative frequency scores of *han* 'he', *honom* 'him' and *hans* 'his' as calculated in Stylo and divided this number by 1 + the relative frequency scores for all the pronouns in the text. The reflexive possessive pronouns *sin*, *sitt* and *sina* were left out of the equation, because they are used to refer to both male and female antecedents. For the she-index (hon-index) I did the same but with *hon* 'she', *hennes* 'her' (object form) and *hennes* 'her' (possessive). Finally for the singular you-index (du-index) I divided the sum of the relative frequencies of *du* 'you' (singular), *dig* and *dej* 'you' (object form in two spelling variants), *din/ditt/dina* 'your' (singular in three inflection forms) by the relative frequencies of all pronouns combined.

Figure 9 shows the results of the indexes in a graph. The ratio of pronouns gives a good indication of the narrative perspective(s) in the texts. Only the results of the Swedish corpus are shown here, because I observed no big differences between the Swedish and the Dutch ratios. Mankell's texts are ordered chronologically from oldest to most recent. The other authors are in random order. The first part of the bars on the bottom left side shows the I-index. In most texts, this index is between 0,10 and 0,30. Clear peaks in the I-index can be detected for the two non-fiction books *Jag dörr, men minnet lever* (2003) and *Kvicksand* (2014), which indeed are mainly written from first person perspective.

Peaks in the I-index can also be observed for *Italienska skor* (2006) and *Svenska gummistövlar* (2015) which are both literary novels with the same main character Fredrik Welin written from an I perspective. Two books by Håkan Nesser: *Maskarna på Carmine street* (2009) and *Kim Novak badade aldrig i Genesarets sjö* (1998) also score high on the I-index. Recall that Nesser's books also appeared close to Mankell's outliers in the PCA.

The second part of the bars in Figure 9 show the han-index (he-index). Most books by Mankell score high on the han-index (he-index) and have indeed a male main character. The third part of the bar show the she-index (hon-index) and it is most interesting to compare these two indexes directly. *Daisy sisters* (1982) is one of the books that score very high on the she-index (hon-index), which makes sense, because it is a novel about three generations of women. The combination of a deviant verb tense (present tense) and a female perspective could very well explain why this particular book appears to be an outlier in the cluster analyses and the PCAs. Two of the children's books (*Eldens gåta* and *Eldens hemlighet*) also score relatively high on the she-index. Some books have a

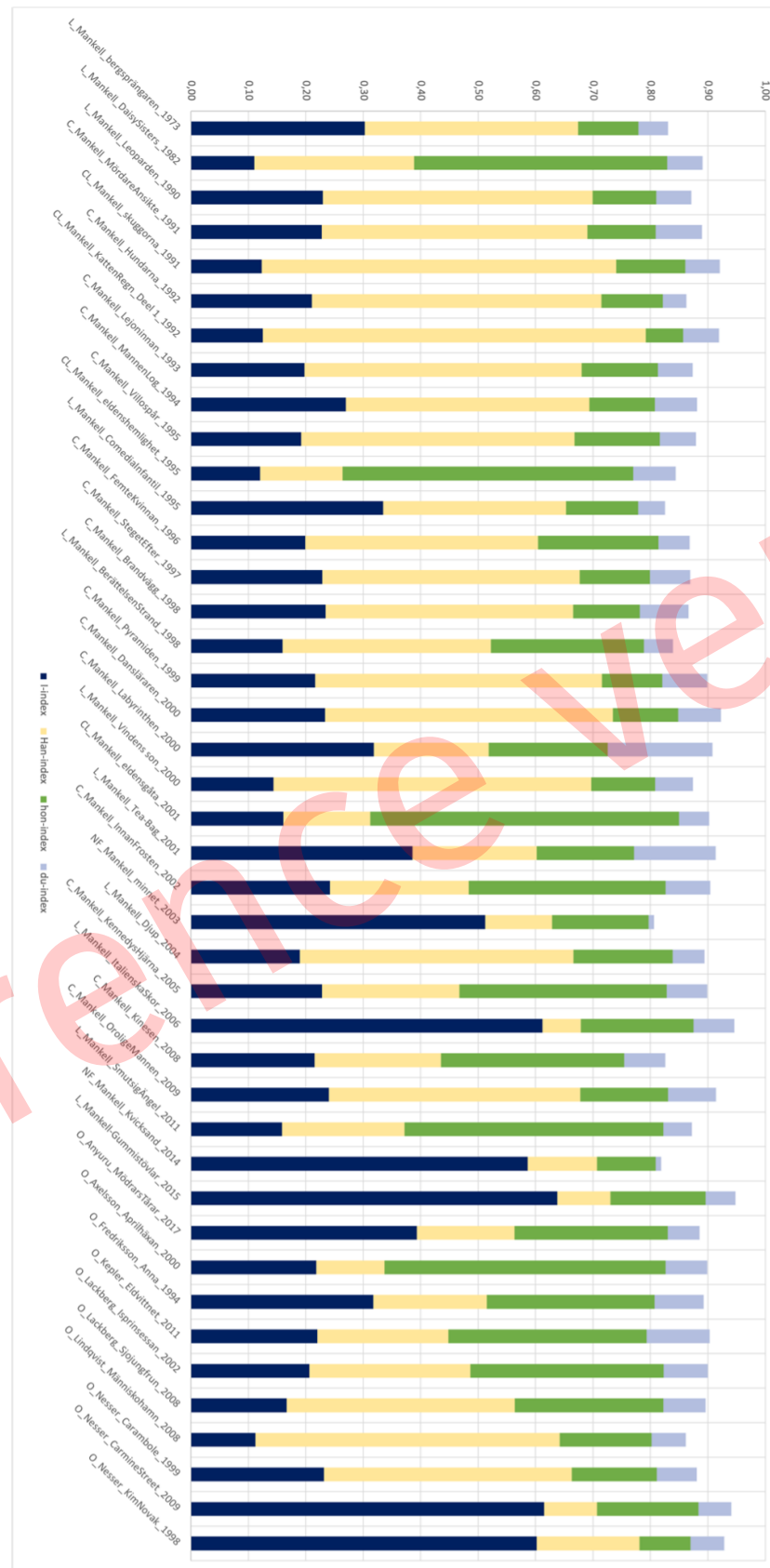


Figure 9: Indication of narrative perspective in the books in the Swedish corpus, measured by I-index (dark blue), Han-index (yellow), Hon-index (green) and du-index (light blue)

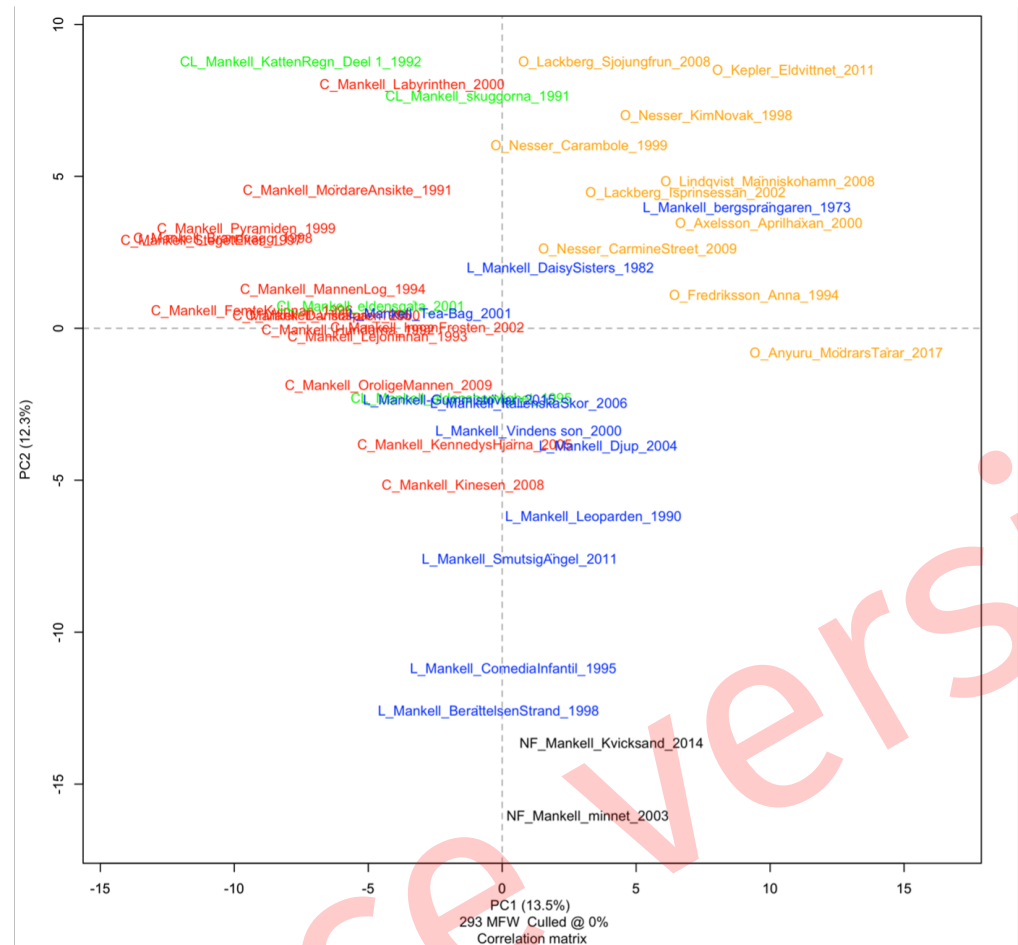


Figure 10: PCA (classic) of the Swedish corpus excluding tense-marked verbs and personal pronouns based on the 1000 MFWs

more evenly divided ratio between pronouns. This is especially the case in *Labyrinthen* (2000) which also was one of the clear outliers in the PCA and cluster analysis together with the earliest Mankell novels. Narrative perspective thus seems to be an important explanatory factor. The analysis of narrative perspective and narration tense leads to useful new observations about what can influence MFW scores for Swedish and Dutch and shows how a novel like *Daisy sisters* differs from other novels by Henning Mankell.

To get more insight into how much of the outcome was influenced by narration tense and narrative perspective, we should only look at the words that are not clearly linked to verb tense and narrative perspective. Stylo has the option to analyze the corpus using an 'existing word list' which enables the researcher to look at specific sets of words. I excluded all verbs marked for tense and all personal pronouns the influence of verb tense was left out of the analysis to better determine how big their influence is on the analyses. I then ran another PCA with the 'loadings' function. The resulting PCA without personal pronouns and verbs indicating tense is shown in Figure 10. This figure clearly shows that now three of the four deviant books are much closer to Mankell's other works, at least on the x-axis, and they no longer form a separate cluster. *Leopardens öga* is also closer to other books in the same genre, but especially *Bergsprängaren* and *Labyrinthen*, and to a lesser extent also *Daisy Sisters*, are still more distant from other works by Mankell.

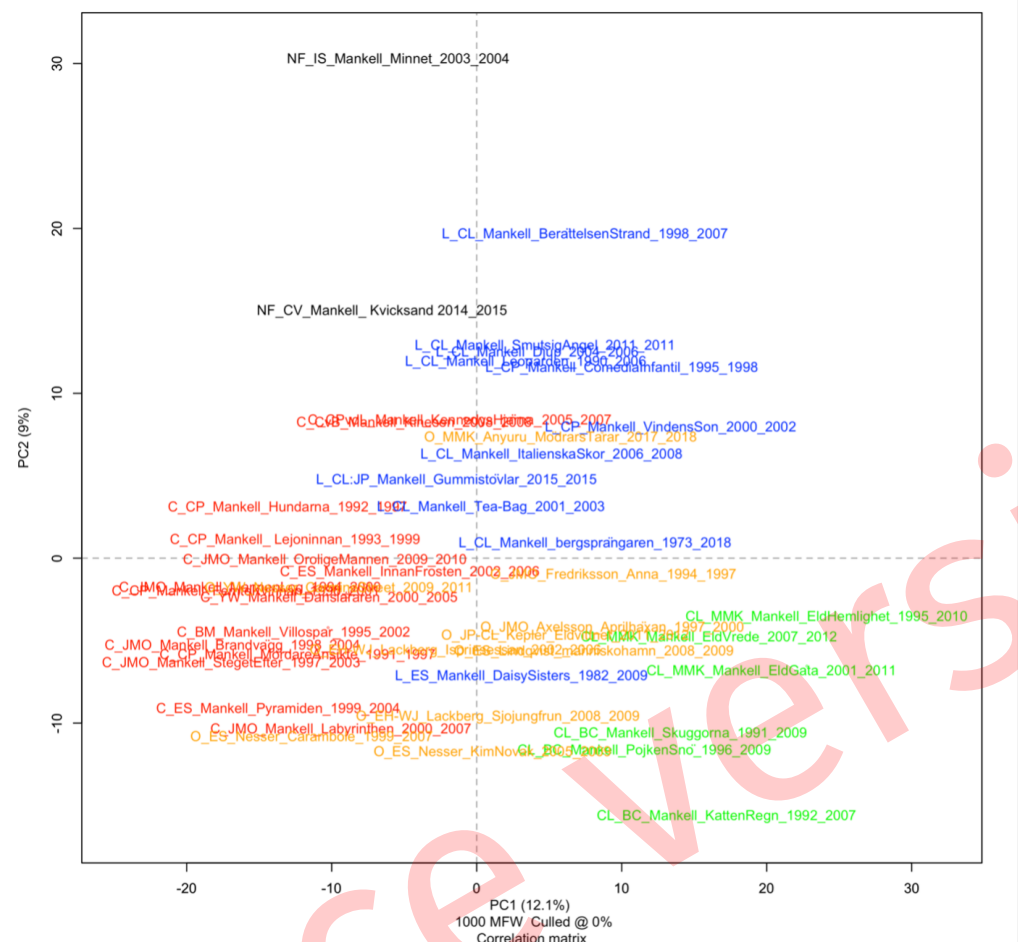


Figure 11: PCA (classic) of the translated Dutch corpus excluding tense-marked verbs and personal pronouns based on the 1000 MFWs

Figure 11 shows the Dutch PCA excluding tense-marked verbs and personal pronouns. In this graph it becomes clear that Daisy sisters (together with Labyrinten) diverges more from other Mankell novels on the Y-axis than Bergsprängaren. So, the translations and the original Swedish texts are different in this perspective. The word frequency patterns of the translations of the other Swedish authors are also much harder to distinguish from the frequency patterns in Mankell's books compared to the results of the analysis of the Swedish texts. This implies that some aspects of style get lost in translation.

5.2 The influence of register

Due to space limitations, I will exclusively focus on one of the earliest Mankell novels *Daisy sisters* in this section. We have seen that this novel is deviant in style, partly because of the use of the present tense and because it is one of the relatively few books by Mankell written from a female third person perspective. A third reason for why *Daisy sisters* has deviant word frequency patterns compared to Mankell novels that were published later, can be detected if we look at Zeta scores.

Zeta was initially introduced by Burrows (2007), and later on improved by Hugh Craig (Craig and Kinney 2009). Burrows's Delta, the method used in this study so far and

the method most often used in stylometry, relies on high frequency words (MFWs). Burrows's Zeta and Craig's Zeta, on the other hand, analyze the middle frequency words and measure distinctiveness or keyness of keywords in a corpus relative to a reference corpus (David L Hoover 2010). Middle frequency words are usually more meaningful than high frequency words, because high frequency words generally are function words (Rybicki 2016, 751). In a Zeta analysis, the texts to be analyzed are first divided into equal segments, then the dispersion of each word in the two separate corpora is registered, by counting how many segments it occurs in at least once (Craig and Kinney 2009).

Stylo can generate wordlists containing the most distinctive keywords in two opposing texts or corpora (Eder et al. 2016). I compiled a primary corpus, consisting of *Daisy Sisters* and a secondary, reference corpus containing all other books in the initial corpus. I did this for both the Swedish corpus and the Dutch translation corpus. I then performed the command `oppose()` in Stylo to analyze *Daisy Sisters* and the reference corpus using Craig's Zeta. Zeta is the sum of the proportions of sections from *Daisy sisters* in which each word occurs and the sections of other works in the corpus in which it does not (David L Hoover 2010). This can point out stylistically interesting characteristics of a text or a corpus. I used samples of 3000 words.

This generates two word frequency lists: one list with words that are preferred (or relatively more frequent compared to the other books in the initial corpus) in *Daisy Sisters* and a list with words that are avoided (or relatively less frequent compared to the other books in the initial corpus). From these lists I selected the twenty most distinctive words, excluding names and verbs, because I was interested in whether there were other stylistic differences besides tense and narrative perspective. The results for the Swedish data are shown in Table 1.

preferred		avoided	
mej	'me'	mig	'me'
dej	'you' (object)	dig	'you' (object)
jo	'yes' (after negation)	genast	'immediately'
fan	'damn'	polis	'police'
herregud	'lord'	mina	'my' (plural)
sej	(reflexive pronoun)	oss	'us'
vadå	'what'	min	'my' (singular)
säjer	'say'	skäl	'reason'
lust	'desire'	polishuset	'police station'
visst	'certainly'	våra	'our' (plural)
jävla	'fucking'	samtalet	'the conversation'
ju	'of course'	telefonen	'the phone'
världen	'the world'	papper	'paper'
omedelbart	'immediately'	nånting	'something'
värre	'worse'	mannen	'the man'
lov	'holidays'/'permission'	sedan	'then'
tåget	'the train'	frågor	'questions'
knappt	'hardly'	din	'your'
morsan	'mom'	rummet	'the room'
full	'drunk'	bland	'among'

Table 1: 20 most distinctive keywords based on Craig's Zeta in *Daisy sisters* compared to other books in the Swedish corpus, excluding verbs and names

Firstly, the results of the Zeta analysis show some clear genre differences. Crime-related words, such as *polis* 'police', *polishuset* 'police station', and possibly words like *skäl* 'reason(s)', *samtale* 'the conversation' and *frågor* 'questions' occur clearly less frequently in *Daisy sisters*. Words related to murder were also on the list. If the books in the reference corpus had only included literary novels, these types of words had probably not been included.

Another obvious difference has to do with spelling conventions and register. *Mej* and *mig* are spelling variants of the same word: 'me'. The variant *mej*, occurring relatively more frequent, in *Daisy sisters*, is the less formal variant which is closer to speech, whereas *mig* is the official variant. The same is true for *dig* and *dej* and *sej* and *sig*. This pattern could also be detected in the spelling of certain verbs, like *säga* 'write', which occurred relatively more frequent in the alternative, informal spelling variant *säja* in *Daisy sisters*. There are other words on the keyness list that confirm the idea that *Daisy sister* is written in a more speech-like, colloquial style. Examples are *jo* 'yes' used after a negation and *ju*, a discourse particle that is especially frequent in spoken language. Similarly, *morsan* is a colloquial form for 'mother' and *vadå* a colloquial form for *vad* 'what'. The keyness list also contains swear words and curse words, which are clearly associated with everyday, informal language, *jävla* 'fucking', *herregud* 'lord' and *fan* 'damn'.

The following example from *Daisy sisters* contains three keywords from the keyness list:

Men vad spelar det för roll att **morsan** är här och säger att hon skäms? Hon kan **ju** inte veta något. Mer än... Ja, **vadå**? Så minns hon allt blod och förstår att det var därför hon måste gå till sjukhuset.

The English translation of this passage is as follows: ³

What does it matter that **mom** is here saying she's ashamed? She can't know anything, **right**? More than.. well **what**? Then she remembers all the blood and realizes that's why she had to go to the hospital.

In both the English translation and the official Dutch translation of this passage the colloquial style is at least partially lost:

Maar wat maakt het uit dat **haar moeder** hier is en zegt dat het een schande is? Ze weet **toch** nergens van. Alleen dat ... Ja, **wat**? Dan herinnert ze zich al het bloed en ze begrijpt dat ze daarom naar het ziekenhuis moest.

Discourse particles in general are very hard to translate, because they can have various meanings depending on context (Aijmer 2008). Here *ju* is translated, but there is no Dutch or English equivalent that is equally frequent and associated with speech as much as the Swedish word. The two other colloquial words in this short passage are translated into standard Dutch, which leads to a loss of this style feature.

A final result from the Zeta analysis is that different synonyms are used in the primary corpus and the reference corpus. In the list of distinctive words, *omedelbart* is preferred in *Daisy sisters* and *genast* is avoided. These words are synonyms and both mean 'immediately' with no difference in register.

3. My translation.

Table 2 shows the list with distinctive words based on Craig's zeta in the Dutch translation corpus. The genre differences are even more obvious in this list compared to the original Swedish list: words like *onderzoek* 'investigation', *politiebureau* and *bureau* 'police station' *vermoord* 'murdered', *waarheid* 'truth' and *lichaam* 'body' are clearly linked to the crime genre. This also confirms the earlier findings in the cluster analyses that genre differences seem to be magnified in the translations.

preferred		avoided	
<i>immers</i>	'after all'	<i>onze</i>	'our'
<i>want</i>	'because'	<i>onderzoek</i>	'investigation'
<i>nou</i>	'well'	<i>ineens</i>	'suddenly'
<i>gewoon</i>	'just'	<i>politiebureau</i>	'police station'
<i>ja</i>	'yes'	<i>dood</i>	'dead'/'death'
<i>opeens</i>	'suddenly'	<i>bureau</i>	'police station'/'desk'
<i>verdomme</i>	'damn'	<i>politieman</i>	'police man'
<i>minder</i>	'less'	<i>vermoord</i>	'murdered'
<i>aardig</i>	'kind/rather'	<i>zee</i>	'sea'
<i>voorbij</i>	'(all) over'/'past'	<i>zeer</i>	'very'
<i>vieze</i>	'dirty'	<i>water</i>	'water'
<i>niks</i>	'nothing'	<i>telefoon</i>	'phone'
<i>kennelijk</i>	'apparently'	<i>vlak</i>	'right'/'flat'
<i>hemel</i>	'heaven'	<i> bezig</i>	'in process'
<i>baan</i>	'job'	<i>aantal</i>	'number'/'amount'
<i>hoekje</i>	'corner'	<i>waarheid</i>	'truth'
<i>raar</i>	'strange'	<i>papieren</i>	'paper'
<i>geluk</i>	'luck'	<i>vervolgens</i>	'then'
<i>nergens</i>	'nowhere'	<i>lichaam</i>	'body'
<i>zij</i>	'she'/'they'	<i>reden</i>	'reason'

Table 2: 20 most distinctive keywords in *Daisy sisters* based on Craig's Zeta, compared to other books in the Dutch translation corpus, excluding verbs and names

The register difference, on the other hand, is not as obvious as in the Swedish list, although *nou* 'well' *gewoon* 'just' *ja* 'yes' *verdomme* 'damn' do point in the direction of register and speech-like language. *Immers*, which is on top of the list of distinctive words in the Dutch translated corpus, is a good example of translationese. It is the translation of the previously mentioned discourse particle *ju*. In terms of meaning, this translation is accurate, but *immers* does not at all belong to the same register. While *ju* is associated with spoken language, *immers* is almost exclusively used in written language and has a somewhat archaic connotation. Again, this indicates that the speech-like, informal style gets partially lost in the Dutch translation. In the Dutch list with distinctive keywords there are also two synonyms both meaning 'suddenly': *opeens* is preferred in *Daisy sisters* whereas *ineens* is preferred in the other books in the corpus. This can likely be explained by the individual preference of the translator. However, more research about the influence of the translator on style is necessary to confirm this.

6. Conclusion

In this paper, 32 books by the Swedish writer Henning Mankell were investigated using stylometric methods, to find out whether his style changed measurably over time, or if some of his books deviate stylistically from his other works for other reasons. 10 books

by other Swedish authors were added to the corpus as a reference. The study also gives more insight into the methods that are frequently used in stylometry, such as cluster analysis and PCA, that basically are black boxes, because they give little information about the stylistic features that differ between texts. For this purpose, the original Swedish texts were also compared to the Dutch translations of the same 42 texts to determine how translation and language influence the results of stylometric analyses.

Cluster analyses and PCAs of the data showed that works were clustered by author in the first place and secondly by genre, although there were a few exceptions. The division into genre was somewhat stronger in the translated corpus. The analyses also seemed to indicate that the factor time explains part of the variance. However, on closer inspection, verb tense rather than year of publication turned out to be the decisive factor: the most deviant books in the corpus were primarily written in the present tense, whereas most other books were predominantly written in the past tense. Moreover, narrative perspective also influenced the results noticeably. An analysis of the pronoun ratios in the works in the corpora indicated that the majority of the novels in the corpus had a dominant third person male perspective. Books that mainly had a first person perspective tended to cluster together, just like books with a third person female perspective. Leaving out pronouns and verbs marked for tense showed a very different picture.

Finally, an analysis of the data based on Craig's Zeta (Craig and Kinney 2009) showed that words most distinctively used in the original Swedish *Daisy sisters* were often colloquial words with a speech-like connotation. Words that were avoided in the novel were associated with the crime genre. However, the most distinctive words in the Zeta analysis for the translated Dutch corpus, were not as clearly related to register. The genre differences seemed magnified in the Zeta results of the translation corpus compared to the list of Swedish keywords. This confirmed the findings in the cluster analyses and the PCAs that books were more clearly clustered by genre in the translated texts. This can be due to the different language and language specific features or due to inherent characteristics of translated texts in general. More research on different languages and translations would be useful to get a better understanding of this process. In a follow-up study I intend to investigate the style differences between translators and how they can be detected and measured.

This study has shown that Zeta analysis and a closer look at word lists in stylometric studies can give useful insights into the specific style features that make texts different from each other instead of only focusing on the fact that they differ.

7. Data Availability

Data can be found here: data.example.edu/data

8. Software Availability

Software can be found here: github.com/something

9. Acknowledgements 579

I would like to thank Karina van Dalen-Oskam for our fruitful conversations and for her valuable feedback on earlier versions of this article. 580
581

10. Author Contributions 582

Martje Wijers: Conceptualization, Writing – original draft, Formal analysis, Investigation 583
584


References 585

- Aijmer, Karin (2008). "Translating discourse particles: A case of complex translation". 586
In: *Incorporating Corpora. The Linguist and the Translator*, 95–116. 587
- Arvas, Paula and Andrew Nestingen (2011). *Scandinavian Crime Fiction*. University of 588
Wales Press. 589
- Berglund, Karl (2013). *Deckarboomen under lupp: Statistiska perspektiv på svensk kriminallit-* 590
teratur 1977–2010. 10.1353/scd.2013.0008. 591
- Burrows, John F. (2002). "'Delta': a measure of stylistic difference and a guide to likely 592
authorship". In: *Literary and linguistic computing* 17 (3), 267–287. 593
- (2007). "All the way through: testing for authorship in different frequency strata". 594
In: *Literary and Linguistic Computing* 22.1, 27–47. 595
- Can, Fazli and Jon M Patton (2004). "Change of writing style with time". In: *Computers* 596
and the Humanities 38, 61–82. 597
- Craig, Hugh and Arthur F Kinney (2009). *Shakespeare, computers, and the mystery of* 598
authorship. Cambridge University Press. 599
- Dalen-Oskam, Karina van (2021). *Het raadsel literatuur: Is literaire kwaliteit meetbaar?* 600
Amsterdam University Press. 601
- Eder, Maciej (Dec. 2015). "Visualization in stylometry: Cluster analysis using networks". 602
In: *Digital Scholarship in the Humanities* 32.1, 50–64. ISSN: 2055-7671. 10.1093/llc/fqv 603
061. eprint: [https://academic.oup.com/dsh/article-pdf/32/1/50/11046630/fqv](https://academic.oup.com/dsh/article-pdf/32/1/50/11046630/fqv061.pdf) 604
061.pdf. <https://doi.org/10.1093/llc/fqv061>. 605
- Eder, Maciej, Jan Rybicki, and Mike Kestemont (2016). "Stylometry with R: A Package 606
for Computational Text Analysis". In: *The R journal* 8 (1). [https://github.com/com](https://github.com/computationalstylistics/stylo) 607
[putationalstylistics/stylo](https://github.com/computationalstylistics/stylo). 608
- Herrmann, J. Berenike, Karina van Dalen-Oskam, and Christof Schöch (Mar. 2015). 609
"Revisiting Style, a Key Concept in Literary Studies". In: *Journal of Literary Theory* 9 610
(1). ISSN: 1862-5290. 10.1515/jlt-2015-0003. 611
- Hoover, David L (2010). "Teasing Out Authorship and Style with T-tests and Zeta." In: 612
DH, 168–170. 613
- (2020). *Modes of Composition and the Durability of Style in Literature*. Routledge. 614
- Jacobsen, Kirsten (2012). *Mankell om Mankell*. Leopard Förlag. 615
- Jautze, Kim (2014). "Measuring the style of chick lit and literature". In: *DH*. 616
- Jautze, Kim, Corina Koolen, Andreas van Cranenburgh, and Hayco de Jong (2013). 617
From high heels to weed attics: a syntactic investigation of chick lit and literature, 72–81. 618
<http://literaryquality.huygens.knaw.nl>. 619

- Jockers, Matthew L (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press. 620
621
- Ríos-Toledo, Germán, Juan Pablo Francisco Posadas-Durán, Grigori Sidorov, and Noé Alejandro Castro-Sánchez (2022). "Detection of changes in literary writing style using N-grams as style markers and supervised machine learning". In: *Plos one* 17.7, 622
623
624
625
626
- Rybicki, Jan (2016). "Vive la différence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies". In: *Digital Scholarship in the Humanities* 31.4, 746–761. 627
628
- Squires, Claire (2007). *Marketing Literature: The Making of Contemporary Writing in Britain*. Palgrave Macmillan. 629
630
- Van Rossum, Lisanne, Joris J. van Zundert, and K.H. van Dalen-Oskam (2020). "I Catching: Computationally Operationalising Narrative Perspective for Stylometric Analysis". In: *DH Benelux*. 631
632
633

Translation-based connotation visualization for classical poetic Japanese vocabulary of the *Kokin Wakashū* ca. 905

Xudong Chen¹ 
Yamamoto Hilofumi¹ 
Hodošček Bor² 

1. School of Environment and Society, Tokyo Institute of Technology , Tokyo, Japan.
2. Graduate School of Humanities, Osaka University , Osaka, Japan.

Citation

Xudong Chen, Yamamoto Hilofumi, and Hodošček Bor (2023). "Translation-based connotation visualization for classical poetic Japanese vocabulary of the *Kokin Wakashū* ca. 905". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-21

Date received 2023-01-31

Keywords

classical poetic Japanese text, parallel corpora, connotation, operationalization

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. To offer an intuitive visualization of connotation in the classical poetic Japanese vocabulary of the *Kokin Wakashū* as a independent supplement for poetic language dictionaries, this paper presents an operationalization to tackle connotations using non-literal elements which is unveiled during the cross-cultural communication process, i.e., the translation process. Grounded on Schramm's communication model, we suggest calculating the set difference between the *Kokin Wakashū* and its ten translation versions to capture the lexical explanatory additions (non-literal elements) in the translations. Methodologically, we apply the set difference in two distinct ways and implement the visualization on the six most frequent poetic flora words in the *Kokin Wakashū*, resulting in various depictions of non-literal elements. The set difference-based approaches to non-literal element visualization have proven to reflect associative images and rhetorical techniques related to queried poetic words, which are two crucial aspects of connotation. While the other aspects of connotation, such as encyclopedic knowledge, sociolinguistic style, and emotion, are not covered by the proposed visualizations.

1. Introduction

Classical Japanese poetry is a representative type of poetry in classical Japanese literature, with a thematic tendency toward natural elements and a strict length limitation. Ki no Tsurayuki, the principal compiler of the first emperor-ordered anthology, the *Kokin Wakashū* (henceforth *Kokinshū* for short), noted in the *Kana Preface*:

The seeds of Japanese poetry lie in the human heart and grow into leaves of ten thousand words [...] all that they think and feel is given expression in description of things they see and hear [...] In the age of the awesome gods, songs did not have a fixed number of syllables [...] By the time of the age of humans, [...] poems of thirty-one syllables were composed. Since then many poems have been composed when people were attracted by the blossoms or admired the birds, when they were moved by the haze or regretted the swift passage of the dew, and both inspiration and forms of expression have become diverse. (translation by Rodd et al. 1996, pp.35–36)

As shown in the *Kana Preface*, classical Japanese poets used nature elements to hint at inner thoughts rather than expressed their thoughts directly. Moreover, the standard form, *tanka* 短歌 (en. short poem), is limited to 31 syllables (see the example below).

- (1) つれづれの/ながめにまさる/なみだがわ/そでのみぬれて/あふよしもなし
tsurezureno **nagameni masaru namidagawa** sodenomi nurete au yoshimo nashi
[Kokinshū 617]
'in idle reverie/I weep tears that overflow like/the long rains of spring/my sleeves
are drenched with the stream/that flows when we cannot meet'
[translated by Rodd et al. (1996)]

Poem (1) employs the ensuing rhetorical stratagems, thereby augmenting the polysemy and conveying rich information within the 31-syllable constraint:

- *kakekotoba* 掛詞, pivot words permitting manifold interpretations;
- *engo* 縁語, affiliated words fabricating a parallel imagery.

In poem (1), *kakekotoba* and *engo* are in bold. "Nagame" is *kakekotoba* having parallel meanings in its string (長雨, en. long rain, or 眺め, en. reverie); "namidagawa" (en. river of tears) and "masaru" (en. rise/overflow) is *engo* for "nagame," since they are conceptually related to the sense of water. When "nagame" is interpreted as "long rain," "masaru" means the "rising" water in the river because of the rain. When "nagame" is interpreted as "reverie," "masaru" means the "overflowing" tears because of "reverie."

Such words are known as typical words in *uta-kotoba* 歌ことば, the vocabulary used in classical Japanese poetry (for the definition, cf., Kubota 1994, p.89). The poetic vocabulary often conveys indirect and non-literal information as the above example. This indirect and non-literal information is connotation in the poetic vocabulary. From this perspective, classical Japanese poetry is a treasure trove of connotation.

The current study aims to visualize such lexical connotation (i.e., non-literal information in the poetic vocabulary) in the *Kokinshū*, using its ten versions of contemporary Japanese translation. The connotation visualization is intended to assist in supplementing classical poetic Japanese dictionaries in a user-interactive way. Technically, we adopted a connotation visualization strategy based on comparison of textual information between original poems and translations, alluded to by the communication model (Schramm 1954). That is, the visualizations reveal connotation as the non-literal part of the original poems that bears explicit explanatory addition in the translations. As a first attempt, we experiment with the most frequent poetic flora words, as they are representative natural elements emphasized in the *Kana Preface*.

We structure the study as follows. In section 2, we introduce the background regarding the concept and the operationalization of connotation. Then we provide a brief view of the classical poetic Japanese dictionary (Katagiri 1983) and the translation of Japanese poetry. Finally, we present the theoretical basis of our method (Schramm's (1954) communication model). In section 3, we introduce the materials, and two implementations of connotation visualization. In section 4, we provide six example visualizations for flora poetic words. Besides, we also compare the two implementations and examine

whether the visualization can reproduce the connotation included in Katagiri (1983), and whether it can reveal some aspects of connotation that do not exist in the dictionary. In section 5, we discuss what aspects of connotation the proposed visualization can present and what the visualization cannot. Also, we will discuss the contributions and the limitations of the current work.

2. Motivations

In section 2, we begin with the inconsistent definitions of connotation and the challenges in its operationalization, clarifying why we use non-literal elements in explanatory materials as the basis for connotation visualization. Next, we assert why we need to use additional information in translations rather than other materials to supplement the connotation description in the dictionary. Finally, we discuss how to address non-literal elements in the comparison between the original poems and the translations from the perspective of the communication model by Schramm (1954).

2.1 Aspects of connotation

In a basic definition, denotation is a word's explicit meaning, whereas connotation encompasses sociocultural and individual affiliations (cf., Chandler 2002, pp.173–174). Nevertheless, the definition and scope of connotation vary among scholars, pertaining to sociolinguistic aspects (gender, style, social class, region, etc.) (e.g., Bloomfield 1933; Hjelmslev 1969), emotional aspects (positive, negative, or neutral emotions) (e.g., Eco 1976; Osgood et al. 1957), and associative aspects (collocates, associations, and figurative language) (e.g., Rössler 1979). According to Eco's (1976) theory, rhetorical techniques, ideology, hyponyms, hypernyms, and antonyms also pertain to the realm of connotation. Moreover, in communication, pragmatic aspects of meaning, such as a speaker's attitude and a listener's value judgment of the received speech are also part of the connotation (Mounin 1976, pp. 159–160). Such connotation can be independent between the speaker and the listener. From denotation without any intended connotation by the speaker, the listener can occasionally perceive unintended/potential connotation (Rössler 1979, p. 101). Moreover, connotation is inseparable from denotation (Stede 1999, p.91; Stubbs 2002, p.198; Voloshinov 1986, p.105) and connotative meanings are inexhaustible (cf., Chandler 2002, p.139). Therefore, operationalizing the connotation is challenging.

On the other hand, almost all aspects of connotation are primarily non-literal, except for collocates or typical phraseology as connotation. The non-literal property is a vital aspect we can use, as such non-literal elements can become literal when explicated for better understanding in cross-cultural communication. Denotation and connotation exist at the conceptual level; they are inseparable and thus nonoperational without access to the speech community to conduct psychological experiments or surveys. Conversely, literal and non-literal elements are physically written/unwritten; they are cleanly separable and thus operational, when explanatory material is available. Hence, we may view non-literal elements as a physical world projection of the concept of connotation, aiding operationalization (table 1). Although the two are not equivalent, non-literal elements can intuitively reflect certain aspects of connotation. Dictionaries, introductory books, philological annotations, and translations can serve as explanatory materials.

	explicit	implicit
physical (operational)	literal (projection of denotation)	non-literal (projection of connotation)
notional (nonoperational)	↑ denotation	↑ connotation

Table 1: Relationship between denotation/connotation and literal/non-literal: dashed lines signify the instability of the division; solid lines signify the stable division.

The operationalization of lexical connotation has been attempted within computational linguistics, and corpus linguistics. These mainly focus on certain aspects such as emotional value, sentiment, valence, impact, and collocates (e.g., Allaway and McKeown 2021; Rashkin et al. 2016; Stubbs 2002), while the intuitive property of connotation, non-literal, is not widely utilized.

2.2 Dictionary and translation as explanatory materials

As we noted above, many materials can serve to turn non-literal elements into literal ones during cross-cultural communication. Among these materials, only dictionaries and translations can provide a relatively systematic approach to classical Japanese poetic vocabulary, as the others typically select limited poems or words to offer specific explanations. Section 2.2 discusses a currently available classical poetic Japanese dictionary of Katagiri (1983) and explains why we need to use additional information in translations to reflect connotation as an independent supplement for the dictionary.

Katagiri (1983) includes 830 lexical entries. The dictionary does not explicitly distinguish between connotation and denotation, yet covers a broad spectrum of meaning, encompassing not only the denotative aspects but also the connotative ones. The dictionary's connotation pertains to collocates, phraseological patterns, and figurative/rhetorical usages. For instance, in its description for “sakura-bana” (en. cherry blossoms), it incorporates the following elements (pp. 172–173): collocates such as “chiru” (en. falling) and associations such as the impermanence of life and the passing of spring.

However, dictionary compilers inevitably apply their conscious knowledge in the dictionary compilation. For example, compilers decide which words to include, which examples and meanings to emphasize. However, when actually translating Japanese poetry, relying solely on the dictionary proves insufficient for producing comprehensive translations. As Masao Takeoka, one of the translators of the *Kokinshū*, noted:

Translation is not solely an introduction or explanation of the original text's “plot.” It is, after comprehending the author's way of perception and feeling from all aspects, entirely transforming the perceptions and feelings into the same or as similar as possible expressions that exist in contemporary language. (Takeoka 1976a, p. 11, translation is by the authors)

To provide a comprehensive understanding of the “perceptions and feelings” of poets in translations, beyond consciously aligning corresponding equivalent words, translators must simultaneously incorporate those perceptions and feelings about classical poetry into their translations. These incorporations are rarely covered in the dictionary, as they remain unconscious for compilers until they are actually translated. The following

example of a parallel text (translation is by (Katagiri 1983))¹ illustrate how additional information in translations serves a role to reveal hidden perceptions in the poem. 133 134

(2) a. つれづれの/ながめにまさる/なみだがわ/そでのみぬれて/あふよしもなし 135
[Kokinshū 617] 136

tsurezureno nagameni masaru namidagawa sodenomi 137
tsurezure=no nagame=ni masar-ru namida+kawa sode=nomi
reverie=GEN long rain=DAT rise-NPST tear+river sleeves=ONLY
nurete au yoshimo nashi
nur-e-te aw-ru yoshi=mo na-shi
wet-THM-SEQ meet-NPST method=ADD negative-NPST

‘the long rains of reverie drives the river of tears to surge. [I can] only dampen my sleeves and have no means to meet [you]’

[literal English translation is by the authors]

b. 長雨のみならず私の物思いの眺めによって水かさが増している涙の川、渡ろう 138
と思っても、涙で袖が濡れるばかりで渡ることができず、逢いに行く手段とて 139
ありません。 [Contemporary translation by Katagiri (1983, p. 318)] 140

nagamenominarazu watashino monoomino nagameniyotte 141
nagame=nominarazu watashi=no monoomoi=no nagame-niyotte
long rain=ONLY.COP.NEG 1.SG=GEN reverie=GEN gaze-DAT.REASON
mizukasaga mashiteiru namidano kawa, wataruto
mizukasa=ga mashi-teir-ru namida=no kawa, watar-u=to
water=NOM rise-PROG-NPST tear=GEN river, cross-INT=QT
omottemo, namidade sodega nurerubakaride
omow-temo, namida=de sode=ga nure-ru=bakari=de
think-ADVS, tear=INST sleeves=NOM wet-NPST=ONLY=INST
watarukotoga dekizu, aini iku shudantote
watar-ru=koto=ga deki-zu, aw-i=ni ik-ru shudan=tote
CROSS-NPST=FN=NOM able-NEG, meet-THM=PURP go-NPST method=ADVS
arimasen.
ar-i-masen
exist-THM-POL.NEG

‘Due to not only the long rains but also my gaze of reverie, the river of tears surges. [I] endeavor to cross the river, but can only dampen my sleeves with tears, unable to cross. There is no means to go to meet [you].’

