



Conference Reader
4th Annual Conference of
Computational Literary Studies
CCLS2025 Kraków
July 3-4, 2025

Venue	Jagiellonian University, Faculty of Philology Al. Mickiewicza 9, 31-120 Kraków
Local Organizer	Jagiellonian Centre for Digital Humanities Jan Rybicki
Web	https://jcls.io/site/ccls2025/
Contact	jchc@uj.edu.pl
Hashtag	#CCLS2025
JCLS Editors	Evelyn Gius, Christof Schöch, Peer Trilcke
JCLS Editorial Assistants	Svenja Guhr, Julian Häußler, Élodie Ripoll, Henny Sluyter-Gäthje
Contact JCLS	info@jcls.io https://jcls.io/site/contact/

Conference Programme

Thursday | July 3, 2025

9:15 a.m. to 9:30 a.m. | Opening

9:30 a.m. to 10:30 a.m. | Session 1

- Fotis Jannidis, Rabea Kleymann, Julian Schröter, Heike Zinsmeister: **Do Large Language Models Understand Literature? Case Studies and Probing Experiments on German Poetry**
- Keli Du, Uygur Navruz, Nazan Sınır, Julian Valline, Christof Schöch, Sarah Ackerschewski: **Reconstructing Shuffled Text. Bad Results for NLP, but Good News for Using In-Copyright Text**

11:00 a.m. to 12:30 a.m. | Session 2

- Maria Levchenko: **Computational Analysis of Literary Communities: Event-Based Social Network Study of St. Petersburg 1999-2019**
- Gilad Aviel Jacobson, Yael Dekel, Itay Marienberg-Milikowsky: **From Readers to Data. Uncertainty in Computational Literary Citizen Science**
- Julia Neugarten: **A Powerful Hades is an Unpopular Dude: Dynamics of Power and Agency in Hades/Persephone Fanfiction**

1:30 p.m. to 3:00 p.m. | Session 3

- Daniil Skorinkin, Boris Orekhov: **The Outward Turn: Geocoding the Expansion of Fictional Space in Russian 19th Century Literature**
- Svenja Guhr, Jessica Monaco, Alexander J. Sherman, Matt Warner, Mark Algee-Hewitt: **Making BERT Feel at Home. Modelling Domestic Space in 19th-Century British and Irish Fiction**
- Eva Eglāja-Kristsons, Anda Baklāne, Valdis Saulespurēns: **Urban Transportation in the Latvian Early Novels or “Why do you use a 19th-century horse-drawn cab when you have a 20th-century taxi?”**

3:30 p.m to 4:30 p.m. | Session 4

- Rongqian Ma, Keli Du, Yiwen Zheng: **Verse within Prose. Annotating and Classifying Narrative Functions of Embedded Poems in Chinese Qing (1644–1912) Vernacular Fiction**
- Natalie M. Houston: **Rhymefindr: An Historical Poetics Method for Identifying Rhymes in Nineteenth-Century English Poetry**

5:00 p.m. to 6:00 p.m. | Keynote

- Maciej Eder: **Text Analysis Made Simple (Kind of), or Ten Years of Stylo**

7:00 p.m. | Conference Dinner

Friday | July 4, 2025

9:15 a.m. to 10:45 a.m. | Session 5

- Katrin Rohrbacher: **Opening Worlds: Narrative Beginnings and the Role of Setting**
- Noa Visser Solissa, Andreas van Cranenburgh, Federico Pianzola: **Event Detection between Literary Studies and NLP. A Survey, a Narratological Reflection, and a Case Study**
- Andrew Piper: **Towards a Moral History of the Novel Using Large Language Models**

11:15 a.m. to 12:45 p.m. | Session 6





- Julia Havrylash, Christof Schöch: **Exploring Measures of Distinctiveness. An Evaluation Using Synthetic Texts**
- Allison Keith, Antonio Rojas Castro, Hanno Ehrlicher, Kerstin Jung, Sebastian Padó: **A Computation Analysis of Character Archetypes in the Works of Calderón de la Barca**
- Yuri Bizzoni, Pascale Feldkamp, Kristoffer L. Nielbo: **Encoding Imagism? Measuring Literary Imageability, Visuality and Concreteness via Multimodal Word Embeddings**

12:45 p.m. to 1:00 p.m. | Closing

Do Large Language Models understand literature?

Case studies and probing experiments on German poetry.

Fotis Jannidis¹ 
 Rabea Kleymann² 
 Julian Schröter³ 
 Heike Zinsmeister⁴ 

1. Institut für Deutsche Philologie, Julius-Maximilians-Universität Würzburg , Würzburg, Germany.
2. Institut für Germanistik und Interkulturelle Kommunikation, Technische Universität Chemnitz , Chemnitz, Germany.
3. Digital Literary Studies, LMU , München, Germany.
4. Institut für Germanistik, Universität Hamburg , Hamburg, Germany.

Citation

Fotis Jannidis, Rabea Kleymann, Julian Schröter, and Heike Zinsmeister (2025). "Do LLMs understand literature? Case studies and probing experiments on German poetry". In: *CCLS2025 Conference Preprints* 4 (1). 10.26083/tuprints-00030139

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-02-08

Keywords

LLM, CLS, Generative AI, interpretation, understanding, probing experiment

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. This paper explores the capabilities of large language models (LLMs) in understanding literary texts, specifically poetry, through a series of qualitative experiments. We define "understanding" in a way which allows us to assess task-specific capabilities while avoiding anthropomorphism. Analyzing two German poems—one very well-known, one unknown—we assess nine textual aspects: meter, rhyme, assonance, lexis, phrases, syntax, figurative language, titles, and meaning. Three levels of interaction— general knowledge, expert knowledge, and abstraction and transfer — guide our evaluation. Our results show LLMs excel in analyzing semantic aspects, including figurative speech, but struggle with formal elements like rhythm and sound. Performance differences exist across textual aspects rather than complexity levels. Notably, LLMs favor established interpretations over original insights and LLMs are relatively inflexible when it comes to shifting cultural perspectives unless explicitly prompted. Thus, we show the extent to which LLMs' performance covaries more with textual aspects and the extent to which it covaries with levels of task complexity.

1. Introduction

The beginning of the discussion about the capabilities and limitations of language models in 2021 was characterized by very general claims. On the one hand, some proclaimed that this was a big step towards Artificial General Intelligence (AGI). On the other hand, members of the linguistically oriented NLP and AI community criticized the language models as "stochastic parrots" (Bender et al. 2021), i.e. they produce language that looks like the language produced by humans, but has severe deficits. These deficits are not explicitly tied to any particular task. While humans share a common ground and "model each other's mental states as they communicate" (ibid. p. 616), "text produced by an LM is said to be not grounded in communicative intent, a model of the world, or a model of the reader's state of mind" (ibid.). The authors therefore claim that textual output produced by machines "has no meaning" (ibid.).

However, neither of these extreme positions really contributed to a better understanding of the real capabilities of LLMs. Thus, they were soon replaced by more limited studies that attempted to experimentally clarify the capabilities of models in a particular domain, from a particular perspective, or for specific tasks, for example, LLM's abilities in logical reasoning (Mirzadeh et al. 2024) or their cognitive abilities to understand other people in terms of theory of mind (Trott 2022; Trott and Jones 2023; Trott et al. 2023). Similarly, the goal of our study is to investigate the ability of LLMs to 'understand' literary texts, especially poetry, i.e., to perform specific tasks that humans can only perform if they have an adequate mental representation of a literary text and possess the knowledge and skills necessary to perform those tasks. AI and the new LLMs have been addressed by researchers interested in literary texts from very different angles: Kirschenbaum 2023 and Gengnagel et al. 2024 looked at the theoretical dimension of the concept of meaning and language that is realized or proved by LLMs. Walsh et al. 2024 explore their ability to generate literature, Bamman et al. 2024 examine which literary texts the models have seen during their training. In many public discussions, LLMs have been seen as a challenge to established teaching practices, or as part of the neoliberal world order hostile to the spirit of critique and reflection in the humanities. Concerns have also been raised about their environmental impact, particularly their high energy consumption and resource-intensive training processes. In this study, we do not contribute to any of these debates. We are interested in the question of what competencies they demonstrate in analyzing literary texts, and what attributes of their internal representation of literary texts we can infer.

For pragmatic reasons, we are focusing on poetry. We use two German-language poems in our investigation, *Hälfte des Lebens* (1804) by Friedrich Hölderlin (Hölderlin 1805) and *Unsere Toten* (1922) by Hans Pfeifer (Pfeifer 1922). Hölderlin's poem is well known, so we can assume that the models have seen it during training. Additionally, there are many interpretations of it, some of which may also have been in the training corpus. We used the well-known interpretations by Strauss 1965a and Schmidt 1982 as reference for most of our text descriptions and interpretations. Pfeifer's poem, like the author, is completely unknown. To our knowledge, it has only been published once in an anthology from 1922 that to our knowledge has not been digitized (Uhlmann-Bixterheide 1922).¹ We chose German poetry because this is our field of philological expertise. However, we also used or made English translations to be able to compare linguistic domains (Hölderlin 1965).²

Since we consider the paradigm of work-immanent interpretation to be the most appropriate basis for the design of the study outlined below, we have selected from the abundance of possible aspects of literary analysis the following nine that we consider particularly relevant:

1. meter
2. rhyme
3. assonance
4. lexis
5. phrases
6. syntax

1. We would like to thank Merten Kroencke, who digitized the anthology and made the poem available to us.
2. See the file definitions.py for all texts and their sources.

7. figurative language	57
8. title	58
9. text meaning	59

As this selection reflects an approach that primarily focuses on the text itself, context integration will have to be included in these nine aspects, respectively. For each of these aspects, we have developed a series of prompts to check how extensive, adequate and knowledgeable the ‘understanding’ of the literary text is. In our view, it is of paramount importance to be able to distinguish and scale different levels of complexity of understanding. These levels will be differentiated in the following sections, starting (in section 3) with the ability to generalize, to (4) more complex and expertise-like reasoning, and (5) the ability to perform more abstract steps of inductive and abductive reasoning. At the first level, we want to see how well the models work at the level of general knowledge, i.e., roughly the knowledge that students have when they leave school with a high school diploma. At the second level, we want to know how well the models can solve problems like experts in literary studies. On the third level, we tested whether the models are able to abstract counterfactual rules from examples and apply them to the poems. The rules are counterfactual in that we invented them and therefore they have probably never been applied to a literary text before.³ Each of these nine aspects mentioned above will be investigated using this distinction between three levels of complexity of understanding. This broad and inclusive use poses some challenges for our study design. Although many related studies have recently specialized in very narrowly defined tasks, such as LLM-based recognition of metaphors (Hicke and Kristensen-McLachlan 2024), our study of understanding a literary text requires a broad, integrative approach that links the different aspects of producing understanding at three different levels of complexity.

Overall, we are taking an exploratory approach in our work and will not present any quantitative results. Although we constantly set tasks for the models, we are not really interested in whether they solve them all flawlessly; rather, we are more interested in how they approach the tasks than in a successful solution. We do not want to test the models; rather, we are interested in what our experiments reveal about the type of representation, argumentative structures and the problem-solving skills shown in their answers. In a preliminary study, we found that smaller models ($\leq 70B$) made too many errors and showed too little of the skills necessary to “understand” poetry. Under the assumption that models develop qualitatively different abilities as the number of parameters increases, we concentrated on the large models: Claude 3.5 Sonnet 3.5 (Anthropic), Gemini 1.5 (Google), and GPT-4o (OpenAI). We usually performed single-run evaluations, meaning that our prompts were run only once across all three models, without systematic repetition. Our paper has the following structure. In section 2, we introduce a theoretical framework that helps to overcome the extreme position on ‘understanding’ outlined above. This is meant to build a philosophically informed basis for the kind of analysis we offer in this paper. In section three to five we report on the experiments on the three levels of (3) general knowledge, (4) expert knowledge and (5)

3. We used Jupyter notebooks for our analysis. The notebooks contain much more detailed information on all our experiments, and we believe them as important as our summary in this study. All the notebooks, scripts and data used in our experiments will be referenced in the following with ‘NB’ and can be found here: https://anonymous.4open.science/r/llms_read_hoelderlin-A11D.

abstraction and transfer. In our conclusion we will summarize our findings and will discuss some follow-up research questions.

2. Understanding

The basic distinction we recommend for an appropriate framework is that between internalist and externalist approaches to the concept of understanding. The internalist perspective is interested in the conditions that have to be fulfilled in the (human) mind and consciousness for understanding to take place. Although there are different approaches, Wilhelm Dilthey's (Dilthey 1974) can be seen as a classic internalist position, which assumes the psychological reproduction of a psychological state of the interpreted utterer or author, and also requires the ability to charge the utterance to be understood with relevance to the interpreter's personal life (Makkreel 2002).⁴ Extreme internalist positions, usually subject to accusations of psychologism, would claim that the criterion for understanding a poem (or anything else) is a completely subjective sense of evidence in the first-person perspective. However, even positions like Bender et al. 2021 are internalist in that they make the notion of the respective ability or property (communication, meaning, and understanding) dependent on some internal requirement, here a grounding human consciousness.

In contrast, according to externalist approaches, often associated with the late Wittgenstein of the *Philosophical Investigations* (Wittgenstein 2001), understanding occurs in the form of practices (Künne 2003; Strube 2003). Proving whether agents have understood an utterance or an action depends on the behavior they can show. From an externalist stance, whether someone has understood a poem or utterance does not depend on a certain subjective quality of experience, but on whether they can show that they have understood that poem or utterance. Understanding is then seen as a *practice* of acquiring understanding and as a kind of rule-following.⁵ The most prominent approach to an externalist strategy of verifying some agents intellectual abilities was Alan Turing's essay on *Computing Machinery*. (Turing 2021).

Although there are internal aspects of understanding that cannot be proven as irrelevant by just reducing understanding to external aspects,⁶ we will take an externalist stance and thus examine the external aspects of understanding. There are several reasons for this approach: (1) As can be seen in Bender et al. 2021, internalist discussions easily lead to a priori arguments about whether understanding *per se* requires a truly human agent.⁷ While relevant in certain areas of philosophical reasoning, such a *a priori* discussion would be a dead end for a deeper understanding of the capabilities of machines. (2) We believe that a comprehensive discussion of the concept of understanding, which includes

4. The so-called continental European tradition of philosophical hermeneutics, with its phenomenological foundations and concepts of the 'pre-structure' of understanding (Gadamer 1965), shows a strong internalist tendency; see the critical analysis in Scholz 2005.