[literal English translation is by the authors]

1. Glossing follows the Leipzig Glossing Rules (<https://www.eva.mpg.de/lingua/resources/glossing-rules.php>) and the gloss list for dialects published by Michinori Shimoji’s Laboratory, Kyushu University (https://docs.google.com/spreadsheets/d/14wKM61WalZ34-Dcj3Q_vFUmrwqu5Do7VeQu-It8wZVU/edit#gid=267816193, accessed on 20 May, 2023). In each example, the first line is the romanized version of Japanese; the second line is the morph-by-morph segmentation; the third line is the morph-by-morph gloss. Abbreviations for used glosses are as follows: 1: first person / -: affix boundary / =: clitic boundary / +: compound boundary / ADD: additive / ADVS: adversative / COP: copula / DAT: dative / DVLZ: deverbalizer / EMP: emphasis / FN: formal noun / GEN: genitive / INST: instrumental / INT: intentional / NEG: negative / NOM: nominative / NPST: non-past / ONLY: only / POL: polite / PROG: progressive / PURP: purposive / QT: quote / REASON: reason / SEQ: sequential / SG: single / THM: thematic vowel.

In this parallel text, the translation has rendered the two parallel images of long rain and reverie of the kakekotoba “nagame” by adding corresponding words “monoomoi no nagame” (en. gaze with reverie). The dictionary (Katagiri 1983, p. 304) also mentions the connotation of kakekotoba “nagame,” whereas in the translation for the part regarding “namidagawa” (en. river of tears), the translator used additional words to highlight the aspiration to “cross the river.” Understanding the dilemma of longing to cross the river to reach a loved one, yet unable to, is challenging through the dictionary alone.

We therefore pay attention to the explanatory addition in translations. Our goal, to provide an independent supplement to the dictionary, necessitates the visualization of the unconscious knowledge exposed by translators. Through a user-interactive visualization system, Japanese poetry readers can query all lexical entries in the *Kokinshū*, not just the consciously crucial lexical entries and usages pre-selected by compilers.

2.3 Non-literal elements in Schramm’s communication model

Based on the above two sections, the operationalization of the concept connotation will be based on the operationalization of non-literal elements revealed in the translations. This section introduces a specific way of addressing non-literal elements in parallel texts with hints from Schramm’s (1954) communication model.

Schramm’s model features five parts: source, encoder, signal, decoder, and destination (fig. 1). The source encodes a message which the destination decodes. In electronic systems, microphones encode and earphones decode. In human communication, speakers encode thoughts into language (the signal), and listeners decode it. (cf., Schramm 1954, p.6).

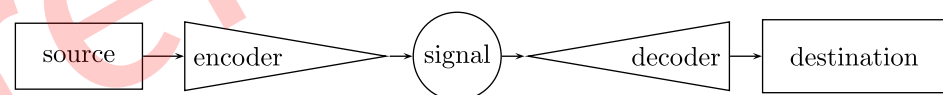
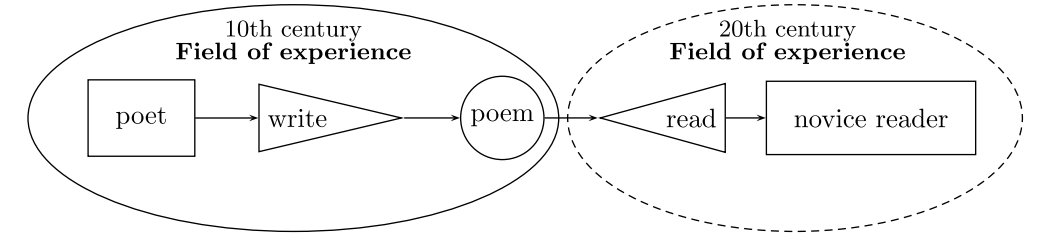


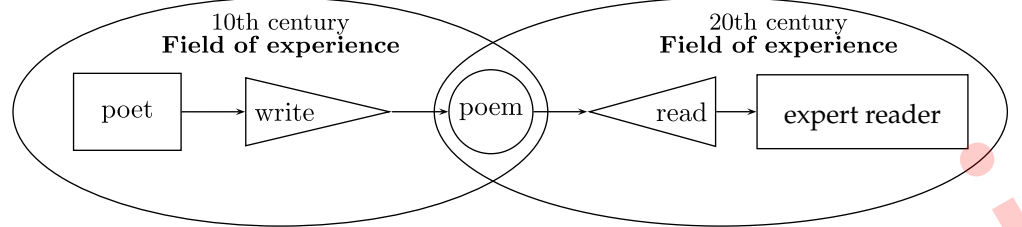
Figure 1: Process of communication based on (Schramm 1954, p.4): a communicative process consists of source, encoder, signal, decoder, and destination.

Yamamoto (2005) proposes that the translation and reading behaviors of classical Japanese poetry be considered communication processes. In the process of reading, the poet is the encoder who transfers thoughts into a poem; the reader (novice/expert reader) is the decoder who processes the information in the poem into their understanding; the poem is a signal in the process (fig. 2). The translation process follows the reading process: after processing the information of the poem into understanding as a decoder, the expert turns into an encoder who transfers the understanding of the poem into contemporary translations.

The difference in the reading process between a novice reader and an expert reader lies in the “field of experience” (Schramm 1954, p.6), which refers to accumulated experience. When a novice reader attempts to understand a classical Japanese poem as shown in fig. 2a, he/she may fail to understand the poem as the novice reader in the twentieth century shares no field of experience with a poet who lived in the tenth



(a) Reading-by-novice as a communicative process



(b) Reading-by-expert as a communicative process

Figure 2: Translation and reading as two kinds of communicative processes based on Yamamoto (2005, p.27): (a) indicates that reading classical Japanese poetry by a novice reader is a communication process where the encoder and the decoder share no common experience; (b) indicates that the reading classical Japanese poetry by an expert is a communication process where the encoder and the decoder share part of the accumulated experience.

century. Conversely, an expert in classical Japanese poetry, having accumulated a wealth 178
of relevant knowledge from numerous ancient books and documents (fig. 2a), shares 179
a broader field of experience. Hence, the expert can decode classical Japanese poetry 180
better than a novice reader. 181

Sharing a similar field of experience means that most of the information can be under- 182
stood non-literally; on the other hand, sharing a different field of experience means 183
most information must be explained literally to be understood. Between the two read- 184
ing processes by experts and novice readers, the process of translation by experts acts 185
as a bridge to communicate the two fields of experience (fig. 3). In the translation 186
process, to share the information in the original poem smoothly with novice readers, 187
experts translate literal elements in the original poem and add non-literal elements into 188
translations. 189

On this basis, we can formalize the non-literal elements (formula 1), where a Poem is a 190
set consisting of literal elements and a Translation is a set consisting of elements in the 191
Poem set and non-literal elements. Such formalization turns our extraction of non-literal 192
elements into a well-defined technical problem of calculating the complement set. 193

$$\text{Non-literal elements} = \text{Translation} \setminus \text{Poem} \quad (1)$$

2.4 Summary and potential issues 194

In section 2, we clarified that through non-literal elements revealed in translations, we 195
can visualize a portion of lexical connotation. Moreover, non-literal elements revealed 196
in translations can provide a heuristic for the unconscious aspect of connotation that 197
is not covered by dictionaries. We can extract the non-literal elements by computing 198
the set difference between translations and original poems, guided by the hint from 199

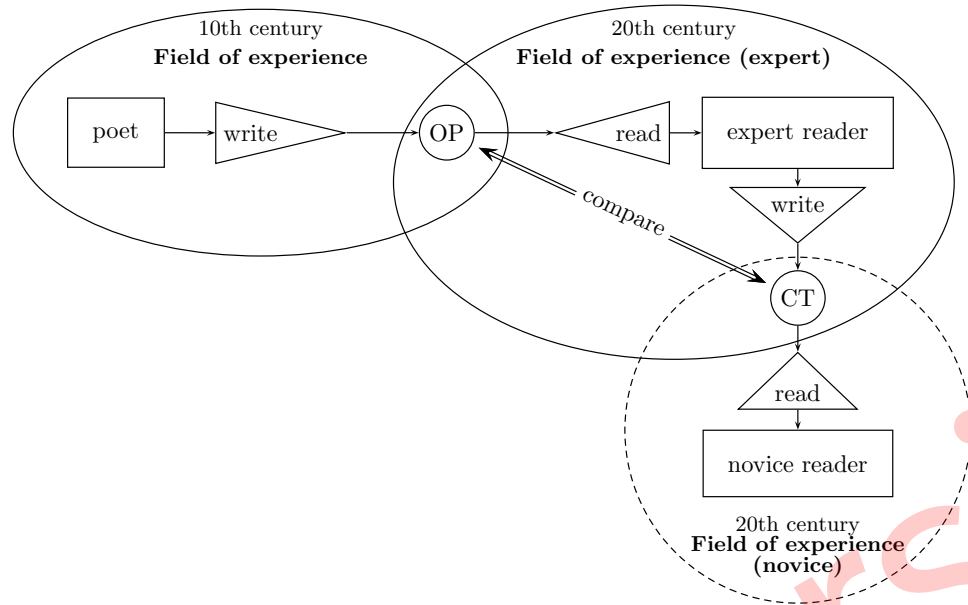


Figure 3: Connotative information retrieval in the schema of Schramm's communication process (reused from Yamamoto 2005, p.28): we simplify non-literal elements as the residual between two kinds of signals - the poem and the translation.

Schramm's communication model.

On the other hand, we must highlight some issues when utilizing non-literal elements for connotation visualization.

The first is the coverage issue, which is unsolvable because “no inventory of the connotative meanings generated by any sign could ever be complete” (cf., Chandler 2002, p.139). That is, through non-literal elements, we cannot visualize all aspects of connotation and cannot reproduce all connotation described in the dictionary (cf., fig. 4).

The second is the “impurity” issue. The impurity mainly refers to the inclusion of function words used to facilitate the translations are also included in the extracted non-

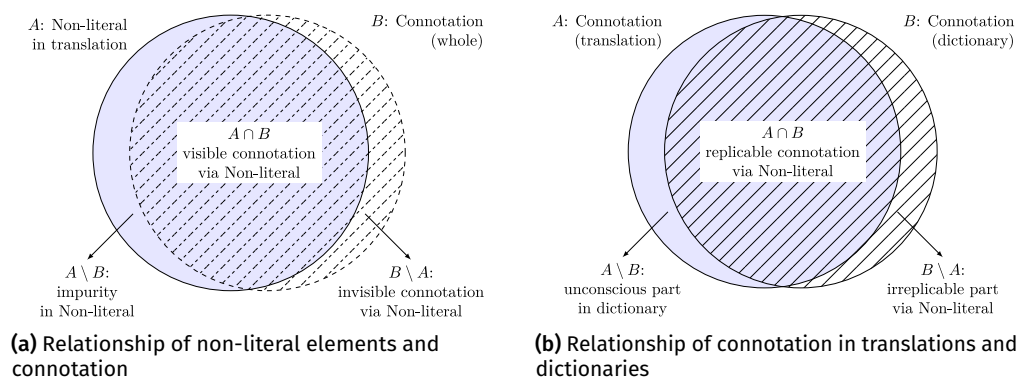


Figure 4: Relationship of connotation and non-literal elements, and relationship of connotation in translations and dictionaries: (a) Non-literal elements can reflect some aspects of lexical connotation, but they can also include non-connotative elements and cannot cover the whole connotation. (b) Connotation in translations covers some of the connotation described in dictionaries and captures the unconscious aspects not explicitly mentioned in the dictionaries. Dashed line represents an open set; solid line represents a closed set.

literal elements (cf., fig. 4a), which represent the syntactic differences between target languages and source languages. In the following sections introducing the methodology, we will explain how we handle these “impurities.”

3. Methods

Section 3 introduce the materials, the *Kokinshū* and its ten translations, and their data formats. Finally, we introduce the two implementations - one based on word misalignment and one based on salient word co-occurrence.

3.1 Materials

3.1.1 The *Kokinshū* text data

The *Kokinshū* comprises 1111 poems. We utilize solely the initial 1000 poems², which are classified as tanka to maintain uniformity in poem length.

We use the *Kokinshū* text data from the Hachidaishu Vocabulary Dataset (Hodošček and Yamamoto 2022)³. The dataset provides both a TEI format version and a space-delimited format version. For compatibility with the translation data format, we choose the space-delimited format (refer to fig. 5 for details). The space-delimited format version is generated by automatically processing poems in the *Hachidaishū* using a domain-specific morphological analysis system for classical Japanese poetry (Yamamoto 2007) and a semantic category code annotation system (Yamamoto 2009). The semantic category codes are based on an older version of the *Word List by Semantic Principles* (WLSP) (Nakano et al. 1994), a collection of words classified and organized by semantic categories. The dataset includes both compound tokens and their constituents (i.e., simplex tokens), and assigns multiple semantic category codes to each polysemous token.

We use semantic category codes during processing instead of lemmas. For compound tokens, we use the compound token itself rather than its breakdown. In cases where a token has multiple semantic category codes, we select the first code. We do not exclude any stop words in the preprocessing since even function words in classical Japanese poems can occasionally function as content words with lexical meaning.⁴ Table 3 provides a summary of the preprocessed data.

3.1.2 Translation text data

We use the ten contemporary Japanese translations (table 2), We tokenize all translation texts into tokens using Chasen (Matsumoto et al. 2002), a Japanese morphological analysis system⁵. We annotate each token with the same semantic category code

2. The remaining 111 poems consist of *choka* (long poem), *sedoka* (head-repeated poem), *azuma-uta* (poem written in Eastern dialect), and others. Some vary in length and form from tanka, some employ dialects, and some touch upon religious matters.

3. The *Hachidaishū* consists of the first eight imperial anthologies of classical Japanese poetry, with *Kokinshū* being the initial anthology.

4. For example, the conjugated form of “tsu,” an auxiliary verb used to express the perfective aspect, is “tsuru.” The reading of “tsuru” also carries the additional meaning of “crane (bird)”.

5. Due to limitations of Chasen, we convert the historical kana orthography in the earliest translation Kaneko (1933), which was widely used before orthographic reforms post World War II.

	database ID	token ID	semantic category	morpheme forms										
possible semantic variants	01:000001:0001	A00	BG-01-1630-01-0100	02 年 とし 年	year									
	01:000001:0001	A10	BG-01-1911-03-1800	02 年 とし 年										
	01:000001:0002	A00	BG-08-0061-07-0100	02 の の の	of (particle)									
				⋮										
possible decompositions	01:000001:0010	B00	BG-01-1950-14-0100	02 一年 ひととせ 一年	one year									
	01:000001:0010	C00	BG-01-1950-01-0300	19 一 いち 一	one									
	01:000001:0010	C00	BG-01-1630-01-0100	02 年 とし 年										
				⋮										
	01:000001:0015	A00	BG-02-3120-01-0100	02 いは 言ふ いふ 言は いは	say									
	01:000001:0016	A00	BG-03-3012-03-2600	02 ん む む む む	auxiliary verb: inference									
	<u>01:000001:0016</u>	<u>A10</u>	<u>BG-09-0010-02-0100</u>	<u>02 ん む む む む</u>										
anthology ID	poem ID	token sequence ID	token type	general ID	POS ID	semantic group ID	semantic field ID	specific ID	POS number	surface form	lemma kanji (if any)	lemma kana	conjugation kanji (if any)	conjugation kana

Figure 5: Format of the Hachidaishu Vocabulary Dataset (Yamamoto and Hodošček 2021): in the hachidai.db format, a line consists of 7 columns separated by spaces. The 1st column 01:000001:0007 consists of 3 fields separated by colons: 1) anthology, 2) poem number, and 3) serial ID of the token. The anthology ID 01 indicates the *Kokinshū*; the 2nd column indicates the type of token: A is a single token; B is a compound token; C is a breakdown of B. A00 indicates a single token; A01 indicates a single token with another meaning; B00 indicates a compound token; B01 indicates a compound token which has another meaning; C00 indicates the first element of the B00/B01.. breakdown; C01 indicates the second element of the B00/B01.. breakdown; 3rd column BG-02-1527-01-0102: classification ID based on semantic categories; 4–9th column indicates respectively: a Part-of-Speech number, a form that appears in literary works, a lemma in kanji script, a lemma in kana script, conjugated form in kanji writing form, conjugated form in kana writing form.

system (i.e., WSLP) as that of the *Kokinshū* data, allowing us to later determine semantic 242
equivalence between tokens in the old and contemporary languages. 243

Although many translators claimed to follow a word-for-word translation style, Ya- 244
mamoto and Hodošček (2019) has demonstrated that regardless of the translation style, 245
every version of the translation includes approximately 50% of tokens that lack semantic 246
equivalence with any token in the corresponding original poems. 247

During the implementation of our methods, we adhere to the same preprocessing steps 248
used for the *Kokinshū* data. The only distinction is that we remove all symbols present 249
in the translation data. A summary of the preprocessed data can be found in table 3. 250
Currently, the translation data cannot be open-sourced due to copyright restrictions. 251

	abbreviation	references	manuscript	translation style
1	KNK	Kaneko (1933)	Teika	word-for-word
2	KBT	Kubota (1960a,b,c)	Teika	word-for-word
3	MTD	Matsuda (1968a,b)	Teika	not mentioned
4	OZW	Ozawa (1971)	Teika	wording changed
5	TKOK	Takeoka (1976a,b)	Teika	word-for-word
6	OKMR	Okumura (1978)	Teika	intention oriented
7	KSJ	Kyusojin (1979)	Teika	word added
8	KMCY	Komachiya (1982)	Teika	not mentioned
9	K&A	Kojima and Arai (1989)	Teika	not mentioned
10	KTGR	Katagiri (1998a,b,c)	Teika	word-for-word

Table 2: Ten contemporary Japanese translations of the *Kokinshū* and their translation style, ordered by year.

abbreviation	# of tokens	# of types	# of texts
KNK	42,439	3,356	1,000
KTGR	36,362	2,882	1,000
K&A	33,867	2,955	1,000
KMCY	30,869	2,692	1,000
KBT	32,210	2,701	1,000
KSJ	34,050	2,770	1,000
MTD	31,860	3,007	1,000
OKMR	32,321	3,153	1,000
OZW	36,173	3,384	1,000
TKOK	29,844	2,861	1,000
total	339,995	8,252	10,000
<i>Kokinshū</i>	16,687	1,496	1,000

Table 3: Details of the preprocessed data

3.1.3 Data summary

Table 3 summarizes the post-preprocessing data. The *Kokinshū* and its ten translations serve as the foundation for a parallel corpus. The dataset includes 10,000 parallel text pairs; the *Kokinshū* side consists of 1,000 sentences each repeated ten times, a scale less than the so-called big data employed in state-of-the-art learning techniques for natural language processing tasks. This lack in parallel historical data complicates the application of state-of-the-art learning techniques (cf., Kalouli et al. 2019, p.109). Hence, we opt for traditional computational methods to attain our objective.

3.2 Implementation A: misalignment-based visualization

Implementation A (Chen et al. 2022) visualizes the explanatory addition in translations (that is, non-literal part in original poems) based on a statistical word alignment model, IBM model 2 (Brown et al. 1993). This method applies the set difference in the guise of misalignment extraction. We trained two word alignment models: one trained with parallel texts in source-to-target order (Model A); the other trained with parallel texts in target-to-source order (Model B). From the alignments⁶ inferred by Model B, we

6. Alignment of a parallel text suggests a set of word pairs, each comprising an original poetic word and a translated word (cf., Koehn 2010, p. 84).

deduct alignments inferred by both Model A and Model B. For the remaining alignment pairs, that is, pairs in misalignment, the translated word isn't precisely semantically equivalent, yet related to the original poetic word. The translated word is added to transmit information that requires verbalizing in the poem to beginner readers. As observed in section 2.4, non-literal information includes functional words, which do not represent connotation. We omit those functional words from the extracted non-literal elements with post-processing. The specifics and process are as follows.

Step A: training of word alignment models We apply IBM Model 2 as the word alignment model to the parallel corpus.⁷ We train two models: a source-to-target model (Model A) and a target-to-source model (Model B). We use the IBM Model 2 from nltk 3.7.0, a natural language processing toolkit, within Python 3.8. The intersection of alignments by Model A and B results in a 69.85% precision while Model B results in a significantly lower precision of 38.25% (when iteration is 8), indicating we can extract numerous misaligned patterns.⁸

Step B: extraction of misalignment For each parallel text containing a queried word, we apply the word alignment models learned in Step A to the parallel text in the source-to-target direction (Model A) and the target-to-source direction (Model B). For a queried poetic word, there are three types of alignments with words in translations inferred: (a) an aligned translation word inferred by Model A; (b) one or more aligned translation words inferred by Model B; (c) an intersection of the above two. For most poetic words, (c) indicates the correct translation. Conversely, several aligned translation words in (b) contain many incorrect but statistically related translation words for the queried poetic word in the translations. To obtain such non-literal elements, we subtract aligned results of (c) from (b); that is, we calculate the set difference of (b) and (a). Figure 6 illustrates the procedure.

Step C: post-processing to filter function words While Chen et al. (2022) do not explicitly include this step, it is implied that without it, the non-literal elements in the results inevitably contain function words for translation. Therefore, Step C removes those functional elements from the extracted non-literal elements, retaining only the connotative elements. The filtering process is based on WLSP semantic category codes. Function words that are not nouns, verbs, or adjectives are excluded prior to visualization.

Step D: network visualization in three phases (fig. 7) We execute steps B and C on each set of parallel texts and build an aggregate visualization by accumulating misalignments, following a bottom-up strategy. This implies that we can deconstruct the visualization, phase by phase: (a) visualize the non-literal elements (misalignment) from a single parallel text containing the queried word; (b) visualize the non-literal elements from ten parallel texts of the same poem (single poem featuring the queried word with its

7. The implementation can also utilize any other word alignment model if the model does not require additional data annotations.

8. Detailed precision, recall, and AER (average error rate) of word alignment models are reported at <https://github.com/nehcx/kokinMisalign>, accessed on 20 May 2023. During these statistics' calculation, we used the consistency between the WLSP semantic category codes of the word pairs in each alignment to judge whether the alignment is a sure alignment, possible alignment, or misalignment. This led to an extremely strict criterion for sure alignment and possible alignment; therefore, the precision and recall reported are considered much lower than actual, and AER is higher than actual.

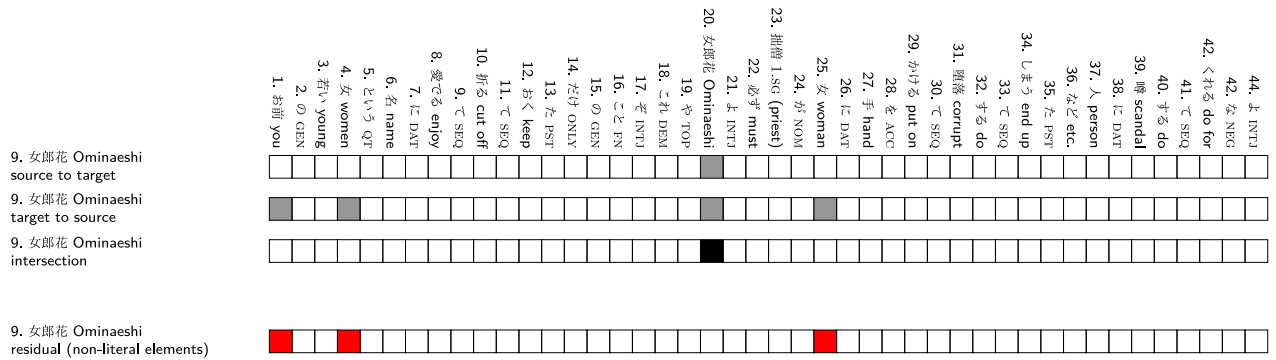


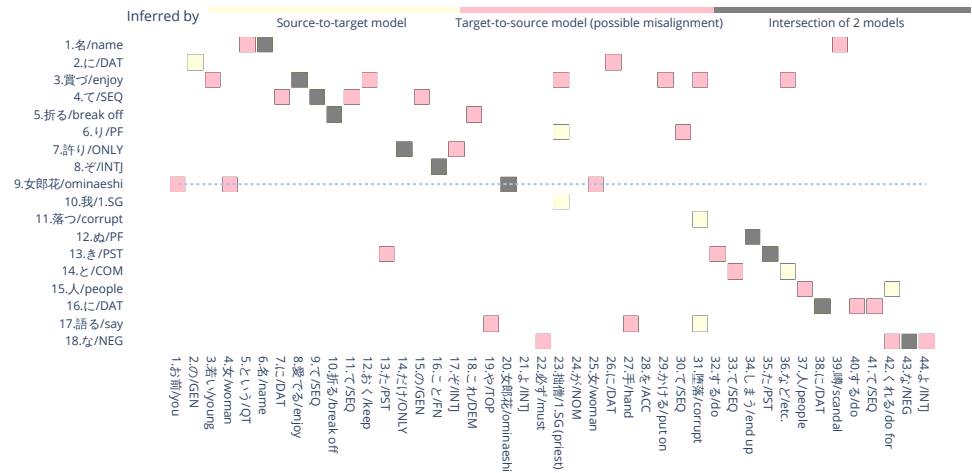
Figure 6: Procedure of the set difference of aligned results of the poetic word “ominaeshi” 女郎花 (en. golden valerian) in the 226th *Kokinshū* poem and its translation by Kaneko (1933) (figure based from Chen et al. 2022): the first row is the aligned results of (a); the second row is the aligned results of (b); the third row is (c) – the intersection of (a) and (b); the final row is the complementary set of (a) in (b) and hence the non-literal information.

ten translation texts); (c) visualize the non-literal elements from all parallel texts that include the queried word. (c) is the final aggregate visualization of the non-literal elements for the queried word. The aggregate visualization can show which non-literal elements of a poetic word are commonly clarified by various translators. Conversely, (a) and (b) depict non-literal elements for a word at the poem level, aiding in contextual checks.

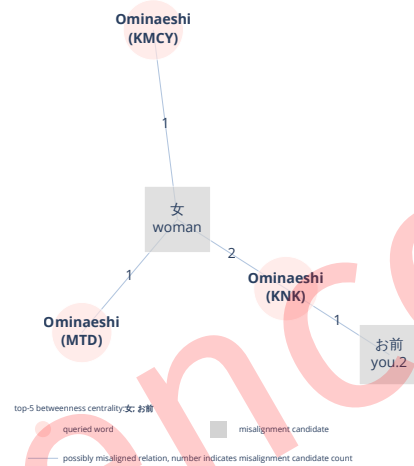
In summary, Implementation A visualizes non-literal elements hidden in poetic words as explanatory additions in translations. Non-literal elements are represented by misaligned translation words for each poetic word. Additionally, Implementation A can trace non-literal elements of a poetic word in detail with its three-phase visualization: it shows from which translator’s version a non-literal element originates, and from which poem it comes. On the other hand, Implementation A is designed to visualize non-literal information with a strict criterion - if there is no explanatory addition in the translations directly for the queried poetic word, the visualization outputs nothing. This does not mean the poetic word has no connotation; rather, it means all ten translators commonly consider that the poetic word is understandable to novice readers in a poem or multiple poems without any explanation. In other words, the connotation for the word is largely shared between poets and contemporary readers. In this case, visualizable connotation via non-literal elements is limited, which can be a drawback. Our Implementation B can address this drawback.

Implementation A is now accessible on Github⁹. The repository provides a dashboard built with Dash 2.7.1 in Python 3.8. This dashboard integrates all stages discussed in Implementation A. Upon entering the dashboard, one can first access the query interface to search for words in classical Japanese poetry. The search outcome is a table composed of each parallel text where the searched word appears. One can also select any version of the translation (the default version is KNK/Kaneko (1933)). One can click on word alignment visualization to see the word alignment outcomes of the present parallel text. Within the misalign network visualization phase, one can view two types of visualizations: one shows what non-literal elements of the searched word are included in different translations of a specific poem (one can visualize by clicking

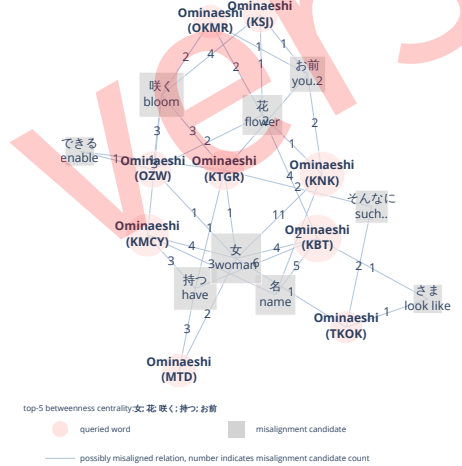
9. <https://github.com/nehcx/kokinMisalign>, reached on 20 May 2023



(a) Phase 1: alignment and misalignment visualization for a specific “ominaeshi” poem with a specific version of the translation



(b) Phase 2: misalignment network visualization for a specific “ominaeshi” poem with multiple versions of translations



(c) Phase 3: overall misalignment network visualization for all “ominaeshi” poems with all versions of translations

Figure 7: Phase-by-phase misalignment visualization for “Ominaeshi” (女郎花, en. golden valerian): (a) displays the detailed supplementary translation words for golden valerian in a specific parallel text; (b) collectively presents the supplementary translation words by different translators for golden valerian in a specific poem; (c) illustrates the overall supplementary translation words for golden valerian.

on the specific poem in the table); the other discloses what has been included for the 334
searched word in all its translations in all poems where it appears. 335

3.3 Implementation B: co-occurrence pattern-based visualization 336

Implementation B builds on (Yamamoto 2005), an information-theoretical-based approach. 337
Unlike Implementation A, the analysis unit of Implementation B is the co- 338
occurrence pattern rather than the word. In Implementation B, both the Poem set and 339
Translation set are sets of co-occurrence patterns. The set difference is, therefore, used 340
to visualize co-occurrence patterns that exist only in the Translation set. 341

In Implementation B, we apply co-occurrence patterns for the subsequent reasons. Ini- 342

tially, if we merely calculate the set difference using two bag-of-words, the remaining words in the complementary set for contemporary translation will forfeit their interrelationships, namely, the visualized non-literal elements will not maintain the minimum context. This diminishes the interpretability of the visualized non-literal elements. Secondly, without co-occurrence, the relation between the visualized non-literal elements and the queried word may be unclear. Unlike the misalignment-based Implementation A, which associates non-literal elements with queried words as misalignment, a single-word-based set difference can fail to indicate which elements link to the queried word and how. Additionally, the co-occurrence-based implementation can tackle the weaknesses of Implementation A: when there is no explanatory addition directly for a queried poetic word, Implementation A may visualize nothing for the word; inversely, Implementation B can still visualize something. In the instance of the co-occurrence pattern, for a poetic word p_a , if it shares a co-occurrence relationship with a word p_b in both the original poem and the translation, the relation (p_a, p_b) will not be visualized directly for p_a . However, when p_b holds a co-occurrence pattern with an additional element c in the translation, p_b can remain as the non-literal element in the visualization for poetic word p_a in the form of co-occurrence pattern (p_b, c) . That is, Implementation B can visualize not only co-occurred words in original poems (e.g., p_b) but also indirectly related textual additions in translation (e.g., c) as non-literal elements¹⁰ for the queried word p_a . Such non-literal patterns are not remnants of the direct co-occurrence with the queried word in original poems, but minimum context indirectly related with the queried word that arose in the contemporary translation.

Furthermore, Implementation B also encounters the issue that non-literal elements may contain non-connotative contamination. To tackle this matter, unlike the post-processing tactic in Implementation A, Implementation B sifts co-occurrence patterns with a pliable keyness threshold of patterns. The specifics and process are as follows

Step A: construction of two sets of co-occurrence patterns for queried words Suppose we query a poetic word. We keep each original poem and each translation that contains the queried poetic word. From the retained poems, we collect all the co-occurrence patterns observed in the poems, regardless of the distance between two co-occurring words, and form the queried Poem set of co-occurrence patterns. The same is carried out on the kept translations to form the queried Translation set.

Step B: co-occurrence pattern keyness calculation We calculate the keyness of each co-occurrence pattern in the queried poems and the queried translations relative to the whole poems/translations respectively. The keyness is for filtering functional words and unimportant words. The relative keyness is modeled by cw (eq. (2)).¹¹

$$cw(t_1, t_2, d) = (1 + \log ctf(t_1, t_2, d)) \cdot \sqrt{idf(t_1) \cdot idf(t_2)} \quad (2)$$

$$idf(t) = \log \frac{N}{df(t)} \quad (3)$$

10. More precisely, p_b is literal in original poems; while c and the minimum context (p_b, c) for p_b in translations is non-literal for the original poems.

11. cw is an extended form of $tfidf$ (Manning et al. 2008, p.109). $tfidf(t, d) = tf(t, d) \cdot idf(t)$, where $tf(t, d)$ is the relative frequency of term t within document d ; $idf(t)$ (eq. (3)) is the inverse value of the frequency of term t (Spärck Jones 1972). $tfidf$ models the importance of a single word in a specific document. In (2), $(1 + \log ctf(t_1, t_2, d))$ and $\sqrt{idf(t_1) \cdot idf(t_2)}$ correspond to the tf part and the idf part in $tfidf$, respectively.

where $\log ctf(t_1, t_2, d)$ indicates the co-occurrence frequency of words t_1 and t_2 in a document d ; $df(t)$ signifies the document frequency of word t . We here opt to use cw because the cw considers not only the keyness of co-occurrence patterns as a whole but also the keyness of each word in each pattern simultaneously. $\sqrt{idf(t_1) \cdot idf(t_2)}$ part emphasizes the keyness of t_1 and t_2 should be high simultaneously for document d ; otherwise, (t_1, t_2) may not be a key pattern. By doing so, we can avoid those unimportant function words from being included in key patterns. Conversely, general keyword extraction methods for single words may ignore the keyness of each word in patterns, leading to high keyness for patterns like high idf content word + low idf function word.¹²

Step C: filtering co-occurrence patterns We filter co-occurrence patterns where cw is lower than a specific threshold to automatically remove “impurity” before the calculation of set difference. According to Yamamoto and Hodošček (2018), cw is normally distributed; after standardization, cw can sample key patterns consisting only of content words by setting the threshold as one standard deviation of the normal distribution. In this study, however, we do not use the threshold of one standard deviation. To simultaneously reduce the overlap of labels in visualization, we adjusted the threshold flexibly. Because of the different sizes (usually a ratio near 1:3 or 1:4) between the queried Poem set and the queried Translation set, the thresholds for the two sets are set differently. The thresholds for the Translation set are often 3 or 4 times higher than that of the Poem set.

Step D: set difference We subtract the intersection of the two filtered set of co-occurrence patterns from the filtered Translation set and the filtered Translation set. Figure 8 shows the procedure of the set difference. Details of the example complementary set can be seen in section 3.3¹³.

Step E: network visualization The visualization uses the *dot* language to visualize important co-occurrence patterns that are added in translations. These co-occurrence patterns form networks and present a global view of additional information and the salient relation with the queried word linked by the additional information.

In Implementation B, the complementary set will contain two types of non-literal elements: (a) explanatory additions that are added directly for translating the queried words (e.g., “break off” for “plum” in translations); (b) collocates with the queried words in original poems, which hold explanatory additions in translations (e.g., “woven hat-hide” for “plum,” in which “hide” is the explanatory addition for the “woven hat” in translations). Different from Implementation A, non-literal elements visualized by Implementation B include not only words but also the relations. From this perspec-

12. Aside from cw , various keyness measures are available (e.g., Burrows 2002; Dunning 1993; Spärck Jones 1972), while comparisons among measures (e.g., Du et al. 2022; Paquot and Bestgen 2009; Schöch et al. 2018) have mainly focused for traditional linguistic units, i.e., words, rather than extended units of meaning such as co-occurrence patterns. The recent systematic comparison (Du et al. 2022) among the measures suggests dispersion-based methods can better differentiate texts in the case of shorter, randomly selected segments. However, because each classical Japanese poem ends within a single sentence, with each word type appearing about 1-2 times, the advantages of dispersion-based methods may not be fully utilized. Comparison of measures specific for such extremely short forms of literary texts and for extended linguistic units is necessary in the future.

13. The full lists, before reducing the translation set by removing the original set, is available on https://github.com/nehcx/kokinMisalign/tree/master/supplementary_material, visited on 27, May, 2023

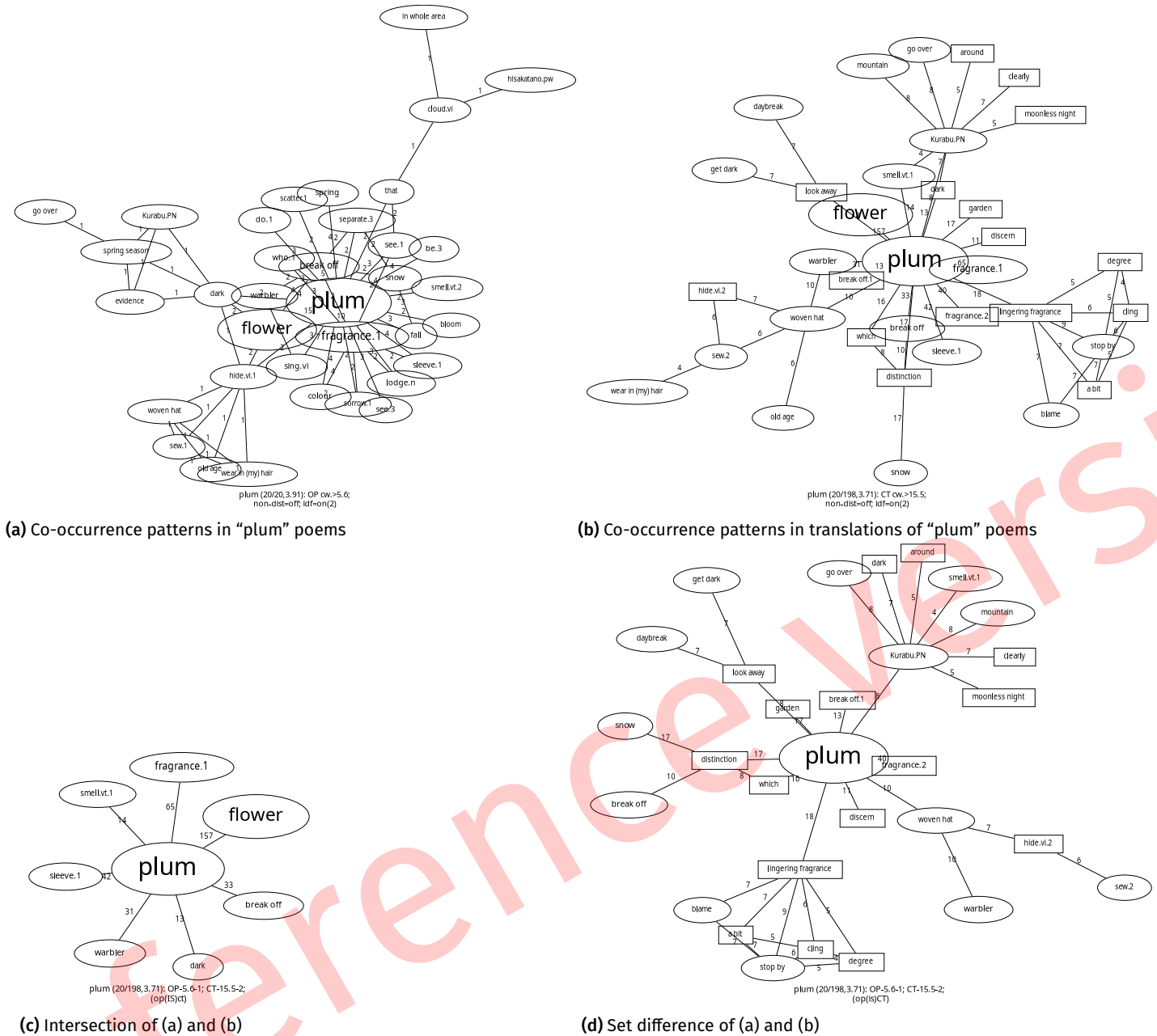


Figure 8: Procedure of set difference for the salient co-occurrence network of “ume” (梅, en. plum) poems and their translations: Co-occurrence patterns are connected by edges; nodes represent words. From the co-occurrence patterns in translations (b), we subtract the intersection (c) of the co-occurrence patterns of the original poems (a) and (b). Square nodes indicate word forms that exist only in contemporary Japanese, while elliptical nodes indicate poetic words

tive, (a) represents the direct relations with the queried word, and (b) represents the relations with poetic words that hold a relation with the queried word. Each relation is salient for the queried word in both original poems and translations, and almost no patterns contain function words, the “impurity,” with the *cw*-based filtering. Such relations, i.e., co-occurrence patterns, are as non-literal information in the poem revealed by translators to the translations for novice readers.

We provide a simple web application to visualize key co-occurrence patterns for classical

	<i>cw</i>	<i>cf</i>	<i>t</i> ₁	<i>f</i> (<i>t</i> ₁)	<i>idf</i> (<i>t</i> ₁)	<i>t</i> ₂	<i>f</i> (<i>t</i> ₂)	<i>idf</i> (<i>t</i> ₂)
1	21.15	5	Kurabu.PN	8	9.21	moonless night	7	7.13
2	19.83	8	Kurabu.PN	8	9.21	go over	10	4.50
3	19.83	7	Kurabu.PN	8	9.21	dark	10	4.92
4	19.57	7	Kurabu.PN	8	9.21	clearly	10	4.79
5	19.37	40	plum	198	3.71	fragrance.2	42	4.59
6	18.12	18	lingering fragrance	19	5.84	plum	198	3.71
7	18.01	8	plum	198	3.71	Kurabu.PN	8	9.21
8	17.97	9	lingering fragrance	19	5.84	stop by	10	5.40
9	17.50	6	cling	7	6.72	lingering fragrance	19	5.84
10	17.21	7	lingering fragrance	19	5.84	blame	7	5.84
11	17.07	5	Kurabu.PN	8	9.21	around	5	4.65
12	16.90	16	which	16	5.40	plum	198	3.71
13	16.83	6	stop by	10	5.40	cling	7	6.72
14	16.82	17	plum	198	3.71	distinction	17	5.18
15	16.80	4	smell.vt.1	14	5.38	Kurabu.PN	8	9.21
16	16.76	7	look away	8	7.01	get dark	10	4.62
17	16.55	7	blame	7	5.84	stop by	10	5.40
18	16.33	7	a bit	9	5.26	lingering fragrance	19	5.84
19	16.32	11	discern	11	6.21	plum	198	3.71
20	16.30	8	distinction	17	5.18	which	16	5.40
21	16.26	5	lingering fragrance	19	5.84	degree	5	6.64
22	15.95	4	cling	7	6.72	degree	5	6.64
23	15.91	7	look away	8	7.01	daybreak	9	4.16
24	15.71	7	a bit	9	5.26	stop by	10	5.40
25	15.71	8	plum	198	3.71	look away	8	7.01
26	15.69	6	sew.2	6	6.72	hide.vi.2	14	4.70
27	15.64	5	degree	5	6.64	stop by	10	5.40
28	15.63	10	distinction	17	5.18	break off	33	4.32
29	15.63	7	hide.vi.2	14	4.70	woven hat	10	5.99
30	15.62	10	woven hat	10	5.99	warbler	40	3.73
31	15.61	17	plum	198	3.71	garden	17	4.46
32	15.58	10	plum	198	3.71	woven hat	10	5.99
33	15.57	17	distinction	17	5.18	snow	40	3.18
34	15.54	13	break off.1	14	5.12	plum	198	3.71
35	15.52	8	Kurabu.PN	8	9.21	mountain	8	2.76
36	15.52	5	cling	7	6.72	a bit	9	5.26

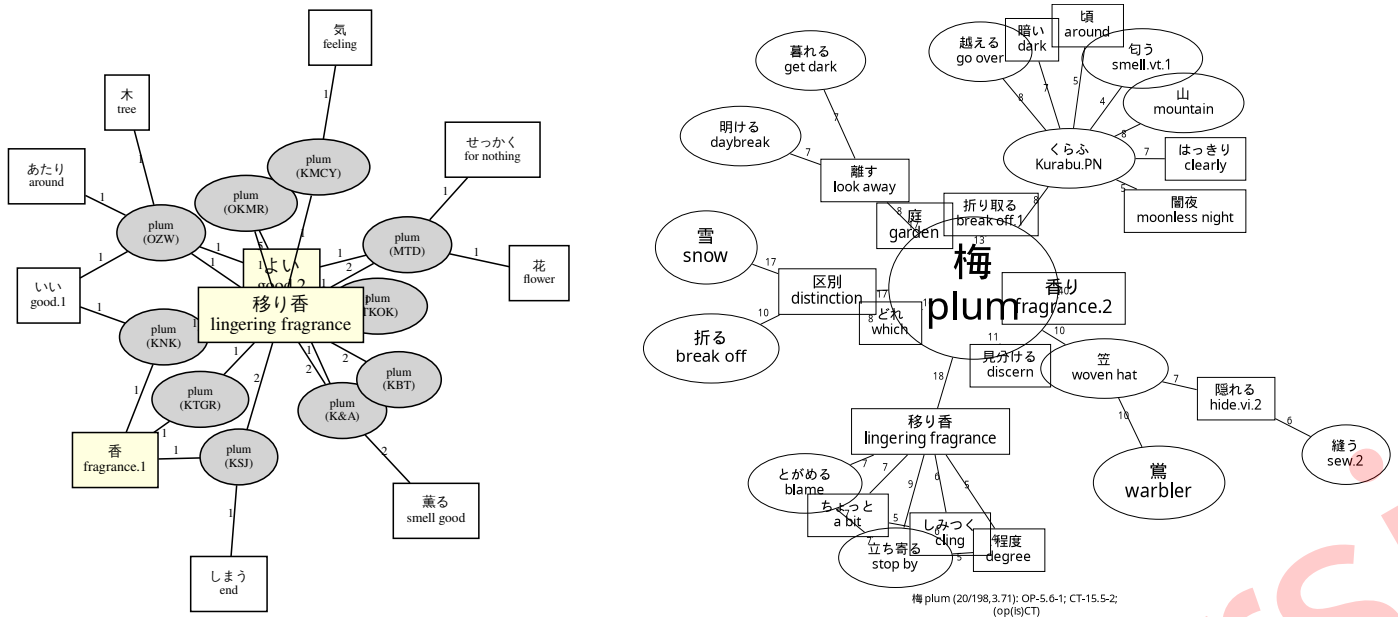
Table 4: Complementary set of co-occurrence patterns for “plum”: thresholds for poems and the translations are 5.6 and 15.5 respectively; 30 nodes (poetic words = 14, translation words = 16) and 36 edges in total.

poetic words¹⁴. The web application can also be applied to translations, but translations are currently protected by copyright. Therefore, the set difference of co-occurrence patterns is not publicly available.

4. Results

We applied the above two methods to the six most frequently used flora poetic words in the *Kokinshū*: “ume” (梅, en. plum), “ominaeshi” (女郎花, en. golden valerian), “kiku” (菊, en. chrysanthemum), “sakura” (桜, en. cherry), ‘matsu’ (松, en. pine), and “yamabuki” (山吹, en. kerria). This section shows the differences between the

14. <https://cuckoo.js.ila.titech.ac.jp/~yamagen/waka/poem.cgi>, accessed on 20, May, 2023. Currently, the visualization does not provide English translations.



(a) Visualization A: grey = queried word, yellow = common misalignment

(b) Visualization B: cw threshold = 5.6, 15.5 for poem/translation

Figure 9: Network visualization of “ume” (梅, en. plum): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

visualizations of the two implementations. Figures 9 to 14 display the visualizations for
each of the flora¹⁵.

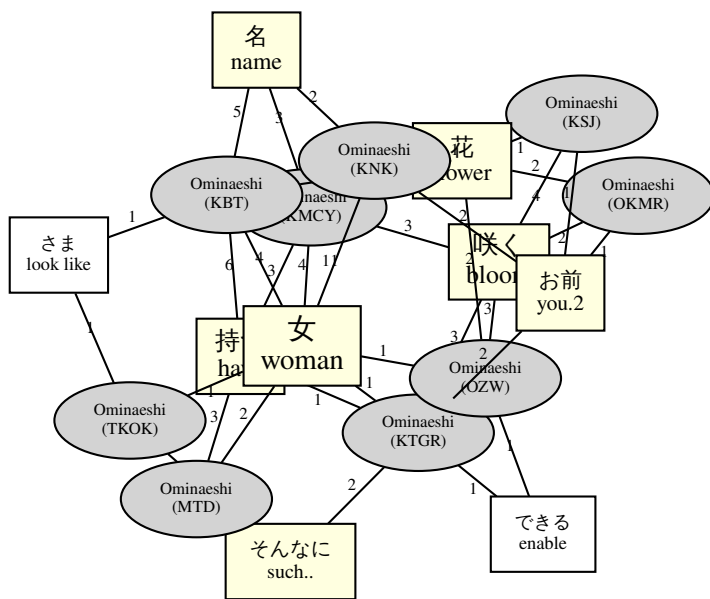
The results showed consistency and inconsistency with the descriptions of the flora poetic in the Dictionary of poetic vocabulary (Katagiri 1983).

4.1 Case-specific observations compared with the dictionary

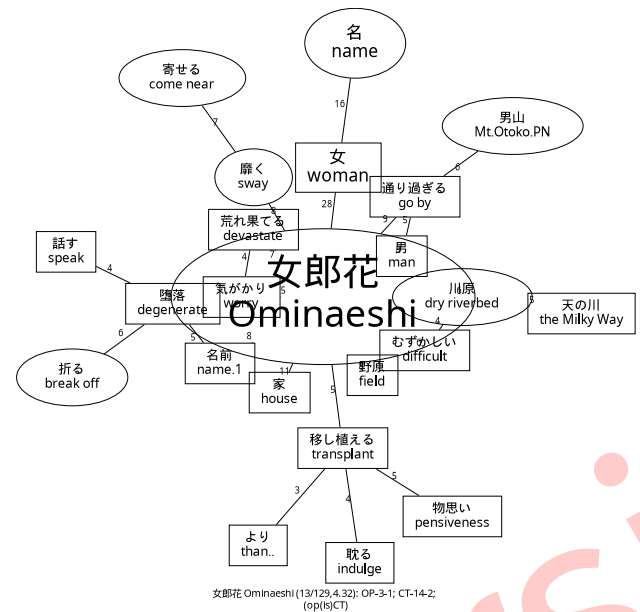
“Ume” (en. plum; fig. 9) We observed that the translation word “fragrance” (ja. 香/ 434
 移り香/薫る) was near the hub of the “plum” network generated by Implementation 435
 A as well as Implementation B. According to Katagiri (1983, pp.82–83), praising the 436
 fragrance of “plum” became common¹⁶ in classical Japanese poetry after the *Kokinshū*. 437
 The results agreed with the core description of the poetic word “plum.” Moreover, 438
 Implementation B presents a potentially significant node, “woven hat,” not directly 439
 mentioned in the “plum” entry of the dictionary. The dictionary contains an entry “ume 440
 no hanagasa” (梅の花笠, en. woven hat of plum blossoms), which often suggests a woven 441
 hat stitched by a warbler (Katagiri 1983, p. 83). Given the connection between “plum” 442
 and “woven hat,” “plum” can also retain a link with “warbler,” a frequently utilized pair 443
 in classical Japanese poetry and Japanese visual arts. Besides, An explanatory addition 444
 “distinction” ties “plum” with “snow” indirectly, from which we might deduce that 445
 the cluster portrays a situation where it is difficult to distinguish between “snow” and 446
 “plum.” Similarly, in fig. 9b, we can deduce from which classical Japanese poem each 447
 cluster of the network originates. For instance, the cluster “snow-break off-distinction- 448

15. For visual consistency with Implementation B and improved readability, we used dot to generate graphs for Implementation A, similar to Implementation B, instead of the original visualization based on the dashboard.

16. More than half of the poems regarding “plum” are about the fragrance (Katagiri 1983, p.83).



(a) Visualization A: grey = queried word, yellow = common misalignment



(b) Visualization B: *cw* threshold = 3, 14 for poem/translation

Figure 10: Network visualization of “ominaeshi” (女郎花, en. golden valerian): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

plum-which” could stem from No. 337.¹⁷ in the *Kokinshū*, and the cluster “plum-woven 449
hat-warbler-hide-sew” could stem from the No. 36¹⁸. Conversely, fig. 9a fails to replicate 450
a context of the original poem. It can display the core explanatory addition (“fragrance”) 451
as a hub for various translation versions. 452

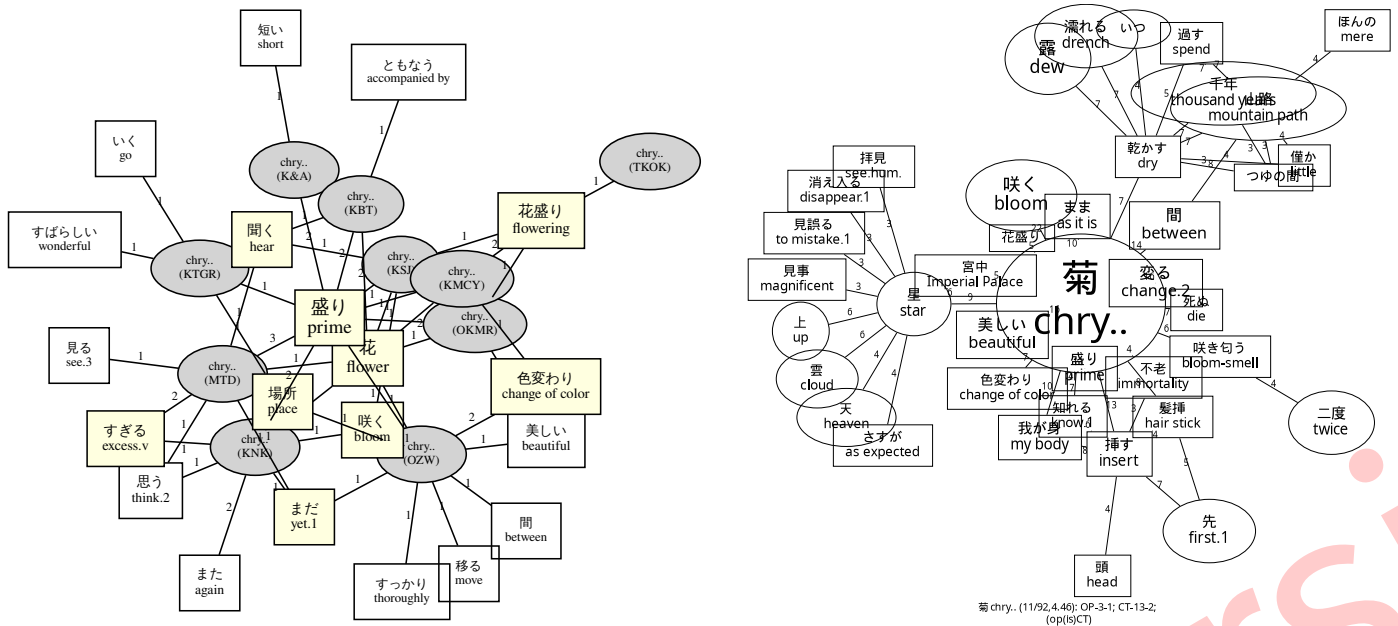
“Ominaeshi” (en. golden valerian; fig. 10) The explanatory “woman” was the hub in the “golden valerian” network formed by Implementation A (fig. 10a). This outcome aligned with (Katagiri 1983, pp.481–482): during the Heian period (794–1185), the majority of poems concerning the the flower depicted the flower as “woman.” The hypernym “flower” is frequently added in translations. Generally, the flower consistently holds an image due to its “name” of “woman.” Figure 10b also included the connotative term “woman” while “woman” was not in a central position in the network. Besides, Figure 10b visualized antonym terms “male” and the “Mountain of Man” (Mt. Otoko). In Eco (1976), hypernyms and antonyms are also connotations. Moreover, each cluster in fig. 10b can mirror the story of corresponding poems, poem No. 226, No. 227, No. 230, respectively.

“Kiku” (chrysanthemum; fig. 11) Both visualizations capture the connotation “color change” (ja. 色変わり/移る/変わる), which was described in Katagiri (1983, pp.127–129): during the Heian period, people admired the gradual change in color of white chrysanthemums to red due to the cold weather. According to Katagiri (1983, p.128), after the revival of the Chongyang Festival¹⁹ in the fifth year of Emperor Saga’s reign (814), “chrysanthemum” came to be used in many classical Japanese poems. Chrysanthemum

17. when snow has fallen/flowers appear on all the/trees clusters of white blooms/from which of them/can I pluck the fragrant plum blossoms (translated by Rodd et al. 1996, p. 143)

18. they say the thrush weaves a rain hat of flowering/plum perhaps I too/may pluck a spray and make a/garland to conceal my age (translated by Rodd et al. 1996, p. 59)

19. In Japan, the festival is also known as the Chrysanthemum Festival. According to Katagiri (1983, p.128), the chrysanthemum was introduced to Japan along with the Chinese Chongyang Festival.



(a) Visualization A: grey = queried word, yellow = common misalignment

(b) Visualization B: cw threshold = 3, 13 for poem/translation

Figure 11: Network visualization of “kiku” (菊, en. chrysanthemum): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

was an essential part of the Japanese Chongyang Festival because it was believed to have the effect of prolonging one’s life. Therefore, “chrysanthemum” was often used in classical Japanese poems to convey the sense of longevity. As for the connotative words regarding longevity, fig. 11a only presented “peak (of blossom season)” (ja. 盛り), which is a common addition at the hub; fig. 11b presented “one thousand years” (ja. 千年), “never get old” (ja. 不老), and “peak (of blossom season)” (ja. 盛り / 花盛り) while these words are not at an important position in the network (fig. 11b). Besides, fig. 11b presented “hear” (parallel image for “chrysanthemum” as kakekotoba) as a common addition. Nevertheless, the two methods failed to capture the relationship between the poetic word “chrysanthemum” and the Chongyang Festival, which should be essential background knowledge unfamiliar to contemporary Japanese novice readers. This is due to the Chongyang Festival not appearing in the translation text.

“Sakura” (en. cherry; fig. 12) The dictionary (Katagiri 1983, pp. 172–173) includes “Sakura” as a compound item “sakura-bana” (桜花, en. cherry blossom). According to Katagiri, most cherry blossom poems in the *Kokinshū* are about falling/scattering (ja. 散る) cherry blossoms which are often related to the transience of human life. We could observe connotative words regarding the sense of falling – “a way of falling” (ja. 散り方) and “to be or to lay scattered about” (ja. 散り乱れる) in fig. 12a. On the other hand, fig. 12a indicated that four translators incorporated “fall” (four added general “fall;” one added 散り果てる, en. all falling) into their translations of all “cherry” poems. Moreover, the terms “season,” “now,” and “later” (ja. 以上 and 後) were common additions in fig. 12a, which might suggest that “cherry” as a flower could have a strong relationship with time, evoking a mood of lamenting life’s impermanence and the passing of spring.(cf., Katagiri 1983, p. 173). Furthermore, fig. 12b illustrates the robust connection between “cherry” and “mountain” (inclusive of “mountain breeze,” “Yoshino,” “mountain cherry”). According to Katagiri (1983, p. 456), “cherry” and “Mt.

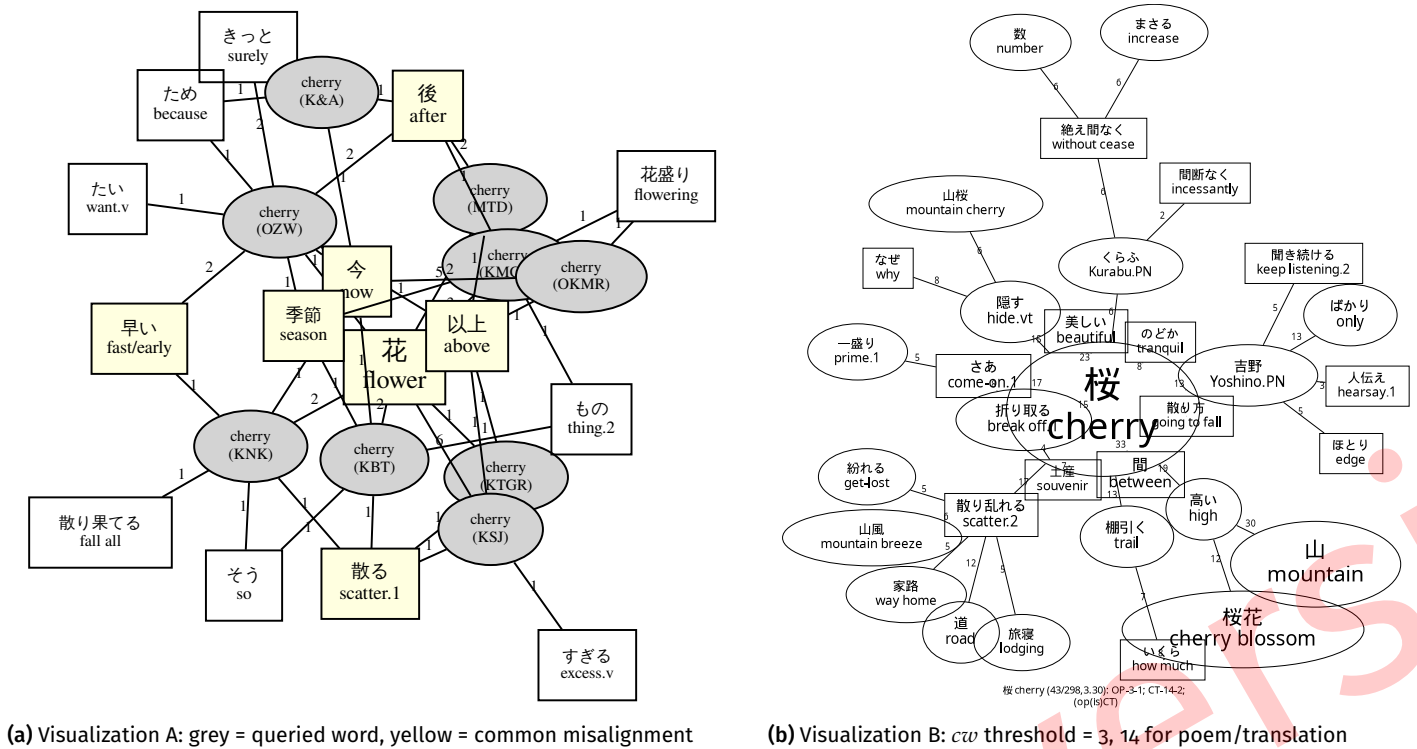


Figure 12: Network visualization of “sakura” (桜, en. cherry): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

Yoshino” have maintained a significant relationship since the *Shin Kokinshū*, the eighth 496
anthology of the *Hachidaishū*, while instances of the duo are scarce in the *Kokinshū*. From 497
the non-literal associations visualized here, we discern that “mountain” is a vital context 498
for cherry blossoms. This could provide a foundation for the forthcoming shifts in the 499
usage of “cherry.” 500

“Matsu” (en. pine; fig. 13) The most salient node in the two networks is “to wait”. 501
The results were consistent with Katagiri (1983, pp.383–384): the word “pine” is a 502
kakekotoba that holds another meaning of “to wait” (ja. 待つ, the same kana character 503
with “pine” in Japanese), which implies “long-awaited,” to celebrate the eternal nature 504
of the pine tree. Poets associated it with the word “one thousand years old” (ja. 千歳) 505
or with “crane” (ja. 鶴) and “wisteria” (ja. 藤), which also symbolize the one-thousand- 506
year longevity (Katagiri 1983, pp.383–384). We identified these terms in fig. 13b. On 507
the other hand, fig. 13a displayed “change” and “color,” as frequently added non-literal 508
elements, from which we can deduce that translators aim to highlight the longevity of 509
pine because its “color” never “changes.” Nevertheless, the both visualizations failed to 510
capture many descriptions in the dictionary. For example, poets use the poetic word 511
pine as “evergreen pine” (ja. 常盤の松) or “pine green” (ja. 松の緑). “Pine” is also 512
known for “kadamatsu (gate pine)” (ja. 門松)²⁰, where the gods resided, and “pine 513
wind” (ja. 松風), the wind that blows through the treetops of pine trees (Katagiri 1983, 514
p.384). They are rarely explained in translations. 515

20. “Kadamatsu” are traditional Japanese decorations made for welcoming ancestral spirits of the harvest in the New Year.

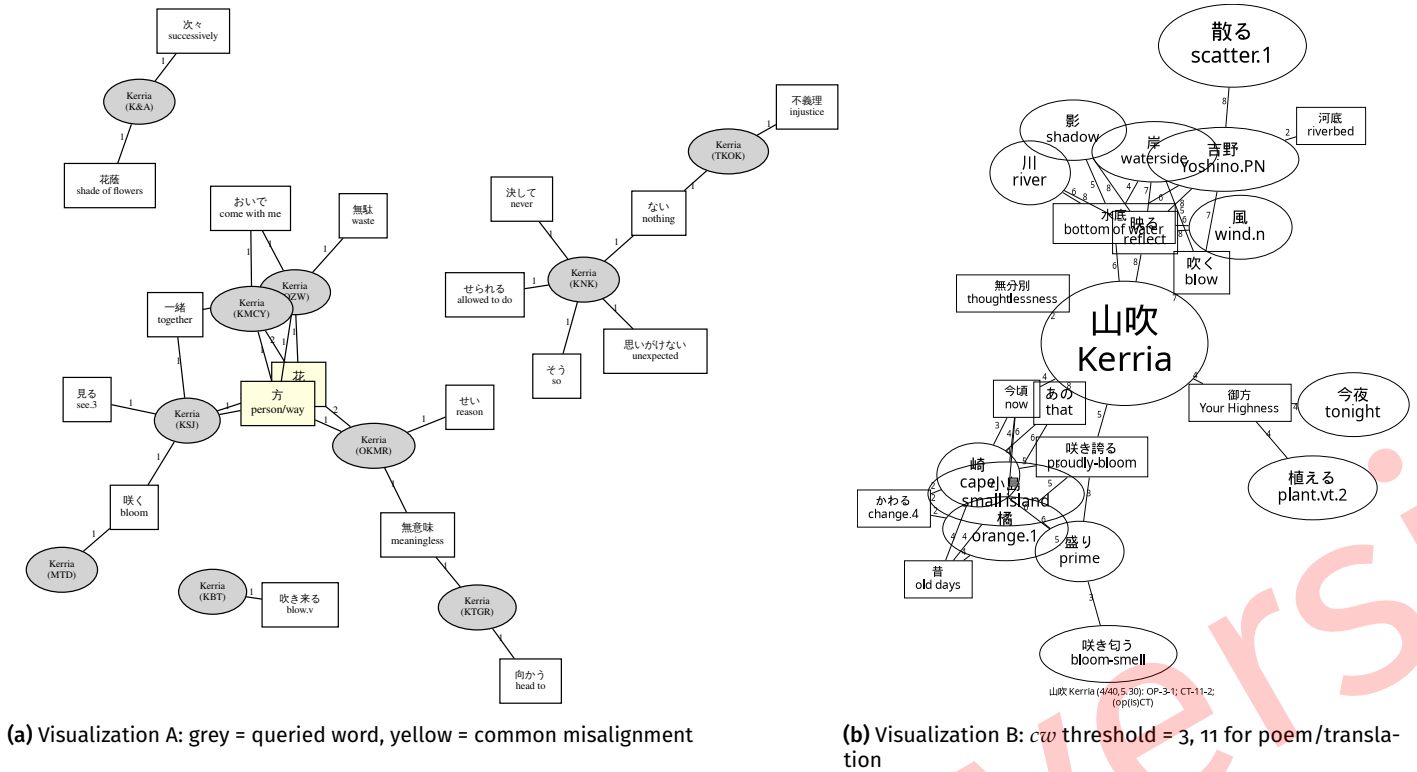


Figure 14: Network visualization of “yamabuki” (山吹, en. kerria): Square nodes indicate additions in translations; elliptical nodes indicate poetic words.

We also found that sometimes Implementation A visualized very common sense as a hub, which is often the hyponymy of queried words, such as “flower” for “cherry”. Since such words were not function words, we could not exclude them. Although the hypernym and hyponym for the queried words are viewed as connotations in some definitions, for our objectives, hypernyms and hyponyms are not the ideal connotation candidates. This is because these semantic relations are domain-general rather than domain-specific knowledge only for poetic words.

5. Discussion

This section discusses what we could learn from the project of connotation visualization for classical poetic Japanese vocabulary, including a summary of the differences between Implementation A and B, as well as the contributions and limitations.

5.1 Differences between two implementations

Both Implementation A and B are based on the schramm’s communication model and visualize non-literal elements by calculating complementary set of original poems and translations. However, they takes different strategies to perform set difference calculation and hence have different pros and cons (see table 5).

Implementation A does not require most of the prior annotations except for POS tags and does not require setting any arbitrary threshold. It provides breakdown visualization options, i.e., visualizes non-literal elements for a poetic word in a specific poem by a specific version (or all versions) of translation. Conversely, Implementation B requires

	Implementation A	Implementation B
visualization strategy		
analysis unit	word	co-occurrence pattern
criterion	strict	flexible
equivalence judgement	word alignment-based	WLSP category code-based
exclusion of impurity	POS-based filtering	cw-based threshold
set difference level	parallel text-based set difference (text-as-set; local level)	sub corpus-based set difference (corpus-as-set; global level)
aggregation strategy	gather local set differences then aggregate visualization	direct aggregate visualization (without breakdown visualization)
pros and cons		
coverage of connotation	low	high
arbitrary threshold	no	yes
breakdown visualization	yes	no
requirement of semantic annotation	no	yes

Table 5: Differences between Implementation A and Implementation B

the adjustment of parameters (the *cw* threshold) and a semantic category annotation 562
shared by two languages. Implementation B calculates set differences at the corpus level 563
and hence could not provide breakdown visualization as Implementation A. However, 564
Implementation B uses co-occurrence patterns as analysis units and therefore could 565
retain most of the important contextual information in the visualization, providing more 566
explainable illustrations. 567

In relation to the aspects of connotation detailed in section 2.1, Implementation A 568
visualizes non-literal elements reflecting rhetorical techniques, hypernyms, core associa- 569
tions; Implementation B visualizes non-literal elements reflecting rhetorical techniques, 570
hyponyms, antonyms, and a wide array of associations (e.g., phraseological pattern, 571
contextual information). Since translation is a communication process, the pragmatic 572
aspects (e.g., translator’s value judgement) of connotation could also be included in 573
the visualization. Conversely, both of Implementation A and B cannot capture soci- 574
olinguistic aspects (properties concerning gender, style, social class, region, etc.) and 575
emotional aspects of connotation. These aspects are also less mentioned in the dictionary, 576
which might be left for further studies using approaches from corpus-based variationist 577
linguistics and stylistics to explore. 578

For our objective of supplementing the descriptions for poetic language dictionaries, Im- 579
plementation B is a more feasible solution, as it can help incorporate minimal contextual 580
information from original poems through co-occurrence pattern clusters. In contrast, a 581
dictionary with selective consciousness may occasionally overlook such information. 582
Meanwhile, Implementation A aims to demonstrate the consistent use of explanatory 583
additions among experts. 584

5.2 Contributions 585

The two visualizations demonstrated how to utilize explanatory additions in the trans- 586
lations of literary texts to visualize the inaccessible part of the connotation of historical, 587
literary languages. We showed practically that not only dictionaries but also transla- 588
tions are feasible resources as a medium to access connotation in a historical, literary 589
language. Employing traditional statistical natural language processing algorithms and 590

information theory-based methods enhanced our interpretation when the historical literary text data is low resource.

Although our operationalization of connotation is only an imperfect simplification, it makes sense to distinguish connotative information from translations physically. Removing all the semantically equivalent elements between the parallel texts and visualizing the resulting leftover additional information is a transparent way of approaching the connotation. Furthermore, using misalignment or co-occurrence can link the additional information in translation with corresponding poetic words.

The misalignment-based implementation, which does not use semantic category coding, can be applied not only to contemporary Japanese translations of classical Japanese poetry but also to translations in other languages. On the other hand, the co-occurrence-based implementation can cover a large part of the connotation described in the dictionary and also provide essential contexts stemming from original poems.

The two implementation may also provide some theoretical considerations as follows.

Firstly, connotation could be considered a relative concept. The misalignment-based Implementation A yields only a minimal number of non-literal elements related to connotation, suggesting that much connotation may still be accessible to contemporary Japanese people, such as some domain-general knowledge. Consequently, the translator may choose not to translate this connotation. In other words, the connotation found at the intersection of contemporary and ancient Japanese people remains unseen in the non-literal set defined by Implementation A. However, we perceive this imperfect visualization as a dynamic feature of Implementation A, where connotation is seen as a relative concept. The connotative information in one group is relative to other groups. This might explain why it is seen as an open set so far. Our aim is to visualize the connotative information that requires supplementation in the communication between these two groups.

Secondly, extended units of non-literal elements are essential for uncovering lexical connotation. When there is no direct explanatory addition for a queried poetic word, Implementation A might not provide a visualization for the word (which does not mean the word has no connotation). In contrast, the co-occurrence patterns in Implementation B can help visualize the unseen aspect of connotation in Implementation A. This is because Implementation B extracts not only the additions in translations but also maintains the relationships in the original poem linked by the addition. In other words, the significant co-occurrence of a queried word in original poems can be visualized through an explanatory addition. The explanatory addition may be not directly for the queried word itself, but for the co-occurrence relation that the queried word holds. This also reflects that connotation exists not only in the poetic word itself but is also implied in the interrelationship between poetic words.

5.3 Limitations

The operationalization and the two implementations leave the following problems:

Firstly, encyclopedic knowledge and sociocultural context cannot be extracted. Kamens (1997, p.64) noted that every classical Japanese poem narrates a story or at least a portion

of it. However, our methods could never visualize intricate narratives and sociocultural contexts behind the poem. In other words, we could only infer the historical and sociocultural context based on the information provided by the poems and translations. For example, both methods could not visualize the connection between the poetic word "chrysanthemum" and the Chongyang Festival, which could be common tacit knowledge among poets during the Heian period. If such knowledge and sociocultural context do not appear in the translation and original texts, we are unable to visualize them.

Secondly, different fields of experience presumption is imperfect. The presumption that novice readers share no field of experience with poets is extreme. In reality, some domain-general connotation is shared between ancient Japanese poets and contemporary novice Japanese readers. Therefore, based on our operationalization, we could only occasionally visualize non-literal elements regarding these parts of connotation. However, if we understand connotation as a relative concept, then this drawback of operationalization can actually help us filter out non-literal elements that are irrelevant to the connotation described in the dictionary.

Thirdly, the signal analogy of literary texts is imperfect. We compared literary texts to a communication system that signals communication (see section 2.3). However, while a signal is continuous, literary language is considered a discrete set of meaning units in this study. Subtracting signals results in another continuous signal with an amplitude. On the other hand, subtracting a poem from its translation with set difference only yields an additional set of meaning units. Theoretically, poems should be treated as a sequence, akin to a signal. In practice, we can only treat them as a set, sacrificing the continuity of the poem as a sequence of words.

6. Conclusion

To offer a user-interactive lexical connotation visualization tool as a supplement to dictionaries of classical poetic Japanese, this paper aims to visualize connotation in the classical poetic Japanese vocabulary in the *Kokinshū* by presenting non-literal elements of each word. The non-literal elements were extracted by calculating the set difference between the *Kokinshū* and its ten translation versions, inspired by Schramm's (1954) communication model.

This paper outlines the motivation for using non-literal elements revealed by translations as a projection of connotation and discusses the relationship between connotation and non-literal elements, as well as the relationship between the dictionary and translations. We argued that although non-literal elements are not equivalent to connotation, when holding an explanatory addition in translations, non-literal elements could reflect part of connotation. Moreover, an explanatory addition of connotation in translation could illuminate the unconscious part of connotation, which is rarely covered by a dictionary.

From a technical standpoint, the paper presents two different implementations for visualizing non-literal elements based on Schramm's model: one is a word-misalignment-based visualization; the other is a co-occurrence pattern-based visualization. We attempted to apply these two visualizations to the six most frequent floral words in

the *Kokinshū*. As a result, word misalignment-based visualization could visualize the common explanatory addition by different translators, reflecting the most robust and essential perception of poetic floral words among the translators. However, the visualization only presented limited aspects of connotation, most of which remain unseen because they are not explained at the single-word level in the translation. On the other hand, co-occurrence pattern-based visualization covered a wide range of connotation descriptions in the poetic Japanese dictionary (Katagiri 1983). Furthermore, from the co-occurrence pattern-based visualization, we could infer essential contexts originating from original poems. Its success may lie in the usage of extended analysis units (co-occurrence in the poems and translations), which implies that some lexical connotations should be revealed at the word co-occurrence level. In other words, for some lexical connotations, they exist if and only if the word is in relation with other words. For those unseen aspects of connotation in our visualization, they regard encyclopedic knowledge beyond the translation texts, the sociolinguistic aspects and emotive aspects, in which sociolinguistic aspects and emotive aspect are also less described part in the dictionary.

Finally, we can draw a temporary conclusion. For our purpose of supplementing a dictionary, co-occurrence pattern-based visualization is a better choice. On the other hand, misalignment-based visualization could show the variation among different translation versions. In future research, we can also apply misalignment-based visualization to translations of the *Kokinshū* in different languages and reflect the connotation unclear to other contemporary language users via the explanatory additions made in the translations.

7. Data Availability

Data can be found here: <https://github.com/nehcx/kokinMisalign/tree/master/data>

8. Software Availability

Software can be found here: <https://github.com/nehcx/kokinMisalign>; <https://cuckoo.js.ila.titech.ac.jp/~yamagen/waka/poem.cgi>

9. Acknowledgements

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2112.