5. Note that externalist approaches (Stekeler-Weithofer 2002) that draw heavily on Wittgenstein's notion of rule-following may make quasi-internalist demands when they claim that understanding in the full sense includes aspects of normativity, personal and social obligation, and other implications. We believe that such aspects will become more important in future discussions when relating AI 'understanding' capabilities to a full sense of understanding in human and social contexts.

6. The most prominent argument against a purely externalist notion of human understanding is Searle's Chinese Room argument (Searle 1980). For a summary of the debate see Cole 2024.

7. Think also of Searle's argument that language use is a sufficient condition for assigning understanding only if the agent is a human being (Searle 1980).

internal as well as external dimensions, should start from a fine-grained and nuanced analysis of the external aspects. (3) The actual challenge for a deeper understanding of LLMs' abilities is to reasonably describe the transitions between the internal and external dimension. In order to make this challenge even more tangible, we consider it fruitful to clarify the distinction between internalism and externalism with a further distinction that Daniel Dennett introduced into the debates on the philosophy of mind and AI in the late 1980s with the difference between 'design stance' and 'intentional stance' (Dennett 1987). When we take the intentional stance, we treat our counterparts as intentional systems - regardless of their internal properties - and interpret their behavior as behavior aimed at achieving the agents' goals (desires) based on their knowledge (beliefs) as more or less rational action. When taking the design stance, in contrast, we describe and explain the behavior shown by our counterparts based on our knowledge of their functioning. The intentional stance requires a radically externalist approach to interpreting the behavior of machines. Internalist approaches such as Bender et al. 2021 implicitly or explicitly require that some capacities, such as consciousness or intentionality, must also be present at the level of the design stance. Thanks to Dennett's distinction, we can clarify that in what follows we take an externalist, instrumental, intentional stance without making any claims about emergent properties of AI at the level of the design stance. The problem of convincing internalists will be to find reasoned explanatory transitions between the design stance and the intentional stance. Due to the *a priori* structure of internalist arguments, we are very skeptical about the prospects of convincing radical internalists in this way.

An externalist stance implies that we are not referring to the internal states of an LLM. Instead, we are referring to the text-processing abilities, the knowledge, that we would have to attribute to a person if that person gave us an answer like the LLMs' answers to our questions. Here and in the following we will use the term 'knowledge' in the broad sense of the word, which includes not only declarative knowledge but also practical and procedural knowledge (Ryle 1945). Thus, when a model produces an answer *A* to a query, our use of 'understanding' refers to the complex attribution that we as humans would make, if we received *A* in response to a similar query from a human. Our use of the notion of 'understanding' as literary scholars and linguists focuses on behavior showing the ability to recognize specific textual features, such as meter or figurative language. It also includes behavior showing the ability to apply specialized knowledge, such as technical terminology or relevant historical knowledge.

The general approach and design outlined in the previous section allows one to address another challenge. Just like probing experiments from the fields of NLP (Chang et al. 2023), psychology (Trott and Jones 2023; Trott et al. 2023), often with a special focus on psychometrics (Chollet 2019), or from a more general interest in the human-like abilities of LLMs (Mirzadeh et al. 2024), CLS research must also address the development of appropriate metrics for measuring the correctness of LLMs' behavior. However, there is a general tension between the need for standardized measurability and the shortcomings of single-point metrics. Since we are dealing with stochastic machines that react on the basis of randomization and probability, future research will have to consider not only individual responses, but also variations of types or patterns and distributions of responses. This will require a more rigorous formalization of the correct versus incorrect output to be evaluated. Therefore, one of the aims of this paper is to lay the groundwork

for future work on robust metrics that can be scaled for iterative or bootstrapping methods of analysis. At level 1 of general knowledge, we can already present metric measures of the comprehension performance that we see in LLMs. At levels 2 and 3 of expert knowledge and abstraction and transfer, however, we are dealing with complex forms of practice. How convincing the respective LLM output is as a (first-order) act of producing understanding is only determined by an act of interpretation (which can thus be seen as a second-order act of understanding) on the part of us researchers. On these levels, the experiments come close to Turing's original idea of an imitation game (Turing 2021, §§ 1–2). This judgment requires philological expertise. The (second-order) acts of interpretation on our part as researchers, however, come only after the (first-order) output produced by the LLM. The question of the evaluation of the LLM output as an act of understanding can therefore only be partially answered by analyzing the concrete probing experiments.

3. General knowledge

At the foundational level of general knowledge, we focus on the one hand on tasks that involve widely taught and culturally accessible information. We assume that the LLMs have already seen much of this information during training. On the other hand, we address straightforward assignment tasks that presuppose the LLMs already possess an understanding of our nine aspects. These first experiments aim to evaluate not just the correctness of their outputs but rather their ability to approach tasks in ways that reflect a meaningful understanding of the poems. Thus, we sought to explore generalization, pattern recognition, and meaning attribution skills in terms of what we would expect from rational actors with rudimentary knowledge and skills.

The exploration of the first level begins with a series of experiments that focus on the formal and structural features of the two poems. Starting with the analysis of the metrical structure of the poems (NB 1) involves two steps: detecting the scansion and reporting it in a summary way. While Sonnet performed this task almost flawlessly, GPT4o and Gemini struggled. For the poem *Hälfte des Lebens*, Sonnet consistently produced accurate scansion, whereas GPT4o and Gemini produced errors. Regarding the second step, a notable observation across all LLMs was their frequent inability to summarize the scansion patterns they identified. They often report more stressed syllables than were detected, but never fewer. This discrepancy is probably related to the inability to count and deal with symbols that do not coincide with token boundaries (Edman et al. 2024; Xu et al. 2024). These findings highlight that while LLMs can simulate scansion recognition, they struggle with tasks requiring metrical abstraction. The analyses of rhyme words (NB 2) and schemes are directly linked to this. For *Hälfte des Lebens*, the absence of rhymes was generally detected, but GPT4o and Gemini occasionally produced false positives. The reason for this is that both counted non-terminal words as rhyming elements. This indicates that the representation of verse structure is not well modeled in these LLMs. In contrast, Sonnet exhibited an almost perfect 'understanding' of German pronunciation and rhyme structure. Regarding the unknown poem *Unsere Toten*, which follows an AABB rhyme scheme with an internal rhyme in the final verse, GPT4o and Gemini correctly identified the rhyme words, but only Sonnet accurately detected the rhyme scheme, while the other two models showed inconsistent results.

While the prompt designs for meter and rhythm consist of simple zero-shot detection tasks, the prompts for the detection of assonance (NB 3) include different definitions of assonance ranging from simple descriptions to technical explanations involving phonemes (Zymner 2007). The LLMs demonstrate very low precision and recall for the German poems, regardless of the definition provided and the poem. By contrast, their precision and recall on the English translations of the poems is markedly higher independently of the definition provided and also for a prompting that does not offer any definition (Table 1).

Model	Recall (de)	Precision (de)	max. precision english	prompt for max recall
GPT4o	0.2	0.2	0.83	Def 1
Gemini	0.2	0.2	0.44	Def 2
Sonnet	0.2	0.35	0.75	Def 1

Table 1: Recall and Precision for Assonances in the German version of Hölderlin’s “Hälfte des Lebens” with n 10 instances of true assonance occurrences.

This indicates that training on phonetic features in German either did not constitute a major role during the development or that such training was not sufficiently effective. To address this, a two-step chain-of-thought (CoT) prompting method is applied, asking the LLMs to first transcribe the poem into International Phonetic Alphabet (IPA) and then identify assonance based on the transcription. Though all LLMs are good at performing the first step of transcribing the poems into IPA, they are not able to efficiently base the second on the first step. Regarding the overall performance, the models do not improve either for English or German. The findings of these experiments on assonances are directly tied to the level of general knowledge because it highlights the difference in how humans and LLMs approach tasks that rely on foundational language skills. At the level of general knowledge, humans are able to intuitively detect and process phonetic patterns like meter, rhyme, and assonance without needing specialized training or explicit systems such as IPA. It can be assumed that this ability stems from a mixture of linguistic capacities and learned experience, which allows humans to recognize phonetic similarity. In contrast, LLMs lack this phenomenological foundation. Their performance on tasks involving phonetic similarity is therefore constrained by the quality of their training data. The difficulty they face in handling these tasks underscores a limitation in their general knowledge capabilities.

The analysis of the lexis show across all models the ability to identify the semantic field of selected nouns and verbs. The identification of the parts of speech on the other hand is partially flawed (NB 4). In the first of our experiments aimed at reconstructing how the LLMs understand the imagery, figurative speech, and meaning of the two poems, the LLMs were first tasked with identifying all instances of figurative speech in the poems and, for each instance, providing reasons for why it was identified as figurative speech (NB 5). At the level of general knowledge, this allows for differentiation between linguistic devices that render an entire text as an overarching image and those that serve as localized elements of illustration within the text (Burdorf 2015). Our particular interest lies in these localized figures of speech, such as metaphors, metonymies, synecdoches, and symbols.⁸ All LLMs work remarkably well, as they identify many instances,

8. The capability of LLMs to understand metaphors in non-literary language has been tested in Wachowiak and Gromann 2023 and more systematically in Tong et al. 2024 using older and smaller LLMs with very mixed results. For metaphors in literary texts see Boisson et al. 2024.

even if none of them covers all of them. Indicators that an expression is supposed to be understood figuratively are, according to the models, that a literal understanding does not make sense, for example "Human qualities are attributed to inanimate objects (walls)." (Gemini on *Hälfte des Lebens*). Interpretations often refer to an established understanding of the symbol, which is then explicitly marked, e.g., "Roses are often symbols of beauty, passion, or transience" (GPT4o on *Hälfte des Lebens*). In order to investigate the relation between figurative speech and literal understanding, a completion task using the "simple suffix prompting" (Liu et al. 2022) method with "that is to say" was conducted (NB 5). The task for the LLMs was to interpret the figurative phrase "Die Mauern stehn sprachlos und kalt" of the poem *Hälfte des Lebens*. In the prompt design, the suffix "that is to say" was inserted between the figurative phrase and its possible literal explanation, signaling the need for interpretation. Only Gemini correctly engaged with the syntactic structure of the prompt and completed the sentence with the literary description "indifferent, uncaring." In a further step, the LLMs were asked to choose the best completion for the figurative phrase from four options, evaluate their choice with regard to the context of the poem, and provide a confidence score for their decision. All three models selected the same completion: "the emptiness echoes within the confines of their silence," assigning it an identical confidence score of 0.9. The chosen completion aligns with traditional interpretative approaches. However, it became clear that the theme of "speechlessness," as discussed in scholarly research on *Hälfte des Lebens*, was not selected (Strauss 1965b). Exploring how the three LLMs engage with figurative speech show that their outputs are primarily shaped by conventional interpretations. All three LLMs draw on culturally entrenched associations rather than generating novel interpretations.

The experiments on syntactic structure (NB 6) show that all models have generalized broad syntactic signals of German well. However, they did not manage to elaborate the difference between the two stanzas of *Hälfte des Lebens* in terms of different types of enjambments (line break at phrase boundary versus line break splitting a phrase).

The experiments on title (NB 8) and text meaning (NB 9) highlight the LLMs' capabilities in interpreting central themes and motifs as well as generating plausible interpretation hypotheses. At the level of general knowledge, the models demonstrated a strong ability to generate conventional interpretations (NB 9). Zero-shot prompting reveals a strong focus on oppositional motifs or theme. In the case of *Hälfte des Lebens*, all three models focused on the central oppositional pairs of "summer and winter" to summarize the poem's thematic elements. However, their interpretations rarely ventured beyond these straightforward dichotomies to address more figurative or nuanced meanings. For the less familiar poem *Unsere Toten*, the models displayed greater diversity in their hypotheses, referencing historical contexts such as the World War I and II. In addition to this, all LLMs provided a range of different interpretations. We then asked the models for handling some of the most salient or most surprising aspects of the poems' titles (NB 8). Processing the title of a work of art is a complex interpretive operation that requires understanding the work itself and relating its meaning to the literary meaning of the title and then to think about the effects of connecting both. For Hölderlin's *Hälfte des Lebens*, the models recognized the title's relevance to the poem's dual structure, connecting the "half" to the juxtaposition of summer and winter imagery. However, interpretations diverged in their reasoning. GPT4o and Sonnet argue that the poem

thematizes both halves of life, while Gemini claimed the poem exclusively addresses the first half, associated with summer. Despite failing to engage with the second stanza's conditional structure ("wo nehm ich, wenn"), Gemini's interpretation framed the title as emphasizing youth and vitality. Meanwhile, GPT4o and Sonnet took a more abstract approach, interpreting "Hälfte" as representing a midpoint or turning point in life, reflecting a moment of awareness about life's contrasting phases. For *Unsere Toten*, all models correctly identified the reference to "German soldiers" and the invocation of national identity as central elements of the title. Sonnet was the only model to explicitly and correctly associate the poem with World War I. The tasks set here can be described, abstractly, as a work-immanent approach in which semantic relations between a title-phrase and the bundled sentence meanings of the respective poems are to be described. As with other – at first glance complex – tasks of linking semantic units (meaning, metaphor), all language models are very good in this respect and at a level that fulfills the requirements of general knowledge.

4. Expert knowledge

On the level of expert knowledge we designed prompts that forced the models to enter more sophisticated output according to what one could call a philological style of reasoning.⁹ On this level, we expected the LLMs to perform tasks that either performed a multi-step inference process of analysis, or forms of philological reasoning that take into account relevant but not self-evident historical or linguistic context. For this, the basically work-immanent tasks for each of the nine aspects often incorporated some trans-textual (i.e., intertextual or contextual) dimension. Particularly challenging at this level is the assessment of the conditions under which an LLM's answers represent a convincing, adequate, or even correct act of understanding. Although the quality of the models' output could be assessed after interpreting the output, we tried to provide some pre-registered conditions for each experiment, respectively. According to theoretical discussion of the different dimensions of the philological concept of understanding (Künne 2003; Strube 2003), these conditions include the ability to integrate historical context-knowledge, the ability not only to provide plausible answers but also to judge on the empirical appropriateness of different explanatory hypotheses, to connect different layers and aspects of the work, finding an appropriate level of abstraction.