10. Author Contributions

Xudong Chen: Conceptualization, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing

Yamamoto Hilofumi: Conceptualization, Project administration, Investigation, Supervision, Resources, Methodology, Software, Writing – original draft, Writing – review &

editing	711
Hodošček Bor: Conceptualization, Project administration, Resources, Data curation,	712
Software, Writing – original draft, Writing – review & editing	713

References 714

Allaway, Emily and Kathleen McKeown (2021). “A Unified Feature Representation for Lexical Connotations”. In: <i>Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume</i> . Online: Association for Computational Linguistics, 2145–2163. 10.18653/v1/2021.eacl-main.184 .	715 716 717 718
Bloomfield, Leonard (1933). <i>Language</i> . New York: H. Holt and Company.	719
Brown, Peter E, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer (1993). “The Mathematics of Statistical Machine Translation: Parameter Estimation”. In: <i>Computational Linguistics</i> 19.2, 263–311.	720 721 722
Burrows, J. (Sept. 1, 2002). “‘Delta’: A Measure of Stylistic Difference and a Guide to Likely Authorship”. In: <i>Literary and Linguistic Computing</i> 17.3, 267–287. ISSN: 0268-1145, 1477-4615. 10.1093/lc/17.3.267 .	723 724 725
Chandler, Daniel (2002). <i>Semiotics: The Basics</i> . London: Routledge. 10.4324/9780203166277 .	726 727
Chen, Xudong, Bor Hodošček, and Hilofumi Yamamoto (2022). “Tango Araitomo no Ayamari Taioo wo motiita Utakotoba no Konoteeshon Kenshutsu/Connotation Detection for Classical Poetic Japanese Vocabulary Using Word Alignment Mismatch[単語アライメントの誤り対応を用いた歌ことばのコノテーション検出]”. In: <i>Proceedings of Symposium for Humanities and Computer 2022 [人文科学とコンピュータシンポジウム 2022 論文集]</i> Vol. 2022.No. 1, 111–118.	728 729 730 731 732 733
Du, Keli, Julia Dudar, and Christof Schöch (2022). “Evaluation of Measures of Distinctiveness: Classification of Literary Texts on the Basis of Distinctive Words”. In: <i>Journal of Computational Literary Studies</i> 1.1. ISSN: 0000-0000. 10.48694/JCLS.102 .	734 735 736
Dunning, Ted (1993). “Accurate Methods for the Statistics of Surprise and Coincidence”. In: <i>Computational Linguistics</i> 19.1, 61–74. https://aclanthology.org/J93-1003 .	737 738
Eco, Umberto (1976). <i>A Theory of Semiotics</i> . Bloomington and Indianapolis: Indiana University Press.	739 740
Hjelmslev, Louis (1969). <i>Prolegomena to a Theory of Language</i> . Rev. Engl. ed., reprinted. Madison, Wisc: Univ. of Wisconsin Pr.	741 742
Hodošček, Bor and Hilofumi Yamamoto (2022). “Development of Datasets of the Hachidaishū and Tools for the Understanding of the Characteristics and Historical Evolution of Classical Japanese Poetic Vocabulary”. In: <i>Digital Humanities 2022 Conference Abstracts</i> . Tokyo: The University of Tokyo, 647–648.	743 744 745 746
Kalouli, Aikaterini-Lida, Rebecca Kehlbeck, Rita Sevastjanova, Katharina Kaiser, Georg A. Kaiser, and Miriam Butt (2019). “ParHistVis: Visualization of Parallel Multilingual Historical Data”. In: <i>Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change</i> . Florence, Italy: Association for Computational Linguistics, 109–114. 10/gh547n .	747 748 749 750 751
Kamens, Edward (1997). <i>Utamakura, Allusion, and Intertextuality in Traditional Japanese Poetry</i> . New Haven, CT: Yale University Press.	752 753


- Kaneko, Motoomi (1933). *Kokinwakashu Hyoshaku: Showa Shimban/An Annotated Kokinwakashu: The New Showa Edition* [古今和歌集評釈: 昭和新版]. Tokyo: Meijishoin. 754
755
- Katagiri, Yoichi (1983). *Utamakura utakotoba jiten zoutei ban/Dictionary of poetic vocabulary additional version*[歌枕歌ことば辞典増訂版]. Tokyo: Kadokawa Shoten. 756
757
- (1998a). *Kokinwakashu zen hyoshaku/A complete annotated edition of Kokinwakashu* [古今和歌集全評釈]. Vol. 1. Tokyo: Kodansha. 758
759
- (1998b). *Kokinwakashu zen hyoshaku/A complete annotated edition of Kokinwakashu* [古今和歌集全評釈]. Vol. 2. Tokyo: Kodansha. 760
761
- (1998c). *Kokinwakashu zen hyoshaku/A complete annotated edition of Kokinwakashu* [古今和歌集全評釈]. Vol. 3. Tokyo: Kodansha. 762
763
- Koehn, Philipp (2010). *Statistical Machine Translation*. 1st ed. New York: Cambridge University Press. 447 pp. 764
765
- Kojima, Noriyuki and Eizo Arai (1989). *Kokinwakashu* [古今和歌集]. Tōkyō: Iwanamishoten. 766
- Komachiya, Teruhiko (1982). *Kokinwakashu: Gendaigo Yaku Taisho/Kokinwakashu: With Modern Japanese Translations* [古今和歌集: 現代語訳対照]. Obunsha Bunko/Obunsha Series [旺文社文庫]. Tokyo: Obunsha. 767
768
769
- Kubota, Utsubo (1960a). *Kokinwakashu Hyoshaku/An Annotated Kokinwakashu* [古今和歌集全評釈]. Vol. 1. Tokyo: Tokyodo. 770
771
- (1960b). *Kokinwakashu Hyoshaku/An Annotated Kokinwakashu* [古今和歌集全評釈]. Vol. 2. Tokyo: Tokyodo. 772
773
- (1960c). *Kokinwakashu Hyoshaku/An Annotated Kokinwakashu* [古今和歌集全評釈]. Vol. 3. Tokyo: Tokyodo. 774
775
- (1994). “Kago no henshen/Transition in poetic vocabulary [歌語の変遷]”. In: *Gekkan Gengo/Laguage monthly*[月刊言語] 267, 58–65. 776
777
- Kyusojin, Hitachi (1979). *Kokinwakashu zen chushaku/Comprehensive annotations of Kokinwakashu* [古今和歌集全注釈]. Vol. 1. Kodansha gakujutsu bunko/Kodansha academic collection of Japanese literature [講談社学術文庫]. Tokyo: Kodansha. 778
779
780
- Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schutze (2008). *Introduction to Information Retrieval*. New York: Cambridge University Press. 781
782
- Matsuda, Takeo (1968a). *Shinshaku Kokinwakashu Hyoshaku: /A New Annotated Edition of Kokinwakashu* [新釈古今和歌集]. Vol. 2. Tokyo: Kazamashobo. 783
784
- (1968b). *Shinshaku Kokinwakashu Hyoshaku: /A New Annotated Kokinwakashu* [新釈古今和歌集]. Vol. 1. Tokyo: Kazamashobo. 785
786
- Matsumoto, Yuji, Akira Kitauchi, Tatsuo Yamashita, Osamu Imaichi, and Tomoaki Imamura (2002). *Morphological Analysis System ChaSen Version 2.2.9 Manual*. 787
788
- Mounin, Georges (1976). *Les problèmes théoriques de la traduction*. Paris: Gallimard. 789
- Nakano, Hiroshi, Ooki Hanashi, Hisao Ishii, Makoto Yamazaki, Masahiko Ishii, Yasuhiko Kato, Tatsuo Miyajima, and Akiko Turuoka (1994). *Bunrui goi hyo furoppi ban/Word List by Semantic Principles, floppy disk version* [分類語彙表 フロッピー版]. Vol. 5. okuritsu Kokugo Kenkyujo Gengoshori datashu/National Language Research Institute Language Resource [国立国語研究所言語処理データ集]. Tokyo: Dainippon Toten. 790
791
792
793
794
795
- Okumura, Tsuneya (1978). *Kokinwakashu* [古今和歌集]. Shincho nippon koten shusei/Shincho collection of classical Japanese literature [新潮日本古典集成]. Tokyo: Shinchosha. 796
797
798
- Osgood, Charles Egerton, George J. Suci, and Percy H. Tannenbaum (1957). *The Measurement of Meaning*. University of Illinois Press. 799
800



- Ozawa, Masao (1971). *Kokinwakashu* [古今和歌集]. Thirteenth. Nihon koten bungaku zenshu/the complete series of classical Japanese literature [日本古典文学全集]. Tokyo: Shogakukan. 801-803
- 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. Amsterdam: Rodopi, 247–269. 804-807
- Rashkin, Hannah, Sameer Singh, and Yejin Choi (2016). “Connotation Frames: A Data-Driven Investigation”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 311–321. 10.18653/v1/P16-1030. 808-811
- Rodd, Laurel Rasplica, Mary Catherine Henkenius, and Tsurayuki Ki, eds. (1996). *Kokinshū: A Collection of Poems Ancient and Modern*. 1st pbk. ed. C&T Asian Literature Series. Boston, MA: Cheng & Tsui Co. 812-814
- Rössler, Gerda (1979). *Konnotationen: Unters. Zum Problem d. Mit- u. Nebenbedeutung*. Zeitschrift Für Dialektologie Und Linguistik : Beihefte n.F., Nr. 29. Wiesbaden: Steiner. 815-816
- Schöch, Christof, Daniel Schlör, Albin Zehe, Henning Gebhard, Martin Becker, and Andreas Hotho (2018). “Burrows’ Zeta: Exploring and Evaluating Variants and Parameters”. In: *Book of Abstracts of the Digital Humanities Conference*. the Digital Humanities Conference. Mexico City: ADHO. <https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/>. 817-821
- Schramm, Wilbur Lang (1954). *The Process and Effects of Mass Communication*. Urbana: University of Illinois Press. 822-823
- Spärck Jones, Karen (1972). “A Statistical Interpretation of Term Specificity and Its Application in Retrieval”. In: *Journal of Documentation* 28.1, 11–21. ISSN: 0022-0418. 10.1108/eb026526. 824-826
- Stede, Manfred (1999). *Lexical Semantics and Knowledge Representation in Multilingual Text Generation*. Boston, MA: Springer US. 10.1007/978-1-4615-5179-9. 827-828
- Stubbs, Michael (2002). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford ; Malden, MA: Blackwell Publishers. 829-830
- Takeoka, Masao (1976a). *Kokinwakashu zen hyoshaku: Kochu nanashu shusei/A complete annotated edition of Kokinwakashu: with seven version of old annotations* [古今和歌集全評釈: 古注七種集成]. Vol. 1. Tokyo: Yubunshoin. 831-833
- (1976b). *Kokinwakashu zen hyoshaku: Kochu nanashu shusei/A complete annotated edition of Kokinwakashu: with seven version of old annotations* [古今和歌集全評釈: 古注七種集成]. Vol. 2. Tokyo: Yubunshoin. 834-836
- Voloshinov, V. N. (1986). *Marxism and the Philosophy of Language*. Trans. by Ladislav Matejka and I. R. Titunik. Cambridge, Mass: Harvard University Press. 837-838
- Yamamoto, Hilofumi (2005). “A Mathematical Analysis of the Connotations of Classical Japanese Poetic Vocabulary”. PhD thesis. Canberra: The Australian National University. 839-841
- (2007). “Waka no Tame no Hinshi Taguzuke Shisutemu/POS Tagger for Classical Japanese Poems [和歌のための品詞タグづけシステム]”. In: *Nihongo no Kenkyu/Studies in the Japanese Language* [日本語の研究] 3.3, 33–39. 10.20666/nihongonokenkyu.3.3_33. 842-845
- (2009). “Bunrui Kodo Tsuki Hachidaishu Yogo No Sisoraasu/Thesaurus for the Hachidaishu (ca. 905-1205) with the Classification Codes Based on Semantic Princi- 846-847

- ples [分類コードつき八代集用語のシソーラス]”. In: *Nihongo no Kenkyu/Studies in the Japanese Language* [日本語の研究] 5.1, 46–52. [10.20666/nihongonokenkyu.5.1_46](https://doi.org/10.20666/nihongonokenkyu.5.1_46). 848–849
- Yamamoto, Hilofumi and Bor Hodošček (2018). “A Study on the Distribution of Cooccurrence Weight Patterns of Classical Japanese Poetic Vocabulary”. In: *Proceedings of the 8th Conference of Japanese Association for Digital Humanities (JADH2018) “Leveraging Open Data”* 2018, 179–182. 850–853
- (2019). “An Analysis of the Differences Between Classical and Contemporary Poetic Vocabulary of the Kokinshu”. In: *The 9th Conference of Japanese Association for Digital Humanities (JADH2019) “Localization in Global DH”*, 68–71. 854–856
- (2021). *Hachidaishu vocabulary dataset*. [10.5281/zenodo.4744170](https://doi.org/10.5281/zenodo.4744170). (Visited on 05/17/2023) 857

conference version

What's that Scary Sound? Ambient Sound in Gothic Fiction

Svenja Guhr¹ 
Mark Algee-Hewitt² 

1. Institute of Linguistics and Literary Studies, Technical University of Darmstadt , Darmstadt, Germany.
2. English Department, Literary Lab, University of Stanford , Palo Alto, U.S..

Citation

Svenja Guhr and Mark Algee-Hewitt (2023). "What's that Scary Sound? Ambient Sound in Gothic Fiction". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-05-09

Date received 2023-01-26

Keywords

sound studies, ambient sound, Gothic fiction, 19th century, literary prose, English

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. This paper presents an approach to operationalizing ambient sound as a literary phenomenon. Applied to a use case in literary studies, we manually and then automatically detect ambient sound markers and use these annotations to analyze chosen 19th century English novels and short stories. Our hypothesis is that Gothic Novels, especially, contain many detailed descriptions of the story's ambient soundscape. We use a classification approach based on a state-of-the-art transfer learning algorithm and a domain-dependent fine-tuned BERT model for English to automatically detect word-level sound indicators and compare their occurrence over the course of the novel and with a comparative view on our corpus texts.

1. Introduction

"The rising **blast sighed** through the towering pines, which rose loftily above Matilda's head: the distant **thunder**, **hoarse** as the **murmurs** of the grove, in indistinct **echoes** mingled with the hollow **breeze**; the **scintillating lightning flashed** incessantly across her path, as Matilda, heeding not the **storm**, advanced along the trackless forest. [...] The **battling elements paused**: an **uninterrupted silence**, **deep**, **dreadful** as the **silence of the tomb**, succeeded. Matilda **heard a noise** – **footsteps** were distinguishable, and looking up, a flash of vivid lightning disclosed to her view the towering form of Zastrozzi." (Shelley, P. *Zastrozzi*)

Thunder, lightning, breezes, and echoes: the narrator describes the battling elements of nature that surround the protagonist Matilda in Shelley's Gothic novel *Zastrozzi*. The reader receives insights into the character's sensory experiences: seeing the lightning, hearing the thunder, feeling the breeze. Some of these descriptions provide the soundscape of the scene.

In this paper, we explore the function of sound in literary fiction. Applying sound studies to literary analysis, we systematically investigate indications of ambient sound in fiction with focus on its use in 19th century British Gothic Novels.

We hypothesize that Gothic Novels, especially, contain many detailed descriptions of the story's ambient soundscape (e.g., the growl of a wild animal, the creaking of a wooden floor), paying particular attention to sounds at either end of the loudness

spectrum – from deep silence to loud screams and clashing thunder. This article offers a new approach to operationalizing sound on the word level and introduces methods to manually and automatically detect ambient sound markers in English literary prose.

Our approach is as follows: First, we offer insight into the subfield of literary sound studies (see 2.1), and embed it in the research context of the Gothic (see 2.2). We then describe our operationalization of sound and detail the method we used to analyze it, drawing from manually and automatically generated annotations of a selected corpus of 19th century English novels and short stories. The methods we adopt for the analysis of sound include a dictionary approach to define a baseline for our analysis (see 3.3.1) as well as a transfer learning approach using the *NEISS TEI Entity Enricher* (Zöllner et al. 2021) (see 3.3). Our ability to accurately detect sound words automatically is evaluated through comparisons against manually annotated data. In the Sections 4 and 5, we examine sound references across a corpus of novels and short stories with particular interest in passages with a high density of sound words that contain particularly loud or low sound indications. Overall this study contributes to the young research field of literary sound studies, reflecting on the rising interest in the operationalization of sensual experiences with focus on sound in fictional prose using current distant reading methods.

2. Theoretical Background

2.1 Sound and Literary Studies

There is a great diversity of approaches towards sound in literary studies. On the one hand, sound can be perceived and analyzed as actually produced sound during the reading aloud of literary texts, where pronunciation, stress, pauses, speech rhythm are phonologically studied (Blohm et al. 2021). In their book chapter "Sound Shape and Sound Effects of Literary Texts" as part of the *Handbook of Empirical Literary Studies*, Blohm et al. (2021) describe an empirical approach to sound in literature. They claim that "reading a word automatically activates its abstract sound representation [...] referred to as phonological recoding [...] of written text" and "experience[d] as an 'inner voice'" (Blohm et al. 2021, 11). On the other hand, onomatopoeia, alliterations and rhymes can be stylistically analyzed, which constitute the sound design of literary texts and thereby shape the reading flow during both reading aloud and silent reading (Hinton et al. 1995, 1–10).

While both listed types are reception-oriented and rest on defining sound as physical vibrations transmitting information; however, in literary texts it can be more pragmatically interpreted. As a third type, sound in fiction itself can be examined in order to find out which sounds are described in the fictional world. We see this, for example, in Schafer (1994) who analyzes diachronically changing soundscapes (composite of 'sound' and 'landscape'). In addition to analyzing real world sonic environments as he did in his *World Soundscape Project*, Schafer (1994) also reads soundscapes through literary texts, claiming that "writers of fiction were reliable 'earwitnesses' whose writings 'constitute the best guide available in the reconstruction of soundscapes past'" (Schafer 1994, 9 in Snaith 2020, 20).

There have been numerous recent attempts in Literary Studies to analyze fictional soundscapes. Scholars focus on the descriptions of sound that aid the reader in imagining the particulars of the fictional world. For audionarratologists, the sound imagination is related to the reader's own memories of described sounds that "create the details of a storyworld in our heads [as] a 'theatre of the mind'" (Verma 2012, Mildorf 2019, 297). Similarly, Picker (2003) analyzes sounds in English literature through Dickens' soundscape descriptions of his fictional London. More recent examples can be found in the essay collection *Literature and Sound* (Snaith 2020).

According to the soundscape analysis of Döblin's *Alexanderplatz* by Bernhart (2008, 61), sounds are most frequently present in dialogue. The most frequent explicit indications of sound in prose occurs through character utterances and their descriptors. By extension, ambient sounds (e.g., those related to machinery or nature) are less often explicitly mentioned (Bernhart 2008, 62). Depictions of sensory experience evidence a hierarchy: standard descriptions of settings in literary texts focus almost exclusively on sight, e.g., describing a landscape or a living-room. Descriptions of sounds beyond dialogue are rare. Narrative seems focused on directing the imaginative gaze of the reader's eyes rather than directing the reader's mental ear with imaginative sounds (Smith 2015, 27–37). Nevertheless, the reader's sound perceptions play a role in their understanding of narrative.

The narrator draws on the reader's knowledge and experience for descriptive techniques. For example, in describing a train entering the station, the moving train is the focus. However, the narrator's description of the world is incomplete, relying on the reader's world knowledge to fill in the missing information about the furnishings of the station building or the color and model of the arriving train. The description of the station's soundscape is similarly omitted. When sounds do occur, it is because they differ from the surrounding default soundscape. As an example, Bernhart (2008, 61–62) emphasizes the sound of bells ringing in Döblin's *Alexanderplatz* by giving an explicit description of sounds that are not part of the reader's **default** experience of the setting.

2.2 Sound in Gothic fiction

Especially in the Gothic, readers are often placed in settings that deviate from the **default**. Known for its "decaying Gothic castles, ruined chapels, underground passages, dark forests and ghostly groaning" and for its "[s]hocks, supernatural incidents and superstitious beliefs", all of which "promote a sense of sublime awe and wonder [...] entwined with fear and elevated imaginations" (Botting 1996, 29, 46), the Gothic sets itself apart from the realistic literature of the 19th century and goes into the sensational, fantastic, and the uncanny (Bacon 2018, 1, Hurley 2002, 191).

Stylistically, it is rich of vocabulary of mystery, uncertainty, terror, horror, fear, and "hyperbolic language [...] [which] attempts to create a brooding, suspenseful atmosphere" (Hurley 2002, 191). In addition to the 'ghosts', 'phantoms', and 'wretches' described, depictions of non-human behavior, supernatural forces, or inexplicable events as in quotation (1) create mystery:

(1) "As I said this I suddenly beheld the figure of a man, at some distance, advancing towards me with **superhuman speed**. He bounded over the

crevices in the ice, among which I had walked with caution; his stature, also, 107
as he approached, **seemed to exceed that of man.**" (Shelley *Frankenstein*) 108

Furthermore, uncertainty is represented in conditional phrases, rhetorical questions or 109
actual questions asked by the characters ("I figured to myself," "I wondered if," "he 110
might") like in quotation (2) that characters try to solve by paying attention to details 111
in their environment. 112

(2) "I preternaturally listened; I figured to myself **what might** portentously 113
be; I **wondered if** his bed were also empty and he too were **secretly** at watch. 114
It was a deep, soundless minute, at the end of which my impulse failed. 115
He was quiet; he **might** be innocent; the risk was hideous; I turned away." 116
(James *The Turn of the Screw*) 117

In addition to the environmental details, Gothic texts also contain minute descriptions 118
of the fictional world through depicted "emotions [...]" by detailing the protagonist's 119
thoughts and feelings" (Ellis 2000, 9) and sensory experiences described by the narrator 120
(sight, smell, taste, touch, hearing) as in quotation (3). 121

(3) "In a few minutes after, I **heard** the creaking of my door, as if some 122
one endeavoured to open it softly. I **trembled** from head to foot; I **felt** a 123
presentiment of who it was and wished to rouse one of the peasants who 124
dwelt in a cottage not far from mine; but I was overcome by the sensation of 125
helplessness, so often **felt** in frightful dreams, when you in vain endeavour 126
to fly from an impending danger, and was rooted to the spot. Presently I 127
heard the sound of footsteps along the passage; the door opened, and the 128
wretch whom I dreaded appeared." (Shelley *Frankenstein*) 129

Quotation (3) is just one sample for the importance of auditory descriptions in the 130
Gothic. In his monograph on *Gothic Voices: The Vococentric Soundworld of Gothic Writing*, 131
Foley (2023) discusses how the Gothic atmosphere of horror, or suspense rely on the 132
soundscape: "creaking floorboards, howling winds and thunder rolling are just some of 133
the acoustic motifs that alert us to a Gothic atmosphere" (Foley 2023, 1). 134

Mysterious sounds and deep silence are common descriptions of the Gothic sound- 135
scape indicated through hearing events¹ introduced by the verbs 'listen' or 'hear' as in 136
quotation (3) or by sound words as 'scream', 'burst', 'cry', 'yell', in quotation (4). 137

(4) "A terrible **scream** — a prolonged **yell** of horror and anguish — **burst** 138
out of the **silence** of the moor. That frightful **cry** turned the blood to ice in 139
my veins." (Doyle *The Hound of the Baskervilles*) 140

In quotation (4) we can also observe an oscillation between especially loud and low 141
sound descriptions that corresponds to Hurley (2002, 9) reasoning on the Gothics contin- 142
uous "confrontations between the low and the high [...]" [or] other opposed conditions 143
— including life/death, natural/supernatural, ancient/modern, realistic/artificial, and 144
unconscious/conscious". 145

Later in this paper, we will show that Gothic texts often contain scenes within a silent 146

1. By using the term 'event' for narratological segments, we refer to the *event I* definition by Hühn (2013) and Gius and Vauth (2022, 3): "event I is any change of state and thus a general type of event without further requirements".

ambience in which sudden sounds appear, see 5.1. Finally, quotation (5) represents a summary of the enumerated particularities:

(5) "It's eleven o'clock **striking** by the **bell** of Saint Paul's. **Listen** and you'll **hear** all the **bells** in the city **jangling**.' Both sit **silent**, **listening** to the metal **voices**, near and distant, **resounding** from towers of various heights, in **tones** more various than their situations. When these at length **cease**, all **seems** more **mysterious** and **quiet** than before. One disagreeable result of **whispering** is that it **seems** to evoke an atmosphere of **silence**, **haunted** by the **ghosts** of **sound**, **strange cracks** and **tickings**, the **rustling** of garments that have no substance in them, and the tread of **dreadful** feet that would leave no mark on the sea-sand or the winter snow. So **sensitive** the two friends happen to be that the air is full of these **phantoms**, and the two **look** over their shoulders by one consent to **see** that the door is shut." (Dickens *Bleak House*)

The sensory experiences of the two characters are described ("listen/ing", "see", "look"). Known ("bell of Saint Paul's") and unknown ("strange cracks") sounds resound in the silent ambience of the city scene ("atmosphere of silence"). A vocabulary of mystery is employed ("mysterious", "ghosts", "phantoms") to trigger an atmosphere of uncertainty ("seems") and fear, prompting the reader's desire to know what happens next.

In the following, we will focus on sounds and show how the operationalization of this phenomenon, and consequently, the systematic annotation and automated detection of sound indications can be used to analyze the soundscape of the Gothic.

3. Method

Traditionally, research on sound in fiction has been conducted through close-reading methods. In our article, we present our distant reading approach as an alternative that can access disparate elements of a soundscape that are invisible to even the most careful reader. Our systematic analysis of ambient sound is based on a corpus of 19th century literary texts. We analyzed the corpus through a combination of common computational literary studies methods, namely manual and semi-manual annotation (Horstmann 2020) as well as automated annotation using a Transfer Learning Named Entity Recognition approach (Zöllner et al. 2021) evaluated on manually annotated data.

3.1 The Research Corpus

For our corpus, we collected 55 texts of different length (short stories, novellas and novels) based on a selection of 28 English novels and short stories that were mentioned in *The Handbook of the Gothic* (Mulvey Roberts 2009) enriched by 27 canonized 19th century English fictional prose.

The corpus texts (original texts as well as in-line sound annotated texts) are accessible as plain txt-files (in UTF-8) and XML files with TEI annotations (TEI Consortium 2022). Additionally, we provide a metadata table (xlsx) containing information on, i.a., text name, author name, author gender, publication year, text length in words, file names,

number of texts	55
number of texts written by female authors	19
number of texts written by male authors	36
shortest text (in words)	981
longest text (in words)	357.469
texts manually sound annotated	14
texts dictionary-based sound annotated (corrected false positives)	7
texts automatically annotated for sound	36

Table 1: Corpus Description.

annotation status, and more). For some statistical information about the corpus, see [Table 1](#).

3.2 Operationalizing Ambient Sound

To operationalize the phenomenon of ambient sound in literature we adopted the proven procedure of reflective text analysis of Pichler and Reiter (2020). Through an iterative manual approach, we systematically annotated ambient sound, at the word level, as a lexical unit.

To detect these lexical units, we distinguished between implicit and explicit sound indications, and refer to explicit sound descriptions as **concrete**: there is detailed information about the sound present in a scene from a semantically loaded sound word. Meanwhile, implicit sound description relies on the reader's imagination and interpretation related to a described event in a scene. See the following sample phrases:

a) Implicit Sound Description:

(6) "The train is entering the station."

The sentence is an example of an implicit sound description. As experienced readers who have heard a train entering a station, we know that the arrival of the train is associated with sounds. Nevertheless, in this sentence, sound is not annotated because it is not explicitly described with a lexical unit.

b) Explicit Sound Description:

The lexical unit becomes an explicit sound description when, for example, the action verb 'enter' is exchanged for the sound-indicating verb 'rattle', as in the sentence:

(7) "The train rattles into the station."

Here, the rattling sound of the arriving train is explicitly indicated on the word level of the literary text, so that the sound can be attached to a lexical unit – here the verb 'rattle'.

This word is then considered the annotation unit. An annotation in TEI would be:

(8) "The train <sound>rattles</sound> into the station."

Particular Annotation Cases

On some occasions, human sounds are also part of the ambient soundscape. This is the case, for example, when the scream of a woman is depicted, or when the sound of a crowd singing or rumbling is mentioned. In these cases the human-made sound does not bear a communicative purpose: it does not convey a verbal message of an identifiable speaker to a specific addressee.

Some ambient sound indications are not annotated. For example, sound descriptions

author	year	title	words	sw	swd	man/dic
Brontë	1847	<i>Jane Eyre</i>	188.598	604	0.32	dic
Brontë	1847	<i>Wuthering Heights</i>	119.475	351	0.29	dic
Byron	1819	<i>Fragment of a Novel</i>	1.977	1	0.05	man
Doyle	1898	<i>The Brazilian Cat</i>	8.148	65	0.80	man
Dickens	1848	<i>A Christmas Carol</i>	29.243	194	0.66	man
Gaskell	1852	<i>The Old Nurse's Story</i>	9.805	66	0.67	man
M.R. James	1895	<i>Canon Alberic's Scrap-Book</i>	4.716	20	0.42	man
M.R. James	1904	<i>The Mezzotint</i>	4.682	0	0.0	man
Kipling	1890	<i>The Mark of the Beast</i>	5.109	37	0.72	dic
Lewis	1808	<i>The Anaconda</i>	18.996	75	0.39	man
Oliphant	1881	<i>The Open Door</i>	18.763	161	0.86	dic
Potter	1902	<i>The Tale of Peter Rabbit</i>	981	10	1.02	man
Shelley, P.	1818	<i>Zastrozzi</i>	30.971	229	0.74	dic
Trollope	1875	<i>The Way we live now</i> (Ch.1-10)	35.895	12	0.03	man
Wells	1897	<i>The Invisible Man</i>	49.808	385	0.77	dic
Wilde	1891	<i>The Picture of Dorian Gray</i>	80.396	288	0.36	dic
Yonge	1853	<i>The Heir of Redclyffe</i> (Ch.1-10)	59.774	61	0.10	man
total			1.323.574	4.139	0 0.31	

Table 2: Texts of the different training sets: Manually annotated or dictionary annotated with manual false positive correction. Indicated are the total number of words, the number of sound words (sw), the calculated swd, and how it was annotated.

referring to iterative events, generalizations and regularities, or references to sounds realized in the past are not included into the annotation (e.g., “the bells always ring at noon time” (generalization, regularity), “they often sang the Requiem at funeral services” (regularity, past), here ‘ring’ and ‘sang’ do not indicate a realized sound in that particular scene). Consequently, only sounds that can be diagetically related to events in the corpus were tagged. This also excludes negated sounds (e.g., “the bell did not ring today”), as well as the pronunciation of wishes and conditional statements (e.g., “Oh, that some encouraging voice would answer in the affirmative!” (Shelley *Frankenstein*). The situation is different, however, for often explicitly described silence (e.g., “there was a peaceful silence over the misty morning landscape”). Generally, silence is treated as absence of loud sounds resulting in a calm soundscape. Our decision to tag these passages does not mean that we simply felt that there were no sounds: rather the language of text flags the complete absence of sounds by referring explicitly to silence or to sounds that are realized at a very low volume level that cannot be perceived by humans.

Manual Annotation

On the basis of our preceding operationalization of ambient sound, we formulated annotation guidelines for the manual annotation of ca. 25% of our corpus. Three annotators (trained in both literary studies and annotations) manually annotated a total of 14 texts of varying length from the corpus following the *Guidelines for Ambient Sound Annotation* (Guhr 2023). The annotation guidelines were developed following an iterative process according to Reiter (2020). In a manual annotation of Lewis’ *The Anaconda* by two annotators, we received an inter-annotator-agreement of 0.80 Cohen’s *kappa* (Scikit-learn Developers 2022), which is considered to be a decent agreement for a manual annotation task of literary phenomena but also indicates the complexity of this task for human readers. Four of the 14 manually annotated texts form the test set.

author	year	title	words	sw	swd
Crookenden	1802	<i>The Vindictive Monk</i>	7.672	16	0.21
Doyle	1913	<i>How it happened</i>	1.429	11	0.77
Doyle	1902	<i>The Hound of the Baskervilles</i>	59.931	125	0.21
Shelley, M.	1818	<i>Frankenstein</i>	75.235	254	0.34
total			144.267	406	0.28

Table 3: Test Set : Manually annotated corpus texts. Indicated are the total number of words, the number of sound words (sw), and the calculated swd.

3.3 Approaches for Automatizing the Annotation of Ambient Sound 247

In order to automate the annotation of ambient sound descriptors, we compared two 248
approaches: a simple dictionary approach that consequently served as the baseline for 249
automated annotation (see 3.3.1), and a classification approach based on a state-of-the- 250
art transfer learning algorithm and a BERT language model (see 3.3.2). 251

3.3.1 Dictionary approach 252

To determine a baseline for automated ambient sound annotation, we applied a simple 253
dictionary approach. After lemmatizing the manually annotated training texts (see 254
Table 2) using the NLTK (Loper and Bird 2002), we extracted the unique sound word 255
lemmas. We then took these lemmas and found matches to them in a lemmatized set 256
of texts resulting in a dictionary with a key-value pair {'lemma' : 'sound annotation'}. 257
After each of three rounds of annotations, the sound word list was refined based on 258
discussions among the annotators and subsequent updates to the guidelines. In each 259
new round, a smaller list of sound words were extracted. Starting from 289 sound words 260
in the first round, only 258 sound word lemmas were left in the second round, and only 261
228 in the third round. 262

Error Analysis of the dictionary approach 263

In a next step, we used the dictionary approach to automatically annotate Doyle's short 264
story *How it Happened* comparing the results with our manual annotation of 11 total 265
sound words. The evaluation results can be found in Table 5. Comparing the first round 266
(19 false positives, 2 false negatives) to the second round (12 false positives, 2 false 267
negatives) and the third round (2 false positives, 9 false negatives), we see a decrease 268
in false positives as the sound word list is revised; however we also see a rise in false 269
negatives. Consequently, we observe that the dictionary approach has high accuracy 270
and recall results due to the generalization of the annotation process that tends to be 271
oversensitive to words that could lexically be sound words but do not indicate actual 272
sounds in the diegesis or are homographs of non-sound words (see *Particular Annotation* 273
Cases in 3.2). Looking at the false negatives across all rounds, which consequently were 274
not part of the sound word lists, we can see the genre and time span dependence of our 275
dictionary approach. For example, in all rounds, "whir" was left out in the annotation, 276
which is not treated explicitly in any other of the 19th century training texts. 277

In summary, the dictionary approach is useful to detect mentions of sound words on the 278
word level; however, it does not take context into account, resulting in a high number of 279
false positives. 280

sound word (lemma)	absolute frequency in training set
sound	44
silence	18
cry	17
voice	15
wept/weep	19
silent	14
loud	13
thunder	10
step	10
scream	10
groan	8
calm	8
stillness	6

Table 4: The 13 most frequent sound words appearing in the training set.

	1 st round	2 nd round	3 rd round
sw lemmas	289	258	228
accuracy	0.98	0.99	0.99
precision	0.32	0.43	0.5
recall	0.81	0.81	0.82
F1-score	0.46	0.56	0.62

Table 5: Table with Evaluation Results of the Dictionary Approach (baseline).

3.3.2 Classification with NEISS NTEE

281

To get context sensitive annotations of ambient sound words in our corpus, we adopted
a classification approach based on a state-of-the-art transfer learning algorithm and a
BERT language model from *NEISS TEI Entity Enricher* by Zöllner et al. (2021).

282

283

284

In our approach, we followed the findings from earlier studies on generalized named
entity recognition for the detection of abstract entities such as places and spaces or
character gender in German language novels (Flüh et al. 2022; Schumacher 2022). Both
approaches employed the open access and open source software *Stanford Named Entity
Recognizer* (*StanfordNER*) (Manning et al. 2014), that was originally trained to detect
named entities in a narrow sense, namely, names of people, organizations or places,
using a conditional random field algorithm (Finkel et al. 2005), and then fine-tuned the
model on manually annotated data (Schumacher 2022, 79–93).

285

286

287

288

289

290

291

292

Recent approaches to entity detection use the software *NEISS TEI Entity Enricher* (Zöllner
et al. 2021) to fine-tune recent pre-trained models with manual annotations (Schumacher
et al. 2022). Based on a transfer learning (Kamath et al. 2019) approach, the tool provides
access to large-scale language models like the *Bidirectional Encoder Representations from
Transformers* (BERT) architecture by the *Hugging Face* (Devlin et al. 2018). Using this
method, Flüh and Lemke (2022) recognized named entities in German language letters
from the 19th and 20th century.

293

294

295

296

297

298

299

In our study, we used the pre-trained BERT model provided in the software *NEISS NTEE*
(originally the English language model (*bert-base-cased*) by the *Hugging Face* (Devlin
et al. 2018)) and fine-tuned it using a ground truth based on our manual annotations of
ambient sound words, see 3.2. For the prediction step, the software takes a non-labeled

300

301

302

303

literary text in XML-format as input and automatically annotates it with the XML-tag `<sound>TOKEN</sound>`. 304
305

After each of five training round evaluations, we adapted the training data and tried 306
different combinations of manually sound-annotated texts. As part of the ground truth 307
building in *NTEE*, the training data was split into training, validation, and test data. 308
Furthermore, one advantage of *NTEE* is the "Shuffle By Sentence" option to train and 309
evaluate the performance of the trained entity tagger independently of the text type 310
(novel or short story), using sentence-by-sentence training and evaluation. Presumably, 311
a set of meaningful segments (like events or scenes) would be more useful for shuffling 312
over the data than shuffling sentences (Lemke 2022); however, such automatic segmen- 313
tation is not yet advanced enough to be integrated into the software training process 314
(e.g., Zehe et al. 2021). 315

Combining the Dictionary Approach with the Automatic Prediction 316

As we saw in the training step, the prediction performance (EF-1 score) rose with the 317
amount of training data; however, manual human annotation is a costly task (Gühr 318
and Gius 2023). We therefore integrated our dictionary approach to aid in generating 319
new training data, calculating the frequency of sound words for each text in the corpus 320
from which we selected seven additional texts with a high incidence of sound words 321
for annotation. We then used our dictionary approach to annotate them (see 3.3.1) and 322
manually corrected the annotations by removing false positives. 323

Through an error analysis of these semi-automatically predicted sounds, we discovered 324
the following groups of false positives: 325

1. Human communication that is not explicitly non-verbal or verbal: 326
"crying to get free". 327
2. Sound related to human communication: "louder voice", or "chattering". 328
3. Sounds related to human communication that were edge cases: 329
"But while she hesitated what to do, she heard a `<sound>voice</sound>` at the 330
door requesting admission [...]." In this sample, it is uncertain whether the 'voice' 331
should be annotated as communication given the ambiguity of what is said or 332
who says it. 333
4. Acceptable annotations missed by human annotators or edge cases: 334
"trampled" is comparable to "stamping". 335
5. Adjectives and adverbs that indicate properties of sounds: 336
"A `<sound>piercing</sound>` `<sound>shriek</sound>` of horror `<sound>` 337
burst`</sound>` from me!" In the sample "piercing" is a false positive, but could 338
be annotated as sound-indicating property of "shriek". One has to distinguish 339
between properties relating to descriptive sound properties (namely, "loud", 340
"calm") and judgmental properties without direct relation to the property of a 341
sound like "beautiful", "charming", "violent" that indicates the perception from a 342
given narrative perspective. In a revision of the annotation guidelines, we explicitly 343
excluded non-sound-indicating properties. 344
6. Negated sounds: "she heaved not one sigh". 345

	texts	words	unl. words	sw	tr. ep.	b.ep.	E-F1
training set 1	6	73.027	72.656	371	4	4	0.6753
training set 2	11	285.745	284.088	1.657	11	8	0.7403
training set 3	13	640.533	637.919	2.614	13	9	0.7007
training set 4	13	640.531	638.103	2.428	12	10	0.7157
training set 5	15	656.077	653.583	2.494	8	6	0.6589

Table 6: Table with Evaluation Results of the training rounds partly with added manually corrected dictionary-based annotated data.

	tr. set 1	tr. set 2	tr. set 3	tr. set 4	tr. set 5
SeqEvalF1	0.6804	0.5470	0.6327	0.7083	0.6061
Precision	0.6226	0.4384	0.5741	0.6538	0.5454
Recall	0.75	0.7272	0.7045	0.7727	0.6818
F1-score	0.6804	0.5470	0.6327	0.7083	0.6060

Table 7: Table with test set evaluation results of the training partly with added manually corrected dictionary-based annotated data.

7. Hypothetical sound, subjunctives, wishes: 346
 "I might have cried, but I didn't.", "I guess, she will cry.", "I wish I could cry." 347
8. Sounds in the past: "Last night I heard a woman screaming." 348
9. Polysemy: 'ring' that can either be the sound of a bell or jewelry. 349

By focusing on the correction of false positives in a dictionary-annotated set, we were 350
 able to reduce the labor of manual annotation while still creating a more robust training 351
 set of true positives. 352

Adding these semi-annotated texts to the training set resulted in higher E-F1 scores on 353
 the split evaluation set and on the test set (see Table 6, training set 2 and 3). Training 354
 set 4 with 13 manually and semi-automatically annotated training texts received the 355
 best results on the split evaluation set (0.7157 E-F1) as well as on the test set (0.7083 F1) 356
 (see Table 7). However, the addition of two additional short stories to that training set 357
 (set 5) did not improve the evaluation results and so we elected to stop adding training 358
 data to our corpus. 359

Evaluation of the Prediction 360

To evaluate the predictions, we used the provided evaluation option of the *NEISS NTEE* 361
 software, namely, a sequence labeling evaluation (E-F1) based on the *seqeval* Python 362
 framework: "Instead of computing a token- or word-wise F1 score, E-F1 evaluates a 363
 complete entity as true positive only if all tokens belonging to the entity are correct" 364
 (Nakayama 2018). Using this framework, we calculated the E-F1 score first on the 365
 basis of the validation set (part of the split training set), second on the chosen test 366
 set additionally indicating precision, recall, F1-score (see Table 7). In comparison to 367
 the dictionary approach (see 3.3.1), F1-score performance improved by 0.1 using this 368
 method. Looking at the F1-score and the E-F1-score calculations of training set 2 and 4, 369
 it is interesting to note that when given more annotated data, training set 4 received a 370
 lower E-F1-score on the split evaluation set and a higher F1-score on the independent 371
 test set. This contradiction, between the better performance of the model overall and 372
 its lower performance on the split training set, may be explained by the semi-manually 373
 annotated texts, which contain many false negatives (see Table 7). After completing the 374

	loudness level	example
0	no annotation	-;-
1	non-audible sounds	<i>silence</i>
2	low sounds	<i>rustling</i>
3	normal indoor volume	<i>snoring</i>
4	loud sound	<i>thunder</i>

Table 8: Loudness Levels.

training and selecting the model with the best predictive performance according to the 375
evaluation, we used the model (best epoch) to automatically annotate the remaining 376
texts in the corpus. 377

3.3.3 Measurement of Sound Word Density 378

To compare several texts by their use of sound words, we adapted a method to measure 379
a text’s sound word density, developed in the dissertation work of the main author for 380
regarding character sounds (comparable to Schumacher (2022, 127)). The calculation 381
normalizes occurrences over text length: the number of annotated sound words *sw* is 382
divided by the total number of tokens *t*, and multiplied by 100. Sound word density 383
(swd) scores can be found in Table 2 and Table 3. 384

$$\text{swd} = \frac{n_{sw}}{n_t} \cdot 100 \quad (1)$$

Using swd measurements, we can compare the tendency to represent ambient sound 385
between, for example texts of different periods, genres, or authors. At finer scales, this 386
measurement can also be used to compare individual segments of text. 387

3.3.4 Loudness Level Labeling 388

In addition to measuring the density of the occurrence of sound words, these descriptors 389
can also be used to approximate loudness levels. To generate these levels, two annotators 390
manually annotated the 228 word dictionary of sound words with loudness levels as 391
follows: 1 for non-audible sounds, e.g. *silence*; 2 for low sounds, e.g. *rustling*, 3 indicating 392
normal indoor volume, e.g. *snoring*, 4 for loud sounds, e.g. *thunder*, *screaming*, *explosions*. 393
Based on the inter-annotator-agreement of this set (0.71 Cohen’s *kappa* (Scikit-learn 394
Developers 2022)) it is clear that for even human readers labeling loudness of context- 395
free sound words is a non-trivial problem. Although the two annotators were able to 396
compromise on an agreed-upon set. We applied the loudness levels to our automatically 397
tagged texts by mapping detected sound words to the annotators’ loudness values. 398

4. Analysis: Loudness in the Gothic 399

4.1 Mapping the Manual Annotations 400

We visualized the manual annotations of sound and those enriched with loudness levels 401
of each token in the text. 402

Lewis *The Anaconda*

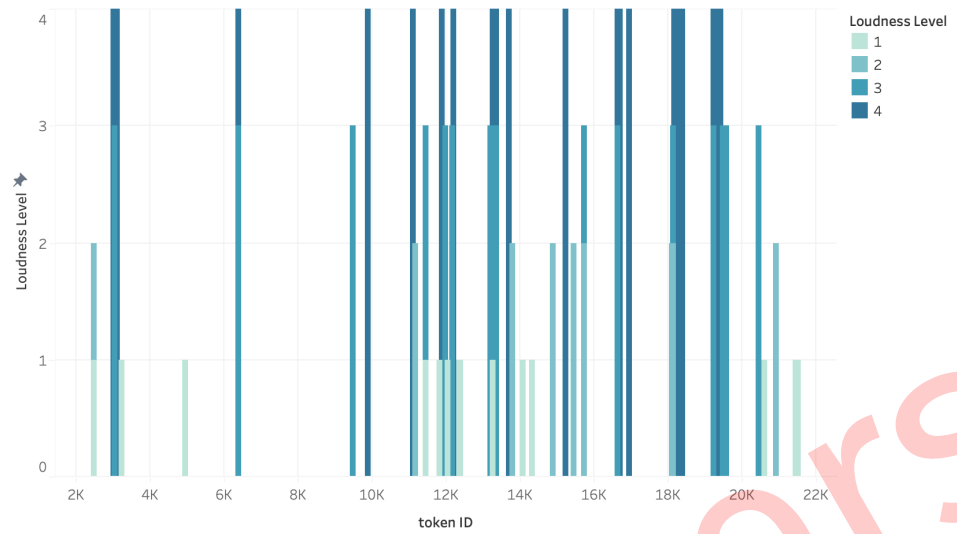


Figure 1: The columns indicate the loudness level of each sound word in Lewis *The Anaconda* (from 1 (silence) to 4 (loud sound)).