When asked to identify instances in both poems where the meter is changed to indicate a semantic aspect, all models identified the change in meter between the first and second stanzas in Hölderlin's poem and associated it with the change in meaning and emotion (NB 1). While this may be due to the fact that this change is mentioned in many interpretations of the text, all models also identified the change in meter in the last two lines of Pfeifer's poem. Additionally, we asked for the verse meter of *Unsere Toten*. The correct answer is 'Knittelvers', which has the rhyme scheme AABB, four stressed syllables. Since it allows free filling of unstressed syllables (variable number of syllables), a simple bottom-up detection from smaller units does not work. Two of the models, Sonnet and GPT4o, answered correctly, but only when asked for a "German

9. For the notion of 'Styles of Reasoning', see Hacking 1994. We can only hint at the new possibilities for distinguishing the different forms of philological reasoning that have been identified through large qualitative and praxeological analysis in Winko et al. 2024.

verse meter”, not when asked for a verse meter in general. Identifying the Knittelvers can thus be considered an expert task because it requires the interpreter to spontaneously take into account an appropriate set of categories. No model was able to assume the culturally appropriate set of verse meters here, but some models were able to find the correct category according to the context specified by our intervention. Since expert-level understanding of poetry essentially involves the ability to spontaneously find the most relevant context for interpretation, we see that the models often fall short of the ability to contextualize. For the analysis of rhyme (NB 2), we focused on Pfeifer’s poem because *Hälfte des Lebens* does not have a rhyme. The models were asked to describe two strategies for relating rhyme to the meaning of the poem or parts of it, and then to determine whether any of them produced interesting insights into the poems. All models made plausible suggestions on a general level, and all applied their proposed approaches to the poem. Before we queried the models, we pre-registered a set of acceptable answers. One being a semantic relation between the rhymed words (semantic rhyme words), another a relation of the vowel structure of the poem or parts of it and the vowel structure of the rhyme words, both as indicators of a specific tone or mood (rhyme and mood). We determined that there is some additional semantic relation in the rhyme words of the second half of the poem, but no interesting finding for the rhyme and mood approach. GPT4o, for example, proposed the rhyme and mood approach and described a relation which, however, was not plausible at all.

Since the models failed at the first level of general knowledge when prompted to detect assonance (NB 3), it might seem questionable to further advance to the level of expert knowledge. Nevertheless, it is enlightening to force the models to perform more complex tasks. The task we designed was to provide an interpretation of the poem that tries to relate the words that are connected by assonance on the level of textual meaning by grouping and contrasting assonance-based semantic fields. Again, all models failed to correctly detect assonance. However, in cases with correct connection, surprisingly good hypotheses were raised – for instance by Gemini, which based its interpretation on the assonance claiming for the second stanza of Hölderlin’s poem that “the ‘e:’ sound links the speaker’s lament (“Weh mir”) with the absence of summer elements (“nehm”, “stehn”). Though not an entirely precise analysis, the connection between lament and privation is reasonable. Here we see that the models perform well at the seemingly more complex tasks of seeing non-trivial and latent semantic relations and even of combining different latent relations. However, since the basic description of the sound characteristics of the poem lacks precision, further interpretive hypotheses on the level of expert knowledge have not yet become convincing.

To investigate the ability of the models to analyze the lexis of the poems they were asked to focus on one semantic contrast that is triggered either by a morphologically complex word or within a phrase, and to elaborate how this contrast contributes to the meaning of the poem. All three models identify “heiligenüchterne” as the most striking example of semantic contrast in Hölderlin’s poem. Only Sonnet makes use of specialized vocabulary describing “heiligenüchtern” as “oxymoronic combination”. No model refers to the classical topos of “sobria ebriatas” (Schmidt 1982) (NB 4).

Regarding the aspect of phrases (NB 5), we explored the ability of LLMs to contextualize ambiguous phrases and translate non-figurative to figurative language. Our focus for

this task was on the phrase "im Winde klirren die Fahnen." As noted by Strauss 1965a, the notion of a fabric flag is very likely to firstly come to mind but must be replaced by the idea of a weather vane to align with the intended meaning. In an initial zero-shot prompt, the LLMs were asked to interpret the meaning of "Fahne." with mixed results. Confronted with very unlikely meanings of the phrase in the next step, GPT4o and Sonnet refused most of them, but selected the military context, while Gemini suggested a new metaphorical interpretation. For the unknown poem *Unsere Toten*, the study examined the models' ability to generate and assign figurative phrases based on their interpretations. Specifically, the task addressed the transition from non-figurative to figurative language, using the phrase "die Füße mühen sich im zitternden Mondenschein". The scenario assumed that part of the text was unreadable, leaving either only an interpretation or a gap-filled text for reference. In both cases, the LLMs demonstrated the ability to generate figurative phrases that thematically align with the poem. However, they did not account for metrical considerations in particular. Regarding the gap-filled text, GPT4o and Gemini both added references to cardinal directions, while Sonnet only added "Süden" (south) as an additional direction. Notably, the word "Schlürfen" (to slurp) was frequently completed with terms such as "wandern" (to wander) and "gehen" (to walk).

Engaging with a literary text on a research level often involves addressing literary theoretical positions. As Köppe and Winko 2013 note, it is impossible to read a text without theory. In examining the interpretative outputs of LLMs with regard to statements about textual meaning (NB 9), we therefore asked to what extent, and in what ways, the models reflect specific literary theoretical approaches. Can we identify latent representations of literary and cultural theories in the interpretations generated by LLMs? Our study focuses on the dimensions of representativeness in these outputs. The starting point was Görner 2016, p.107 and his thesis for *Hälfte des Lebens*: "Postcolonial literary studies do not take us far in understanding [Hölderlin's] work. By contrast, (post-)structuralists and deconstructionists appear – albeit unintentionally – to have prepared the way for interpreting Hölderlin." Notably, all the LLMs produced postcolonial interpretations for both poems, incorporating key terms central to the theory. However, when ranking the literary theoretical positions, GPT4o and Sonnet indicated that the poem *Hälfte des Lebens* "lacks overt colonial references" or "lacks specific markers of colonization." In contrast, for the unknown poem *Unsere Toten*, Gemini suggested that the "poem can be read as an allegory for the lasting impact of colonialism." The recourse to interpretation-theoretical framework assumptions thus in no case went beyond very superficial remarks.

Regarding figurative Language (NB 7), we tested for both poems the ability of the LLMs to change its understanding of figurative language when additional information was given about a specific term which was used figuratively. In the case of *Hälfte des Lebens* we added the information that its author was a great admirer of classic antiquity (which is common knowledge) and that in classic literature swans are often a metaphor of the poet, the latter being specialized knowledge applied to this poem first by Schmidt 1982. All models provided a before and after interpretation and used the information to change or deepen the interpretation. For example, the interpretation changes from "swans can be initially read as representing a harmonious connection between nature's elements" to "swans become a symbol of the poet in his ideal state: connected to nature, inspired, and capable of creating beautiful and meaningful art". Their level of 'understanding'

is quite impressive, because they explain how this additional information changes not only the meaning of swan in itself, but how the situation of the swans and their acts have now a new or additional meaning. And this works up to the level of the text meaning, when one model summarizes it now as a "meditation on the poet's role and the crisis of modern poetry versus ancient ideals". In the case of *Unsere Toten* we added the information, that the poem was first printed in 1922 (we added the whole bibliographic information, but the models concentrated on the date). Though all models described the specific situation after World War I in Germany, only one understood that the returning people are the dead soldiers.

With regard to processing the poems' titles we expected the models to operate with context, here with intertextual resonance in the titles (NB 8). For *Hälfte des Lebens*, we deliberately provided an anachronistic and thus irrelevant but similar title: *Mitte des Lebens*, a novel from 1978 by Luise Rinser. For *Unsere Toten*, we offered a potentially relevant context by mentioning that there was a journal, *Jahrbuch der Schiffbautechnischen Gesellschaft*, (1914) which had a section "Unsere Toten". Our expectation was that the models would warn of potentially anachronistic interpretation, are able to abstract from the journal section "Unsere Toten" some potential genre-like rules of commemoration that are applicable to the poem. The risk of anachronism (with referring to Rinser's novel) has been raised by no model. None of the models was able to refer to the content of the intertexts that were mentioned as a context. For *Unsere Toten* the models were able to infer the genre-like function of commemoration. Gemini connected this function with the poem's phrase "Nur nicht vergessen! Uns nicht vergessen!" and thus highlighted the commemorative function of this part of the poem, without addressing the obvious differences. We conclude that when provided with potentially relevant context information, all models reason on the level of merely semantic surface relationships by extracting semantic information from the information that is available for text and context/pretext. It is particularly striking that the models process all the information provided as actually relevant and create associations between text and context. The relevance check of the context offered, which is specific to philologically sophisticated interpretation, does not take place.

We can thus summarize that the abstraction that is performed by the models when they try to applicate some aspect of a given context to the text being interpreted works well on a purely semantic level but not on the level of relating works, events, objects, and persons as historical positivities. This result aligns well with the finding of other studies that do not see any complex world model included in the language model, which is, as its name says, a language model without anything beyond the semiotic relations of language itself.¹⁰

5. Abstraction and Transfer

As mentioned above, in this section we want to discuss our probing experiments which are supposed to show the abilities of the models to infer rules or relations from given textual data and apply them to the two poems. Our main focus was to test whether

10. With regard to an analysis of the theory of language that is realized by LLMs and the thesis that LLMs largely actualize post-structuralist notion of language, see Underwood 2023.

the models were able to abstract counterfactual rules from examples and apply them 480
 to the poems. As a preliminary step, and in cases where the main task could not be 481
 solved successfully, we asked the models to apply counterfactual rules that we had 482
 explicitly articulated. By either giving the models counterfactual, i.e. made-up, rules or 483
 having them infer such rules inductively, we can assume that we are asking for analytical 484
 processes that the models could not have seen during training. For the analysis of meter 485
 (NB 1), our plan was to give the models two tasks: First, we wanted to test their ability 486
 to apply arbitrary rules about meter to given data and analyze the results. Secondly, 487
 we wanted it to abstract these rules from given example data. The rules we define are 488
 simple. They mix information about the emphasis of words with the ability to detect 489
 vowels and change the emphasis, when a specific combination is met (accented syllables 490
 containing the vowel 'i' were changed to non-accented syllables). The task was then 491
 to apply this rule to the poems. No model was able to solve this task. Based on this 492
 result we decided to skip the second experiment, because counterfactual information 493
 on the level of meter seems to be a challenge even in a simple setting. For the analysis 494
 of assonance (NB 3), we used a similar task. We defined a made-up phenomenon 495
 called 'eusionance': when the vowel sounds of the stressed syllables in two consecutive 496
 words in the same line are a i-sound and a German e-sound (e.g. "Ich" and "esse" in 497
 "Ich esse Kuchen"). With the two-step CoT-prompting we asked the models firstly to 498
 translate the poems to IPA and then to give an analysis of all eusionances. The problems 499
 were, on a structural level, of the same types as with the basic detection of assonance. 500
 Most severely, the models assumed words that were not even in the poem. Then, false 501
 phonemes were claimed as generating eusionance (i.e. a-, o- and u-sounds). Also, very 502
 simple aspects of the definition were disregarded (for instance, combinations of two 503
 e-sounds were also included in the answer on a regular basis). The second task was 504
 to abstract and infer a rule of interpretation based on an interpretation provided for a 505
 different poem (here, Eichendorff's *Waldgespräch*, 1815). The models are surprisingly 506
 good at inferring the intended rule from the interpretation. The difference between 507
 a well-known poem likely to be part of the training material and a poem previously 508
 unknown to the model is insignificant when it comes to sound qualities in the domain 509
 of German language. The seemingly most difficult tasks of inferring and abstracting 510
 rules and of performing meaning and association-based operations between the lexical 511
 units of the poem is relatively easy for all models. It is mostly the level of detecting the 512
 basic properties of sound quality that is still largely missing in all LLMs. 513

To test their abstraction and transfer ability with lexis, the models received the very 514
 challenging task to interpret attributive adjectives (in contrast to predicatively used 515
 ones) in terms of their antonym or semantic opposite (attributed to a made-up poet). In 516
 the case of Hölderlin, two of the models refused the task because they determined that 517
 the rules were inadequate for Hölderlin. Only one model (Sonnet) behaved like this in 518
 case of Pfeifer. Gemini only understood part of the rule, the mapping to antonyms, but 519
 it did not realize that whether the rule should be applied depended on the exact part of 520
 speech (NB 4). Gpt4o couldn't identify the rule in the case of 'Unsere Toten'. 521

Our experiments probing the ability of the LLMs to process poems at the phrases 522
 level (NB 5) started from the assumption that LLMs may lack a multimodal horizon 523
 of experience and perception, which can be crucial for interpreting certain phrases, 524
 particularly in poetry. To address this limitation, we introduced a new rule of meaning 525

that posits: only the onomatopoeic level of a phrase carries significance. And we gave two examples for this new rule. In both cases, the LLMs recognized and applied the new rule of meaning to their interpretations. Notably, the onomatopoeic translations produced by all three LLMs displayed striking similarities across the two poems. This consistency might suggest that the LLMs are capable of engaging with the onomatopoeic layer of meaning in literary texts, even though they lack direct sensory or experiential input.

In our experiments on the ability to process the meaning of texts, we introduced a non-referential rule of meaning to guide the overall interpretation of the poems (NB 9). According to this rule, the meaning of an expression lies solely in its function within a communicative act, independent of any direct connection to an extralinguistic reality. This rule was applied to two case studies (Helga M. Novak's *HÄUSER* (1982) and Nikolaus Lenau's *Einsamkeit* (1834)). In explaining the new rule of meaning, all three LLMs emphasized that the "communicative act" is central to the assignment of meaning. For both poems, the LLMs' proposed interpretations focused primarily on the lyric first person, but the notion that extralinguistic reality does not play a meaningful role remained somewhat vague in their results. This suggests that while LLMs can conceptually engage with the idea of communicative meaning, they struggle to fully articulate its implications when detached from concrete referents.

Our experiment on the use of figurative language only used the poem by Pfeifer (NB 7). It started with the counterfactual claim that there was a group of authors in the Weimar Republic working hidden references to the new medium film into their texts. The models were then tasked with identifying these references in Pfeifer's poem and to change the interpretation of the text as a whole. We expected the models to detect the unusual phrase "the trembling moon's light" which is hardly consistent with a realistic moon light. (The movement of the bodies in the same poem which may remind modern readers of Zombies would be an anachronism, as the first Zombie movie was made in 1932.) All models successfully identified the phrase and connected the poem to the medium film changing the interpretation of the text accordingly. In this case we did not ask the models to identify the rule (there are references to the medium film) themselves, but we did not specify the relation, so the model had to apply this very general pattern to the elements of the text themselves.