Gaskell *The Old Nurse's Story*

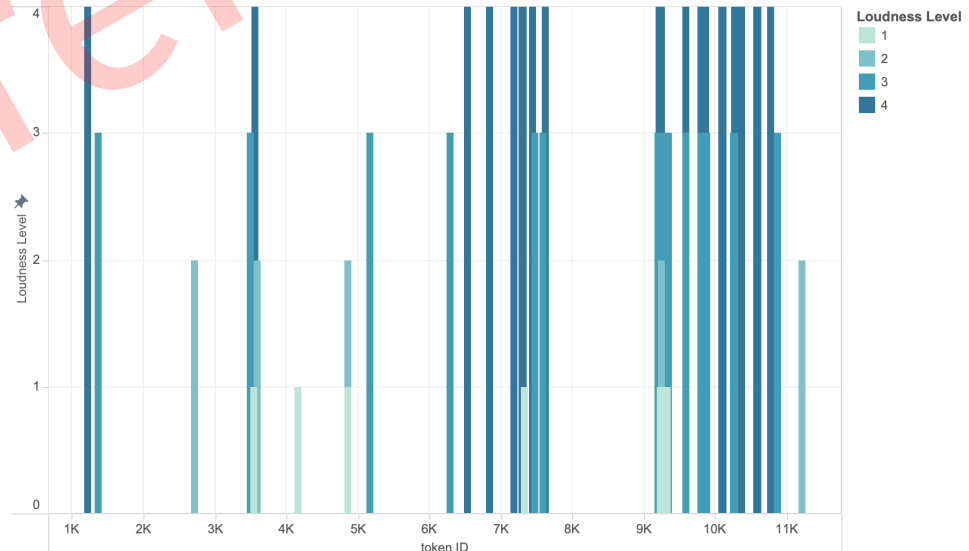


Figure 2: The columns indicate the loudness level of each sound word in Gaskell *The Old Nurse's Story* (from 1 (silence) to 4 (loud sound)).

In Figure 1, we visualized the manual annotations of Lewis' Gothic short story *The Anaconda*. Of note are the clusters of loud sounds (represented by the bars), which tend to occur together, interspersed with periods of quiet or references to absolute silence. The majority of the story describes an encounter with an anaconda, which the characters attack at discrete moments of the text. These conflicts are captured by the two clusters of loud sounds towards the end of the text and represent the final attempt to kill the creature:

"But on a sudden a loud and rattling rush was heard among the palms, and with a single spring the snake darted down like a thunder-lap and twisted herself with her whole body round her devoted victim. [...] We all at once attacked her, and she soon expired under a thousand blows" (Lewis *The Anaconda*).

In Figure 2, the distribution of the sound words over the course of the short story is even more clustered. Interestingly, loud sound indications are particularly frequent in the second half of the story and correlate with scenes² that are particularly suspenseful.

(9) "One fearful night, just after the New Year had come in, when the snow was lying thick and deep; and the flakes were still falling – fast enough to blind any one who might be out and abroad – there was a great and violent noise heard, and the old lord's voice above all, cursing and swearing awfully, and the cry of a little child, and the proud defiance of a fierce woman, and the sound of a blow, and a dead stillness, and moan and wailing dying away on the hill-side!" Gaskell *The Old Nurse's Story*

Quotation (9) is the beginning of the cluster (around tokens 9.000- 11.000) that we can identify in 2 indicating a detailed description of the fictional soundscape. The implied silence of the snowfall is interrupted by "a great and violent noise," "cursing," "swearing," and a "cry". These punctuated loud sounds within a short interval increase the suspense of the scene bringing the story's plot to a climax. The characters, and reader, however, have already encountered these loud sounds earlier in the text (around tokens 6.500- 7.500), where they may foreshadow the catastrophe.

Finally, the visualization of Doyle's *The Hound of the Baskervilles* shows relatively few sound words over the course of the plot. When they do occur, however, they appear clustered around passages that also provide Gothic features like vocabulary of mystery and fear. This is particularly the case in the passages with the highest density of sound words: ((a) swd 3.13: 23 sw on 734 words in the penultimate cluster; (b) swd: 2.9: 18 sw on 621 words in the final cluster). These are both significantly higher than the average swd of the entire text (Ø-swd: 0.21). The actual passages reveal that the clusters of sound words coincide with two scenes that bracket the climax of the story. The last passage, containing a high density of sound words, contains the novel's denouement: the appearance and killing of the mysterious hound.

2. By using the term 'scene' for narratological segments, we refer to the scene definition by Zehe et al. (2021): "From a narratological point of view, a scene can be defined by reference to a set of four dimensions: time, space, action and character constellation. Using these dimensions, a scene is a segment of the *discours* (presentation) of a narrative which presents a part of the *histoire* (connected events in the narrated world) such that (1) time is equal in *discours* and *histoire*, (2) place stays the same, (3) it centers around a particular action, and (4) the character constellation is equal."

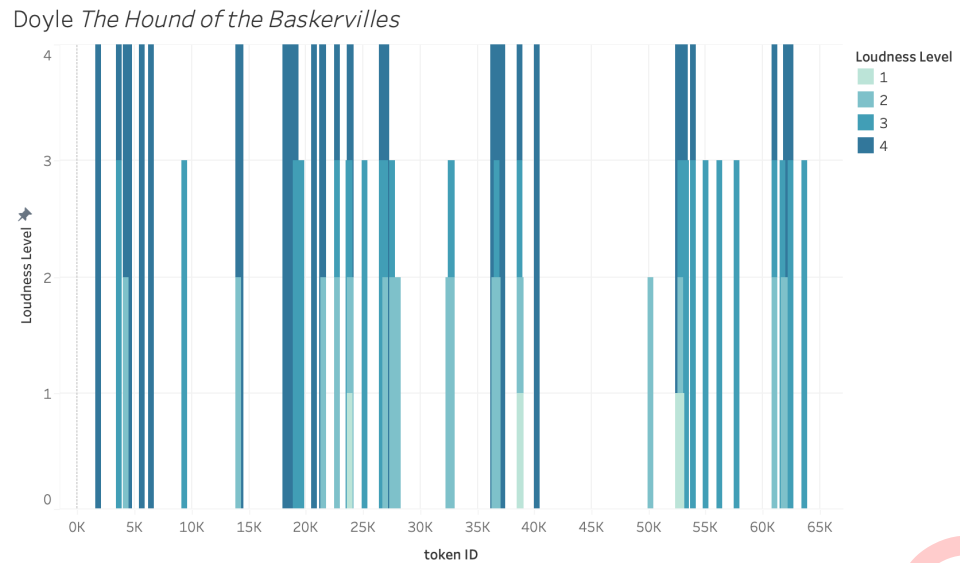


Figure 3: The columns indicate the loudness level of each sound word in Doyle *The Hound of the Baskervilles* (from 1 (silence) to 4 (loud sound)).

4.2 Comparing Sound Word Density

We compared the sound word density (see 3.3.3) in the automatically annotated corpus texts to determine if Gothic texts had a significantly higher swd than texts labeled as “other”, such as city or romance novels.

Despite the normalization of the measure, there was still a strong negative correlation (-0.4685) between swd and length at the extremes of the corpus size (an increase of 10 sound words has a much larger effect on a text of 1.000 words than one of 1.000.000 words). To counteract this bias, we divided the corpus into two subcorpora based on whether the texts were longer or shorter than 100.000 words, which lowered the correlation per subcorpus to between -0.17 and -0.19.

The plots of swd in the corpus texts demonstrate the categorical difference between the density of sound words in longer and short texts, with a mean swd of 0.471 in short texts and a mean swd of 0.26 for longer texts (note the difference in axis scale).

In the plot of longer texts (see Figure 4), one can recognize that texts labeled as “other” have a lower sound word density than the Gothic novels. Romance and city novels, like Austen’s *Sense and Sensibility* or Trollope’s *The Way we live now*, have an especially low sound word density with More’s *Coelebs in Search of a Wife*, whose subtitle promises “Observations on Domestic Habits and Manners, Religion and Morals,” having the lowest value. The two long texts with the highest sound word density are Dickens’ *Bleak House* - an originally serialized novel, and Corelli’s *The Sorrows of Satan* - a late 19th century horror novel. Both are labeled as Gothic texts. Interestingly, Mitford’s *Atherton and Other Tales* has a high sound word density, although this could be because it contains a series of shorter tales and consequently may be more comparable to the texts represented in Figure 5, where its sound word density of 0.35 is less than the swd mean.

The short texts do not show such a clear difference in swd between Gothic and “other”

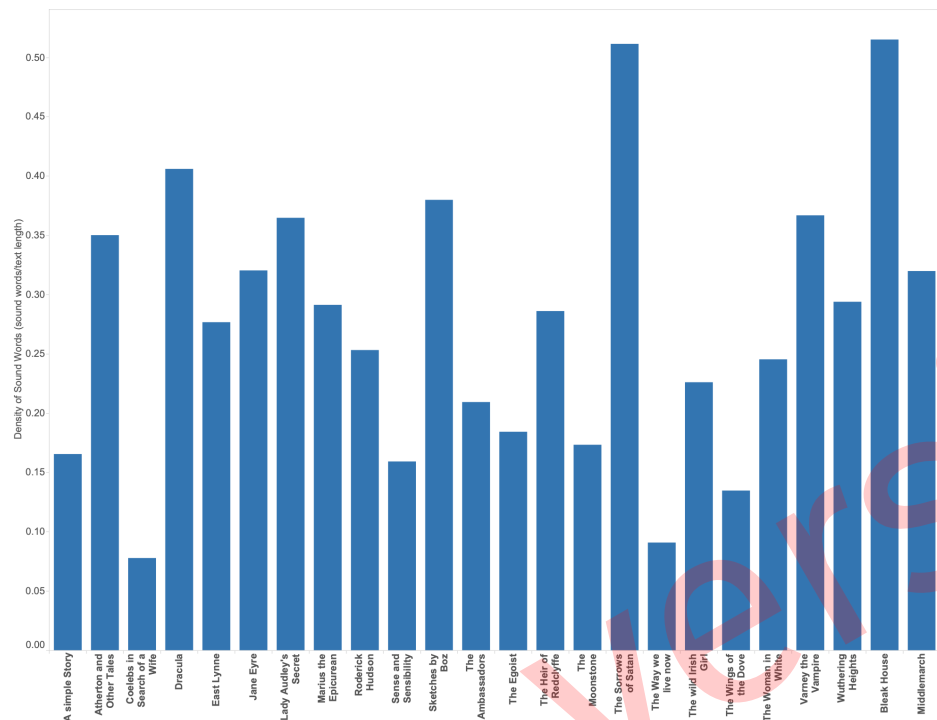


Figure 4: Swd of the long corpus texts (> 100.000 word tokens). The scale differs from the one in Figure 5.

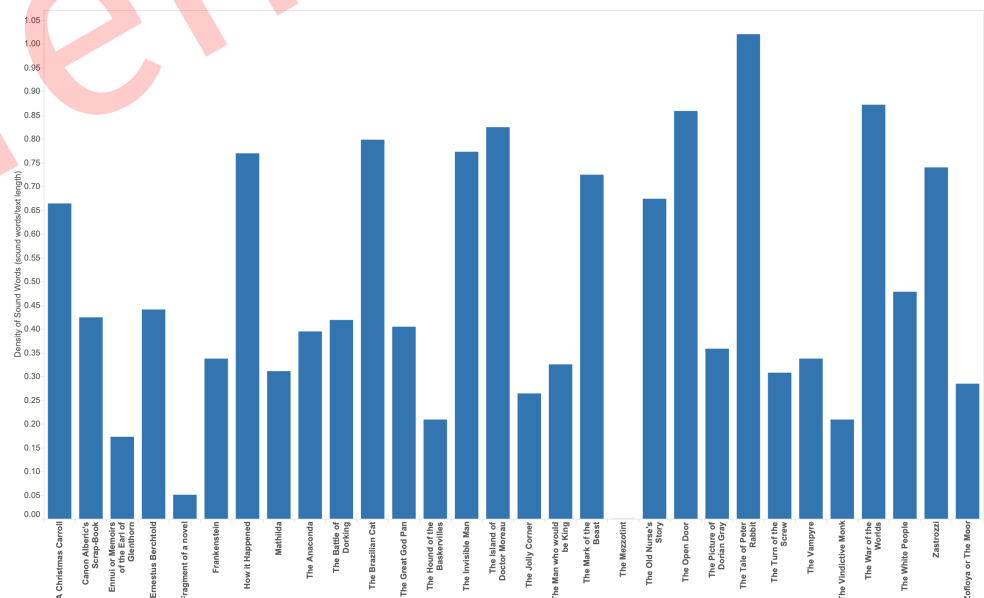


Figure 5: Swd of the short corpus texts (< 100.000 word tokens). The scale differs from the one in Figure 4, because shorter texts have the tendency to have a higher sound word density that is related to its text length.

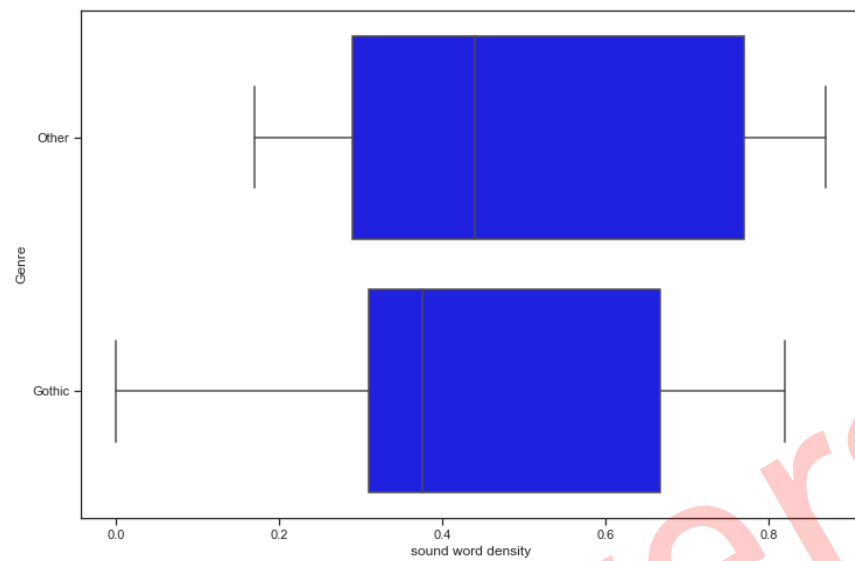


Figure 6: Swd of the short corpus texts (< 100.000 word tokens) ordered by genre "Gothic" or "other".

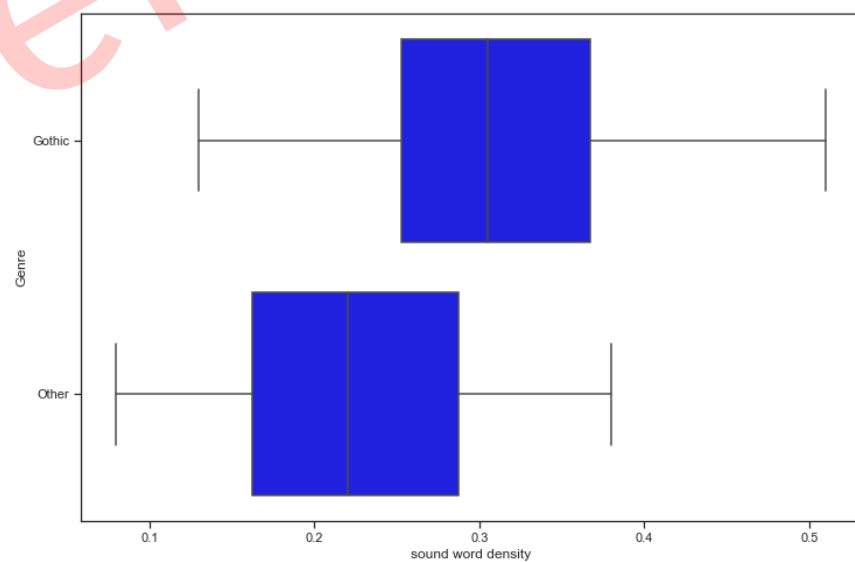


Figure 7: Swd of the long corpus texts (> 100.000 word tokens) ordered by genre "Gothic" or "other".

texts. Nevertheless, there are outliers with a particularly high sound word density such as Oliphant's *The Open Door*. However, also Well's *The War of the Worlds* that is labeled as "other" (Science Fiction) shows a high sound word density. There are, however, also Gothic texts with only one detected sound word like Byron's vampyre story *Fragment of a Novel* or even no sound word at all as in M.R. James' *The Mezzotint*). In contrast, Potter's children's story *The Tale of Peter Rabbit* shows the highest sound word density of all corpus texts and simultaneously is also the shortest corpus text followed by Doyle's *How it happened* – the second shortest text that has a slightly lower sound word density than the other short texts.

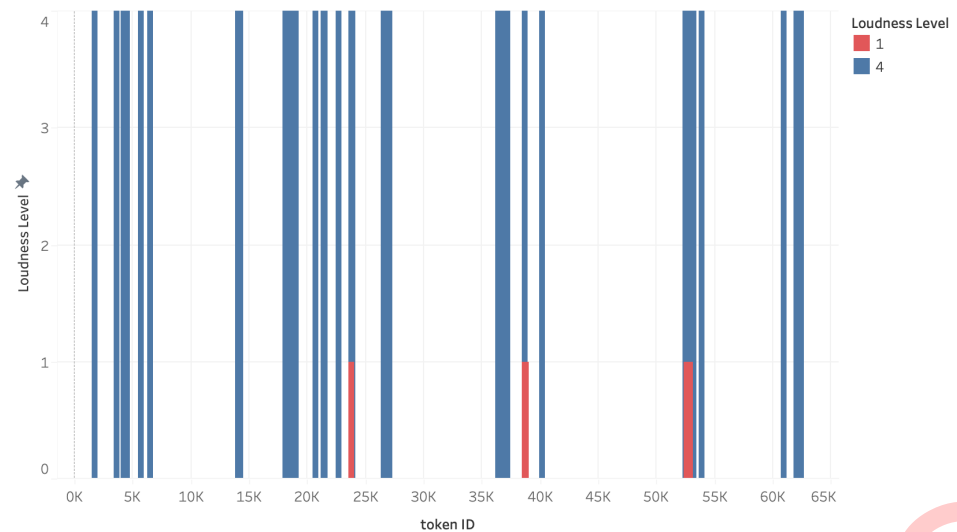
5. Discussion

As we stated in the introduction (see Section 1), we are particularly interested in whether Gothic texts have more detailed ambient sound descriptions than texts labeled as "other", e.g. city or romance novels. As our sound word density analysis in 4.2 shows, there does seem to be a relationship between the Gothic genre and a high density of ambient sound words. Similarly, we found that passages that have a higher sound word density indicate important passages for the plot as we could see in the sample scenes mentioned in 4.1 in which we could detect the climax of the plot by looking at the sound word distribution. In the following section, we will discuss these findings with an eye towards the distribution of loud sounds, as well as the role that silence plays in Gothic fiction.

5.1 Silence versus Loud Sounds

As we explained in the discussion on the operationalization of sound in literary texts (see 3.2), *silence* is a particular sub-phenomenon of the ambient soundscape. There must be a differentiation between the explicit indication of absolute silence and the absence of any sound indication because the latter does not indicate the absence of sound in the fiction. Rather, it plays with the imagination of the reader, offering gaps in the narration that trigger the reader to fill them with world knowledge of an expected soundscape according to the given scene setting. Consequently, only explicitly indicated *silence* denotes the explicit absence of sounds that can be received by the characters.

From these results, we might conclude that silence often sets a scene, as it involves a sustained period featuring the explicit absence of sounds perceptible to humans. Loud sounds, by contrast, flag events that occur spontaneously and irregularly, interrupting this state of silence. The contrast between silence and loudness amplifies the effects of sound, thereby increasing its effects on the reader. An effect of this pattern is that several events together convey a loud ambiance, while the silent initial state is often mentioned only once and therefore also has only a small effect on a passage's mean loudness (see Figure 8). Doyle's *The Hound of the Baskervilles* offers an important example of this scene-setting process. The denouement of the novel starts at token 52.548 with "A terrible scream – a prolonged yell of horror and anguish – burst out of the **silence** of the moor". Here, the silence of the moor is interrupted by a "scream" that is described with words typical for the Gothic vocabulary (like "terrible", "horror", "anguish"). The interruptions oscillate quickly between loudness levels, disorienting both the characters and the reader, and mixing the uncertainty of the passage with moments of surprise in

Doyle *The Hound of the Baskervilles*Figure 8: Doyle *The Hound of the Baskervilles* – silence and loud sounds.

a constant play of tension and release.

510

Gaskell's *The Old Nurse's Story* has few moments of silence but, the twice that it does occur, it sets up a sequence of suspenseful scenes that begin with general silence or at least a quiet ambiance. Instead of explicit representations of silence, Gaskell's text works through absence and negation, describing the missing sounds from an environment. These do not indicate absolute silence, but have a similar effect on the soundscape of a scene. For example, in a key scene, it is the *absence* of the expected sounds that should occur when a character is crying and battering her hands against the window-panes that creates the uncertainty and tension:

511

512

513

514

515

516

517

518

(10) "[A]ll of a sudden, she cried out, "Look, Hester! look! there is my poor little girl out in the snow!" I turned towards the long narrow windows, and there, sure enough, I saw a little girl [...] crying, and beating against the window-panes, as if she wanted to be let in. [...] all of a sudden, and close upon us, the great organ pealed out so **loud** and **thundering**, [I]t fairly made me tremble; and all the more, when I remembered me that, even in the stillness of that dead-cold weather, I had heard **no sound** of little battering hands upon the windowglass, although the phantom child had seemed to put forth all its force; and, although I had seen it wail and cry, **no faintest touch of sound** had fallen upon my ears." (Gaskell *The Old Nurse's Story*)

519

520

521

522

523

524

525

526

527

528

Although in most examples, it is loud sounds that punctuate a silent atmosphere, the reverse is also possible. These occasions can be more unsettling given that the expected sound is replaced by an unexpected silence. In Rymer's *Varney the Vampire*, we have a low scratching noise. However, rather than a crash of an invader, we have a sudden silence and only then does the vampire appear. The low sound is interrupted by the silence rather than the silence interrupted by the sound.

529

530

531

532

533

534

"Mrs. Bannerworth [...] heartily regretted she had not rung the bell, for, before, another word could be spoken, there came too perceptibly upon their

535

536

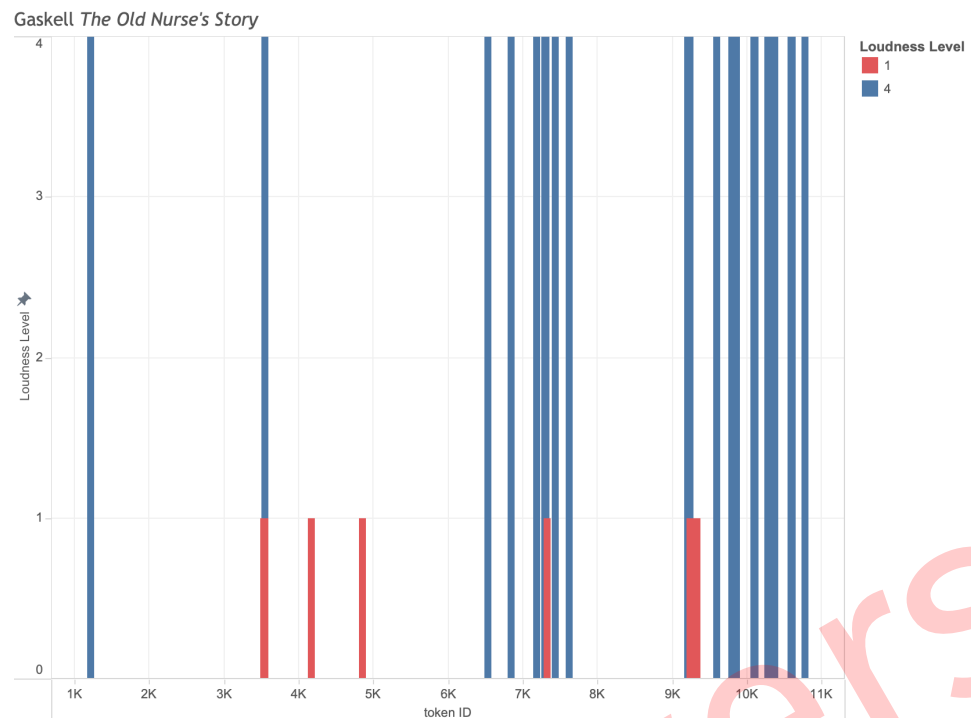


Figure 9: Gaskell *The Old Nurse's Story* – silence and loud sounds.

ears for there to be any mistake at all about it, a strange scratching noise upon 537
the window outside. A faint cry came from Flora's lips [...]. **The scratching** 538
noise continued for a few seconds, and then altogether ceased. [...] When 539
the scratching noise ceased, Flora spoke in a low, anxious whisper, as she 540
said,— 'Mother, you heard it then?'" (Rymer *Varney the Vampire*) 541

5.2 Sound and Suspense 542

With regard to the sound analysis results of our 19th century English fiction corpus, we 543
suggest that sound plays an important role in the Gothic as well as in other suspenseful 544
genres of the period, such as science fiction, or detective stories. However, there are 545
different types of ambient sound to distinguish which cannot be covered in this article. 546
E.g., sounds that are indicative of a character's uncertainty about a current state of affairs, 547
e.g., 'rustling', 'crackling', or even as in quotation (11) the 'sound of wheels', serve a 548
different purpose than, e.g., sounds that are effects of dangerous events (e.g., 'thunder', 549
'explosive blasts'). The distinction between further subcategories of ambient sound and 550
their effects should be investigated in a further study. Still, we could recognize that 551
sounds appear to be much less frequent in romance or city novels in which even the 552
background soundscape is rarely referenced: 553

(11) "At that moment the **sound of wheels** was heard and Charlotte flew off 554
to her private post of observation." (Yonge *The Heir of Redclyffe*) 555

When representations of sound do intrude into these novels, they display the deviation 556
from the default soundscape and suggest action, but do not offer enough information for 557
the reader to interpret it, creating questions and uncertainty. The sudden interruption 558
by explicit sound references in an implicit soundscape, serves like an unexpected event, 559

and has a surprising and suspenseful effect on the reader by interrupting the standard 560
conception of the scene. 561

With regard to Gothic texts, especially mysterious, inexplicable sounds amplify the 562
affect of suspense, creating uncertainty and driving the reader's desire to resolve a given 563
mystery: 564

(12) "'That is the story. **Whatever the sound is, it is a worrying sound,**' 565
says Mrs. Rouncewell, [...] 'and what is to be noticed in it is that it **MUST** 566
BE HEARD. My Lady, who is afraid of nothing, admits that when it is there 567
[...]'" (Dickens *Bleak House*) 568

For the analysis of the fictional soundscape however, it is not sufficient for a sound 569
word to simply lexically appear in the text like in quotation (12). As we argue in 3.2, 570
the word must represent the presence of the sound itself in the scene. Conditional 571
statements about possible sounds, descriptions of eagerly awaited sounds, comparisons 572
to known sounds, or reports of sounds that happened in the past do not effect the fictional 573
soundscape. Consider, for example, the atmosphere created by the conversation on the 574
mysterious sobbing of a woman the characters, in Doyle's text, have heard the night 575
before: 576

(13) "'And yet it was not entirely a question of imagination,' I answered. 577
'Did you, for example, happen to hear someone, a woman I think, sobbing 578
in the night?' 'That is curious, for I did when I was half asleep fancy that I 579
heard something of the sort. [...]' (Doyle *The Hound of the Baskervilles*) 580

The mystery of the sound is undercut by the rational conversation that contextualizes 581
it: it is not experiential but recollected and so does not trigger suspense for either the 582
reader or character. 583

6. Conclusion and Outlook 584

In this article, we have demonstrated the great potential of sound studies to literary 585
analysis. Our analyses, which combined distant reading methods with close readings, 586
offer evidence for our hypothesis that Gothic texts contain more detailed descriptions 587
of the story's ambient soundscape than our corpus texts labeled as "other". Our op- 588
erationalization of ambient sound, and the prediction model that we subsequently 589
trained from the data it produced, enabled us to explore sound from a computational 590
perspective to reveal new facets of the soundscape of fiction. The distinction between 591
represented and implicit or hypothetical sounds, however, presented a challenge to 592
our model. The distinction between represented and implicit or hypothetical sounds 593
presented a challenge to our model. Consequently, context is crucial for understanding 594
the role that sound words play as demonstrated by the difference in success of our 595
dictionary model versus the transfer-learning classifier. Despite the relatively high num- 596
ber of false positive predictions, the model trained on the manual and semi-automatic 597
annotations performed surprisingly well at detecting ambient sounds. 598

Our results argue for increased scholarly attention to sound in novels, and, in particular, 599
for the ways in which such automated approaches to the analysis of sound could be 600

harnessed to provide a deeper understanding of the role sound plays in narrative across 601
 a much broader period. E.g., the systematic analysis of the relation between sound and 602
 suspense could be interesting in future work. Similarly, as we close read the passages 603
 surfaced by our study, we also found evidence of other sensory descriptors. In a future 604
 project, the analysis of olfactory or haptic sensations could extend our study, as well as 605
 open up new affective representations for analysis. 606

7. Planned Revisions for the JCLS Journal Paper 607

The following revisions are planned for the Journal version, but could not be imple- 608
 mented for the Conference Reader version so far: 609

- better explanation of the domain adaptation of NER for Sound detection 610
- revision of the visualizations 611
- ideas for a more robust measure of sound word density / sound intensity 612

8. Data Availability 613

Data can be found here: https://github.com/SvenjaGuhr/Sound_and_Suspense 614

9. Software Availability 615

Used Software can be found here: [https://github.com/NEISSproject/tei_entity_e](https://github.com/NEISSproject/tei_entity_enricher) 616
[nricher](https://github.com/NEISSproject/tei_entity_enricher) and https://github.com/SvenjaGuhr/Sound_and_Suspense 617

10. Acknowledgements 618

We thank the reviewers for their detailed comments on our manuscript and constructive 619
 feedback that helped to refine and focus this article. 620

This article is the result of Svenja Guhr's research stay at the Stanford Literary Lab as a 621
 visiting research student in Autumn Term 2022, which was financially supported by the 622
 following three parties: 623

- Department of Digital Philology, Institute of Linguistics and Literary Studies at 624
 the Technical University of Darmstadt, 625
- Support of Women in Academia as part of the Equal Opportunity Commission of 626
 the History and Social Sciences at the Technical University of Darmstadt, 627
- Fellowship of the German Academic Exchange Service (DAAD Program for visit- 628
 ing doctoral students). 629

Special thanks go to our student assistant Alina Klein who supported the iterative 630
 process of annotation guideline creation and the annotation of the training data as a 631
 third annotator. 632

11. Author Contributions 633

Svenja Guhr: Conceptualization, Coding, Writing – original draft 634

Mark Algee-Hewitt: Supervision, Writing – review & editing 635

References 636





- Bacon, Simon, ed. (2018). *The gothic: a reader*. OCLC: on1030967356. Oxford: Peter Lang. 637
ISBN: 9781787072688. 638
- Bernhart, Toni (2008). "Stadt hören: Auditive Wahrnehmung in Berlin Alexanderplatz von Alfred Döblin". In: *Zeitschrift für Literaturwissenschaft und Linguistik* 38.1, 51–67. 639
10.1007/BF03379955. 640
- Blohm, Stefan, Maria Kraxenberger, Christine A. Knoop, and Mathias Scharinger (2021). *Sound Shape and Sound Effects of Literary Texts*. De Gruyter. 10.1515/9783110645958- 642
002. 643
- Botting, Fred (1996). *Gothic. The new critical idiom*. London ; New York: Routledge. 645
ISBN: 9780415132299 9780415092197. 646
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR abs/1810.04805*. 10.48550/arXiv.1810.04805. 647
- Ellis, Markman (2000). *The history of gothic fiction*. OCLC: ocm45236901. Edinburgh: Edinburgh University Press. ISBN: 9780748611959. 648
- Finkel, Jenny Rose, Trond Grenager, and Christopher Manning (2005). "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, 363–370. 10.3115/1219840.12 649
- Flüh, Marie, Jan Horstmann, and Mareike Schumacher (2022). "Genderaspekte in Fantasy-Jugendromanen von 2008 bis 2020: Distant Gender Reading". In: *Gender in der deutschsprachigen Kinder- und Jugendliteratur*. Ed. by Weertje Willms. De Gruyter, 457–482. 10.1515/9783110726404-025. 650
- Flüh, Marie and Marc Lemke (2022). "An experimental attempt to use Transfer Learning for Named Entity Recognition in letters from the 19th and 20th century". In: *Book of Abstracts*. <https://dh2022.dhii.asia/dh2022bookofabsts.pdf> (visited on 12/28/2022). 651
- Foley, Matt (2023). *Gothic Voices: The Vococentric Soundworld of Gothic Writing*. 1st ed. Cambridge University Press. 10.1017/9781009162579. 652
- Gius, Evelyn and Michael Vauth (2022). "Towards an Event Based Plot Model. A Computational Narratology Approach". In: *Journal of Computational Literary Studies* 1.1. 10.48694/jcls.110. 653
- Guhr, Svenja (2023). *Sound and Suspense*. GitHub Repository. GitHub. https://github.com/SvenjaGuhr/Sound_and_Suspense. 654
- Guhr, Svenja and Evelyn Gius (2023). "Maschinen als Erzähltheoretiker". In: *Kongressakten IVG 2020. Internationales Jahrbuch für Germanistik*. Peter Lang. 655
- Hinton, Leanne, Johanna Nichols, and John Ohala (Jan. 1995). "Introduction: Sound-symbolic processes". In: *Sound Symbolism*. Ed. by Leanne Hinton, Johanna Nichols, 656

- and John J. Ohala. 1st ed. Cambridge University Press, 1–12. ISBN: 9780521452199 676
9780521026772 9780511751806. 10.1017/CB09780511751806.001. [https://www.ca 677](https://www.cambridge.org/core/product/identifier/CB09780511751806A008/type/book_part)
[mbridge.org/core/product/identifier/CB09780511751806A008/type/book_part 678](https://www.cambridge.org/core/product/identifier/CB09780511751806A008/type/book_part)
(visited on 05/29/2023). 679
- Horstmann, Jan (2020). “Undogmatic Literary Annotation with CATMA”. In: *Annotations 680*
in Scholarly Editions and Research. Ed. by Julia Nantke, Frederik Schlupkothen, and 681
Jan Horstmann. De Gruyter, 157–176. 10.1515/9783110689112-008. 682
- Hühn, Peter (2013). “Event and Eventfulness”. In: ed. by Peter Hühn, John Pier, Wolf 683
Schmid, and Jörg Schönert. [https://www-archiv.fdm.uni-hamburg.de/lhn/node 684](https://www-archiv.fdm.uni-hamburg.de/lhn/node)
[/39.html](https://www-archiv.fdm.uni-hamburg.de/lhn/node) (visited on 10/18/2022). 685
- Hurley, Kelly (2002). “British Gothic fiction, 1885–1930”. In: *The Cambridge companion 686*
to gothic fiction. Ed. by Jerrold E. Hogle. Cambridge companions to literature. Cam- 687
bridge: Cambridge University Press, 189–207. ISBN: 9780521791243 9780521794664. 688
- Kamath, Uday, John Liu, and James Whitaker (2019). “Transfer Learning: Domain 689
Adaptation”. In: *Deep Learning for NLP and Speech Recognition*. Springer International 690
Publishing, 495–535. 10.1007/978-3-030-14596-5_11. 691
- Lemke, Marc (2022). *NEISS NTEE. User Interface. Documentation*. [https://github.com 692](https://github.com/NEISSproject/tei_entity_enricher/wiki/user-interface)
[/NEISSproject/tei_entity_enricher/wiki/user-interface](https://github.com/NEISSproject/tei_entity_enricher/wiki/user-interface). 693
- Loper, Edward and Steven Bird (2002). “NLTK: The Natural Language Toolkit”. In: 694
10.48550/ARXIV.CS/0205028. 695
- Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and 696
David McClosky (2014). “The Stanford CoreNLP Natural Language Processing 697
Toolkit”. In: *Proceedings of 52nd Annual Meeting of the Association for Computational 698*
Linguistics: System Demonstrations. Association for Computational Linguistics, 55–60. 699
10.3115/v1/P14-5010. 700
- Mildorf, Jarmila (2019). “Can Sounds Narrate? Prosody in Sound Poetry Performance”. 701
In: *CounterText* 5.3, 294–311. 10.3366/count.2019.0167. 702
- Mulvey Roberts, Marie, ed. (2009). *The handbook of the gothic*. 2nd ed. OCLC: ocn318534311. 703
New York: New York University Press. ISBN: 9780814796016 9780814796023. 704
- Nakayama, Hiroki (2018). *sequeval: A Python framework for sequence labeling evaluation*. 705
<https://github.com/chakki-works/sequeval>. 706
- Pichler, Axel and Nils Reiter (2020). “Reflektierte Textanalyse”. In: *Reflektierte algorithmis- 707*
che Textanalyse. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. De Gruyter, 43–60. 708
10.1515/9783110693973-003. 709
- Picker, John M. (2003). *Victorian soundscapes*. Oxford University Press. 710
- Reiter, Nils (2020). “Anleitung zur Erstellung von Annotationsrichtlinien”. In: *Reflek- 711*
tierte algorithmische Textanalyse. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. De 712
Gruyter, 193–202. 10.1515/9783110693973-009. 713
- Schafer, R. Murray (1994). *The soundscape: our sonic environment and the tuning of the 714*
world. Destiny Books ; Distributed to the book trade in the United States by American 715
International Distribution Corp. 716
- Schumacher, Mareike (2022). *Orte und Räume im Roman*. Digitale Literaturwissenschaft. 717
J.B. Metzler. 10.1007/978-3-662-66035-5. 718
- Schumacher, Mareike, Marie Flüh, and Marc Lemke (2022). “The model of choice. Using 719
pure CRF- and BERT-based classifiers for gender annotation in German fantasy 720
fiction”. In: *Book of Abstracts*. [https://dh2022.dhii.asia/dh2022bookofabsts.pdf 721](https://dh2022.dhii.asia/dh2022bookofabsts.pdf)
(visited on 12/28/2022). 722

- Scikit-learn Developers, x (2022). 3.3. *Metrics and scoring: quantifying the quality of pre-* 723
dictions. https://scikit-learn/stable/modules/model_evaluation.html (visited 724
on 12/17/2022). 725
- Smith, Mark (2015). *Listening to nineteenth-century America*. The University of North 726
Carolina Press : Made available through hoopla. 727
- Snaith, Anna, ed. (2020). *Sound and literature*. Cambridge University Press. 728
- TEI Consortium (2022). "TEI P5: Guidelines for Electronic Text Encoding and Inter- 729
change". In: 10.5281/ZENODO.3413524. 730
- Verma, Neil (2012). *Theater of the Mind: Imagination, Aesthetics, and American Radio Drama*. 731
University of Chicago Press. 10.7208/9780226853529. 732
- Zehe, Albin, Leonard Konle, Svenja Guhr, Lea Dümpelmann, Evelyn Gius, Andreas 733
Hotho, Fotis Jannidis, Lucas Kaufmann, Marcus Krug, Frank Puppe, Nils Reiter, 734
and Annekea Schreiber (2021). "Shared Task on Scene Segmentation (STSS). Task 735
Description Paper". In: *Proceedings of the 17th Conference on Natural Language Processing* 736
(KONVENS). [http://lsx-events.informatik.uni-wuerzburg.de/files/stss202](http://lsx-events.informatik.uni-wuerzburg.de/files/stss2021/proceedings/stss.pdf) 737
[1/proceedings/stss.pdf](http://lsx-events.informatik.uni-wuerzburg.de/files/stss2021/proceedings/stss.pdf) (visited on 12/21/2022). 738
- Zöllner, Jochen, Konrad Sperfeld, Christoph Wick, and Roger Labahn (2021). "Optimiz- 739
ing Small BERTs Trained for German NER". In: *Information* 12.11, 443. 10.3390/info 740
12110443. 741

Connecting the Dots

Variables of Literary History and Emotions in German-language Poetry

Leonard Konle¹ 
Merten Kröncke² 
Simone Winko² 
Fotis Jannidis¹ 

1. Institut für Deutsche Philologie, Julius-Maximilians-Universität Würzburg, Würzburg, Germany.
2. Seminar für Deutsche Philologie, Georg-August-Universität Göttingen, Göttingen, Germany.

Citation

Leonard Konle, Merten Kröncke, Simone Winko, and Fotis Jannidis (2023). "Connecting the Dots. Variables of Literary History and Emotions in German-language Poetry". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-21

Date received 2023-02-17

Keywords

Bayesian hierarchical generalized linear model, literary history, German-language poetry, emotion

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. In this study, we will take the first steps toward a quantitative literary history, attempting to identify factors relevant to the history of literature and to model assumptions about the relations between them. We use a case study of German-language poetry in the transition from realism to early modernism to approach our methodological goal. Using a Bayesian hierarchical generalized linear model we focus on one aspect relevant to the history of poetry, the emotions represented in the poems, and also include period, author profession, author gender, rhyme, and verse length in the model. We can confirm the important role of thematic genres and find unexpectedly high values for rhyme and author profession. We also discuss some of the methodological problems of our attempt to model this entangled network of variables involves.

1. Introduction

Any association between an outcome and a predictor can be nullified or reversed when another predictor is added to the model. And the reversal can reveal a true causal influence or rather just be a confound.¹

Most literary histories are based, at least implicitly, on the assumption that the history of literature is determined, on the one hand, by a more or less strong dynamic of its own and, on the other hand, by a number of external factors. These factors, be they the history of the media, the history of reading, the history of social structures, developments in the history of ideas, or the history of mentality, such as the change in ideas about man, etc., are often presented in separate chapters, and their literary-historical relevance may be illustrated by individual texts. However, it is notoriously difficult to show that even one of these factors is relevant to all literary texts, even if there are examples for each of them that make it impressively clear that certain texts can only be understood if one takes the respective focused factor into account. This sometimes leads to the description of literary history as a whole as an autonomous process. This view is further supported by the fact that the connection between factors external to literature and the literary text and

1. McElreath 2020: 345.

its specific form can only be plausibly modeled in a few exceptional cases. In addition, literary history hardly has access to methods for the systematic aggregation of individual observations in such a way that the frequency of a phenomenon is taken into account and exceptions are not regarded as refutations. This is where literary-historical projects that use digital corpora and statistical methods come in. They can make digitally based statements about relationships between factors modeled in the analysis and literary texts. These statements are limited to the texts and factors under investigation: one cannot see what one does not see. The complexity of the subject matter complicates the study: there are numerous factors that must be taken into account; we do not know at this point exactly what the role of each of these factors is. Many of the factors clearly interact with each other, making their statistical analysis particularly difficult. Our work should be seen as a first step towards a quantitative literature history that attempts to identify the relevant factors. As noted above, this can only be done on a specific corpus. Whether the results are transferable to other corpora representing other genres and the literature of other periods, languages, and cultures is a separate research problem that can only be addressed as more studies of this kind become available. Maybe the important factors can only be identified in the context of a genre or a period or other groups.

At the same time, the moment is particularly opportune: for the first time, literary texts are available in digital form in sufficient numbers and methods to analyze complex situations are available. On the other hand many aspects, especially extra-literary factors, cannot yet be integrated into the model because the data is not yet available or not available in digital form. This is another reason why these first results are very preliminary. But without these first steps and without having learned from them, subsequent research would lack a foundation.

We conduct our experiments and methodological considerations on a case study: on the transition from the poetry of realism to early modernism. Our corpus consists of about 6000 poems from these periods. In what follows, we focus on the question of whether, and to what extent, factors contribute to one aspect of the texts, namely the emotion that is thematized. Our focus is on emotion because there is ample evidence in literary history and contemporary accounts that the difference between the periods is evident in the literary representation of emotions. In this paper we will look only at a small set of factors and how they are related to the emotions of poetry in a regular way. We have chosen the factors we focus on because they have been deemed relevant by research and because we can represent them as data. Some of them are external factors like gender or profession of the author, others are text aspects like rhyme or the thematic genre. We do not assume that the same kind of real-world causality underlies all these factors. The gender of the author may have an influence on the selection of emotions depicted in a poem, but not the other way around. The same is probably not true in the same way for thematic genre or rhyme.

2. Research

In literary studies, the relationships between literary texts and other literary or extra-literary phenomena are often discussed under the term "context." We highlight two

related aspects of the debate on context that are relevant to our work:	61
1. One strand of discussion focuses on the extent to which context is relevant to understanding and/or explaining literature and literary change in general. Some approaches, e.g., Marxism, argue for strong contextual influences or even determinacy, while other approaches assume that literature is either autonomous and independent of external factors or should at least be treated as such (cf. Kalliney 2019, Ladegaard and Nielsen 2019).	62 63 64 65 66 67
2. For those approaches assuming that contexts play at least some role for literary texts, another much-debated question is how to select and weight particular contexts (cf. King and Reiling 2014; Borkowski 2015; Engel 2018; Thomsen 2019: 207) – a question of great importance for this paper. The discussions on context selection and context weighting are also related to the issue of how to explain literary change (cf. Gittel 2016 for a distinction between different types of explanations). Some positions, often influenced by particular literary theories, name specific contexts they consider relevant. For example, feminist literary theory typically assumes that the gender of the author and/or the role and status of women in society, culture, economics, and politics are important contexts for understanding literature and literary change. Other positions do not focus on specific contexts but on <i>criteria</i> for contexts. For example, King and Reiling 2014 propose “relevance,” “representativeness,” and “usefulness” as criteria for context selection.	68 69 70 71 72 73 74 75 76 77 78 79 80
When discussing the relevance of context in general or the selection and weighting of specific contexts, literary scholars mostly rely on theoretical considerations, on examples from literary history, or on individual text analysis. While these types of arguments are certainly valuable, what seems to be missing is a practical method for comparatively analyzing the relevance of context(s) that is suitable for our purposes. At the very least, we are not aware of any method proposed in literary studies that is (a) not limited to very specific text corpora or time periods, (b) somewhat independent of theory-specific presuppositions (such as „X is always the most important context”), and, most importantly, (c) computationally operationalizable in a reasonably clear, practicable, and intersubjective way. For example, the criteria of King and Reiling 2014 are so abstract that they immediately raise the question of how to figure out what is “relevant,” “representative,” or “useful.” There are no obvious answers. King and Reiling 2014: 21 themselves write that their suggestions are rather a “heuristic framework” and “rough criteria” that complement the “methodological evidence of the individual case”.	81 82 83 84 85 86 87 88 89 90 91 92 93 94
It seems possible to use the context selection criteria mentioned in literary studies as helpful heuristics that can narrow down the set of potential contexts worth investigating. Beyond that, however, we need to draw on other resources to achieve the methodological goals of this paper. At best, this will also allow us to contribute to the ongoing debates about “context” in literary studies.	95 96 97 98 99
Identifying relevant factors in historical processes is a comparatively new perspective in Computational Literary Studies, but there is field which has produced a series of studies with this goal, and its unifying characteristic is a view on culture as an evolutionary process. In this context single works are not important but historical trends and shifts (see for example section VII in Barrett and Dunbar 2007). In recent years more and more	100 101 102 103 104

studies appeared using the framework to analyze larger data collections, for example Sobchuk and Tinitis 2020 who analyze the development of anachronies in mystery film between 1970 and 2009. But many of them, like Sobchuk and Tinitis, describe a trend in the data quantitatively and discuss at the end on a qualitative level possible predictor variables. This is also the case in most studies from Computational Literary Studies. Usually they identify an interesting trend in their data and propose reasons for the observed pattern. Discussion and controversies concentrate on the question, whether the trends are really there, like Langer et al. 2021 and Piper 2022. But this begins to change slowly. Using corpora in three languages Šeĭa et al. 2021 can show that metrical forms and the semantic features of corresponding poems are systematically linked. Underwood et al. 2022 seems to be the first one to discuss a causal factor in literary based on the analysis of a large corpus. They find that cohort succession plays an important role in literary history.

3. Resources

3.1 Corpus

Our corpus includes poems from two periods: from realism and the period ‘around 1900’ (or early modernism). The poems in question were published in those anthologies that aim to provide their audience with an overview of contemporary poetry (in part also of ‘the best’ contemporary poetry). Thus, the corpus texts have been labeled, so to speak, by contemporary poetry experts. The ‘realism’ sub-corpus consists of eight anthologies published between 1859 and 1882, the ‘around 1900’ sub-corpus of 12 collections published between 1885 and 1911. Thus, the entire corpus contains 6.249 poems from 20 anthologies (for a description of the corpus, see Winko et al. 2022).

3.2 Emotion

Before analyzing the association of different features with the distribution of emotions in poetry, it is necessary to determine how often which poems represent which kind of emotions in the first place. For this purpose, we rely on a machine learning setup that has been trained and evaluated on manual annotations, which we have already described in Konle et al. 2022:

We annotated the representation of emotions in 1352 corpus poems. The goal was not to annotate readers’ emotions, but rather the emotions represented in the text itself, e.g., whether the speaker or a character is happy, sad, in love, etc. The annotators used a list of 40 discrete emotions (e.g., love, joy, surprise, envy, regret, fright), the selection of which was based both on existing emotion models (e.g. Ekman 1992, Ekman 1999; Plutchik 1980a, Plutchik 1980b, Plutchik 2001) and on the emotions that were regularly represented in the poems of our corpus. Because a substantial number of emotions were only infrequently annotated (e.g., disgust), we categorized the emotions after annotation into 6 major groups, inspired by the emotion hierarchy in Shaver et al. 1987: agitation, anger, fear, love, joy, sadness. First, each poem was annotated independently by two annotators, who then manually merged the annotations into a consensus annotation. Their agreement, measured with γ (Mathet et al. 2015), was 0.6445 for individual emotions and 0.7491 for the emotion groups (annotation guidelines:

Kröncke et al. 2022), where 0 indicates agreement no better than chance and 1 indicates perfect agreement.

In order to detect emotions automatically, we performed an emotion classification task which we modeled as a series of binary classifications to avoid the complexity of a multi-labeling task. Basis of our classification is the german BERT (Devlin et al. 2018) model gbert-large (Chan et al. 2020). Because gbert is trained on contemporary webtext, we continue its pre-training² with poetry to adapt to our target domain. Subsequently we perform fine-tuning on the binary emotion classification tasks. To overcome the class imbalance we apply undersampling by randomly sampling examples from the majority class in every epoch. While the classification of single emotions leads to a large spread in predictive quality³, the grouped emotions lead to more stable performance at an acceptable level of uncertainty (Table 1).

Table 1: Quality of Emotion Classification.

Emotion	Joy	Love	Sadness	Anger	Fear	Agitation
f1 (macro)	0.73	0.77	0.74	0.71	0.79	0.62

For all further analyses, we focus on the six emotion groups, using consensus annotations where possible and relying on model predictions otherwise.

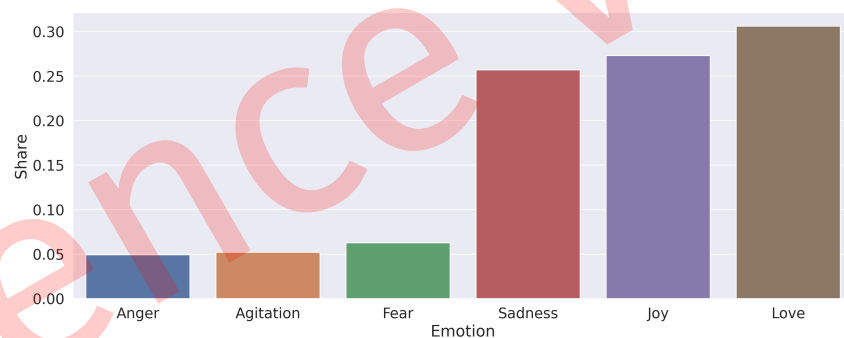


Figure 1: Share of predicted emotions.

The poems represent love, joy, and sadness most often; anger, fear, and agitation occur much less frequently.

3.3 Features

We cannot rule out the possibility that a myriad of features affect the distribution of emotions in poems, including various aspects of the author's biography (e.g., their gender or attitude toward religion) and the poem's content, form, and style (e.g., its themes or stylistic register). However, for reasons of data availability alone, it is not possible to include all conceivable features in our study. Therefore, we confine our analysis to a limited set of features selected according to several criteria: (1) Other researchers consider the features to be relevant to the literary representation of emotions and/or

2. Hyperparameter: 500 steps, batchsize 30, learningrate 2e-5 (see Konle and Jannidis 2020, Gururangan et al. 2020).

3. Very frequent emotions like longing (f1: 0.73) or suffering (f1: 0.72) yield sufficient classifiers, but less frequent ones like calmness or desire lead to results similar to a random baseline.

the features appear to be important given the context selection heuristics mentioned 171
 in literary studies (see section 2) and/or studies from our working group indicate that 172
 there is a relationship between the features and the emotions represented in our corpus; 173
 (2) the features cover a wide range of textual (e.g., theme) and extratextual phenomena 174
 (e.g., author gender); and (3) respective data is available for the entire corpus. As 175
 indicated, the restriction to a limited set of features means that our analysis certainly 176
 cannot encompass all features that might affect the representation of emotions in poetry. 177
 Nevertheless, we are confident that by applying the criteria mentioned above, we have 178
 selected features that, taken together, can provide relevant insights. 179

For each feature, the following sections justify why we include it in our analysis and 180
 how we collected the data. 181

Period Literary periods like realism or modernism have traditionally been among 182
 the most important categories for organizing and characterizing literary texts. When 183
 examining various aspects of literature, the literary period is almost always considered 184
 as a possible factor of difference. Likewise, researchers assume that the literary rep- 185
 resentation of emotions differs significantly depending on the period (e.g. Andreotti 186
 2014: 319). Previous studies by our own working group on the relationship between 187
 emotions and periods point in a similar direction and indicate that poems from early 188
 modernism represent fewer and less positive emotions than poems from realism, even 189
 though the differences are not huge (Konle et al. 2022). 190

An alternative to including literary change via period as a categorical variable (realism, 191
 modernism) would be to use the year of writing/publication as a numerical variable in 192
 order to allow for finer-grained analyses of time series. However, since we have not yet 193
 collected the writing/publication dates for all poems, we will stick with the periods for 194
 now. 195

For categorizing the poems into periods, we use the corpus anthologies and the division 196
 into subcorpora described above. The 3367 texts from the anthologies published between 197
 1859 and 1882 are assigned to the period ‘realism’ and the 2882 texts from the anthologies 198
 published between 1885 and 1911 to the period ‘modernism’. Again, it is important to 199
 keep in mind that the period labels are provided by the anthologists, and that their 200
 views on what counts as ‘realist’ or ‘modernist’ do not always have to align with today’s 201
 research. 202

Gender The gender of the author provides another basic category for differentiating 203
 poems. When literary scholars analyze text corpora, gender is regularly brought into 204
 view as a possible factor of difference. Hypotheses that assume gender-specific differ- 205
 ences also exist with regard to the representation of emotions. For example, one study 206
 argues that until the second half of the twentieth century sexual desire was represented 207
 more frequently, or at least more openly, by male than by female authors (Härle 2007: 208
 118f, 135). And an early twentieth-century anthologist named Margarete Huch, to give 209
 another example, claims that in her time, “love for man takes up by far the largest space 210
 in all women's poetry” (Huch 1911: 12), which might lead to the expectation that female 211
 authors are especially likely to represent love, i.e. more likely than men. 212

We obtained data on the author’s gender from the database of the German National 213

Library (<https://dnb.de/>), by manual research, e.g. in biographical dictionaries, and/or 214
 on the basis of the author's first name. Since the multiplicity of possible gender positions 215
 was not yet generally recognized during the period under study, and the resources 216
 mentioned above identify the corpus authors exclusively as "male" or "female," the 217
 feature 'gender' is limited to these two options. We were able to collect data for 78% of 218
 the corpus authors (who wrote 85% of the poems). 85% of the identified authors are 219
 male, 15% are female. 220

Profession As another feature, we include the profession of the authors in our analysis. 221
 Since numerous corpus authors are not only writers, but also, for example, politicians, 222
 historians, or philosophers, the feature 'profession' allows to differentiate the authors 223
 and their poems in a meaningful way. Certainly, there are not many previous studies that 224
 demonstrate a general correlation between profession and emotion in poetry. However, 225
 there is at least some research on individual texts or authors that assumes such a 226
 connection, and in this respect suggests that a more systematic study of the relationship 227
 between profession and emotion might be informative. For example, with respect to 228
 Gottfried Benn, it has been argued that there is a link between his profession as a 229
 physician and his sober and unemotional style in early poems such as *Kleine Aster* (e.g., 230
 Hiebel 2005: 218-21). Moreover, the inclusion of profession allows us to integrate aspects 231
 of the author's socioeconomic background in the analysis which would otherwise not 232
 be considered at all. Still, including profession as a feature must be understood as an 233
 experiment that may or may not prove relevant to the representation of emotions. 234

To obtain data on the author's profession, we relied on the authority files of the German 235
 National Library, the GND, which stores this type of information for each person in its 236
 database and which takes into account that a person may have more than one profession 237
 at the same time.⁴ However, the resulting labels are very diverse: according to the 238
 database, the corpus authors practice 346 different professions, with 190 professions 239
 being assigned to only one author. It was obvious that the number of classes had to be 240
 reduced. Therefore, we divided the professions into 8 groups based on the thematic 241
 genres (people working in the field of history, people working in the field of politics, 242
 etc.), supplemented by a group for 'other professions'. Certainly, using thematic genres 243
 to categorize professions is only one of many possibilities, and it may be useful to 244
 experiment with other categorizations in the future.⁵ 245

That a large number of authors have been assigned to the field of 'poetology' is un- 246
 understandable, since the category includes all those who are designated by the German 247
 National Library as 'writers,' 'poets,' etc., which is true of most corpus authors. 248

Thematic Genre Just like period or author gender, genre is a fundamental category 249
 for characterizing literary texts. Among other things, the attribution of thematic genres 250
 such as love poetry or nature poetry provides basic information about the content of 251
 the poems. Furthermore, studies by our own working group have shown that there 252
 is a relevant relationship between thematic genre and the representation of emotions 253

4. <https://www.dnb.de/gnd>.

5. Furthermore, there is still some work to be done on the data drawn from the German National Library, as samples have shown that they are not always reliable (e.g., the professions 'writer, lawyer, librettist' are assigned to Franz Kafka, <https://d-nb.info/gnd/118559230>).

Table 2: Number of Authors per Profession.

Profession	Examples	Count
Love		0
Nature	natural scientist, botanist	18
Poetology	writer, poet	1497
History	historian, archivist	54
Politics	politician, diplomat	101
Philosophy	philosopher	24
Religion	priest, bishop	87
Culture	actor, painter	251
Other	lawyer, physician	899

in poems (Kröncke et al. 2023). For all these reasons, we include thematic genre as a feature in our analysis.

As described in more detail in Kröncke et al. 2023, we manually annotated 8 thematic genres in 1412 poems (the themes were love, nature, philosophy, religion, poetology, politics, culture, and history). It was possible to assign exactly one, but also none or several genres to a poem. The annotators reached an agreement of 0.69 (krippendorff's alpha). We continued with training binary classifiers for each thematic genre with the exception of political poetry, for which we had too few annotations. As with the automatic detection of emotions, the classification is based on the gbert-large language model adapted to our corpus.

Table 3: Quality of Genre Classification.

	Love	Nature	Poetology	History	Politics	Philosophy	Religion	Culture
Acc.	.88	.833	.632	.872	-	.732	.815	.470
Std.	.022	.024	.224	.026	-	.028	.059	.169

The performance of the classifiers seems to be sufficient, with only cultural poetry and poetological poetry being detected less reliably than the other genres. For this reason, we exclude cultural and poetological poetry (as well as political poetry) from the analysis i.e. treat texts of these three genres in the same way as texts without a genre label.

Form Including formal aspects in the analysis seems instructive, since the term 'form' encompasses basic features of general importance, for poems especially meter and rhyme. Moreover, several studies have indicated that there is a connection not only between form and content (e.g., Carper and Attridge 2003, and, using a computational approach, Šeĭa et al. 2021), but also between form and emotion (e.g., Obermeier et al. 2013; Tsur 2017; Haider 2021).

To collect data on formal aspects automatically, we used the Metricalizer tool (<https://metricalizer.de/de/>), which we evaluated with manually annotated poems from our corpus. Three annotators analyzed 100 poems and annotated for each verse, among other things, the number of stresses and the binary distinction of whether the verse is rhymed or unrhymed. Regarding the number of stresses, the annotators reached an agreement of 0.91; for rhyme, an agreement of 0.95. After the initial annotation, the annotators discussed all instances of disagreement and created a consensus annotation that we used as the basis for evaluating the Metricalizer. The tool achieved an accuracy

score of 0.91 or an alpha agreement of 0.96 for the number of stresses and an F1 score of 0.97 for the binary distinction ‘rhymed/unrhymed’.⁶

Since the performance of the Metricalizer seems good enough, we let the tool analyze all corpus texts and extracted the following features for every poem to be used in our analysis:

- the *proportion of rhymed verses*, since texts with few or no rhymes may represent different emotions than poems that are fully rhymed (cf. Obermeier et al. 2013). According to the Metricalizer, 5% of the corpus poems contain no rhymes at all; 40% include both rhymed and unrhymed verses, while in the remaining 55% all verses are part of a rhyme;
- the *average verse length* of the poem, measured by the number of characters per line, a feature that is not directly based on the results of the Metricalizer, but is strongly correlated with what the Metricalizer yields as the number of stresses ($r = 0.9$). Research suggests that in certain contexts, longer verses tend to convey a more serious mood, while shorter verses are associated with lighter themes (Šeĭa et al. 2021). If this is true, then verse length might also be related to the representation of emotions. In our corpus, the average verse length is 34 characters (SD = 12).

In addition to the formal features mentioned above, it would have been instructive to include the verse foot (iambic, trochaic ...) as well. However, due to technical difficulties and limitations associated with the Metricalizer, we have to postpone the addition of the verse foot for the time being.

4. Methods

4.1 Formal Modeling of the Factors

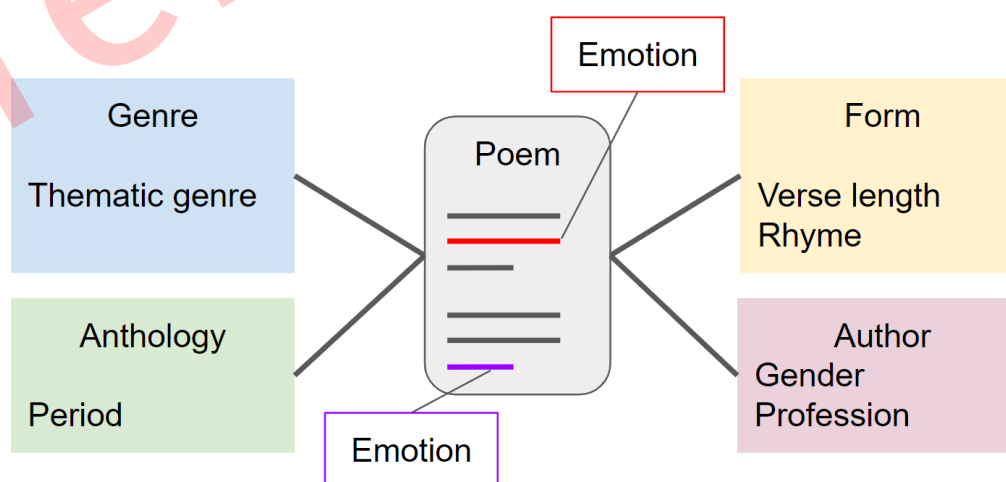


Figure 2: Datamodel.

Our analyses focus on emotions in poems. Both the annotations and the machine learning models for emotion detection operate on spans within poems. As a result,

6. In addition, we annotated and evaluated the role of the verse in the rhyme structure, using a letter system (a, b, c ...), and (for the first two verses of each poem) the verse foot (iambic, trochaic ...). Ultimately, however, we did not use these data in our analysis.

several emotions are attributed to a poem in different, partly overlapping passages of text. The other factors, however, are at the document level (see Fig. 2). It would facilitate the analysis if all factors were on the same level together with the target variable (emotion). This can be achieved by transforming the data. For example, the emotions could be weighted according to the length of the annotation span. However, we decided not to use this or any other transformation, as they require assumptions that we do not take for granted. In order to weight emotions according to the length of their passage, we would have to assume that this actually allows us to make a statement about the relevance of the emotion in the whole poem. However, it is quite conceivable that an emotion that is only present in the last verse of a poem, for example, could be at the center of the spectrum of emotions. The disadvantage of not transforming, and thus working at the level of emotions rather than poems, is that a poem is modeled as multiple data points. Or, to put it another way, we create an additional layer of inter-correlation in our data as we take multiple 'measurements' from one poem. In summary, we model individual emotions in poems with the global properties of *thematic genre*, *proportion of rhymed verses*, and *average verse length*. These poems come from anthologies that can be assigned to a *literary period*, and were written by authors of different *genders* whose *professions* we fetched and grouped.

4.2 Bayesian Generalised Hierarchical Linear Model

For the statistical analysis of our dataset, we decided to use a Bayesian hierarchical⁷ generalized linear model. The basis for this decision is the nested structure of our data. To determine whether an emotion is overrepresented within a thematic genre, we could simply count how often we find it there and whether the probability is greater than for poems that do not belong to that group. However, this approach does not take into account the other characteristics of the poems within the genre. If we, for example, assume that they tend not to rhyme and that this favors the occurrence of the same emotion, we cannot say whether our findings are valid or not. The use of a generalized linear model (GLM) allows us to take these contingencies into account, at least to the extent that we control for the variables (factors) present in our data set. This is an improvement, but still not enough.

This will become clearer, when we look at an example: Given we see that the probability of an emotion does not change, if the poem belongs to the category religious poetry. On further analysis, however, we find that the probability of the emotion decreases in modern religious poetry and increases in religious poetry from realism. This exciting information escaped us, because we averaged over the genre independently of its literary-historical context. However, there is the possibility, and literary history gives us reason to believe that this is not unlikely, that genres evolve with the change of literary periods (in our case exclusively with regard to the representation of emotions). To take this possibility into account, we model hierarchy in our GLM. From a purely statically motivated point of view without subject-related theoretical assumptions, we would have to assume that this kind of relationship can exist between all factors. This would result in groups such as unrhymed poems by natural scientists or historical poetry with low verse length, all of which must be treated individually. To avoid this scenario and

7. also called: mixed, mixed effects or multi-level

an unnecessarily complex model, we should ask which of the combinations do not have a purely additive relationship in terms of their effect on the distribution of emotions. For example, it is not very plausible to assume that the effect of rhymes on emotions changes or even reverses when they are used in a poem about nature. The factor with the greatest plausibility for producing non-additive effects with other factors is the literary period. The genre is also a valid option, but we have decided against using it as a grouping factor, since the greatest variation can be assumed in the interplay with the period and we already cover this by grouping by period itself. The combination of profession and genre also seems plausible, for example when a natural scientist writes a poem about the subject matter of his or her discipline, but our dataset is too small here to allow for certain conclusions anyway.

The hierarchical GLM follows the notation⁸ below:

$$\begin{aligned}
 h_f &\sim \text{Normal}(0, 1) && [\text{for each factor } f] \\
 c_f &\sim \text{Normal}(0, 1) && [\text{for each factor } f] \\
 s_{f,e} &\sim \text{Normal}(0, 1) && [\text{for each factor } f \text{ and period } e] \\
 \text{logit}(p) &= x_f(h_f + s_{f,e}) + c_f \\
 y &\sim \text{Bernoulli}(p) && [\text{prob. of emotion is present}]
 \end{aligned}$$

First, we define slopes h and intercepts c for each factor (gender, rhyme, etc.) individually. Then a second slope parameter s is added, this time for each factor and each literary period, resulting in three parameters for each factor: global intercept c , global slope c and grouped slope s . The slopes are summed and multiplied by the actual factor from our data set and added to the global intercept. In this way, we model the slope of a factor as being normally distributed around a mean and then corrected by an offset with respect to the literary period. This has the advantage over modeling the effect of, say, modernist religious poetry on emotions directly, because we allow the model to 'share' information between periods through h and c while allowing for variation by s . The limitation of the hierarchisation to slopes (free intercept and free slope and free intercept models are also conceivable) follows from results of a preliminary study with a free intercept free slope model, which shows only negligible variance in intercepts.

Since we use a Bayesian model, we need to set priors according to our expectations. But as can be seen from the notation of the model, we initialize all parameters with so-called flat priors. Flat here means that we do not make any strong assumptions about the effect of factors on emotion or, in other words, we assume that the parameters follow a normal distribution that has its mean at zero and a standard deviation of one and leave the rest to the data.

The left plot in Fig. 3 visualizes the initial slope parameter h_{love} for the emotion love in love poems. Each line shows a randomly drawn value from the distribution defined by our priors before the model was fitted to our data set. There are few extreme values indicating a strong negative or positive relationship and more values leaning towards

8. We use an offset approach instead of classic hyperparameter to improve computational efficiency (see Wiecki 2017; Betancourt and Girolami 2015).

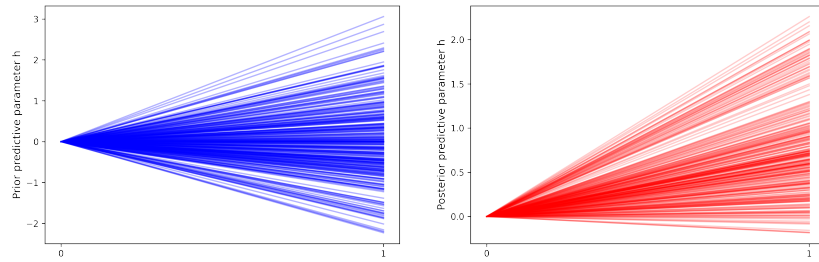


Figure 3: 300 random samples of parameter h_{love} from prior and posterior distributions.

zero. After the model is fitted, the parameter distribution changes. The right plot contains almost only slope values above zero and the highest density of lines in the range between 0.5 and 0.7. The exact mean of this posterior distribution is 0.77, this equals an odds ratio (e^h) of 2.13 stating that if every other factor is fixed the chance to observe the emotion love in love poems is more than twice as high as in non-love poems.

The mean intercept c_{love} from the posterior distribution is -1.5, this translates to a chance of 0.19 ($e^c/(1+e^c)$) that an emotion span from a non-love poem contains the love emotion. Since we already know the effect of a poem being a love poem we can calculate the chance of 40% for emotion spans in love poems containing love. Up to this point, we have not yet considered the hierarchical structure of the model. If we want to know whether the influence of the genre love poem changes with the passing of literary periods, we must additionally sample from the posterior distribution of the parameters $s_{love,realism}$ and $s_{love,modernsim}$. Their mean values are 0.252 and 0.249 and 1.286 and 1.282 odds ratios, respectively. Thus, the model indicates a stable effect across both epochs. If we want to know the absolute effect of the genre love poem in realism we just need to add $s_{love,realism}$ to c_{love} and calculate the odds ratio (2.77).

We estimate the parameters for each emotion individually in separate models. Each model is fitted with 4000 sampling, 2000 tuning steps and 4 chains.

5. Results

In the paragraphs that follow, we present four selected findings on the relationships between individual features and emotions, before taking an overarching perspective and analyzing more general aspects. The selected results should be understood as examples that on the one hand illustrate what kind of analyses the model facilitates and on the other hand deepen the understanding of the methodology. Additionally we selected results that seemed interesting in the light of the assumptions of literary history about these variables.

Figure 4 shows the probability density of slope parameters from posterior distribution of the model fitted to the emotion anger. The bell shaped curves mark the amount of probability for a certain parameter value. The plot on the left side shows diminishing probability for $h_{philosopher}$ values beyond -2 and 2. The most likely value is located where the curve denotes the highest density. This point (around 0.01) is additionally

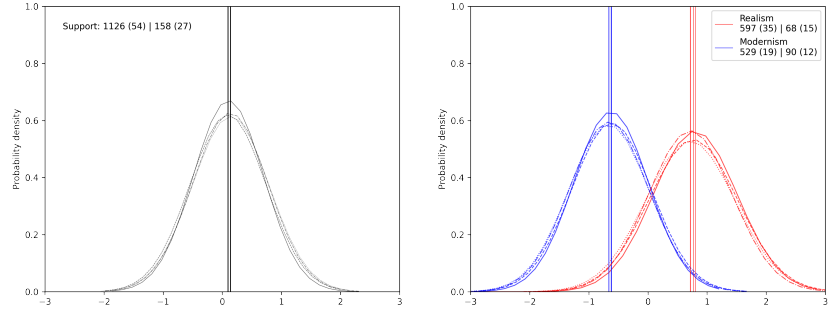


Figure 4: Posterior distribution of $h_{philosopher}$ (on the left side) and $s_{philosopher,realism}$ and $s_{philosopher,modernism}$ (on the right side) for the emotion anger. Support notation: Emotions in group (with anger) | Poems in group (with anger).

highlighted with a vertical line for visual convenience. The more pointed the bell is 416
shaped, the more probability is distributed among fewer parameter values. For example, 417
the blue curve for modernist poems in the plot on the right side is flatter than that for 418
realist poems, which means that the parameter $s_{philosopher,modernism}$ can be estimated less 419
confidently. Each plot contains multiple lines, those come from multiple chains or runs 420
of the same model with the same data. If these lines of the same color vary little we can 421
interpret this as an indicator for model stability. 422

The support information in the left plot reads as follows: In our dataset, there are 1126 423
emotion spans (54 contain the emotion anger) from 158 poems written by authors 424
working in the field of philosophy (27 contain the emotion anger at least once). Realism 425
accounts for 597 of these spans and in 35 contain anger and so on. 426

The analysis of the two plots allows the following statements: If we know nothing about 427
a poem except it is written by a philosopher, we can say that it is just a little more likely 428
(1.1 times), that one of its emotion spans contains anger. But if we get information about 429
the literary period the poem belongs to that changes drastically. Since the relationship of 430
period grouped slopes s and fixed slope h is additive (see Section 4.2) the odds increase 431
to 2.4 in realism and if it is modernist they decrease to 0.6. Despite this being a strong 432
effect, this finding should be taken with care due to the low support numbers. 433

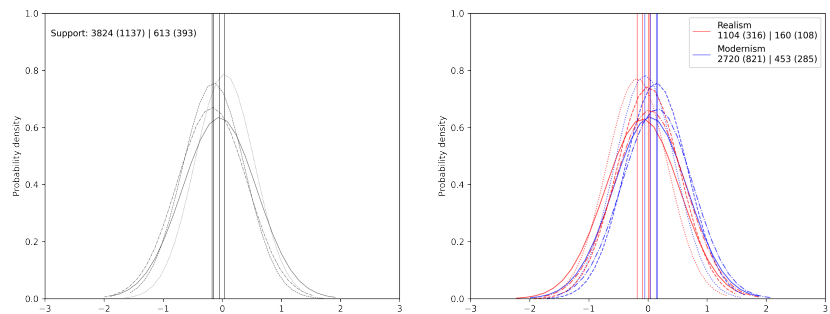


Figure 5: Posterior distribution of h_{gender} , $s_{gender,realism}$ and $s_{gender,modernism}$ for the emotion love. Support notation: Emotions in group (with love) | Poems in group (with love). To belong to the group, a poem needs to be written by a female author.

Figure 5 shows that there is no pronounced difference between male and female authors in the probability with which they represent love in their poems. Nor does this change from period to period ($s_{\text{gender,realism}}$ and $s_{\text{gender,modernism}}$ vary around zero). Note that the support for this finding is stronger than in the case of the “anger/philosopher” example, since our corpus contains a lot more poems and representations of emotions by female authors than by authors working in the field of philosophy. We show this lack of effect because it exemplifies the numerous cases in which we find no substantial relationship between features and emotions. Moreover, this plot shows that the data do not support the assumption that female authors would be more inclined to represent love in their poems (see above, 3.3).

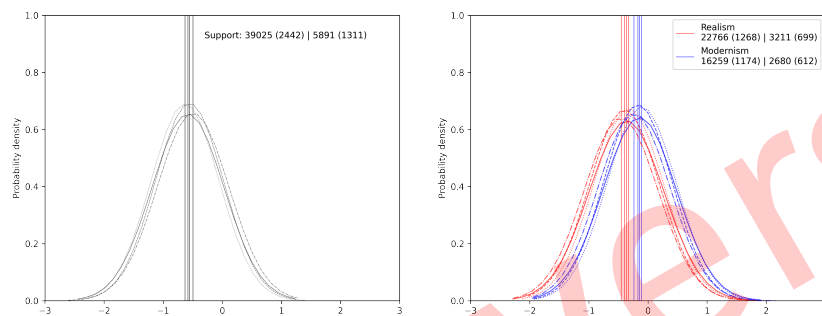


Figure 6: Posterior distribution of h_{rhyme} , $s_{\text{rhyme,realism}}$ and $s_{\text{rhyme,modernism}}$ for the emotion fear. Support notation: Since proportion of rhyme verses is not a categorical variable, every poem belongs to its group.

Fig. 6 shows the relatively strong relationship of rhymed verses in a poem and the emotion of fear. The mean from the left plot states that, if a poem contains only rhymed verses the chance to observe the emotion fear is nearly halved (0.56 odds ratio). The right plot shows negative mean values as well, which decreases the odds further for both realism and modernism. This shift is strong in Realism, but the difference is small and the variation between chains (varying curves of the same color) indicates a low confidence for this small difference.

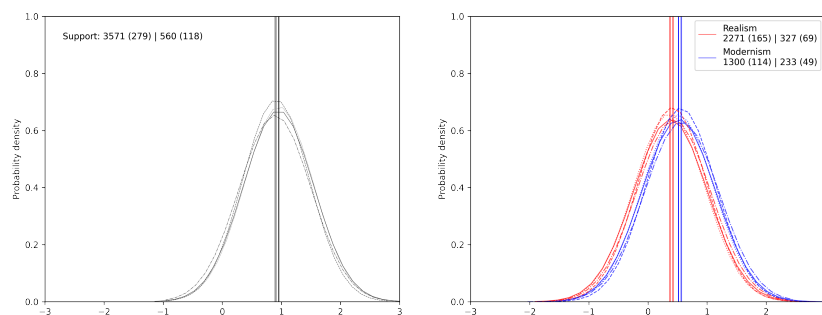


Figure 7: Posterior distribution of $h_{\text{historical}}$, $s_{\text{historical,realism}}$ and $s_{\text{historical,modernism}}$ for the emotion anger.

Our final example (Fig. 7) shows that historical poems, e.g., poems about ancient kings or medieval battles, are (2.59 times) more likely to represent anger than poems in other thematic genres. This effect seems to be relatively constant over time, as we

see no strong difference between realist and modernist poetry in this regard. With 560
historical poems in our corpus, 118 of which contain the emotion “anger,” the support 455
for this relationship is substantial enough to draw the conclusions mentioned. 456

The result illustrates the relationship between thematic genre and emotion. In this 457
case, a possible explanation might be that historical poetry, as a thematic genre, often 458
depicts (violent) conflicts, such as wars, revolutions, or power struggles, and that these 459
themes make it all the more likely that negative, adversarial emotions, such as anger, 460
are represented (on historical poetry, cf. Detering and Trilcke 2013). 461

We limit ourselves to the relationships between individual factors and emotions shown 462
to this point and instead provide an overview of the connections in Table 4. Its content 463
provides an overview of the absolute mean values of the h parameter across all emotions 464
and chains (The value which can be obtained from the left plot in Fig.5-8). Since 465
profession and thematic genre are split into multiple factors, we average their h values. 466

Table 4: Absolute mean value of h from posterior distribution for each factor and emotion.

Feature / Emotion	Love	Sadness	Joy	Anger	Fear	Agitation
Profession	0.16	0.12	0.08	0.33	0.27	0.32
Thematic Genre	0.18	0.19	0.18	0.40	0.42	0.25
Rhyme	0.02	0.24	0.27	0.58	0.56	0.55
Gender	0.08	0.07	0.08	0.08	0.14	0.15
Verse Length	0.07	0.10	0.07	0.03	0.02	0.06

Although insight into the relationship between individual factors and emotions is 467
informative and, more importantly, useful for forming new hypotheses, we want to 468
provide a more summary view of the results of our experiment. However, the simple 469
question of which factor or groups of factors have the greatest influence on emotion raises 470
convoluted methodological and theoretical issues. Let us first concretize the question 471
with the following scenario: Given that we find a new poem, which information should 472
we determine first in order to get a guess about its emotions. The mean posterior 473
distribution of the h parameter (or its transformation into odds ratios) allows us to 474
see how much the odds to observe an emotion changes, given the associated factor is 475
actually present (e.g. is a love poem). While the pure rate of change is exciting in its own 476
right, we need to look at the intercept c as well, because quadrupling a small probability 477
may be less informative than doubling a large one. Assume a factor is characterized 478
by a small intercept c and a very large slope h . The information gained would be great 479
if we found that our poem has the properties of this factor, if not we get not much 480
more certainty about the emotion distribution. That means, we should factor into our 481
consideration how likely our poem is to have a feature before we actually determine it. 482
The only way to get this chance is to look into the distribution of factors in our corpus. 483
By doing so, we assume that the poems outside our corpus follow a similar distribution 484
of factors. Unfortunately, although our corpus is relatively large, its genesis is far from a 485
random drawing from the set of all poems of a literary epoch. However, not including 486
the frequency distribution of the factors in the corpus in the calculation is also not an 487
option, since this would mean assuming an equal distribution of the factors within 488
poetry in general, which is even less likely. The same applies to the distribution of the 489
emotions themselves. A slight tendency against the occurrence of an otherwise very 490
common emotion (e.g. love) can tell us more than a strong correlation with an emotion 491

that is very rare anyway. By taking these considerations into account, we propose the following equation to capture this information:

$$\sum_i (|p(f)h_{if} + c_{if}|)p(em^i) \quad [formula1] \quad (1)$$

It represents the sum of the absolute value of slope h , normalized by the probability that a factor f is present in a poem and intercept c for every emotion i . Afterwards the result is weighted by the probability of observing an emotion in our corpus $p(em^i)$. The values for $p(em^i)$ are depicted in Figure 1. The values for the factor groups Job and Genre are averaged analogously to the values in Table 4. The result is shown in Table 5, which state, that we should first determine a poem's genre to gain the most information about its potential emotions. The profession of the author comes second and after that rhyme, gender and verse length.