To test the abilities of the models to reflect on the relation between the text of a poem and its title (NB 8), another poem (Eichendorff's *Waldgespräch*, 1815) was provided. Given an intuitively straight-forward title based on a first reading ("Lorelei"), the difference between the actual and the intuitively more straight-forward title was used to find some aspect of the poem that is highlighted by this difference. In the final run, counterfactual (but here claimed to be real) titles "Des Dichters Leben" for *Hälfte des Lebens* and "Gefallene Geister" for *Unsere Toten* were provided. In a first run (3a), the real titles were provided, in a second run (3b), made-up titles were used. Although the general problem of induction makes it relatively hard to sort out false rules inferred from one example, it is very interesting to see that the models tend to slightly over- or undergeneralize. But our general impression was that the models work well for this task, probably because it involved a lot of summarization. This semantic condensation is a task the models are very well trained on. At the same time, we see, however, some

arbitrariness in the level of abstraction that the models chose. It remains a matter of
human judgment to decide in which cases the models have reached a hermeneutically
appropriate level of abstraction.

6. Conclusion

Our paper contributes to a better understanding of the abilities of LLMs to analyze
poetry. In general, we saw that all models performed surprisingly well. The models
were good at tasks that we believe are difficult for humans, such as processing non-literal
meaning and combining different levels of semantics by finding non-literal association,
equivalence, and opposition. In this respect, the performance of the LLMs often covaries
more strongly with the nine differentiated aspects from meter to meaning than with
the three levels of complexity (general knowledge to abstraction and transfer). In
some aspects that we believe are comparatively easy tasks for humans such as counting
syllables or recognizing meter, all models struggle. One problem is counting. Another
problem seems to be the lack of a stable representation of these qualities. So, the models
can solve some tasks that rely on one pass through the text, such as detecting a relation
between formal and semantic patterns, but have problems with tasks that demand
repeated access to the formal features. Probably, this cannot be explained by a lack of
training data alone. We assume that humans have a capability of processing formal
aspects based on phonetic information and an understanding of phonetic similarity
which is not available to the same extent to language models yet.

When we tried to overcome the deficiencies, for example, in assonance detection by
adding an intermediate layer of phonetic transcription into the circuit, we found that
while the phonetic transcription into IPA worked well, the overall detection of assonance
did not improve. Testing the lexis we saw that the model recognized the semantic change
but not its dependence on the part of speech. A preliminary conclusions can be drawn
from these observations. We assume that humans have a holistic understanding of the
'whole' artifact combining all the different layers of a poem (prosodic, phonetic, parts
of speech, rhythmic, semantic) as constitutive of the whole work. Returning once again
to the internalist dimension of the concept of understanding, which concerns all aspects
that take place within the 'understanding agent', we can at least suggest at this point
that human beings represent this aspects a parts of a whole in a way LLMs are not yet
able to, instead the level of meaning is the most dominant and stable one.

We generally observe a relatively high quality of the *prima facie* more difficult meaning
related tasks arising not only in the domains of metaphor, meaning, and title but which
were evoked also when more complex inferences should be drawn in the domain of
assonances etc. Processing semantic comparison, contrast, association, and also finding
aspects in works that are connected through multi-step semantic abstraction are more
at the core of the LLMs capabilities. Even on the level of abstraction and transfer,
the models were very efficient at inferring and applying interpretive rules. Providing
historical and empirical arguments by giving good reason for the correctness of specific
historically explanatory hypotheses is not among the abilities that LLMs are at the
moment particularly good at. This can have to do with the training material, but also
with the way historical world knowledge is or rather is not (yet) modeled in LLMs.

Beyond the covariation between the nine aspects, we are also able to identify significant performance problems in terms of complexity of understanding, as reflected by the three levels in sections 3 to 5, especially when focusing on expert knowledge and abstraction and transfer. First, all three LLMs draw on culturally established associations rather than producing novel or surprising connections and interpretations. They produce the expected rather than the original: LLMs follow the path of highest probability and expectability with little surprise, whereas sophisticated interpretations in literary studies at the level of expert communication are expected to look for new and unusual ideas. Second, it is well known that prompting is very important for communicating with LLMs. We observed three interesting behaviors, which we propose to summarize as the problem of culturally sensitive context integration:

- If there is information in the prompt, the models consider all information to be relevant, regardless of whether it is actually relevant. Most of the time, they do not rely on a stable representation of the world that would allow them to reject some given information as obviously nonsensical or useless. Rather, they show an attitude that attributes a high level of competence to the user in selecting the information given in the prompt.
- Even if they are able to employ complex arguments when the prompt already engages them on this level, it is not their standard level of interaction. The language and the argumentative structure of the prompt seem to be far more important for this kind of framing than any explicitly described roles.
- If an answer can be found in a cultural context other than the main English one, the models work better if that context is explicitly stated in the prompt.

We found that ‘understanding’ includes calling on the most appropriate rules (in terms of abduction). For instance, when the correct verse meter is to be inferred, the space of potential and culturally relevant verse meters has to be considered. We saw that LLMs often use the domains and cultural spaces coming from the language they are most thoroughly trained on. According to prompt engineering, we usually expect the users to intelligently control these cultural biases by adding information on the relevant context. If we ask for LLMs’ interpretive capabilities, we can observe that this is one of the most interesting shortcomings to infer the ‘correct’ contexts. This shortcoming affects both the level of expert knowledge (when describing the task as a matter of culturally and historically sensitive context selection) and the level of abstraction and transfer (when describing the abductive inferential reasoning need for such tasks). However, and more interestingly, we can conclude that whenever relevant results are produced with LLMs for producing ‘understanding’, such results are to be considered as a result of human-computer-interaction. Since assessing the choice of the “right” context and the appropriate level of abstraction in a first-order interpretive response given by an LLM requires a thorough second-order interpretive judgment on our part, we believe we were justified in basing this work only on qualitative experiments. Future research will need to clarify how comprehension can also be measured quantitatively.

Towards the end of our work, several new reasoning models (GPTo1¹¹ or DeepSeek-R1, DeepSeek-AI et al. 2025) have been released that may change the models’ performance to

11. <https://openai.com/index/learning-to-reason-with-llms/>

such an extent that the styles of analyzing the models' output, the prompt design itself, 657
 and the ways of measuring the performance will have to be revised. We are planning 658
 to include some of these aspects in the journal version of this contribution. What will 659
 be stable, however, is the way we propose to think about the concept of understanding. 660
 Adherents of internalism, who make consciousness or other aspects they claim are 661
 exclusive to human beings a necessary condition for understanding to occur, will never 662
 be impressed by any observable performance of a machine. However, once one has 663
 come to terms with dealing with the observable aspects of understanding, one will focus 664
 on the externalist dimension and realize that many of the intriguing problems concern 665
 how understanding behaves at different levels of complexity and between the different 666
 aspects of text understanding, from sound patterns to metric and semantic patterns, 667
 from recognition and vague meaning attribution to complex and multi-step reasoning 668
 performances requiring inductive and abductive inferences. These are the questions to 669
 which we wanted to prepare answers. 670

7. Data Availability 671

Data can be found here: https://github.com/cophi-wue/llms_read_hoelderlin 672

8. Software Availability 673

Software can be found here: https://github.com/cophi-wue/llms_read_hoelderlin 674

9. Author Contributions 675

Fotis Jannidis: Conceptualization, Writing – original draft, Formal Analysis, Writing – 676
 review & editing, Methodology, Project administration 677

Rabea Kleymann: Conceptualization, Writing – original draft, Formal Analysis, Writing 678
 – review & editing, Methodology, Project administration 679

Julian Schröter: Conceptualization, Writing – original draft, Formal Analysis, Writing – 680
 review & editing, Methodology, Project administration 681

Heike Zinsmeister: Conceptualization, Formal Analysis, Writing – review & editing, 682
 Methodology, Project administration 683

References 684

Bamman, David, Kent K. Chang, Li Lucy, and Naitian Zhou (2024). *On Classification* 685
with Large Language Models in Cultural Analytics. arXiv: 2410.12029 [cs.CL]. <https://arxiv.org/abs/2410.12029>. 686
 687

- 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*. FAccT '21. event-place: Virtual Event, Canada. New York, NY, USA: Association for Computing Machinery, 610–623. ISBN: 978-1-4503-8309-7. [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922). <https://doi.org/10.1145/3442188.3445922>.
- Boisson, Joanne, Zara Siddique, Hsuvas Borkakoty, Dimosthenis Antypas, Luis Espinosa Anke, and Jose Camacho-Collados (Dec. 2024). *Automatic Extraction of Metaphoric Analogies from Literary Texts: Task Formulation, Dataset Construction, and Evaluation*. arXiv:2412.15375 [cs]. [10.48550/arXiv.2412.15375](https://arxiv.org/abs/2412.15375). <http://arxiv.org/abs/2412.15375> (visited on 02/07/2025).
- Burdorf, Dieter (2015). *Einführung in die Gedichtanalyse*. Stuttgart: J.B. Metzler. ISBN: 978-3-476-02227-1 978-3-476-05422-7. [10.1007/978-3-476-05422-7](http://link.springer.com/10.1007/978-3-476-05422-7). <http://link.springer.com/10.1007/978-3-476-05422-7> (visited on 01/28/2025).
- Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie (2023). *A Survey on Evaluation of Large Language Models*. eprint: 2307.03109.
- Chollet, François (Nov. 2019). *On the Measure of Intelligence*. arXiv:1911.01547 [cs]. <http://arxiv.org/abs/1911.01547> (visited on 11/18/2024).
- Cole, David (2024). "The Chinese Room Argument". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2024. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2024/entries/chinese-room/> (visited on 02/04/2025).
- DeepSeek-AI et al. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv: 2501.12948 [cs.CL]. <https://arxiv.org/abs/2501.12948>.
- Dennett, D. C. (1987). *The intentional stance*. «A» Bradford book. Cambridge, Mass. u.a: MIT Press. ISBN: 978-0-262-04093-8.
- Dilthey, Wilhelm 1833-1911 (1974). *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften*. 4.-5. Tsd. Theorie. Frankfurt/M.: Suhrkamp. ISBN: 978-3-518-06329-3.
- Edman, Lukas, Helmut Schmid, and Alexander Fraser (Nov. 2024). "CUTE: Measuring LLMs' Understanding of Their Tokens". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, 3017–3026. [10.18653/v1/2024.emnlp-main.177](https://arxiv.org/abs/2024.07.177). <https://aclanthology.org/2024.emnlp-main.177/>.
- Gadamer, Hans-Georg 1900-2002 (1965). *Wahrheit und Methode : Grundzüge einer philosophischen Hermeneutik*. 2. Auflage, durch einen Nachtrag erweitert. Tübingen: J. C. B. Mohr (Paul Siebeck). http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=001502940&sequence=000002&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.
- Gengnagel, Tessa, Fotis Jannidis, Rabea Kleymann, Julian Schröter, and Heike Zinsmeister (Feb. 2024). "Bedeutung in Zeiten großer Sprachmodelle (Panel)". In: *Book of Abstracts - DHd2024*. Ed. by Joëlle Weis, Thomas Haider, and Estelle Bunout. Publisher: Zenodo. Passau: Zenodo, 81–85. <https://doi.org/10.5281/zenodo.10686565> (visited on 01/25/2025).



- Görner, Rüdiger (2016). *Hölderlin und die Folgen*. Stuttgart: J.B. Metzler Verlag. ISBN: 978-3-476-02651-4. 735-736
- Hacking, Ian (1994). "Styles of Scientific Thinking or Reasoning: A New Analytical Tool for Historians and Philosophers of the Sciences". In: *Trends in the Historiography of Science*. Ed. by Kostas Gavroglu, Jean Christianidis, and Efthymios Nicolaidis. Boston Studies in the Philosophy of Science. Dordrecht: Springer, 31–48. ISBN: 978-94-017-3596-4. 10.1007/978-94-017-3596-4_3. https://doi.org/10.1007/978-94-017-3596-4_3 (visited on 06/09/2022). 737-742
- Hicke, Rebecca M M and Ross Deans Kristensen-McLachlan (2024). "Science is Exploration: Computational Frontiers for Conceptual Metaphor Theory". In: *CHR Conference 2024*. 743-745
- Hölderlin, Friedrich (1805). "Hälfte des Lebens". In: *Taschenbuch f'ur das Jahr 1805. Der Liebe und Freundschaft gewidmet*. Frankfurt am Mayn: Friedrich Wilmans, 85. 746-747
- (1965). "Halves of Life". In: *An Anthology of German Poetry from Hölderlin to Rilke in English Translation*. Ed. by Angel Flores. Trans. by Kate Flores. Gloucester, Mass.: Peter Smith, 26–27. 748-750
- Kirschenbaum, Matthew (June 2023). *Again Theory: A Forum on Language, Meaning, and Intent in the Time of Stochastic Parrots*. <https://critinq.wordpress.com/2023/06/>. 751-752
- Köppe, Tilmann and Simone Winko (2013). *Neuere Literaturtheorien: Eine Einführung*. 2nd ed. Stuttgart, Weimar: Metzler. 753-754
- Künne, Wolfgang (2003). "Verstehen und Sinn: ein sprachanalytische Betrachtung". In: *Hermeneutik : Basistexte zur Einführung in die wissenschaftstheoretischen Grundlagen von Verstehen und Interpretation*. Ed. by Axel Bühler. Kolleg Synchron. Heidelberg: Synchron, 61–78. 755-758
- Liu, Emmy, Chenxuan Cui, Kenneth Zheng, and Graham Neubig (2022). "Testing the Ability of Language Models to Interpret Figurative Language". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 4437–4452. 10.18653/v1/2022.naacl-main.330. <https://aclanthology.org/2022.naacl-main.330> (visited on 06/18/2024). 759-764
- Makkreel, Rudolf A. (2002). "Pushing the Limits of Understanding in Kant and Dilthey". In: *Grenzen des Verstehens: philosophische und humanwissenschaftliche Perspektiven*. Ed. by Gudrun Kühne-Bertram and Gunter Scholtz. Göttingen: Vandenhoeck & Ruprecht. ISBN: 978-3-525-30138-8. 765-768
- Mirzadeh, Iman, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar (Oct. 2024). *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*. arXiv:2410.05229 [cs]. <http://arxiv.org/abs/2410.05229> (visited on 10/22/2024). 769-772
- Pfeifer, Hans (1922). "Unsere Toten". In: *Die deutsche Balladen-Chronik. Ein Balladenbuch von deutscher Geschichte und deutscher Art*. Ed. by Wilhelm Uhlmann-Bixterheide. Dortmund: Ruhfus, 290. 773-775
- Ryle, Gilbert (1945). "Knowing How and Knowing That: The Presidential Address". In: *Proceedings of the Aristotelian Society* 46, 1–16. 776-777
- Schmidt, Jochen (1982). "Sobria ebrietas. Hölderlins 'Hälfte des Lebens'". In: *Hölderlin-Jahrbuch* 23, 182–190. 778-779
- Scholz, Oliver (2005). "Die Vorstruktur des Verstehens. Ein Beitrag zur Klärung des Verhältnisses zwischen traditioneller Hermeneutik und 'philosophischer' Hermeneu- 780-781