Table 5: Score of the factor groups calculated with formula 1.

Profession	Genre	Rhyme	Gender	Verse length
0.16	0.18	0.13	0.12	0.10

6. Discussion

Are the results surprising? What strength of association between features and emotions might we expect according to previous research in (traditional) literary studies? As argued in Section 3.3, for virtually all features (perhaps except for profession), we find at least some statements claiming that the features are relevant for differentiating literature in general and/or the representation of emotions in particular (while hardly anyone claims that any of the features have no relevance at all). Our results broadly confirm these research statements in that for each feature, at least some association with one or more emotions was found. This is true as well for the extra-literary features 'profession' and 'gender', showing that the representation of emotions in poetry is, at least in part, related to factors outside the literary text itself. On a more specific level, however, some qualifications are in order, since we did not find clear associations between *all* features and *all* individual emotions. For example, Figure 5 has shown that there is no strong connection between gender and the representation of love, although some statements, such as the one cited in Section 3.3 by Margarete Huch, might suggest the opposite.

Beyond that, explicit hypotheses that *rank* or *compare* the relevance of multiple features (e.g., "For literary phenomenon X, genre is more important than gender"), are extremely rare in literary studies. However, to have at least some point of reference, we can take a look at the *practice* of literary studies and analyze whether our results correlate with it. To this end, we surveyed literary histories (McInnes and Plumpe 1996, Fähnders 1998, Sprengel 1998, Mix 2000, Sprengel 2004, Aust 2006, Ajouri 2009, Stockinger 2010, Beutin et al. 2013, Willems 2014, Willems 2015, Sprengel 2020) as well as all results of a search for the keyword "emotion" in a major bibliography of German literary studies (<https://www.bdsl-online.de/>). Mainly on the basis of prefaces and outlines (literary histories) or titles (research on emotion according to the bibliography), we assessed which of the selected features literary scholars most prominently focus on,

use for delimiting their objects of study and/or claim to be relevant. It is important 528
to keep in mind that, in addition to assumed relevance, numerous other (reasonable) 529
factors are likely to influence which features literary scholars highlight and/or use to 530
organize or name their studies. It is for this reason, among others, that our purpose is 531
not to 'evaluate' or 'refute' the practice of literary scholars; rather, we use their work as a 532
reference point to frame our findings. 533

Of all the features analyzed in our study, literary scholars are by far the most likely 534
to focus on (thematic) *genre*. Literary histories are regularly organized by genre, and 535
research is quite often concerned with how emotions are represented in particular 536
(thematic) subgenres. Consistent with this, we observe that, according to our model, 537
the representation of emotions is most strongly associated with genre (Table 5). The next 538
most common focus of researchers seems to be *gender*. At least some literary histories 539
include individual chapters on gender and there are multiple research contributions 540
that investigate the relationship between gender and the representation of emotions, 541
even though they do not always focus on the gender of the author. In any case, gender is 542
a more prominent theme in the surveyed contributions than the other three features 543
'profession', 'rhyme', and 'verse length'. However, our results show that the relationship 544
between author gender and the represented emotions, while present, is comparatively 545
limited. The association of gender with emotion is about as strong as that of rhyme and 546
verse length, and weaker than that of profession. As already indicated, the features 547
'profession', 'rhyme,' and 'verse length' receive much less attention in literary histories 548
and research on emotions in literature. Although these features (especially rhyme) are 549
occasionally mentioned, e.g. in the context of single-text analyses, there are hardly any 550
studies on literary emotions or chapters in the surveyed literary histories that focus 551
primarily on one of these factors. Regarding our results, it is fitting that verse length is 552
relatively weakly related to the representation of emotion. But the comparatively strong 553
association of rhyme, and especially profession, is quite surprising. We certainly did 554
not expect the connection of emotion with rhyme to be as strong as that with thematic 555
genre, and the connection with profession to be as strong as that with gender. 556

We explained in our introduction that we don't assume a causal relationship between 557
these factors and the emotions in the poems, and that we don't assume that these factors 558
have the same kind of relationship to the emotions. But it is important to understand the 559
limitations of our findings. If we look at the relationship between rhyme and agitation, 560
for example, we see a rather high value of 0.55 in Table 4. This needs to be taken with 561
a grain of salt, because this analysis is based on the automatically labeled verses, and 562
the F1 score for detecting agitation is quite low, only 0.62. In addition, the distribution 563
of support for different values of rhyme is relevant. If we divide this variable into four 564
groups, all lines are rhymed, none, more than 50% of the lines (but less than 100%), 565
less than 50% (but more than none), we see that it is quite unevenly distributed. And 566
even if we accept this result with all these caveats, we are still only talking about a 567
connection where the nature of the connection and the reasons for its existence are 568
unknown. Perhaps, from the contemporaries' point of view, the disruption of the rhyme 569
or its complete suspension was perceived as a textual strategy, which, in the context of 570
an ideal congruence between content and form, was seen as a good option for expressing 571
something like agitation. Or the expression of agitation is often linked to the expression 572
of abstract themes, and the complexity of expressing them precisely is particularly 573

difficult given the limits of rhyme. Or the preference for unrhymed metrical forms of Greek-Roman antiquity is particularly high in relation to the expression of agitation. All of these, and a combination of them, could be causal reasons for this relationship.

These considerations also point in the direction of our future work. We will continue to identify interesting ideas about the connections between external factors and aspects of literary texts in literary studies and try to systematically integrate them into a more formal model. We will try to find and digitize more data about these external factors. And we will try to provide a richer representation of the textual features we include in our model, for example include more features like style, rhetorical devices like figurative speech etc. Finally, we want to take the first steps toward a model that offers not just descriptions of relationships, but causal explanations - understood in the broadest possible sense.

7. Data Availability

Data can be found here: <https://anonymous.4open.science/r/Connecting-the-Dots-615F>

8. Software Availability

Software can be found here: <https://anonymous.4open.science/r/Connecting-the-Dots-615F>

9. Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft as part of the SPP 2207 Computational Literary Studies in the project The beginnings of modern poetry – Modeling literary history with text similarities.

10. Author Contributions

Leonard Konle: TBD-Conceptualization, Writing – original draft

Merten Kröncke: TBD-Conceptualization, Writing – original draft

Simone Winko: TBD-Methodology, Project administration

Fotis Jannidis: TBD-Conceptualization, Writing – original draft

References

- Ajouri, Philip (2009). *Literatur um 1900: Naturalismus, Fin de Siècle, Expressionismus*. Akademie Studienbücher Literaturwissenschaft. OCLC: 254323074. Berlin: Akad.-Verl. 253 pp. ISBN: 978-3-05-004536-8.
- Andreotti, Mario (2014). *Die Struktur der modernen Literatur: Neue Wege in die Textanalyse. Einführung Epik und Lyrik*. 5th ed. Vol. 1127. Wien/Köln/Weimar. 294 pp.

- Aust, Hugo (2006). *Realismus*. Lehrbuch Germanistik. Stuttgart: J. B. Metzler. ISBN: 9783476018649. 607 608
- Barrett, Louise and Robin Dunbar, eds. (Apr. 5, 2007). *Oxford Handbook of Evolutionary Psychology*. 1st ed. Oxford University Press. ISBN: 9780198568308. 10.1093/oxford 609 610
 hb/9780198568308.001.0001. <https://academic.oup.com/edited-volume/28165> 611
 (visited on 01/31/2023). 612
- Betancourt, Michael and Mark Girolami (2015). "Hamiltonian Monte Carlo for Hierarchical Models". In: *Current Trends in Bayesian Methodology with Applications*. Chapman 613
 and Hall/CRC. ISBN: 9780429172373. 614 615
- Beutin, Wolfgang, Matthias Beilein, Klaus Ehlert, Wolfgang Emmerich, Christine Kanz, 616
 Bernd Lutz, Volker Meid, Michael Opitz, Carola Opitz-Wiemers, Ralf Schnell, Peter 617
 Stein, and Inge Stephan (2013). *Deutsche Literaturgeschichte*. Achte, aktualisierte und 618
 erweiterte Auflage. Stuttgart: J.B. Metzler. ISBN: 9783476008138. 619
- Borkowski, Jan (2015). *Literatur und Kontext. Untersuchungen zum Text-Kontext-Problem 620*
 aus textwissenschaftlicher Sicht. Münster: Mentis. 621
- Carper, Thomas and Derek Attridge (2003). *Meter and meaning: an introduction to rhythm 622*
 in poetry. OCLC: ocm52347475. New York: Routledge. 156 pp. ISBN: 9780415311748. 623
- Chan, Branden, Stefan Schweter, and Timo Möller (Dec. 3, 2020). "German's Next 624
 Language Model". In: *arXiv:2010.10906 [cs]*. arXiv: 2010.10906. <http://arxiv.org> 625
 /abs/2010.10906 (visited on 07/15/2021). 626
- Detering, Heinrich and Peer Trilcke, eds. (2013). *Geschichtsliteratur. Ein Kompendium*. 2 vols. 627
 OCLC: ocn861177035. Göttingen: Wallstein. 2 pp. ISBN: 9783835312937. 628
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (Oct. 11, 2018). 629
 "BERT: Pre-training of Deep Bidirectional Transformers for Language Understand- 630
 ing". In: <https://arxiv.org/abs/1810.04805v2> (visited on 06/24/2021). 631
- Ekman, Paul (May 1992). "An argument for basic emotions". In: *Cognition and Emotion* 632
 6.3, 169–200. ISSN: 0269-9931, 1464-0600. 10.1080/02699939208411068. [https://www.](https://www.tandfonline.com/doi/full/10.1080/02699939208411068) 633
[tandfonline.com/doi/full/10.1080/02699939208411068](https://www.tandfonline.com/doi/full/10.1080/02699939208411068) (visited on 08/03/2022). 634
- (1999). "Basic Emotions". In: *Handbook of Cognition and Emotion*. Ed. by Tim Dalgleish 635
 and Mick J. Power. Chichester, UK: John Wiley & Sons, Ltd, 45–60. 10.1002/0470 636
 013494.ch3. <https://onlinelibrary.wiley.com/doi/10.1002/0470013494.ch3> 637
 (visited on 08/03/2022). 638
- Engel, Manfred (2018). "Kontexte und Kontextrelevanzen in der Literaturwissenschaft". 639
 In: *KulturPoetik* 18.1, 71–89. 640
- Fähnders, Walter (1998). *Avantgarde und Moderne 1890-1933*. Lehrbuch Germanistik. 641
 Stuttgart/Weimar: J. B. Metzler. 642
- Gittel, Benjamin (2016). "Lässt sich literarischer Wandel erklären? Struktur, Gültigkeits- 643
 bedingungen und Reichweite verschiedener Erklärungstypen in der Literaturgeschichts- 644
 chreibung". In: *Journal of Literary Theory* 10.2, 303–344. 645
- Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug 646
 Downey, and Noah A. Smith (July 2020). "Don't Stop Pretraining: Adapt Language 647
 Models to Domains and Tasks". In: *Proceedings of the 58th Annual Meeting of the 648*
Association for Computational Linguistics. Online: Association for Computational Lin- 649
 guistics, 8342–8360. 10.18653/v1/2020.acl-main.740. [https://aclanthology.org](https://aclanthology.org/2020.acl-main.740) 650
 /2020.acl-main.740. 651
- Haider, Thomas (2021). "Metrical Tagging in the Wild: Building and Annotating Poetry 652
 Corpora with Rhythmic Features". In: *Proceedings of the 16th Conference of the European 653*



- Chapter of the Association for Computational Linguistics: Main Volume*. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, 3715–3725. [10.18653/v1/2021.eacl-main.325](https://aclanthology.org/2021.eacl-main.325). <https://aclanthology.org/2021.eacl-main.325> (visited on 01/31/2023).
- Härle, Gerhard (2007). *Lyrik, Liebe, Leidenschaft: Streifzug durch die Liebeslyrik von Sappho bis Sarah Kirsch*. Göttingen: Vandenhoeck und Ruprecht. ISBN: 9783525208502.
- Hiebel, Hans H. (2005). *Das Spektrum der modernen Poesie. 1: 1900 - 1945*. Würzburg: Königshausen & Neumann. 332 pp. ISBN: 9783826032004.
- Huch, Margarete, ed. (1911). *Frauenlyrik der Gegenwart. Eine Anthologie*. Leipzig: Fritz Eckardt.
- Kalliney, Peter (2019). "Introduction: Literary History after the Nation?" In: *Modern Language Quarterly* 80.4. Ed. by Peter Kalliney and Simon Gikandi, 359–377.
- King, Martina and Jesko Reiling (Jan. 1, 2014). "Das Text-Kontext-Problem in der literaturwissenschaftlichen Praxis: Zugänge und Perspektiven". In: *Journal of Literary Theory* 8.1, 2–30. ISSN: 1862-8990, 1862-5290. [10.1515/jlt-2014-0001](https://www.degruyter.com/document/doi/10.1515/jlt-2014-0001/html). <https://www.degruyter.com/document/doi/10.1515/jlt-2014-0001/html> (visited on 01/31/2023).
- Konle, Leonard and Fotis Jannidis (2020). "Domain and Task Adaptive Pretraining for Language Models". In: *Computational Humanities Research*. Amsterdam. <http://eur-ws.org/Vol-2723/short33.pdf>.
- Konle, Leonard, Fotis Jannidis, Merten Kröncke, and Simone Winko (2022). "Emotions and Literary Periods". In: *DH Conference Abstracts*. Digital Humanities. Tokyo.
- Kröncke, Merten, Fotis Jannidis, Leonard Konle, and Simone Winko (Feb. 9, 2022). "Annotationsrichtlinien Emotionsmarker und Emotionen". In: Publisher: Zenodo Version Number: 1.1. [10.5281/ZENODO.6020616](https://zenodo.org/record/6020616). <https://zenodo.org/record/6020616> (visited on 02/12/2022).
- Kröncke, Merten, Leonard Konle, Fotis Jannidis, and Simone Winko (2023). "Gattungen und Emotionen in der Lyrik des Realismus und der frühen Moderne". In: *DHd 2023*. Trier.
- Ladegaard, Jakob and Jakob Gaardbo Nielsen (2019). "Introduction: the question of context". In: *Context in Literary and Cultural Studies*. Comparative Literature and Culture. London: UCL Press, 1–13.
- Langer, Lars, Manuel Burghardt, Roland Borgards, Katrin Böhning-Gaese, Ralf Sepelt, and Christian Wirth (2021). "The rise and fall of biodiversity in literature: A comprehensive quantification of historical changes in the use of vernacular labels for biological taxa in Western creative literature". In: *People and Nature* 3.5. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/pan3.10256>, 1093–1109. ISSN: 2575-8314. [10.1002/pan3.10256](https://onlinelibrary.wiley.com/doi/abs/10.1002/pan3.10256). <https://onlinelibrary.wiley.com/doi/abs/10.1002/pan3.10256> (visited on 01/31/2023).
- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier (Sept. 2015). "The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment". In: *Computational Linguistics* 41.3, 437–479. ISSN: 0891-2017, 1530-9312. [10.1162/COLI_a_00227](https://direct.mit.edu/coli/article/41/3/437-479/110.1162/COLI_a_00227). https://direct.mit.edu/coli/article/41/3/437-479/110.1162/COLI_a_00227 (visited on 01/31/2023).

- McElreath, Richard (2020). *Statistical rethinking: a Bayesian course with examples in R and Stan*. Second edition. Texts in statistical science series. Boca Raton London New York: CRC Press, Taylor & Francis Group. 593 pp. ISBN: 9780429639142.
- McInnes, Edward and Gerhard Plumpe, eds. (1996). *Bürgerlicher Realismus und Gründerzeit, 1848–1890*. Hansers Sozialgeschichte der deutschen Literatur vom 16. Jahrhundert bis zur Gegenwart 6. München/Wien: Carl Hanser.
- Mix, York-Gothart, ed. (2000). *Naturalismus, Fin de siècle, Expressionismus (1890-1918)*. München.
- Obermeier, Christian, Winfried Menninghaus, Martin von Koppenfels, Tim Raettig, Maren Schmidt-Kassow, Sascha Otterbein, and Sonja A. Kotz (2013). "Aesthetic and Emotional Effects of Meter and Rhyme in Poetry". In: *Frontiers in Psychology* 4. ISSN: 1664-1078. 10.3389/fpsyg.2013.00010. <http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00010/abstract> (visited on 01/31/2023).
- Piper, Andrew (2022). "Biodiversity is not declining in fiction". In: *Journal of Cultural Analytics* 7.3. 10.22148/001c.38739. <https://culturalanalytics.org/article/38739-biodiversity-is-not-declining-in-fiction>.
- Plutchik, Robert (1980a). "A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION". In: *Theories of Emotion*. Elsevier, 3–33. ISBN: 978-0-12-558701-3. 10.1016/B978-0-12-558701-3.50007-7. <https://linkinghub.elsevier.com/retrieve/pii/B9780125587013500077> (visited on 08/03/2022).
- (1980b). "A psychoevolutionary theory of emotions". In: *Social Science Information* 21.4, 529–553. ISSN: 0539-0184, 1461-7412. 10.1177/053901882021004003. <http://journals.sagepub.com/doi/10.1177/053901882021004003> (visited on 08/03/2022).
- (2001). "The Nature of Emotions". In: 89.4, 344–350.
- ŠeĽa, Artjoms, Petr Plecháč, and Alie Lassche (Sept. 15, 2021). "Semantics of European poetry is shaped by conservative forces: The relationship between poetic meter and meaning in accentual-syllabic verse". In: *arXiv:2109.07148 [cs]*. arXiv: 2109.07148. <http://arxiv.org/abs/2109.07148> (visited on 11/08/2021).
- Shaver, Philipp, Judith Schwartz, Donald Kirson, and Cary O'Connor (1987). "Emotion Knowledge: Further Exploration of a Prototype Approach". In: *Journal of Personality and Social Psychology* 52.6, 1061–1086.
- Sobchuk, Oleg and Peeter Tinitis (Aug. 26, 2020). "Cultural Attraction in Film Evolution: the Case of Anachronies". In: *Journal of Cognition and Culture* 20.3. Publisher: Brill, 218–237. ISSN: 1568-5373, 1567-7095. 10.1163/15685373-12340082. https://brill.com/view/journals/jocc/20/3-4/article-p218_3.xml (visited on 01/30/2023).
- Sprengel, Peter (1998). *Geschichte der deutschsprachigen Literatur 1870-1900. Von der Reichsgründung bis zur Jahrhundertwende*. München.
- (2004). *Geschichte der deutschsprachigen Literatur 1900–1918. Von der Jahrhundertwende bis zum Ende des Ersten Weltkriegs*. Vol. 2. Geschichte der deutschen Literatur von den Anfängen bis zur Gegenwart 9. 924 pp. ISBN: 9783406521782.
- (2020). *Geschichte der deutschsprachigen Literatur 1830–1870. Vormärz-Nachmärz*. Geschichte der deutschen Literatur von den Anfängen bis zur Gegenwart 8. München: C.H. Beck. 781 pp. ISBN: 9783406007293.
- Stockinger, Claudia (2010). *Das 19. Jahrhundert. Zeitalter des Realismus*. Berlin.
- Thomsen, Mads Rosendahl (2019). "From data to actual context". In: *Context in Literary and Cultural Studies*. Comparative Literature and Culture. London: UCL Press, 190–209.

- Tsur, Reuven (Aug. 7, 2017). "Metre, rhythm and emotion in poetry. A cognitive approach". In: *Studia Metrica et Poetica* 4.1, 7–40. ISSN: 2346-691X, 2346-6901. [10.12697/smp.2017.4.1.01](https://ojs.utlib.ee/index.php/smp/article/view/smp.2017.4.1.01). <http://ojs.utlib.ee/index.php/smp/article/view/smp.2017.4.1.01> (visited on 01/31/2023).
- Underwood, Ted, Kevin Kiley, Wenyi Shang, and Stephen Vaisey (May 2, 2022). "Cohort Succession Explains Most Change in Literary Culture". In: *Sociological Science* 9, 184–205. ISSN: 2330-6696. [10.15195/v9.a8](https://sociologicalscience.com/article/s-v9-8-184/). <https://sociologicalscience.com/article/s-v9-8-184/> (visited on 05/02/2022).
- Wiecki, Thomas (2017). *Why hierarchical models are awesome, tricky, and Bayesian*. While My MCMC Gently Samples. <https://twiecki.io/blog/2017/02/08/bayesian-hierarchical-non-centered/> (visited on 01/31/2023).
- Willems, Gottfried (May 14, 2014). *Geschichte der deutschen Literatur. Band 4: Vormärz und Realismus*. 1st ed. Stuttgart, Deutschland: utb GmbH. ISBN: 9783838538747. [10.36198/9783838538747](https://elibrary.utb.de/doi/book/10.36198/9783838538747). <https://elibrary.utb.de/doi/book/10.36198/9783838538747> (visited on 01/31/2023).
- (Oct. 28, 2015). *Geschichte der deutschen Literatur. Band 5: Moderne*. 1st ed. Stuttgart, Deutschland: utb GmbH. ISBN: 9783838542492. [10.36198/9783838542492](https://elibrary.utb.de/doi/book/10.36198/9783838542492). <https://elibrary.utb.de/doi/book/10.36198/9783838542492> (visited on 01/31/2023).
- Winko, Simone, Leonard Konle, Merten Kröncke, and Fotis Jannidis (2022). *Korpusbeschreibung der Lyrikanthologien 1850-1910*. <https://doi.org/10.5281/zenodo.6053972>.

Computational approaches to opera libretti

An experiment on DraCor corpora

Luca Giovannini¹ 
Daniil Skorinkin² 

1. Institute for German Studies, University of Potsdam, Potsdam, Germany.
2. Digital Humanities Network, University of Potsdam, Potsdam, Germany.

Citation

Luca Giovannini and Daniil Skorinkin (2023). "Computational approaches to opera libretti. An experiment on DraCor corpora". In: *Journal of Computational Literary Studies* (conference reader 2023). tbc

Date published 2023-06-09

Date accepted 2023-04-21

Date received 2023-02-17

Keywords

opera, drama, libretto, genre

License

CC BY 4.0 

Reviewers

tbc

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. The paper offers a first computationally-informed look at German and French opera libretti by modelling them on the basis of their structural features. On one side, it strives to assess whether libretti – a relatively recent genre, born in the early 1600s – exhibit peculiar formal properties which set them apart from contemporary comedies and tragedies from the same linguistic environment. On the other side, it explores the structural development of the genre across history, evaluating how its relationship with traditional genres changed in different periods. Results confirm known challenges of modelling dramatic texts while pointing out to some structural patterns which help identify libretti, particularly non-comic ones.

1. Introduction

Antonio Salieri's one-act piece *Prima la musica e poi le parole* (1786) opens with a Poet and a Composer rushing to put together an opera within four days. According to the Composer, the task should be pretty easy: the score is ready, and now his collaborator should just adapt some words to it. The Poet protests:

Questo è l'istesso, / che far l'abito, e poi / far l'uomo a cui s'adatti.
("That would be the same, as first designing a dress, and then creating a man who would fit it.")

The Composer, however, immediately retorts:

Voi signori poeti, siete matti. / Amico, persuadetevi; chi mai / credete che dar voglia attenzione / alle vostre parole? / Musica in oggi, musica / ci vuole.
("You poets are crazy. My friend, be persuaded: who do you think would pay attention to your words? Music is what we need nowadays.")

By the end of the play, the Poet has begrudgingly come to accept this argument, and thus the piece's title ("First the music, and then the words") seems vindicated. Nonetheless, Salieri's elegant formula was not, by all means, the last word on the issue, nor was it the comment by his Salzburg colleague that "bey einer Oper muß schlechterdings die Poesie der Musick gehorsame Tochter seyn" ("in an opera poetry must absolutely be the music's obedient daughter", qtd. in Kesting 2005: 21),

Indeed, the dispute on the relative weight and importance given to music and words

within the symbiotic construction of operas both predates and goes on after Mozart and Salieri ¹. As Gorlée 1997: 237 points out, one could recognise among opera theorists and practitioners an ongoing confrontation between a “musicocentric” and a “logocentric” approach, with the first one being somehow prevalent throughout the centuries. As a reaction, the second half of the XX century actually saw the birth of a new discipline, librettology, which aimed at investigating operatic texts from a literary point of view ².

In this paper, we offer a first contribution towards a ‘digital librettology’ by investigating libretti ³ through modern computational techniques. We believe indeed that quantitative methods hold some promise in exploring the libretto as an autonomous dramatic form, insofar as they allow to efficiently follow its evolution over time and to gauge its changing relation with other established genres.

In the following, we will therefore try to discover whether operatic texts possess a peculiar “genre signal” which sets them apart from contemporary comedies and tragedies within the same linguistic environments. At the same time, we will analyse the structural development of the genre across history, documenting how its relationship with the traditional genres changed in different periods. More specifically, we will work on a sizable corpus of French- and German-language drama and explore which features best describe their libretti in structural terms, employing computational methods such as dimensionality reduction algorithms, statistical significance tests and feature importance analysis within a classification procedure.

2. Related literature

The ‘digital turn’ affecting the humanities in recent decades seems to have had a limited impact on opera studies. While there have been substantial efforts to build digital databases on operatic materials and performances ⁴, the implementation of computational methods to investigate libretti appears still quite underdeveloped. Among existing literature, Muñoz-Lago et al. 2020 proposed a layered graphical visualisation of operas’ structures, working on texts by Pietro Metastasio, while Jeong and Yoo 2022 applied *k-means* clustering to confirm the validity of traditional periodisation frameworks. Furthermore, Bonora and Pompilio 2021 have worked on the automatic extraction of descriptions of operatic characters through lexical and syntactic patterns. There have also been some attempts to employ sentiment analysis on libretti, especially non-Western ones (Jin et al. 2022, Jeong 2021); a sentiment analysis of arias on the basis of linked recitatives was also proposed by Gervás and Torrente 2022.

Comprehensive computational analyses of opera libretti, however, are still lacking, but such an endeavour might certainly profit from the experience accumulated in the cognate

1. The same topic is discussed, for example, in E. T. A. Hoffmann’s short story *Der Dichter und der Komponist* (1813) and in Richard Strauss’ last opera, *Capriccio* (1941).

2. Pioneers in this sense were, among others, Patrick J. Smith (*The Tenth Muse: A Historical Study of the Opera Libretto*, 1971) and Albert Gier (*Oper als Text: Romanistische Beiträge zur Libretto-Forschung*, 1986; *Das Libretto: Theorie und Geschichte einer musikoliterarischen Gattung*, 1998).

3. This is an admittedly rough moniker we employ throughout the paper to designate modern dramatic texts where music plays a central role. Although dramatic forms had some sort of musical accompaniment since Antiquity, with music being one of the components of tragedy already in Aristotle (*Poet.* 1450a, 10), the first integration of the two aspects in an art form (retrospectively) perceived as new is normally recognised in the early seventeenth-century Italian melodrama – not without some uncertainties (cf. Leopold 2003).

4. An inventory of these sources would go beyond the scope of this paper and is therefore not provided.

field of quantitative drama analysis, where early approaches such as those by Markus 57
1970 or Reichert 1964 eventually opened the way for scores of studies with different 58
methodologies, ranging from stylometry to topic modelling or networks analysis (see 59
e.g. Cuéllar 2023, Lehmann and Padó 2022, Estill and Meneses 2018, Fischer et al. 2017a, 60
Algee-Hewitt 2017) ⁵. This last technique has proven particularly useful for capturing 61
structural patterns within large corpora and modelling literary concepts like plot or 62
characters' system (Fischer et al. 2017b, Trilcke et al. 2022). 63

3. Corpus building 64

For our research we employed the German- and French-language corpora from the 65
DraCor project ⁶, an open-access platform for hosting, accessing and analysing theatrical 66
texts (Fischer et al. 2019). All plays in the DraCor collections are encoded in a semanti- 67
cally rich TEI-XML format, with specific annotation of character speech (which allows 68
in turn to generate co-presence networks) and additional metadata on the texts and 69
their authors. 70

Crucially for our purposes, the DraCor markup contains a `textclass` element with a 71
descriptive genre tag ("Tragedy", "Comedy", etc.) and the genre's Wikidata entity, as in 72
the following example: 73

```
1 <textClass> 74
2 <keywords> 75
3 <term type="genreTitle">Tragedy</term> 76
4 </keywords> 77
5 <classCode scheme="http://www.wikidata.org/entity/">Q80930</classCode> 78
6 </textClass> 79
```

The four Wikidata-linked genres currently present in the DraCor markup are *tragedy* 80
(Q80930), *comedy* (Q40831), *tragicomedy* (Q192881), and *libretto* (Q131084); if no genre is 81
given, the text class element is empty. It is important to note a major difference between 82
GerDraCor and FreDraCor: while the first corpus treats any genre label as exclusive, 83
the second one allows "libretto" to coexist with other tags (for example, Chabanon's *Le* 84
Toison d'Or ⁷ is marked *both* as a tragedy and as a libretto). While this heterogeneity likely 85
stems from the corpora's different sources ⁸, it also underlines the blurred boundaries 86
of genre attribution in different cultural contexts. 87

To ease our operationalisation, however, we decided to normalise the French genre 88
column by having only one genre per play, i.e. marking all libretti with additional genre 89
tags only as 'libretti'. This methodological choice had several reasons. On one side, we 90
considered that genre labels in FreDraCor markup were seemingly auto-generated from 91
the *Théâtre Classique* ones, sometimes through non-transparent patterns, and thus were 92
not fully reliable. Furthermore, some additional uncertainty was due to the conventions 93
of the genre itself: as Senici 2014: 38 points out, "[p]erhaps the weakest contribution to 94

5. A good overview on the state of the field was offered by the recent *Workshop on Computational Drama Analysis: Achievements and Opportunities* (Cologne, 14-15 September 2022).

6. <https://dracor.org/ger>; <https://dracor.org/fre>.

7. <https://dracor.org/api/corpora/fre/play/chabanon-toison-d-or/tei>.

8. GerDraCor was (mainly) derived from the TextGrid repository, (<https://textgridrep.de>) while FreDraCor originates from the *Théâtre Classique* database (<https://theatre-classique.fr/index.html>).

the genrification of opera [came] from the discursive space where genre is normally explicitly named, that is, the generic indicator on published librettos".

In other words, especially at the beginning of opera history, the choice of (sub)titling a work "tragedy" or "comedy" was a deeply rhetorical one, and had more to do with the perception the author intended to convey (e.g. that of an elevated, serious work) than with the text's actual properties. Eventually, our first guess was that the intended usage of a libretto (as component of an operatic staging) would have been more 'distinctive' than its broader thematic alignment along the comic/tragic axis, and thus we choose to keep it as the primary label.

Another issue in the corpora was the large number of texts without any genre label. Again, we tried to address it by exploiting DraCor's Linked-Open-Data capabilities, since the plays' markup often contains a link to the Wikidata item of the work itself (the following example is from Wagner's *Der fliegende Holländer*⁹)

```

1 <standOff>
2 ...
3 <listRelation>
4 <relation name="wikidata" active="https://dracor.org/entity/ger000245"
   passive="http://www.wikidata.org/entity/Q114640"/>
5 </listRelation>
6 </standOff>

```

Scraping the plays through the Python library BeautifulSoup, we recursively accessed all Wikidata items and checked if they contained the Wikidata property P136, designating "genre", and used it (after some manual disambiguation of the results) to assign a genre to unlabelled plays. Unfortunately, the information gain was limited (18 new labels for German plays, 2 for French ones).

On another note, we had also to take into account that not all libretti in our corpora might have been properly marked as such, owing to the profusion of different terminology for designating operatic texts. It has been indeed noted that even in the cradle of opera, Italy, a "plethora of terms" for indicating such works "circulated freely" for decades before one label, *dramma per musica* ("music drama"), emerged in the Venice milieu and eventually became dominant (Senici 2014: 38). A similar situation was therefore to be expected in other areas as well, where various translations of the Italian loanword *opera* (*Oper*, *opéra*) long coexisted with local, often quite diverse (sub)generic denominations.

To ensure no possible libretto was neglected, we searched GerDraCor and FreDraCor for all plays which contained in their title or subtitle at least one of the German or French genre tags that the authoritative *New Grove Dictionary of Music and Musicians* associates, to various degrees, with opera¹⁰. After manually cleaning up the results and removing some false positives, we grouped the newly found libretti under the "libretti (attributed)" label. As a last step, we excluded all texts still without an assigned genre, which would have marred the visualisation without providing any added value for the

9. <https://dracor.org/api/corpora/ger/play/wagner-der-fliegende-hollaender/tei>.

10. Searched terms included: *ballet de cour*, *ballet-héroïque*, *burlesque*, *comédie-ballet*, *divertissement*, *drame lyrique*, *entrée*, *grand opéra*, *intermède*, *Lehrstück*, *Liederspiel*, *Märchenoper*, *masque*, *Monodrama*, *opéra-ballet*, *opéra bouffon*, *opéra comique*, *opéra-féerie*, *pantomime*, *pastorale-héroïque*, *Posse*, *Schuldrama*, *Schuloper*, *Singspiel*, *Spieloper*, *tragédie en musique*, *vaudeville*, *Zauberoper*, *Zeitoper* (see Brown et al. 2001 s.v.)

interpretation. Table 1 shows the final composition of our two research samples.

Genre	German sample	French sample
Tragedies	140	312
Comedies	156	692
Tragicomedies	8	82
Libretti (marked up)	55	58
Libretti (attributed)	28	34
Total	387	1178
Percentage of libretti	21.4%	7.8%

Table 1: Final French and German drama samples.

4. Experiments

In our investigation, we programmatically choose to avoid formulating a rigid initial hypothesis on how a standard libretto would have looked like, i.e. which features would have been more distinctive of the genre as compared to the other ones. Rather, in the vein of John Tukey’s exploratory data analysis (Tukey 1977), we preferred to let data speak for themselves, and we thus organised our research as a series of concatenated ‘experiments’ which tried to exploit the epistemological potential of feature analysis to its fullest extent.

First, we focused our attention on the relation between libretti and the other two main historical genres (comedy and tragedy), with the goal of mapping it on a multidimensional space. A similar, informal attempt at examining the interplay between different genres in an early version of GerDraCor corpus was already made by Trilcke et al. 2015, who measured the evolution of some features in the different genres. By looking at two significant network metrics, i.e. size and density, they argued that the “evolution [of drama] over two centuries shows [...] clearly the proximity of comedy and libretto and the persistent distance from the tragedy”.

To further test this hypothesis and explore more thoroughly the topology of German and French drama, we decided to adapt a method recently proposed by Szemes and Vida 2022 to cluster dramatic genres according to “asemantic” (i.e. content-independent) properties. In our case, however, we did not aim to perform a classification task, but were rather interested in finding out which features (if any) set libretti apart from comedies and tragedies.

The method developed by Szemes and Vida relied on a number of measures, mostly related to network properties and speech distribution corpora, which were developed for the study and/or obtained from DraCor metadata through the API. Based on these metrics, they carried out a supervised classification procedure on DraCor’s German and Shakespeare¹¹ corpora to label comedies and tragedies using the Support Vector Machine (SVM) method. They found no “striking difference” between the two genres insofar as structural features were concerned, but were nonetheless able to single out some properties which are highly predictive of generic alignment – thus “confirm[ing]

11. <https://dracor.org/shake>.

the existence of a “genre fingerprint” that shapes the dramatic structure of tragedies and comedies” (10), and which authors may (un)consciously choose to adhere to or depart from.

Following a similar approach, we also took as starting point the array of features provided by the DraCor API through the corpora’s metadata tables. Out of the 41 features available, we employed almost all the count-based ones, i.e. size-related metrics or measures deriving from social network analysis. Unlike Szemes and Vida, however, we chose a leaner implementation and did not compute additional speech- or distribution-related measures. We considered the following features:

- num_of_segments: number of subdivisions (scenes or acts) in the plays
- num_of_speakers: number of characters with at least one utterance
- num_of_person_groups: number of characters marked as groups (e.g. “Women”)
- word_count_sp: total word count in characters’ utterances
- word_count_stage: total word count in stage directions
- average_degree: average number of nodes to which a node is connected
- density: ratio between the number of actual edges and the maximum number of possible edges
- average_clustering: average of the local clustering coefficients of all the vertices
- max_degree: maximum number of nodes to which a node is connected
- num_of_connected_components: number of independent subgraphs within the network
- diameter: the longest shortest path between two nodes
- average_path_length: average length of the shortest path that can be drawn between any two nodes

4.1 Naïve visualisation

As a first step in our exploratory analysis, we plotted the plays (i.e. their feature vectors) on a two-dimensional plane. To this aim, we tested various unsupervised (i.e. class-blind) techniques for dimension reduction such as Principal Component Analysis (PCA), T-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) (see Maaten et al. 2009, Burges 2010, Waggoner 2021 etc.). Standard PCA – a non-random linear mapping algorithm which tends to capture global structure better than local similarities, as t-SNE and UMAP do – was eventually chosen for plotting the various timeframes and implemented through the sklearn Python library. This first attempt, however, led to somehow disappointing

12. This feature is available only for GerDraCor.

13. Cf. Watts and Strogatz 1998.

outputs, with both corpora displaying a substantial structural homogeneity between all four genres across most timeframes.

Eventually, one had to come to terms with the fact that, at least as our selection of formal variables was concerned, it seemed not possible to confidently detect a process of progressive “genrification” of libretti. While critical consensus tends to present the first two centuries of opera in terms of its “crystallization into a specific, identifiable genre” (Campana 2012: 206), the absence of meaningful clusters seemed however to suggest that the novelty of libretti was difficult to capture through size- and network-related features only – especially if one kept envisioning the libretto as a *unitary* genre with clear-cut properties such as the traditional ones (cf. Szemes and Vida 2022: 16).

Accordingly, we decided to move away from this perspective and tried instead to account for the libretti’s heterogeneity by re-labelling them with respect to their proximity to the main genres. Instead of using the traditional *comic/tragic* dichotomy, however, we chose to employ a binary *comic/non-comic* tagging, because qualitative analysis of genre descriptors (i.e. subtitles) in our sample showed how labels referring to comedy in an explicit (e.g. *Komödie für Musik*, *comédie-ballet*) or implicit manner (*divertissement*, *vaudeville*, *Posse*) were more frequent than the ones related to tragedy (e.g. *Trauerspiel*, *tragédie en musique*) or without a transparent reference (e.g. *opéra*, *Oper*)¹⁴. This led us to the conclusion that, at least as genre assignment was involved, comic-like libretti were somehow easier to identify and model as a group as against non-comic ones.

Accordingly, we semi-automatically extracted the new labels from the subtitles through keywords and run again the PCA algorithm, with the results being presented in Figures 1 and 2. While the model performed better than before, validating to some extent our refining of the libretti labels, it still failed to produce clear-cut clusterings of operatic texts; on top of that, as research by Szemes and Vida 2022 already showed, even the fundamental distinction between comic and tragic zones remained often blurred.

On the other hand, however, the choice of splitting our data in different timeframes represented a valuable improvement on previous attempts, insofar as it highlighted some topological idiosyncrasies which would have got lost in a catch-all visualisation. It is the case, for example, of the second French data frame (plays between 1670 and 1719), where a pattern is indeed visible: while comic libretti, as expected, mostly follow the structural model of comedies, non-comic libretti are clearly distinguishable from all other genres and build a definite, albeit sparse aggregation on the right side of the graph.

The relatively more pronounced clustering of French dramatic genres as against German ones, which the PCA plots show, could be explained by the corpora’s different size and, even more, by their temporal coverage. While the French corpus mostly spreads over a period of normative aesthetics, where texts like d’Aubignac’s *Pratique du théâtre* (1657) or Boileau’s *Art poétique* (1674) set the rules for theatre-writing, the German corpus starts at a time in which Classicism was already losing ground and French theatrical conventions were being actively repudiated or deconstructed (cf. Lessing’s *Hamburgische Dramaturgie*, 1767-1769) – leading to more pronounced interferences between genres

14. The German sample has 39 clearly “comic” libretti and 41 unassigned libretti, while the French one has 45 clearly “comic” libretti, 18 clearly “tragic” libretti, and 27 unassigned libretti.

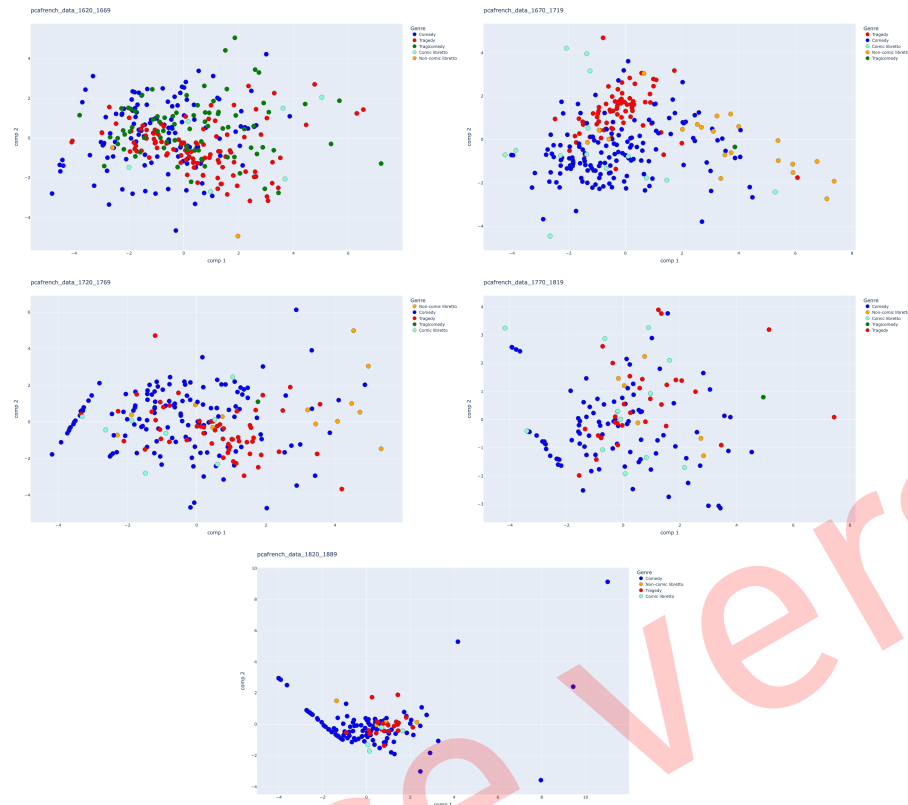


Figure 1: Evolution of French drama, 1626-1889, visualised through PCA.

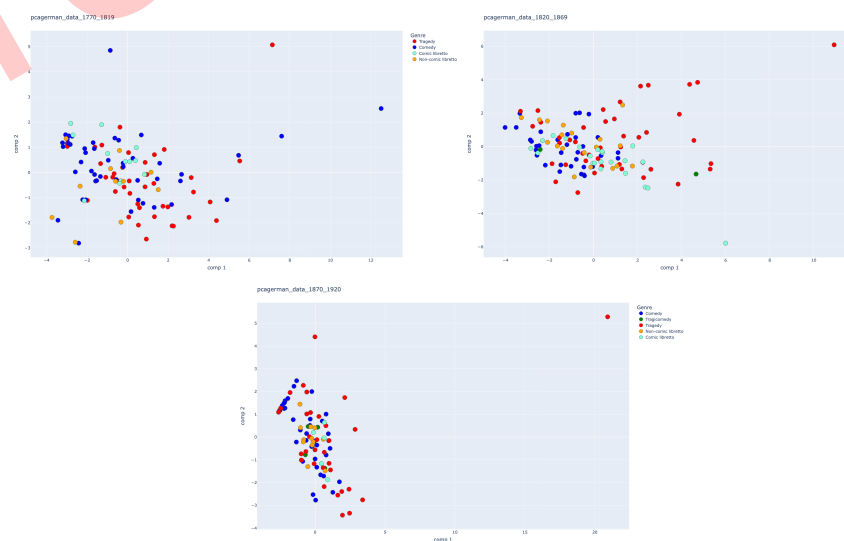


Figure 2: Evolution of German drama, 1770-1920, visualised through PCA.

Features	German sample	French sample
num_of_segments	0.45	0.01868097637357296
num_of_speakers	0.2	4.313551673714e-08
num_of_person_groups	5.8599304523887e-07	n/a
word_count_sp	2.706883066487e-08	4.5e-19
word_count_stage	0.09	1.1e-19
average_degree	0.017723062602512586	0.16
density	0.72	6.27730611604e-09
average_clustering	0.18	0.82
max_degree	0.08	1.529018559478078e-05
num_connected_components	0.21	1.1645e-16
diameter	0.21	0.00916968964547866
average_path_length	0.38	0.002316750707412838

Table 2: Statistical significance (p-value) of features in the two samples according to the Wilcoxon Rank Sum test. Numbers in bold are the ones below the significance threshold ($\alpha = 0.05$), i.e. the most significant.

which the graph seems to capture.

4.2 Statistical testing and classifier

While the PCA yielded some first insights into the composition of our corpora, it also pushed us to develop a stronger understanding of the individual features composing our textual vectors. In order to pinpoint which measures were most useful for the analysis, we began by measuring statistical significance in the features' differences, with the aim of assessing how such parameters performed in telling apart libretti and non-libretti¹⁵. After applying the Shapiro-Wilk test for checking the normality of distributions – which as expected were almost always not normal – we implemented the non-parametric Wilcoxon Rank Sum test to check if the differences between the distributions were substantial (Table 2).

As an additional benchmark, we also run a binary random forest classifier tasked to differentiate between libretti and non-libretti; the number of estimators was optimised through iterative hyperparameter tuning based on five-fold cross-validation. At the beginning of the procedure, we calculated correlation coefficients to assess interdependence between variables. We did it separately for the two corpora since their features are not perfectly overlapping (e.g. FreDraCor does not contain encoding for the collective characters and thus no num_of_person_groups).

As the matrices (Figures 3 and 4) show, some features displayed an excessive correlation ($>\pm 0.75$), and keeping both of them might have misled the classifier. In each pair or triplet of correlated features we therefore chose to keep the ones easier to interpret and discard the other ones: this translated into dropping average path length, diameter, maximum degree, and number of connected components (while keeping density and number of speakers) for the German corpus and dropping number of segments, average path length and maximum degree (while keeping word count of speeches, diameter, number of speakers and average degree) for the French one.

15. From here onwards, we removed tragicomedies from our sample because of their irregular chronological distribution and globally low number.

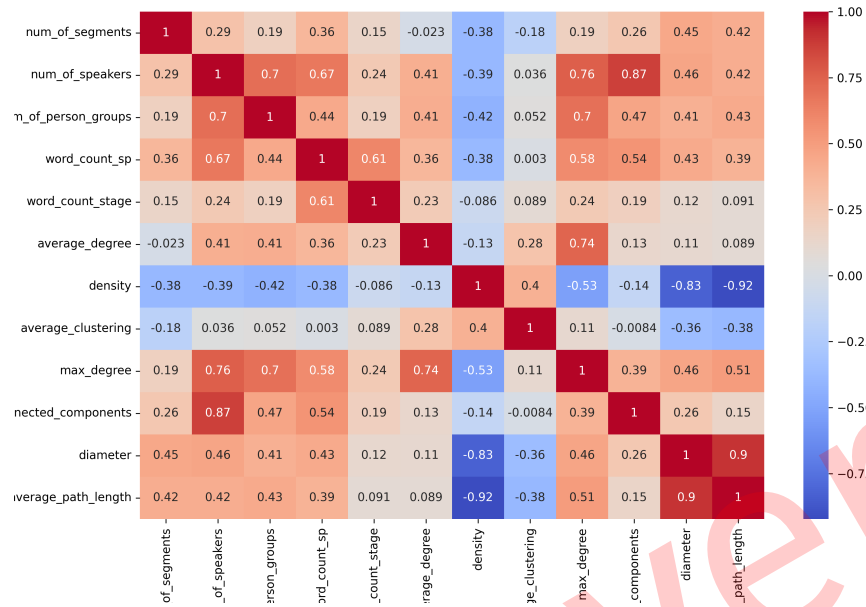


Figure 3: Correlation matrix for the German corpus.

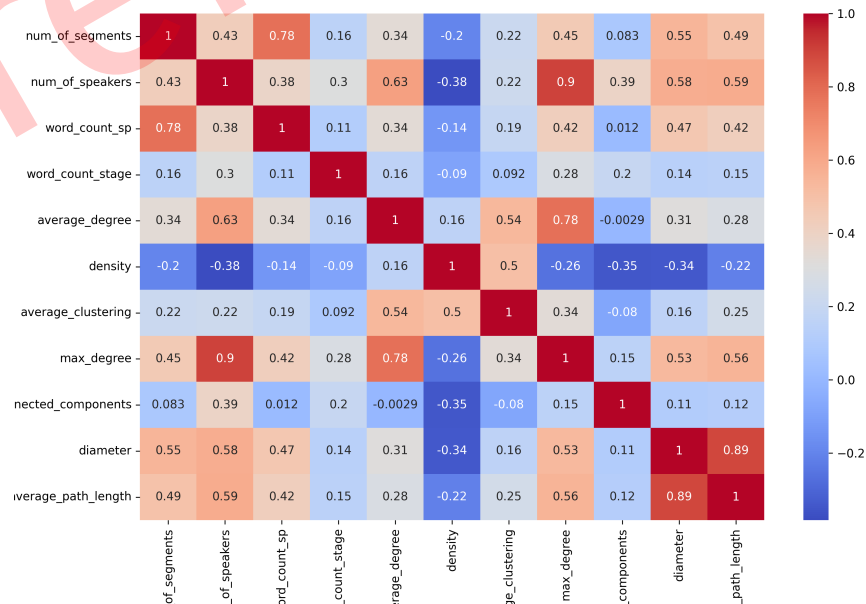


Figure 4: Correlation matrix for the French corpus.

Measures	German sample	French sample
overall accuracy	84.3%	93.1%
precision for class <i>Non-libretto</i>	85.4%	94.1%
recall for class <i>Non-libretto</i>	96.6%	98.7%
f1 for class <i>Non-libretto</i>	90.7%	96.3%
precision for class <i>Libretto</i>	75.6%	68.2%
recall for class <i>Libretto</i>	38.7%	31.1%
f1 for class <i>Libretto</i>	51.2%	42.7%

Table 3: Values for the best performing random forest classifier.

Results after running the model were unsatisfying: while high overall accuracies were expected on such an imbalanced dataset, the binary classifier also showed mediocre performance in identifying libretti (i.e. the minority class) in both corpora (see Table 3). Our intention, however, was not finding a method to efficiently automate libretti classification, but rather ascertaining whether the most relevant features for the classification task (presented in Figure 5) were the same whose variance was statistically significant according to our tests.

Crossing the results of the two pipelines (statistical significance tests and classifier), one could eventually argue that three features (*num_of_person_groups*, *word_count_sp*, *average_degree*) were particularly helpful for distinguishing German libretti from non-libretti, while six of them proved useful for sorting out the French data (*num_of_speakers*, *word_count_sp*, *word_count_stage*, *density*, *diameter*, *num_connected_components*).

4.3 Scatterplots

At this point, we charted some of the most interpretable features as scatterplots¹⁶, using them as hermeneutical tools for discovering which traits were distinctive for libretti at different stages of history. In order to better capture the granularity of the process, we switched again to a four-class visualisation (comedies/tragedies/comic libretti/non-comic libretti) and plotted each play individually; we also applied a local regression algorithm (LOWESS, cf. Cleveland 1979) to draw a smooth curve between the data points and help visualise the distances between the genres and their evolution.

While the unequal distribution of texts across the investigated timespan suggests caution in speaking of *longue-durée* evolutionary phenomena, trends emerging from the pictures give a glimpse into some long-lasting relations between genres. One of the patterns seemingly common to both German and French data, for instance, is the relative independence of non-comic libretti from the other genres in terms of several structural features.

To start with, such phenomenon can be observed when comparing the two most discriminative features for the mapping of German data, i.e. the total word count for utterances and the number of collective entities (Figure 6). The two curves show indeed how non-comic libretti (the yellow line) chart quite an autonomous path, while comic libretti often adhere more to the structural model set by comedies, as seen especially in the *word_count_sp* graph.

16. We removed outliers which were more than three standard deviations away from the mean.

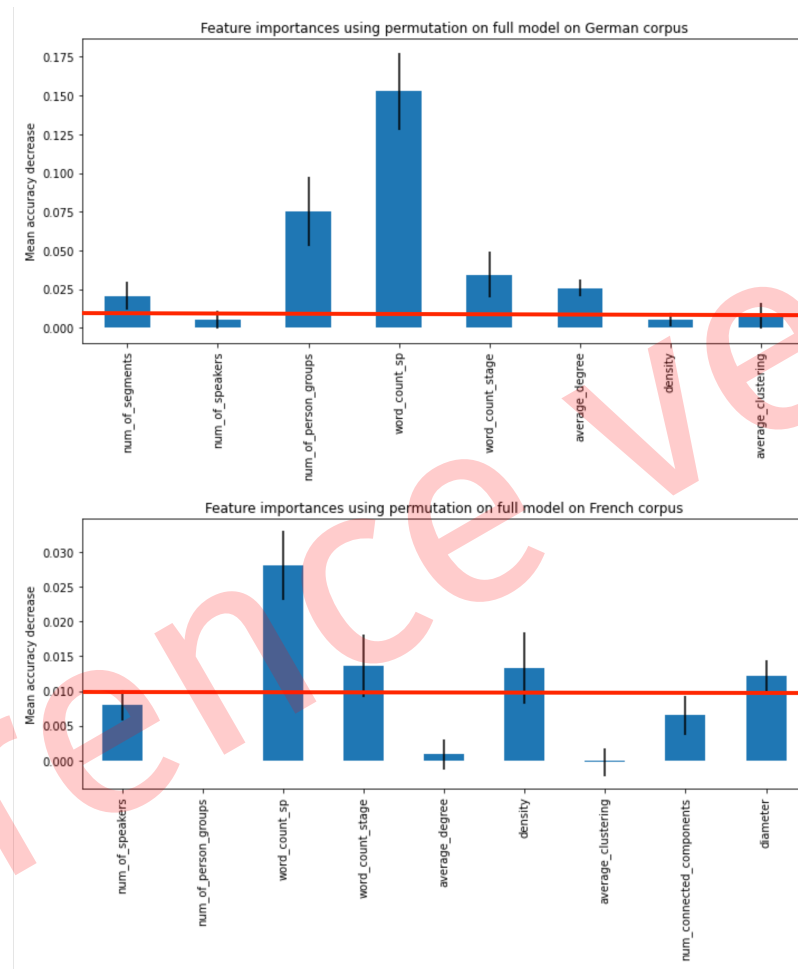


Figure 5: Relative features importance according to the random forest classifier. The barplot indicates which features, if left out, lead to the biggest accuracy decrease. We deemed relevant features which cause an accuracy decrease equal or superior to 0.01 (the red line), with confidence intervals taken into account.

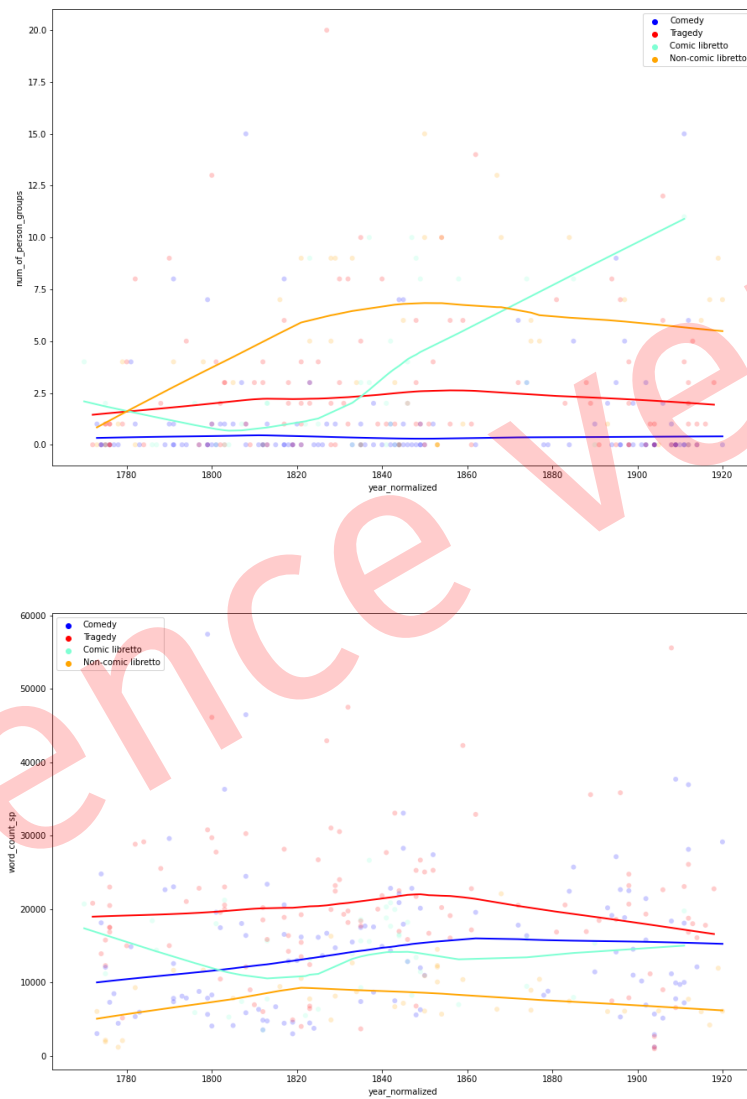


Figure 6: Evolution of selected features in German data: *word_count_sp* (below) and *num_of_groups* (above). The numbers in the graph indicate the number of plays in this timeframe.

Such pattern, which the sparse nature of the German data space makes somehow less evident, emerges with more force in the French data. Figure 7, for example, shows scatterplots for the network-related features `num_of_speakers` and `density`, which are often found as inversely correlated (the higher the number of the characters in a play, the lower the chance they interact with all the other ones). While all genres broadly follow this pattern, comic libretti tend to structurally resemble comedies in having somehow tighter plots, with fewer characters which are highly interconnected, while non-comic libretti display an unusual dissimilarity from any other genres.

A supplementary confirmation of the non-comic libretti's peculiar status comes from another classification experiment we conducted, where we asked the random forest algorithm to sort plays into our four genres. As the confusion matrices in Figure 8 and Figure 9 illustrate, the classifier struggled to distinguish comic libretti and comedies more often than in separating non-comic libretti from tragedies.

Some prominent spikes which are especially visible in the non-comic libretti curves, coupled with the clustering evidence emerging from the PCA, concur in underscoring the particular relevance of the second half of the XVII century – something one cannot explain only on account of the higher number of texts available in that timeframe. Critics indeed consider it as a pivotal moment in the diffusion of opera beyond Italy and into France, where new hybrid forms of theatre, music, and dance (such as the *comédie-ballet*, best exemplified by Molière's *Le Bourgeois gentilhomme*) were soon joined by truly "operatic" (i.e. completely sung) genres such as the *tragédie lyrique*.

The merit of popularising this last mode of expression, which has been considered "the definitive form [of French opera], capable of rivalling the spoken theatre" (Norman 2009: 17), is to be shared between composer Jean-Baptiste Lully and librettist Philippe Quinault, whose operas account indeed for almost one third of the plays written between 1670 and 1719. Closer inspection of some of these operas, such as *Cadmus et Hermione* (1673)¹⁷ or *Persée* (1682)¹⁸, confirms the aforementioned quantitative findings: they feature in fact large ensemble casts whose characters often appear together, thus resulting in well-connected social networks.

Due to higher text availability, French scatterplots are also useful for empirically verifying traditional assumptions on opera by literary critics and musicologists. This is the case, for example, of the relation between diegetic and non-diegetic elements within a dramatic text. As Ulrich Weisstein has argued in his seminal essay *The Libretto as Literature*, for example, "music lacks the speed and verbal dexterity of language, [and] fewer words are needed in opera than would be required in a play of comparable length"; therefore, "librettos are usually shorter than the texts of ordinary dramas, and often to the point of embarrassing the listener or reader" (Weisstein 1961: 19). On the other side, one would expect operatic texts to have a greater share of stage directions, due to the necessity of setting the stage for musical numbers or dances (something along the lines of "Enter five dancers, dressed as knights..."). As Figure 10 illustrates, these trends seem indeed confirmed in both kinds of libretti: comic and non-comic operatic texts follow similar

17. <https://dracor.org/fre/quinault-cadmus-hermione>.

18. <https://dracor.org/fre/quinault-persée>.

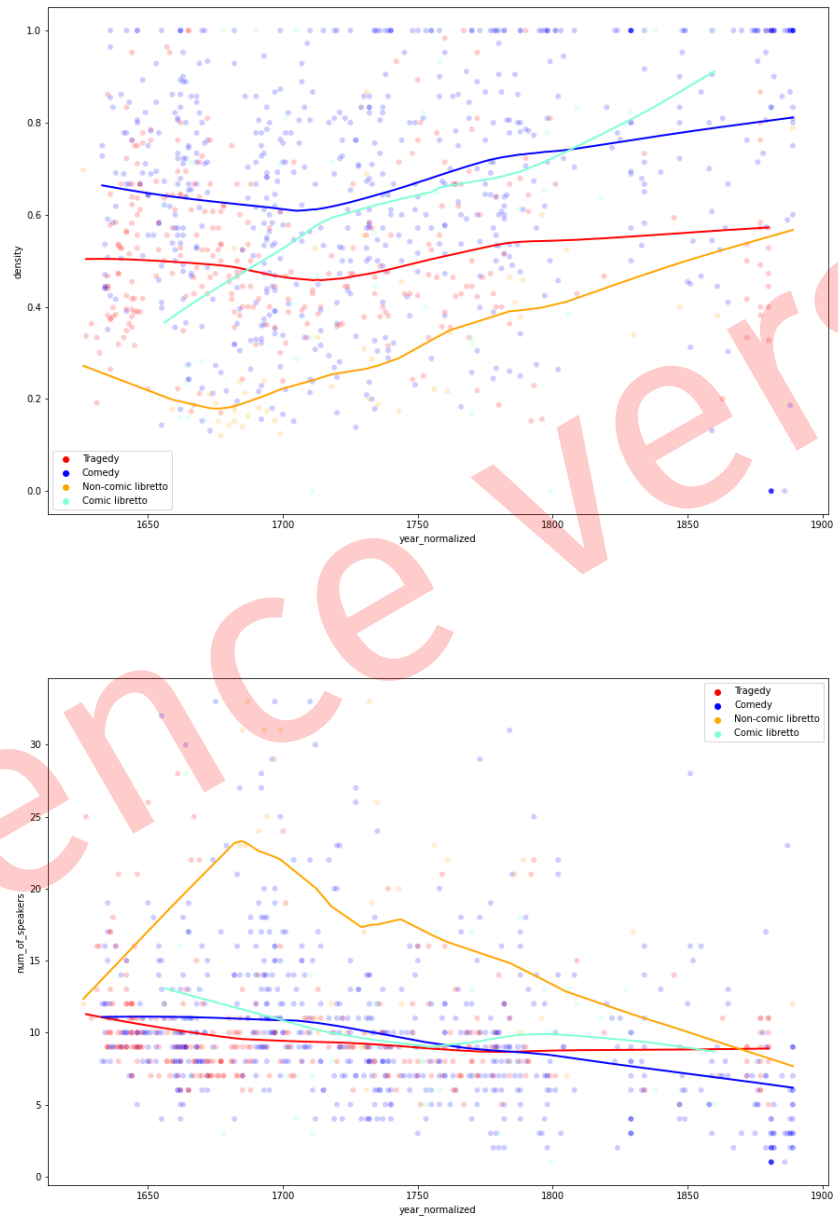


Figure 7: Evolution of selected features in French data (I): *density* (above) and *num_speaker* (below).

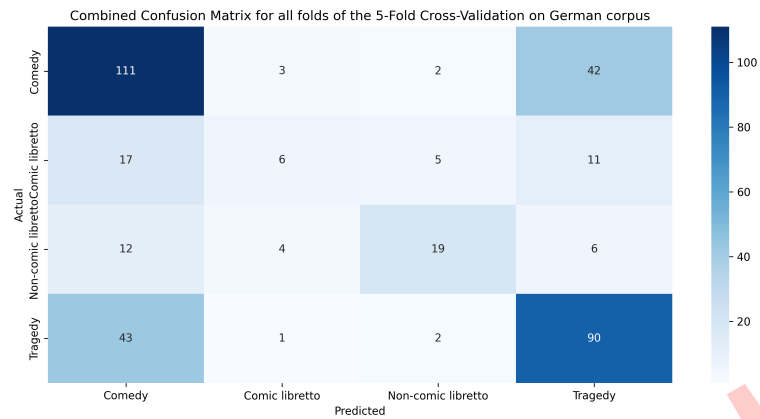


Figure 8: Confusion matrix for a four-class classifier trained and tested on the German sample.

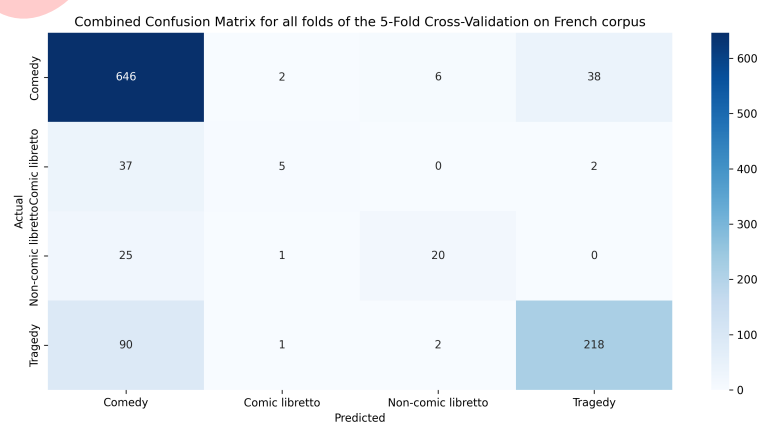


Figure 9: Confusion matrix for a four-class classifier trained and tested on the French sample.

paths in having less character speech and (sizeably) more stage directions.¹⁹

343

5. Conclusion

344

As in many CLS projects, a major limitation of this investigation was represented by the relatively small size of the corpora employed. By relying on DraCor, one of the largest scholarly databases of dramatic texts available, we tried to collect a sample big enough to draw meaningful conclusions, but the results were sometimes mixed. While some timeframes were populated enough to give some actual insight on the dynamics of the age, the results obtained in others might be radically changed by any corpora overhaul, be it in terms of corpus enlargement or markup refining.

351

Reduced text availability also forced us to focus on only two cultural milieus (German and French) and forsake any further comparative attempt. The absence of texts from Italy, one of opera's major playfields, is particularly lamentable, but since DraCor's Italian corpus²⁰ contains so far only a handful of libretti (mostly early *melodrammi* by Metastasio), adding them would have not dramatically improved the quality of our findings while posing several challenges in the implementation. Future studies, however, might exploit the wealth of freely accessible Italian libretti²¹ to perform more encompassing analyses, while possibly enlarging ItaDraCor as well (through the onboarding procedure described in Börner et al. 2023).

360

Eventually, much more work is needed in order to achieve a satisfying diachronic picture of opera in its relationship with other dramatic genres. Nonetheless, our first computational foray into operatic texts yielded some insights on libretti as part of the wider dramatic system. On one side, our mapping of operatic texts clearly showed how "[o]pera spread throughout Europe both as an 'Italian' product and also as a 'native' musical theatre" (Campana 2012: 207, emphasis added), with its local iterations possessing idiosyncratic features which make them stand apart. Specifically, the findings revealed how the two types of libretti often display different behaviours, with comic libretti mainly aligning with comedies but non-comic libretti manifesting a definite distance from both tragedy and the other genres.

370

Furthermore, analysis of the different timeframes through PCA clusterings and feature lineplots also suggested that it is more difficult to discriminate effectively between the two kinds of libretti in German data, while such distance is more substantial within the French dramatic space. This difference is even clearer if one plots through PCA only German and French operatic texts (see Figure 11).

375

On the other side, our clustering attempts on a selection of purely formal (size- and network-based) features failed to identify libretti as a genre possessing a strong degree of formal independence. Such outcomes actually play into the established critical narrative which sees opera as a Protean art form, whose generic essence is continuously contested:

379

19. Coincidentally, the general increase in the share of stage directions for all genres – except for non-comic libretto – supports Trilcke et al. 2020's argument about a "tendency to epification" in the history of drama even beyond their original case study (GerDraCor).

20. <https://dracor.org/ita>.

21. Many libretti with simple or no markup, mostly in HTML or PDF format, are available in online databases such as <https://opera-guide.ch>, <https://opera.stanford.edu>, <https://librettidopera.it>, <https://www.operalib.eu>, etc.

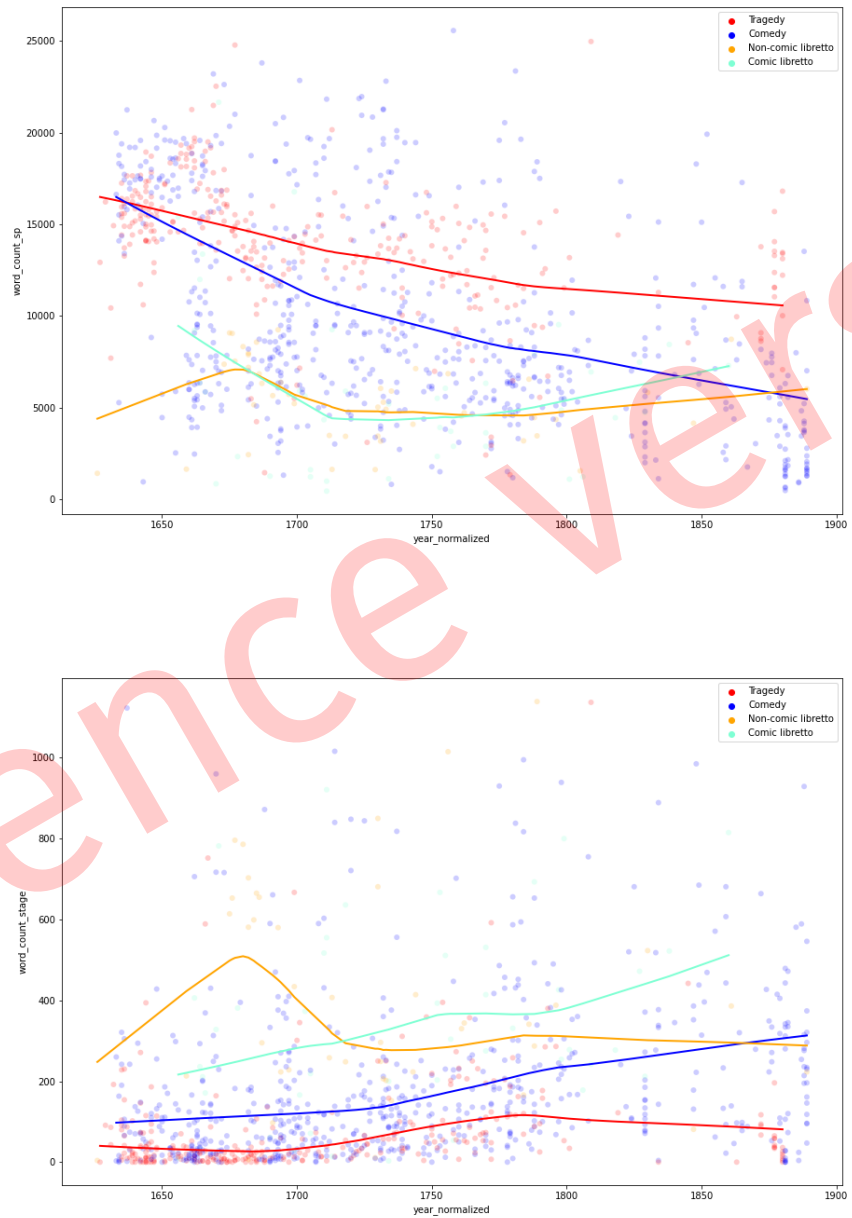


Figure 10: Evolution of selected features in French data (II): *word_count_sp* and *word_count_stage*.



Figure 11: Principal Components Analysis for German and French libretti, with centroids (X).

Opera's identity as a genre [...] relies from the start on mixing, as a contamination of music and theatre, music and word, singing and acting, showing and telling. Thus, it defines itself historically and systematically as a hybrid, challenging at the outset the foundational law of genre discourse. [...] Even more so than literary genres, the hybridity of opera, and its dialogue with the demands of production and performance, contains the possibility of genre being disrupted. (Campana 2012: 205)

Nonetheless, our operationalisation did show that one could identify, to some extent, traits which are clearly distinctive of comic and non-comic libretti, and that such traits do not always align with the ones characterising comedies and tragedies – thus pointing to a complex relationship of imitation/departure from the spoken theatre models.

Ultimately, this paper showcases once again the complexity of modelling the relationship between different dramatic genres – and more generally, the concept of dramatic text itself – on account of formal features. On one side, it could be argued that the array of features employed here was not sufficient to construct a meaningful and productive representation of the plays. The adequateness of network-related metrics, for example, can be called into question, since they don't appear particularly useful in telling libretti apart from other genres. On the other side, however, we believe that one cannot achieve major performance optimisations without a more radical rethinking of operationalisation patterns – be it for this specific task (a contrastive definition of libretti in the German and French milieus) or for a broader formal analysis of drama. Eventually, further experiments will be needed to approach the clustering effectiveness and explanatory power of other state-of-art CLS techniques ²².

22. See e.g. Schöch 2017's successful attempt to apply topic modelling to French Classical and Enlightenment theatre, based on texts now contained in FreDraCor.

6. Data Availability 403

Data and scripts employed can be found here: <https://github.com/DanilSko/opera> 404
a. 405

7. Acknowledgements 406

The authors would like to thank Artjoms Šeļa, Henny Sluyter-Gäthje, and Peer Trilcke 407
for their helpful suggestions. 408

8. Author Contributions 409

Luca Giovannini: Conceptualisation, Formal Analysis, Methodology, Writing 410

Daniil Skorinkin: Software, Visualisation, Methodology, Formal Analysis 411

References 412

- Algee-Hewitt, Mark (2017). “Distributed character: Quantitative models of the English stage, 1550–1900”. In: *New Literary History* 48.4, 751–782. [10.1353/nlh.2017.0038](https://doi.org/10.1353/nlh.2017.0038). 413
414
- Bonora, Paolo and Angelo Pompilio (2021). “Estrazione automatica delle caratteristiche del personaggio d’opera attraverso pattern lessico-sintattici”. In: *Umanistica Digitale* 5.10, 193–210. [10.6092/issn.2532-8816/12426](https://doi.org/10.6092/issn.2532-8816/12426). 415
416
417
- Börner, Ingo, Frank Fischer, Luca Giovannini, Christopher Lu, Carsten Milling, Daniil Skorinkin, Henny Sluyter-Gäthje, and Peer Trilcke (2023). “Onboard onto DraCor: Prototyping Workflows to Homogenize Drama Corpora for an Open Infrastructure”. In: *Dhd 2023 Conference Abstracts*. University of Luxemburg and Trier, Luxemburg/Germany. [10.5281/zenodo.7711513](https://doi.org/10.5281/zenodo.7711513). 418
419
420
421
422
- Brown, Howard Mayer, Ellen Rosand, Reinhard Strohm, Michel Noiray, Roger Parker, Arnold Whittall, Roger Savage, and Barry Millington (2001). “Opera (i)”. In: *Grove Music Online*. Oxford University Press. [10.1093/gmo/9781561592630.article.4072](https://doi.org/10.1093/gmo/9781561592630.article.4072) 423
424
425
6. 426
- Burges, Christopher J. C. (2010). *Dimension Reduction: A Guided Tour*. Boston and Delft: now. 427
428
- Campana, Alessandra (2012). “Genre and poetics”. In: *The Cambridge Companion to Opera Studies*. Ed. by Nicholas Till. Cambridge: Cambridge University Press, 202–224. [10.1017/CCO9781139024976.013](https://doi.org/10.1017/CCO9781139024976.013). 429
430
431
- Cleveland, William S. (1979). “Robust Locally Weighted Regression and Smoothing Scatterplots”. In: *Journal of the American Statistical Association* 74.368, 829–836. [10.1080/01621459.1979.10481038](https://doi.org/10.1080/01621459.1979.10481038). 432
433
434
- Cuéllar, Álvaro (2023). “Stylometry and Spanish Golden Age Theatre: An Evaluation of Authorship Attribution in a Control Group of Undisputed Plays”. In: *Digital Stylistics in Romance Studies and Beyond*. Ed. by Robert Hesselbach, José Calvo Tello, Ulrike Henny-Krahmer, Christof Schöch, and Daniel Schlör. Forthcoming. Heidelberg: Heidelberg University Press. 435
436
437
438
439

- Estill, Laura and Luis Meneses (2018). "Is Falstaff Falstaff? Is Prince Hal Henry V?: Topic Modeling Shakespeare's Plays". In: *Digital Studies/Le champ numérique* 8.1. [10.16995/dscn.295](https://doi.org/10.16995/dscn.295).
- Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtel, Christopher Kittel, Carsten Milling, and Peer Trilcke (2019). "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama". In: *Proceedings of DH2019: "Complexities"*. University of Utrecht, The Netherlands. [10.5281/zenodo.4284002](https://doi.org/10.5281/zenodo.4284002).
- Fischer, Frank, Gilles Dazord, Mathias Göbel, Christopher Kittel, and Peer Trilcke (2017a). "Le drame comme réseau de relations : une application de l'analyse automatisée pour l'histoire littéraire du théâtre". In: *Revue d'historiographie du théâtre*. <https://hal.science/hal-01811799>.
- Fischer, Frank, Mathias Göbel, Dario Kampkaspar, Christopher Kittel, and Peer Trilcke (2017b). "Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts". In: *Proceedings of DH2017: "Access/accès"*. URL: <https://dh2017.adho.org/abstracts/071/071.pdf>. McGill University, Montreal, Canada.
- Gervás, Pablo and Álvaro Torrente (2022). "Emotional Interpretation of Opera Series: Impact of Specifics of Drama Structure". In: *14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. <http://nil.fdi.ucm.es/sites/default/files/KDIR2022-opera-CRC.pdf>.
- Gorlée, Dinda (1997). "Intercode Translation: Words and Music in Opera". In: *Target. International Journal of Translation Studies* 9.2, 235–270. [10.1075/target.9.2.03gor](https://doi.org/10.1075/target.9.2.03gor).
- Jeong, Harim (2021). "Study on sentiment analysis for Opera". In: *Proceedings of APIC-IST 2021*, 89–91.
- Jeong, Harim and Joo Hun Yoo (2022). "Opera Clustering: K-means on librettos datasets". In: *Journal of Internet Computing and Services* 23.2, 45–52. [10.7472/jksii.2022.23.2.45](https://doi.org/10.7472/jksii.2022.23.2.45).
- Jin, Cong, Zhen Song, Jiaqi Xu, and Huiyue Gao (2022). "Attention-Based Bi-DLSTM for Sentiment Analysis of Beijing Opera Lyrics". In: *Wireless Communications and Mobile Computing*. doi.org/10.1155/2022/1167462.
- Kesting, Hanjo (2005). *Der Musick gehorsame Tochter: Mozart und seine Librettisten*. Göttingen: Wallstein.
- Lehmann, Jörg and Sebastian Padó (2022). "Classification of comedies and tragedies written in Calderón de la Barca's Comedias Nuevas". In: *Zeitschrift für Digitale Geisteswissenschaft* 7. [10.17175/2022_012](https://doi.org/10.17175/2022_012).
- Leopold, Silke (2003). "Die Anfänge von Oper und die Probleme der Gattung". In: *Journal of the Seventeenth Century Music* 9. <https://sscm-jscm.org/v9/no1/leopold.html>.
- Maaten, Laurens van der, Eric O. Postma, and Jaap van den Herik (2009). *Dimensionality Reduction: A Comparative Review*. Preprint. https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf.
- Markus, Solomon (1970). *Poetica matematică*. Bucharest: Academiei.
- Muñoz-Lago, Paula, Nicola Usula, Emilia Parada-Cabaleiro, and Álvaro Torrente (2020). "Visualising the Structure of 18th Century Operas: A Multidisciplinary Data Science Approach". In: *24th International Conference on Information Visualisation*, 530–536. [10.1109/IV51561.2020.00091](https://doi.org/10.1109/IV51561.2020.00091).
- Norman, Buford (2009). *Quinault, librettiste de Lully: le poète des grâces*. trans. by Thomas Vernet and Jean Duron. Wavre: Mardaga.

- Reichert, Waltraud (1964). "Kybernetische Methoden der Dramenforschung". In: *Grundlagenstudien aus Kybernetik und Geisteswissenschaften* 5.3-4, 115–120. 487
488
- Schöch, Christof (2017). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". In: *Digital Humanities Quarterly* 11.2. <http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html>. 489
490
491
- Senici, Emanuele (2014). "Genre". In: *The Oxford Handbook of Opera*. Ed. by Helen M. Greenwald. Oxford University Press. 10.1093/oxfordhb/9780195335538.013.002. 492
493
- Szemes, Botond and Bence Vida (2022). "Tragic and Comical Networks: Clustering Dramatic Genres According to Structural Properties". In: *Workshop on Computational Drama Analysis: Achievement and Opportunities*. University of Cologne, Germany. 494
495
496
10.48550/arXiv.2302.08258. 497
- Trilcke, Peer, Frank Fischer, Mathias Göbel, and Dario Kampkaspar (July 2015). *Comedy vs. Tragedy: Network Values by Genre*. <https://dina.github.io/Network-Values-by-Genre>. 498
499
500
- Trilcke, Peer, Christopher Kittel, Nils Reiter, Daria Maximova, and Frank Fischer (2020). "Opening the Stage – A Quantitative Look at Stage Directions in German Drama." In: *Proceedings of DH2020: "carrefours/intersections"*. URL: https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html. University of Ottawa, Ottawa, Canada. 501
502
503
504
505
- Trilcke, Peer, Evgeniya Ustinova, Frank Fischer, Carsten Milling, and Ingo Börner (2022). "Detecting Small Worlds in a Corpus of Thousands of Theater Plays: A DraCor Study in Comparative Literary Network Analysis". In: *Workshop on Computational Drama Analysis: Achievement and Opportunities*. Forthcoming. University of Cologne, Germany. 506
507
508
509
510
- Tukey, John (1977). *Exploratory Data Analysis*. New York: Pearson. 511
- Waggoner, Philip D. (2021). *Modern Dimension Reduction*. Cambridge: Cambridge University Press. 512
513
- Watts, Duncan and Steven Strogatz (1998). "Collective dynamics of 'small-world' networks". In: *Nature* 393, 440–442. 10.1038/30918. 514
515
- Weisstein, Ulrich (1961). "The Libretto as Literature". In: *Books Abroad* 35.1, 16–22. 516
10.2307/40115290. 517