- tik". In: *Geschichte der Hermeneutik und die Methodik der textinterpretierenden Disziplinen*. 782
- Ed. by Jörg Schönert. *Historia Hermeneutica*. Berlin u.a.: de Gruyter, 443–461. 783
- Searle, John R (1980). "Minds, brains, and programs". In: *Behavioral and Brain Sciences* 784
- 3-3, 417–457. 785
- Stekeler-Weithofer, Pirmin (2002). "Sind Sprechen und Verstehen ein Regelfolgen? Prob- 786
- leme konventionalistischer und intentionalistischer Theorien der Sprache". In: *Gibt 787*
- es eine Sprache hinter dem Sprechen?* Ed. by Sybille- Krämer and Ekkehard König. 788
- Orig.-Ausg., 1. Aufl. Suhrkamp-Taschenbuch Wissenschaft. Frankfurt am Main: 789
- Suhrkamp, 190–225. 790
- Strauss, Ludwig (1965a). "'Hälfte des Lebens'". In: *Interpretationen, Band 1: Deutsche 791*
- Lyrik von Weckherlin bis Benn*. Ed. by Jost Schillemeit. Vol. 1. S. Fischer Verlag, 113–134. 792
- (1965b). "Friedrich Hölderlins "Hälfte des Lebens"". In: *Interpretationen: Deutsche 793*
- Lyrik von Weckherlin bis Benn*. Ed. by Jost Schillemeit. Vol. 1. Frankfurt (am Main): 794
- Fischer, 113–134. 795
- Strube, Werner (2003). "Analyse des Verstehensbegriffs". In: *Hermeneutik : Basistexte zur 796*
- Einführung in die wissenschaftstheoretischen Grundlagen von Verstehen und Interpretation*. 797
- Ed. by Axel Bühler. Kolleg Synchron. Heidelberg: Synchron, 79–98. 798
- Tong, Xiaoyu, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova (2024). "Metaphor 799
- Understanding Challenge Dataset for LLMs". In: Publisher: arXiv Version Number: 800
1. 10.48550/ARXIV.2403.11810. <https://arxiv.org/abs/2403.11810> (visited on 801
- 02/07/2025). 802
- Trott, Sean (Oct. 2022). *How could we know if Large Language Models understand language?* 803
- Substack newsletter. [https://seantrott.substack.com/p/how-could-we-know-if-](https://seantrott.substack.com/p/how-could-we-know-if-large-language) 804
- [large-language](https://seantrott.substack.com/p/how-could-we-know-if-large-language) (visited on 02/25/2024). 805
- Trott, Sean and Cameron Jones (Sept. 2023). *Do Large Language Models have a "theory of 806*
- mind"*? Substack newsletter. <https://seantrott.substack.com/p/do-large-langu> 807
- [age-models-have-a-theory](https://seantrott.substack.com/p/do-large-langu) (visited on 02/25/2024). 808
- Trott, Sean, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen (July 809
- 2023). "Do Large Language Models Know What Humans Know?" In: *Cognitive 810*
- Science* 47:7, e13309. ISSN: 0364-0213, 1551-6709. 10.1111/cogs.13309. [https://onli](https://onlinelibrary.wiley.com/doi/10.1111/cogs.13309) 811
- [nelibrary.wiley.com/doi/10.1111/cogs.13309](https://onlinelibrary.wiley.com/doi/10.1111/cogs.13309) (visited on 11/08/2023). 812
- Turing, Alan M. (2021). *Computing Machinery and Intelligence / Können Maschinen denken?* 813
- 1950 Aus dem Englischen übersetzt und herausgegeben von Achim Stephan und Sven 814
- Walter Unter Mitarbeit der Mitglieder des Turing-Studienprojektes. Stuttgart: Reclam. 815
- [https://www.reclam.de/detail/978-3-15-961827-2/Turing__Alan_M_/Computin](https://www.reclam.de/detail/978-3-15-961827-2/Turing__Alan_M_/Computing_Machinery_and_Intelligence___Koennen_Maschinen_denken___EPUB_) 816
- [g_Machinery_and_Intelligence___Koennen_Maschinen_denken___EPUB_](https://www.reclam.de/detail/978-3-15-961827-2/Turing__Alan_M_/Computing_Machinery_and_Intelligence___Koennen_Maschinen_denken___EPUB_) (visited 817
- on 02/04/2025). 818
- Uhlmann-Bixterheide, Wilhelm, ed. (1922). *Die deutsche Balladen-Chronik. Ein Balladen-* 819
- buch von deutscher Geschichte und deutscher Art*. Dortmund: Ruhfus. 820
- Underwood, Ted (2023). "The Empirical Triumph of Theory | In the Moment". In: *Critical 821*
- Inquiry: Again Theory. A Forum on Language, Meaning, and Intent in Time of Stochastic* 822
- Parrots*. [https://critinq.wordpress.com/2023/06/29/the-empirical-triumph-o](https://critinq.wordpress.com/2023/06/29/the-empirical-triumph-of-theory/) 823
- [f-theory/](https://critinq.wordpress.com/2023/06/29/the-empirical-triumph-of-theory/) (visited on 01/25/2025). 824
- Wachowiak, Lennart and Dagmar Gromann (2023). "Does GPT-3 Grasp Metaphors? 825
- Identifying Metaphor Mappings with Generative Language Models". In: *Proceedings* 826
- of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: 827*
- Long Papers*). Toronto, Canada: Association for Computational Linguistics, 1018–1032. 828

- 10.18653/v1/2023.acl-long.58. <https://aclanthology.org/2023.acl-long.58> 829
(visited on 02/07/2025). 830
- Walsh, Melanie, Anna Preus, and Elizabeth Gronschi (Dec. 2024). "Does ChatGPT Have a Poetic Style?" In: *Proceedings of the Computational Humanities Research Conference 2024*. Ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. CEUR Workshop Proceedings, 1201–1219. <https://ceur-ws.org/Vol-3834/paper122.pdf>. 831
832
833
834
- Winko, Simone, Stefan Descher, Urania Milevski, Merten Kröncke, Fabian Finkendey, Loreen Dalski, and Julia Wagner (2024). "Praktiken des Plausibilisierens". In: Accepted: 2024-12-06T09:17:48Z Artwork Medium: Print Interview Medium: Print. 835
836
837
10.17875/gup2024-2639. <https://univerlag.uni-goettingen.de/handle/3/isbn-978-3-86395-641-7> (visited on 02/05/2025). 838
839
- Wittgenstein, Ludwig (2001). *Philosophische Untersuchungen*. 1. Aufl. Frankfurt am Main: Suhrkamp. ISBN: 978-3-518-58312-8. 840
841
- Xu, Ruoxi, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han (May 2024). "AI for social science and social science of AI: A survey". In: *Information Processing & Management* 61.3, 103665. ISSN: 0306-4573. 842
843
844
10.1016/j.ipm.2024.103665. <https://www.sciencedirect.com/science/article/pii/S0306457324000256> (visited on 02/25/2024). 845
846
- Zymner, Rüdiger (2007). "Assonanz". In: *Reallexikon der deutschen Literaturwissenschaft*. Ed. by Klaus Weimar, Harald Fricke, and Jan-Dirk Müller. Vol. 1. Berlin, New York: de Gruyter, 156–157. 847
848
849

Reconstructing Shuffled Text

Bad Results for NLP, but Good News for Using In-Copyright Texts

Keli Du¹ 
 Sarah Ackerschewski²
 Uygur Navruz²
 Nazan Sinir²
 Julian Valline² 
 Christof Schöch² 

1. Trier Center for Digital Humanities, University of Trier , Trier, Germany.
2. Department for Computational Linguistics and Digital Humanities, University of Trier , Trier, Germany.

Citation

Keli Du, Sarah Ackerschewski, Uygur Navruz, Nazan Sinir, Julian Valline, and Christof Schöch (2025). "Reconstructing Shuffled Text. Bad Results for NLP, but Good News for Using In-Copyright Texts". In: *CCLS2025 Conference Preprints 4* (1). [10.26083/tuprints-00030140](https://doi.org/10.26083/tuprints-00030140)

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-01-30

Keywords

Derived text formats, copyright, reconstructibility, evaluation

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. Existing copyright laws in the European Union, the United States, and many other jurisdictions worldwide impose limitations on Text and Data Mining that affect the storage, publication, and reuse of datasets built from in-copyright texts. This issue directly affects researchers in CLS, a field in which work on contemporary materials is desirable and in which Open Science principles are quite strongly established. As a solution, derived text formats (DTFs) have been proposed. One important aspect of DTFs regarding copyright law is the reconstructibility of the source text from its corresponding DTF. In this paper we present the first of a series of experiments we plan to conduct on this issue. For this experiment, we have fine-tuned a large language model to reconstruct source texts from DTFs. The results of the reconstruction vary depending on various conditions, but on the whole are not very successful. This suggests that reconstructing text in DTFs is not as simple as is sometimes assumed and we believe that this result may encourage scholars to convert their in-copyright texts to DTFs and publish them as research data.

1. Introduction

In many Digital Humanities (DH) projects, texts are being digitized, collected and/or enriched in order to be used as research data. However, existing copyright laws in the European Union, the United States, and many other jurisdictions worldwide impose several limitations on Text and Data Mining (TDM) that affect the storage, publication, and reuse of datasets built from in-copyright texts. This undoubtedly has a negative impact on the reproducibility of published research results and on the spirit of open science. As a potential solution to this problem, scholars have proposed and are currently utilizing derived text formats (DTFs), also known as extracted features, for non-consumptive research (see e.g. Y. Lin et al. 2012, Bhattacharyya et al. 2015, Jett et al. 2020, Schöch et al. 2020, Organisciak and Downie 2021). The 'Hathi Trust Extracted Features' (Jett et al. 2020), for example, might be the most widely-used set of DTFs in the digital humanities. However, beyond the specific design choices of this particular DTF, many other kinds of

DTF exist or could be envisioned. 14

The key idea behind DTFs is to selectively remove specific copyright-relevant information 15
from in-copyright texts by applying various transformations to them, so that these texts 16
are no longer readable by humans and do not contain copyright-relevant features. If 17
done in a suitable manner, the publication of such texts as research data is unlikely to 18
affect the rights of copyright holders. At the same time, they remain suitable for (at 19
least some of the) TDM tasks in the digital humanities, such as authorship attribution, 20
topic modeling, or sentiment analysis (see e.g. Kocula 2021, Du 2023). 21

There are several types of transformations that can be used to create text in DTF: removal, 22
exchange, and replacement. For example, the sequence information in text can be 23
removed, that is, as the example in Table 1 shows, the order of the words can be shuffled. 24
To convert a novel to this format, it is first split into chunks (for example 1000-words 25
chunks or 500-words chunks). Then, the sequence information, i.e. the order of the 26
words in each chunk, is removed by randomizing their order. Note that the sequence 27
of the chunks within each whole text is maintained. This allows the text to become 28
less readable while roughly preserving the main structure of the original text. Another 29
possibility is to reduce the information about individual tokens by replacing a certain 30
percentage of word forms with their corresponding Part-of-Speech (PoS) tags, without 31
affecting word sequence information. Furthermore, since the goal of transforming text 32
into DTFs is to keep the textual information for different TDM tasks, word embeddings 33
(both static and contextualized) are also a promising candidate for information-rich 34
DTFs (Schöch et al. 2020). 35

Table 1: An example of a text and its two variants in DTFs.

source text	Sherlock Holmes took his bottle from the corner of the mantel-piece and his hypodermic syringe from its neat morocco case. With his long, white, nervous fingers he adjusted the delicate needle, and rolled back his left shirt-cuff.
word order shuffled	his bottle from of mantel-piece With his syringe from its the neat case. His white, took the fingers he and hypodermic Sherlock the Morocco delicate needle, and nervous corner rolled back his left shirt-cuff. Holmes long, adjusted
50% of words replaced by POS tags	Sherlock PROPN VERB his NOUN from DET corner of the NOUN-piece CCONJ PRON hypodermic NOUN ADP its ADJ morocco case. ADP his long, ADJ, ADJ NOUN he VERB the delicate NOUN, CCONJ rolled ADV his left shirt-NOUN.

Technically speaking, DTFs are actually text that contains noise. It is not a difficult task to 36
convert text data into such. In contrast, DTFs are currently facing more controversy at the 37
legal level. For example, there is the view that converting texts to DTFs and publishing 38
them does not constitute a copyright infringement in itself, only reconstructing the 39
source texts from DTF texts does; however, there is another view that even if the texts 40
are converted to DTFs, these texts could still be protected by copyright law and therefore 41
cannot be made public. Against this background, the legal status of DTFs is discussed 42
in detail in Iacino et al. 2025. Among other points, the article discusses the attitude of 43
German courts towards the relationship between text length and copyright protection. 44
As long as the DTF text does not contain text fragments longer than 11 words that are not 45

sufficiently different from the original work, then such DTF text are unlikely to infringe copyright law.

One important aspect of DTFs regarding copyright law is the reconstructibility of texts in DTFs. If we want to prepare the in-copyright texts in a DTF and make them available to others, we have to be careful that the source texts cannot be easily reconstructed. On this point, Raue and Schöch 2020 stress that the original texts should not be reconstructable with reasonable effort, for example on the basis of position information of the text sequences or other sequence information.¹ Of course, the definition of “with reasonable effort” here is very vague. Therefore, it is essential to demonstrate how easy or difficult it is to reconstruct text in DTFs through practical experiments. In the following, we first outline the motivation of our research in detail (section 2), then we describe our data and methodology (section 3) and provide a discussion of the relation between our research and the memorization issue in LLMs (section 4). After that, we will present and discuss the results of the reconstruction experiments (section 5), before we conclude (section 6).

2. Motivation

Since we are not experts in copyright law, we are focused on evaluating DTFs from the perspective of Natural Language Processing (NLP). Our goal is to share our knowledge in order to provide some arguments for legal experts when the legal status of DTFs is discussed. In recent years, technologies related to large language models (LLMs) have developed rapidly. BERT, for instance, is trained using two tasks: one where the model learns to predict a masked word from context, and one where it learns whether two sentences directly follow each other or not (Devlin et al. 2019). In contrast, BART is trained on texts with sentences in random order, learning to reconstruct the original sequence during training (Lewis et al. 2019). The textual data used to train LLMs is very similar to the text in DTFs, and the task of training LLMs is analogous to reconstructing text in DTFs. Therefore, it can be assumed that LLMs may be capable of reconstructing the original text from DTFs. And indeed, Kugler et al. 2023 demonstrated that the publication of the encoder together with the contextualized embeddings makes it possible to generate data for training a decoder that exhibits a reconstruction accuracy that is very likely sufficient to violate copyright. But their test is not the same as the usage scenario we aim to investigate. Their study was to infer the data used to train LLMs, while we focus on reconstructing text from DTFs similar to that in Table 1.

In fact, most of the NLP experts we have encountered agree that it should be possible to reconstruct text in DTFs using LLMs. Even if it is not possible now, it will be realized in the future as NLP technology advances. However, to the best of our knowledge, there are no relevant studies addressing this issue yet. Our preliminary test with ChatGPT (free version) showed that reconstructing text in DTFs is not impossible, even though

1. Quote in the original German: “Neben der Nichterkennbarkeit wird man als zweite Anforderung von einem urheberrechtsfreien, abgeleiteten Textformat verlangen müssen, dass die ursprünglichen Texte nicht aufgrund von Positionsangaben der Textsequenzen oder sonstiger Sequenzinformationen mit verhältnismäßigem Aufwand rekonstruierbar sind.” (Raue and Schöch 2020).

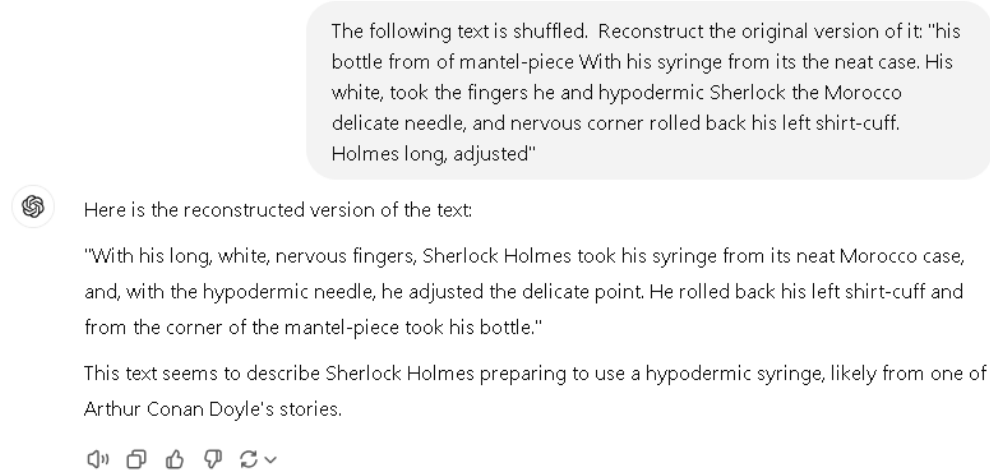


Figure 1: Reconstructing shuffled text using ChatGPT (The source text is in Table 1.)

ChatGPT is not trained for such a specific task (see Figure 1).² So we can assume that it may be possible to reconstruct the source text from different DTFs with the help of specially trained LLMs, even though from the point of view of making in-copyright textual data publicly available, we would prefer that the text reconstruction experiments will result in an *unsuccessful* outcome. Therefore, we are planning to conduct a series of experiments in order to review the degree of difficulty in reconstructing text in different DTFs. In the research reported here, and as a first step, we have tested the reconstruction of the text where the word order was shuffled.

3. Data and Methodology

The primary aim of our research is to investigate whether LLMs can reconstruct shuffled literary texts. For our experiments, we utilized textual data from two datasets: IMDb reviews (non-literary texts) and a subset of the Gutenberg corpus (literary texts).³ Non-literary texts are generally considered less complex than literary texts. Therefore, it is reasonable to hypothesize that reconstructing IMDb reviews would be easier and more successful than reconstructing texts from the Gutenberg corpus. By comparing the results of these two datasets, we can test this hypothesis and gain a deeper understanding of the model's ability to reconstruct shuffled text.

3.1 IMDb-reviews

In the experiments of reconstructing IMDb-reviews, each review was used as one data point. To transform the IMDb-reviews into the DTF format, we only shuffled the word order of each sentence. The order of the sentences in each review was not altered. Three sets of data containing 25000, 50000 and 75000 reviews were prepared as training data, while the validation and testing data contained 5000 unseen reviews in each case. By varying the amount of the training data, we can test the hypothesis that more training

2. Surely, as we can see from ChatGPT's answer, this relatively successful text reconstruction is most likely due to the fact that the model has already seen the original text during training. Therefore, the output text may possibly be "memorized" rather than "reconstructed". The issue of memorization will be discussed in more detail in section 4.

3. All the textual data are published here: <https://github.com/dkltimon/reconstruction>.

data leads to better reconstruction results. 108

3.2 Gutenberg texts 109

In order to ensure that the literary genre does not become a confounding factor in the test results, we randomly selected Gutenberg novels from four different genres: detective fiction, historical fiction, love stories, and science fiction. Two datasets were created for evaluating the impact of the amount of training data. One consisting of 3 novels from each genre (12 novels in total) and the other consisting of 15 novels from each genre (60 novels in total). All the novels were split into chunks, and the words within each chunk were randomly shuffled, while the order of the chunks was not altered. These chunks were then used as data points for model training, validation, and testing in the ratio of 80%, 10%, 10%. The chunk lengths were set to 50, 100, and 500 words. By varying the chunk length, we can also test the hypothesis that reconstructing shorter chunks/texts will be more successful. Since we used either 15 or 60 novels as the dataset, when these novels are divided into chunks and the chunk length is set differently, the total number of divided chunks is different. An overview of the number of chunks can be found in Table 2. 110 111 112 113 114 115 116 117 118 119 120 121 122 123

Table 2: Number of chunks used as training, validation and testing data.

	chunk length: 50		chunk length: 100		chunk length: 500	
	15 novels	60 novels	15 novels	60 novels	15 novels	60 novels
training data	14258	89915	7130	44971	1428	9013
validation data	4753	11240	2377	5622	477	1127
testing data	4753	11239	2377	5621	476	1127

3.3 Method 124

We treated the reconstruction of texts in DTF as an automatic translation task and used the translation pipeline from Huggingface.⁴ Automatic translation converts a sequence of text from one language to another. In the context of our research, these two languages are DTF text and the original text. This task can be formulated as a sequence-to-sequence problem and therefore requires using a sequence-to-sequence large language model. We used the pre-trained T5-base model and fine-tuned the model using DTF texts as the input and the unaltered source texts as the target text of the translation. The “Text-to-Text Transfer Transformer” (T5) model is a framework that treats the tasks of translation, question answering and categorization as the same process: The model takes text as input and generate target text as output. In this way, the same model, loss function, hyperparameters, etc. can be used for different tasks (Raffel et al. 2020). After fine-tuning, the model was evaluated on the unseen testing data. For both fine-tuning and evaluation, six measures were used to compare the predicted text with the target text. They have been proposed to compare the similarity of strings and often used to evaluate the results of automatic translation. 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139

4. See: <https://huggingface.co/docs/transformers/tasks/translation>.

- WER: The word error rate (WER) is derived from the Levenshtein distance, working at the word level. It indicates the average number of errors (substitutions, deletions and insertions) per reference word. The smaller the value is, the higher the similarity (Woodard and Nelson 1982, Morris et al. 2004).
- ROUGE scores: ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation and it is a set of metrics focusing on comparing texts from different perspectives. ROUGE-1 is a unigram (1-gram) based scoring and compares the maximum number of unigrams occurring both in the predicted text and the reference text. ROUGE-2 is a bigram (2-gram) based scoring and compares the maximum number of bigrams occurring both in the predicted text and the reference text. ROUGE-L focuses on the longest common subsequence between predicted and the reference text and the ROUGE-Lsum first splits the text into sentences, then performs ROUGE-L calculations for each sentence individually. For all the ROUGE scores, higher scores indicating higher similarity between the predicted and the reference text (C.-Y. Lin 2004).
- SacreBLEU: SacreBLEU is a tool for calculating shareable, comparable, and reproducible BiLingual Evaluation Understudy (BLEU) scores (Papineni et al. 2002) and reports scores between 0 and 100. 0 being zero resemblance, 100 means identical sentences (Post 2018).

4. Memorization in LLMs

In the test using ChatGPT to reconstruct the text in Figure 1, we see that ChatGPT recognizes the source of the input text. It shows that the model has a memory for the text it has seen in the pre-training. If the training data can be reproduced verbatim, this phenomenon is called Memorization in LLMs (see e.g. Lee et al. 2022, Biderman et al. n.d.). This issue has been examined by inverting the BERT pipeline (Kugler et al. 2023), through name cloze inference (Chang et al. 2023), or by asking the LLMs to complete a passage extracted from a book and measuring the overlap of the first ten tokens it produces with the real text in the book (Zhang et al. 2024). The latter two studies mentioned above prompted generative LLMs to examine which books exists in the training data. In the case of the present study, the datasets used are both publicly available and it is almost entirely certain that the data was used for the pre-training of most of the LLMs, meaning the LLMs have seen these texts already. From a technical perspective, our experiments are therefore also an examination of memorization in LLMs. However, our study is different from memorization-focused studies in the following aspects:

1. The memorization-focused studies looked at inferring the training data of LLMs or proving that certain data was used for training of LLMs. Thus, the LLMs are the object of their study. In comparison, our study use LLMs to reconstruct source texts from DTFs. Therefore, the object of our study are DTFs and the LLMs are used as research tools.
2. Although our experiments have used data that LLMs are highly likely to have seen during training, which makes our experiments fit the scope of research on

memorization, the real-world application of our approach is to reconstruct in-copyright texts that are not available online and less likely to have been included in the pre-training of LLMs. That’s a different task from examining memorization in LLMs.

3. The motivation for our study is law and practice: Our ultimate goal is to enable scholars to make in copyright texts publicly available as research data and DTFs are our solution to this problem. Therefore, regardless of whether or not LLMs have seen the original text, as long as the DTF text is used as input data to LLMs and the original text is reconstructed as a result, then DTF cannot be considered as a solution for making in-copyright texts public.
4. Since the goal of our research is not to examine memorization in LLMs, questions such as the correlation between the memorization of books in LLMs and the appearance frequency or popularity of the same books on the web (Zhang et al. 2024) are not central to us.

5. Results

5.1 Reconstructing IMDb-reviews

The results for reconstructing the unseen 5000 reviews in the testing data is presented in Figure 2, which is a comparison of the three trained models’ performance across six evaluation metrics. The “scrambled_baseline” in the figure represents the string similarity between the text in DTF and the source text. This baseline allows us to examine the extent to which the reconstruction has brought the DTF text back to the original. The “25000_model”, “50000_model” and “75000_model” labels represent the scores achieved with models trained with 25000, 50000 and 75000 reviews, respectively. Since WER is different from all other scores in that higher values represent poorer results, to make it easier to understand the results, the results for 1-WER are shown here. Also, the sacreBLEU scores are scaled down by a factor of 100 for visualization convenience. All six measures show very similar results: models trained with more data have better results in reconstructing text. The model trained using 75000 reviews gets the best scores in all tests, except for the ROUGE-1 score of the “scrambled_baseline”. This is because the ROUGE-1 is a unigram based scoring. Since we’re only disrupting the order of the words in the text, it’s no surprise that the baseline has a perfect score, 1.0. The other ROUGE-1 scores indicate that in the process of reconstructing the text, the model is not simply putting all the input scrambled words in the correct original order, but “rewriting” the text given the information provided by the input text. This is very likely due to the fact that we are treating the reconstruction as an automatic translation task and the model is not given direct instructions to use all the words in the input text during training. Overall, judging by the scores, even in the best cases, the similarity between the reconstruction results and the original text remains limited.

Obviously, these numerical assessments are not sufficient to let us see the full picture of the test results. We therefore selected three examples including the source text, their reconstructed texts, and their scores in order to provide readers with a more intuitive

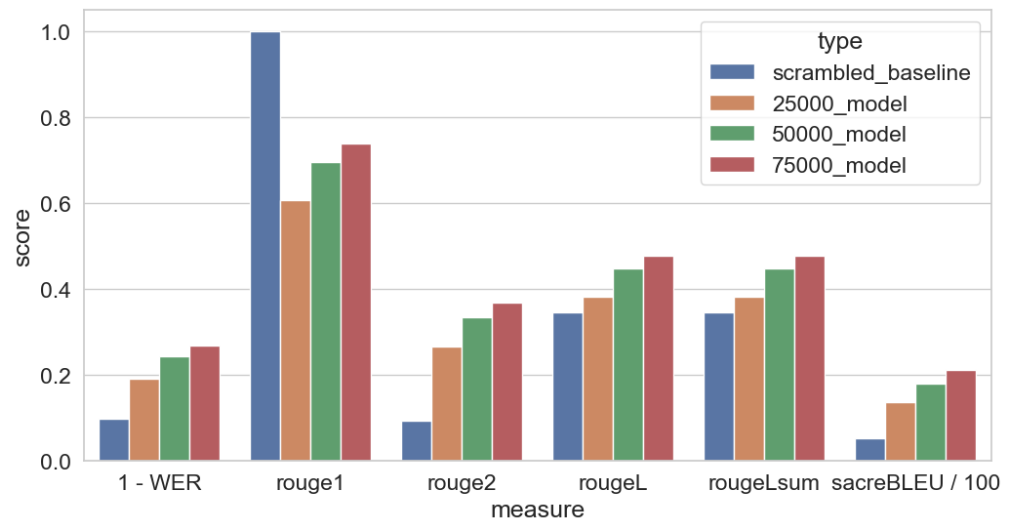


Figure 2: Average similarity scores achieved by different models in reconstructing IMDb-reviews. (Higher values represent better results.)

idea of the results of the reconstruction.⁵ All three examples were selected from the text 223
reconstructed using the best model, “75000_model”. The similarity scores between the 224
source texts and their reconstruction are presented in Figure 3, while the three reviews 225
and their reconstruction are listed in Table 3. In terms of similarity scores, review No. 226
4691 is the best reconstruction result, No. 4320 is the second best, and No. 4758 is the 227
worst. The sacreBLEU score for No. 4758 is 0.00085 after dividing by 100. It’s so low 228
that it’s barely visible in the visualization. 229

As we can see in Table 3, the reconstruction of the review No. 4691 is indeed more 230
successful. Two of the four sentences are identical to the source text, but the other two 231
are not. The last sentence, in particular, expresses the opposite meaning and sentiment of 232
the source text. In comparison, the reconstructed text of the review No. 4320 is different 233
in length from the source text (and its DTF version), and most of the reconstructed text 234
is inconsistent with the source text. Only one or two sentences or phrases are identical 235
to the source text. It is worth noting that the difference between No. 4691 and No. 4320 236
for the ROUGE-1 score is not particularly large, but the difference in the quality of the 237
reconstructed text is very obvious. This suggests that a bigger overlap of unigrams 238
between reconstructed text and the original texts is relatively easy to achieve. In contrast, 239
there is much less overlap between the longer sequences (bigrams etc.) of reconstructed 240
text and the source text. The reconstruction of review No. 4758 can be described as a 241
complete failure. Although the first and last sentences are the same as in the source text, 242
the longer part of the text in the middle has been replaced by the multiple repetition of 243
the string “pseudo-”. In the results of reconstruction of the 5000 unseen reviews, this 244
multiple repetition of the same string can be observed quite often. We assume that this 245
might be caused by Greedy sampling, which is fairly common in tasks that use LLMs to 246
generate text (see: e.g. Welleck et al. 2019, Holtzman et al. 2020, Fu et al. 2021). 247

To determine how many reconstructed texts can achieve a level of similarity comparable 248
to the review No. 4691, we can refer to Figure 4. This figure presents the distribution of 249
similarity scores for all 5000 unseen reviews reconstructed using the “75000_model.” 250

5. All the results of the reconstruction are available online. Please see section 7.

Table 3: Three IMDb-reviews and their reconstruction

Review	type	text
No. 4691	source	This movie starts with a lot of promise. The opening scene, featuring Sean Connery, is very entertaining. However, Connery disappears for most of the rest of the movie along with any talent that anyone else may have exhibited. The movie jumps from place to place with no coherent story. There is no sense of time. The editing is laughable. After the first 5 minutes there is nothing worth watching in this film.
No. 4691	reconstructed	This movie starts with a lot of promise. The opening scene featuring Sean Connery, is very entertaining. However, the rest of Connery may have exhibited any talent along with the rest of the movie. The movie jumps from place to place with no coherent story. There is no sense of time. The editing is laughable. After watching this film there is nothing worth watching.
No. 4320	source	I remember seeing this on T.V. in the early '80's, and even though I was still kind of young, I thought it was awful. Rock Hudson should really have been more selective of the scripts he accepted. some of his films are really good, and others like "embryo" and this piece of drek should have been left to the next generation of actors.now for the scene that I thought was the funniest of the whole movie. it happens at the very end as the camera is pulling away and the screen starts to fade to black. If I remember the scene correctly, a group of people are still in either a wrecked hotel or a cave and some guys wife has just been declared dead. as the camera pulls back you can clearly see the "dead woman" stand up and walk off set.
No. 4320	reconstructed	I remember seeing this on T.V., and even though it was in the early 80's, I still thought it was awful. Rock Hudson should have been more selective for the scripts. I thought that this was the "embryon" of the next generation of actors, and the whole piece of drek was really good and funniest. I think some of the films should have been the funniest and now the scene starts to fade away as the camera starts pulling away at the end. If the scene has been declared dead, or a group of guys are still in a cave or a wrecked hotel.
No. 4758	source	I couldn't make it through the whole thing. It just wasn't worth my time. Maybe one-fourth of the dialogue would have been worth listening to (or reading – since I don't understand French) if the pseudo-profundity and pseudo-wittiness of the other three-fourths of the film were deleted. ... [Here, around 230 words from the source text have been omitted.] ... At least these films are interesting and enjoyable, which is much more than I can say about IN PRAISE OF LOVE (Éloge de l'amour). I give this film 2 out of 10 stars. Not quite offensive enough to rate 1 for "awful" (such as "The Devils" with Oliver Reed and Vanessa Redgrave). If you still want to watch it, go ahead. But don't say I didn't warn you!!!
No. 4758	reconstructed	I couldn't make it through the whole thing. It just wasn't worth my time. Maybe - since the fourths of the French dialogue were deleted (if the pseudo-wittiness of the other three) or - if the pseudo-pseudo-pseudo ... [Here, "-pseudo" repeated 52 times.] ... -pseudo-pseudo-pseudo. But don't say I didn't warn you!!!

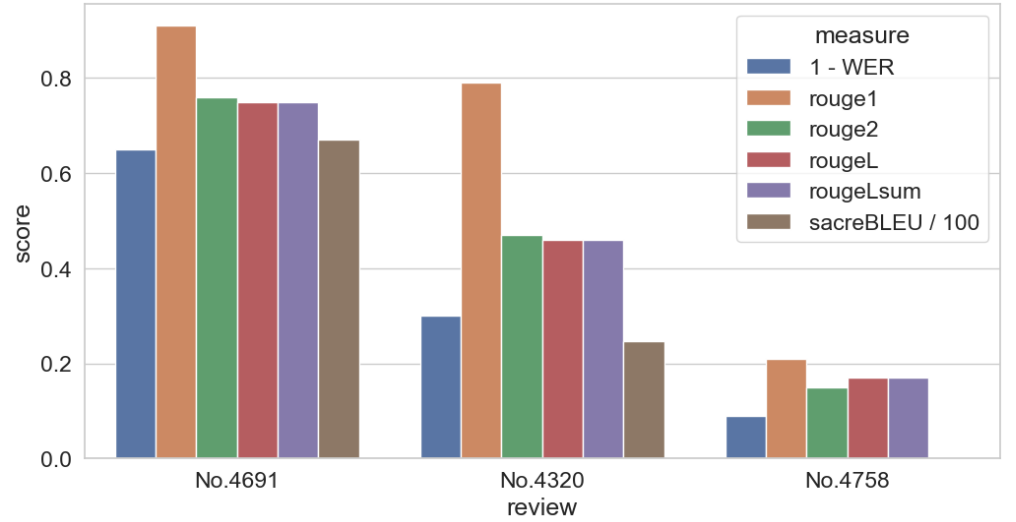


Figure 3: String similarity of three reconstructed IMDb-reviews.

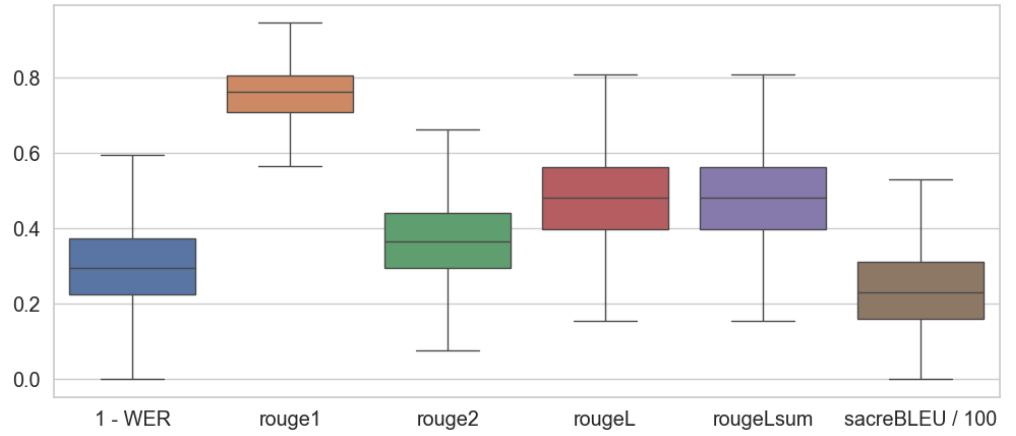


Figure 4: Similarity score distribution of reconstructed IMDb reviews using the “75000_model” (The outliers are not visualized).

The data indicates that a very clear minority of these reconstructed texts (significantly less than 25%) attain a ROUGE-2 or ROUGE-L score greater than 0.7. In comparison, although the ROUGE-1 scores are much higher overall, the majority (around 75%) of the ROUGE-1 scores are also lower than 0.8. Thus, we can conclude that the reconstruction of the IMDb-reviews is not successful.

5.2 Reconstructing Gutenberg texts

The reconstruction results of Gutenberg text chunks are presented in Figure 5. The top plot shows the result using 12 novels and the bottom plot is the similarity scores achieved using 60 novels as data. In the top plot, all evaluation measures have relatively low scores across the three chunk lengths. As in the previous test, ROUGE-1 has slightly higher values compared to other measures, but overall, the scores are low. In comparison, the scores in the bottom plot improve significantly. Compared to other evaluation metrics, ROUGE-1 has the highest scores, especially for smaller chunk lengths (50 and 100). The results suggest that both corpus size and chunk length have an impact on the reconstruction, with larger corpora and smaller chunk lengths generally yielding better

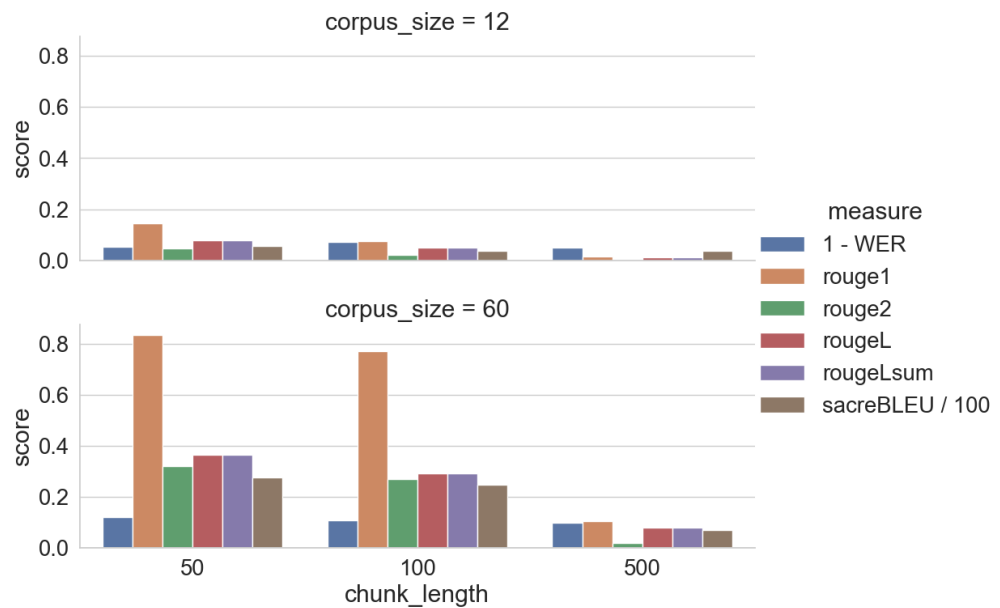


Figure 5: Average similarity scores achieved in reconstructing Gutenberg text chunks.

reconstruction results. Altogether, this test confirms the observation in reconstructing the IMDb-reviews in the previous test. After reviewing all the reconstructed texts, we found that the multiple repetition of the same string caused by Greedy sampling are more frequent when reconstructing text chunks of 500 words. This problem is clearly related to the length of the text that must be generated.⁶

To provide an idea of the quality of the reconstructed text, we selected three examples from the reconstruction results of 60 novels that were segmented into 50-words chunks, as the reconstruction was most successful for shorter chunks using more data. Figure 6 shows the similarity scores for the three reconstructed chunks, while their source and reconstructed texts are provided in Table 4. Chunk No. 3536 achieved the highest scores, including a perfect ROUGE-1 score. Its reconstructed text differs very little from the original text in general, especially the first two sentences, which are nearly identical to the original text. However, the later sentences differ significantly in meaning due to the confused placement of the personal pronouns. In contrast, the scores for Chunk No. 1368 were much lower, and it is quite difficult to infer the source text from the reconstructed text. Chunk No. 5481 had the least successful reconstruction, with minimal scores. Its reconstructed text consisted only of a series of dots and three words.

Figure 7 provides an overview of the distribution of similarity scores for all unseen 50-words chunks. Although more than 75% of the ROUGE-1 scores are over 0.8, over 75% of the other similarity scores (for WER even almost all of the scores) are lower than 0.4. This means that the majority of the reconstructed texts have a high degree of unigram overlap with the source text and they are of moderately poorer quality than No. 1368. In comparison, very successful reconstructions like No. 3536, or very unsuccessful reconstructions like No. 5481, are in a very small minority.

Considered together, the test results of the two datasets show that reconstructing DTF texts is quite challenging, especially for longer texts. Even for texts as short as 50 words,

6. All the reconstructed 50-words, 100-words and 500-words chunks are available online. Please see section 7.



Review	type	text
No.3536	source	David ! " she cried,—“my dear David — ! ” Then she broke off . “ What is it ? ” she asked , in a different tone . He showed her the headlines of the newspaper he was carrying . “ Tragedy ! ” he answered hoarsely . “
No.3536	recon- structed	! " she cried,—“my dear David ! ” He answered hoarsely . “ What is it ? ” she asked . Then he broke off in a different tone . “ David ! ” he showed her the headlines of the newspaper . “ Tragedy ! ” she was carrying
No.1368	source	"else , so that I had not so much as a glimpse of her face . But I knew that it was Mary . "" Come , "" said my lord , pleasantly . "" We will go to her . It may be , she will not have the"
No.1368	recon- structed	". "" Come , my lord , "" she said pleasantly . "" We may not have so much as a glimpse of her face . But it was so , as I knew , that Mary will not go . It will be so , that I will not"
No.5481	source	be humble . The thought had mingled with the sea 's rhythmic lullaby as it hushed her restless soul to sleep last night . He had offered her a new God who was Love,—his God . One who gave him happiness and content . Why should she resist ? Was
No.5481	recon- structed One who had

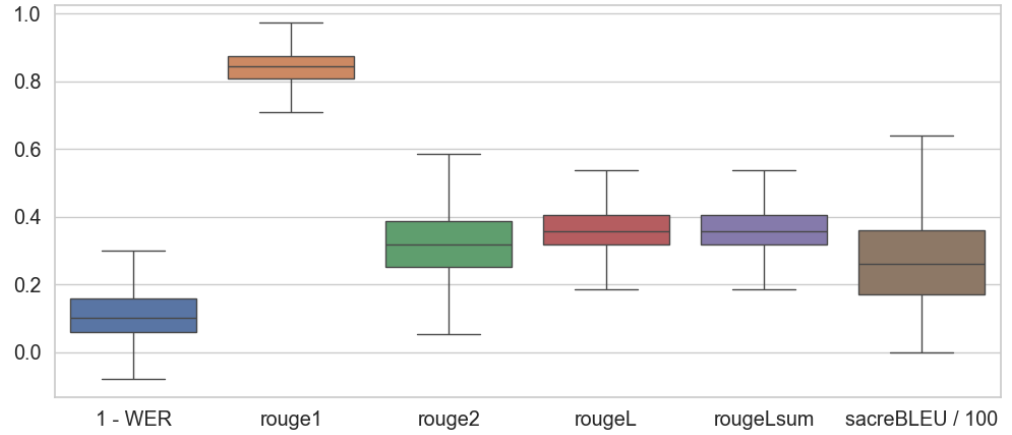


Figure 7: Similarity score distribution of reconstructed Gutenberg 50-words chunks (The outliers are not visualized).

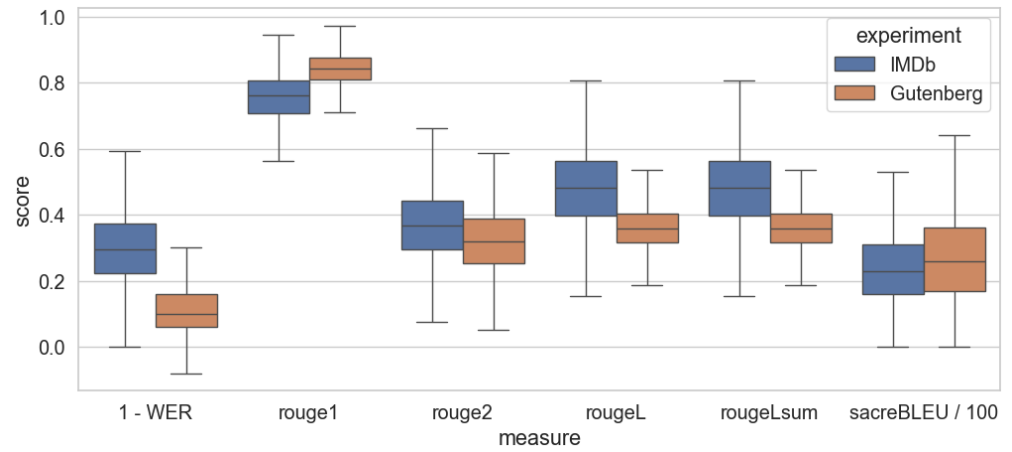


Figure 8: Comparison of the score distributions of the most successful reconstructions on the two datasets (The outliers are not visualized).

the reconstructed texts are, especially at the semantic and content level, still far from the original. By comparing the score distributions of the most successful reconstruction on the two datasets (Figure 8), we can conclude that it is indeed more difficult to reconstruct literary texts. This is because the experiments on Gutenberg dataset scored lower on e.g. ROUGE-2 or ROUGE-L, which can better reflect the results of the reconstruction.

6. Conclusion

In this paper we presented our experiments on reconstructing text in a DTF using fine-tuned LLM. In order to gain a preliminary understanding of the ability of LLMs to reconstruct text, we fine-tuned the T5-base model using text with scrambled word order and used it for the reconstruction of unseen text.

The results of the reconstruction are mixed, but on the whole not very successful. What is clear is that this task of text reconstruction can very likely be improved by optimizing the technical aspects, e.g. by using different model training strategies, more powerful models, more training data, setting the length of the reconstructed text to be the same as the length of the source text, choosing a different sampling mechanism from greedy

sampling, and so on. On the other hand, if the text to be reconstructed is more complex — 307
 such as in-copyright, less well-known literary works that are not available on the Internet, 308
 which aligns more closely with real-world applications of DTFs —, then the task becomes 309
 more difficult. Additionally, if the shuffling of word order goes beyond the level of 310
 sentences or 50 words (for example extending to the level of paragraphs or whole texts), 311
 or combining different DTF methods for transforming texts (for example replacing 10% 312
 of random words with their corresponding PoS tags in addition to shuffling the word 313
 order), this will undoubtedly make the reconstruction significantly more challenging. 314
 Also, if the same book is converted into different DTFs and all these DTF texts are 315
 publicly available, it might be easier to reconstruct the text by combining these DTFs. 316
 All of these aspects remain to be studied and we will keep working on this topic with 317
 more experiments in order to determine exactly how complex it is to reconstruct text in 318
 different DTFs, and what factors this depends on. 319

As a possible reference for defining “with reasonable effort” (mentioned in the “Intro- 320
 duction”), we would also like to briefly report on the resources used to accomplish 321
 this work. This work has been conducted as a collaboration between four NLP Masters 322
 students and a DH postdoctoral researcher, and in close consultation with an established 323
 DH researcher. Our experiments show that reconstructing text in just one DTF is not a 324
 simple task for someone without sufficient expertise in NLP, as we needed to implement 325
 and test a custom-built reconstruction pipeline for this task. This task also requires 326
 considerable resources, in the sense that to train the model, we used a workstation 327
 equipped with an Nvidia GeForce RTX 4090 GPU, which costs several thousand euros 328
 and consumes considerable amounts of power during the training and inference process. 329
 In addition, the process requires time, as depending on the size of the dataset, training 330
 the model and inference on unseen data can take several hours to several days. In 331
 contrast, anyone can obtain digitized text with much better quality by taking photos of 332
 a printed book and running OCR on the page images (even the iPhone, for example, 333
 has OCR software integrated), which is much cheaper, faster and easier. 334

Finally, we believe that the bad results of our experiments is good news for using in- 335
 copyright text as research data. We hope, at the very least, that the results presented here 336
 can be encourage DH scholars to convert their in-copyright texts to DTFs and publish 337
 them as research data, which is very valuable for transparent and sustainable research 338
 and access to large reference corpora. 339

7. Data Availability 340

Data can be found here: <https://github.com/dkltimon/reconstruction> 341

8. Software Availability 342

Software can be found here: <https://github.com/dkltimon/reconstruction> 343

9. Acknowledgements

344

This publication was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

10. Author Contributions

352

Keli Du: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing - original draft, Writing - review and editing.

Sarah Ackerschewski: Resources, Methodology, Software.

Uygar Navruz: Data curation, Methodology, Software.

Nazan Sinir: Data curation, Methodology, Software.

Julian Valline: Resources, Methodology, Software.

Christof Schöch: Funding acquisition, Resources, Supervision, Writing - review and editing.

References

361

- Bhattacharyya, Sayan, Peter Organisciak, and J. Stephen Downie (2015). “A Fragmentizing Interface to a Large Corpus of Digitized Text: (Post)humanism and Non-consumptive Reading via Features”. In: *Interdisciplinary Science Reviews* 40.1, 61–77. [10.1179/0308018814Z.000000000105](https://doi.org/10.1179/0308018814Z.000000000105).
- Biderman, Stella, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Shivan-shu Purohit, and Edward Raff (n.d.). “Emergent and Predictable Memorization in Large Language Models”. In: ().
- Chang, Kent, Mackenzie Cramer, Sandeep Soni, and David Bamman (2023). “Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Association for Computational Linguistics, 7312–7327. [10.18653/v1/2023.emnlp-main.453](https://doi.org/10.18653/v1/2023.emnlp-main.453).
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *arXiv:1810.04805 [cs]*. [http://arxiv.org/abs/1810.04805](https://arxiv.org/abs/1810.04805) (visited on 05/30/2025).
- Du, Keli (2023). “Understanding the impact of three derived text formats on authorship classification with Delta”. In: [10.5281/zenodo.7715299](https://doi.org/10.5281/zenodo.7715299).
- Fu, Zihao, Wai Lam, Anthony Man-Cho So, and Bei Shi (2021). “A Theoretical Analysis of the Repetition Problem in Text Generation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.14, 12848–12856. [10.1609/aaai.v35i14.17520](https://doi.org/10.1609/aaai.v35i14.17520).

- Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi (2020). *The Curious Case of Neural Text Degeneration*. <http://arxiv.org/abs/1904.09751> (visited on 05/30/2025).
- Iacino, Gianna, Paweł Kamocki, Du Keli, Christof Schöch, Andreas Witt, Philippe Genêt, and José Calvo Tello (2025). “Legal status of Derived Text Formats”. In: (submitted).
- Jett, Jacob, Boris Capitanu, Deren Kudeki, Timothy Cole, Yuerong Hu, Peter Organisciak, Ted Underwood, Eleanor Dickson Koehl, Ryan Dubnick, and J. Stephen Downie (2020). *The HathiTrust Research Center Extracted Features Dataset (2.0)*. [10.13012/R2TE-C227](https://doi.org/10.13012/R2TE-C227).
- Kocula, Martin (2021). “Volltext vs. abgeleitetes Textformat: Systematische Evaluation der Performanz von Topic Modeling bei unterschiedlichen Textformaten mit Python”. PhD thesis. [10.5281/zenodo.5552487](https://doi.org/10.5281/zenodo.5552487).
- Kugler, Kai, Simon Münker, Johannes Höhmann, and Achim Rettinger (2023). “InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline”. In: *Conference Reader: 2nd Annual Conference of Computational Literary Studies*. [10.5281/zenodo.8093598](https://doi.org/10.5281/zenodo.8093598).
- Lee, Katherine, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini (2022). “Deduplicating Training Data Makes Language Models Better”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 8424–8445. [10.18653/v1/2022.acl-long.577](https://doi.org/10.18653/v1/2022.acl-long.577).
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. [10.48550/arXiv.1910.13461](https://arxiv.org/abs/1910.13461).
- Lin, Chin-Yew (2004). “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81. <https://aclanthology.org/W04-1013/> (visited on 05/30/2025).
- Lin, Yuri, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov (2012). “Syntactic Annotations for the Google Books NGram Corpus”. In: *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 169–174. <https://aclanthology.org/P12-3029> (visited on 05/30/2025).
- Morris, Andrew, Viktoria Maier, and Phil Green (2004). “From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition.” In:
- Organisciak, Peter and J. Stephen Downie (2021). “Research access to in-copyright texts in the humanities”. In: *Information and Knowledge Organisation in Digital Humanities*. Routledge, 157–177.
- 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*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Association for Computational Linguistics, 311–318. [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135).
- Post, Matt (2018). “A Call for Clarity in Reporting BLEU Scores”. In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, 186–191. <https://www.aclweb.org/anthology/W18-6319> (visited on 05/30/2025).

- Raffel, Colin, Noam 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: *Journal of Machine Learning Research* 21.140, 1–67. <http://jmlr.org/papers/v21/20-074.html> (visited on 05/30/2025).
- Raue, Benjamin and Christof Schöch (2020). "Zugang zu großen Textkorpora des 20. und 21. Jahrhunderts mit Hilfe abgeleiteter Textformate – Versöhnung von Urheberrecht und textbasierter Forschung". In: *RuZ - Recht und Zugang* 1.2, 118–127. [10.5771/2699-1284-2020-2-118](https://doi.org/10.5771/2699-1284-2020-2-118).
- Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke (2020). "Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen". In: *Zeitschrift für digitale Geisteswissenschaften* 5. [10.17175/2020_006](https://doi.org/10.17175/2020_006).
- Welleck, Sean, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston (2019). *Neural Text Generation with Unlikelihood Training*. [10.48550/arXiv.1908.04319](https://arxiv.org/abs/1908.04319).
- Woodard, J.P. and J.T. Nelson (1982). "An information theoretic measure of speech recognition performance". In:
- Zhang, Xinhao, Olga Seminck, and Pascal Amsili (2024). "Remember to Forget: A Study on Verbatim Memorization of Literature in Large Language Models*". In: *Proceedings of the Computational Humanities Research Conference 2024*. <https://ceur-ws.org/Vol-3834/paper96.pdf> (visited on 05/30/2025).

Computational Analysis of Literary Communities: Event-Based Social Network Study of St. Petersburg 1999-2019

Maria Levchenko¹ 

1. Department of Classical Philology and Italian Studies, University of Bologna , Bologna, Italy.

Citation

Maria Levchenko (2025). "Computational Analysis of Literary Communities: A Social Network Study of St. Petersburg 1999-2019". In: *CCLS2025 Conference Preprints 4* (1). [10.26083/tuprints-00030141](https://doi.org/10.26083/tuprints-00030141)

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-02-07

Keywords

social network analysis, cultural events, community detection, Saint Petersburg, contemporary Russian literature

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. This paper presents a computational analysis of literary networks in St. Petersburg from 1999 to 2019, using data from the SPbLitGuide newsletter and exploring cultural connections through event co-participation. By processing 15,012 cultural events with 11,777 participants in 862 venues, we reveal the structure and evolution of the literary network in post-Soviet Russia. Our methodology combines network, spatial, and temporal approaches, demonstrating how systematic event recording can capture patterns of literary community formation typically invisible to traditional literary history. The study covers the last decades of St. Petersburg's predominantly offline literary life before its digital and geopolitical disruptions, providing both a historical record and a methodological framework applicable to other cultural contexts. Our findings show a complex ecosystem characterised by dense local clusters, influential bridge figures, and distinct community boundaries, while documenting crucial shifts in the city's literary infrastructure over two decades.

1. Introduction

Literary communities can be understood through multiple analytical lenses — aesthetic movements, stylistic affiliations, publication networks, institutional memberships, translation flows, or interpretive strategies. This study examines literary community formation through the material practices and embodied experiences of literary life: event co-participation, venue selection, and the situated social interactions that constitute the lived reality of literary culture.

Cultural events are pivotal sites for both the formation of literary communities and the circulation of cultural meanings. Here individual actors coalesce into recognizable communities, and exposure to dialogue, diverse voices, aesthetic positions, and creative practices shapes personal literary development. These gatherings serve as spaces where collective memory — shared understandings of literary tradition, influential figures, and aesthetic values — is performed and transmitted. Attending particular readings, discussions, or festivals reflects not only social affiliation but also intellectual curiosity and aesthetic preferences, creating communities bound together by both personal relationships and shared creative influences.

These patterns of shared participation in readings, discussions, book launches, and festivals both reflect existing relationships and create new ones, forming complex networks

of cultural association where aesthetic alignments manifest through social interaction. 19
Yet these crucial patterns of literary life often remain invisible to historical analysis. 20

This study presents a computational framework for mapping these networks through 21
event participation, drawing on a unique dataset of cultural events in St. Petersburg from 22
1999 to 2019. By combining network analysis with spatial and temporal approaches, we 23
describe the structure of literary life as it manifests in physical spaces and evolves over 24
time, and the patterns of community formation in the cultural capital of post-Soviet 25
Russia. 26

This approach offers a distinct perspective that complements text-based analyses by 27
exploring how communities are actively constituted and sustained through patterns of 28
direct engagement in specific urban spaces and temporal rhythms. It captures ephemeral 29
interactions that leave few textual traces, maps the concrete geographies and temporal 30
rhythms of literary engagement, and brings to light the “hidden figures” — event 31
organizers, moderators, and facilitators — who function as essential nodes in literary 32
networks despite their absence from traditional publication metrics. 33

The literary ecosystem of St. Petersburg presents an optimal case study for this com- 34
putational approach to cultural network analysis. As a metropolis with a historically 35
rich tradition of literary salons and public readings, St. Petersburg has always been the 36
perfect place to explore literary communities. Our framework shows who participates 37
in literary life and how, and generates spatio-temporal mappings of cultural interaction 38
and offer new approaches to geocultural evolution. 39

Significantly, our data covers a transformative period in Russian cultural life. The years 40
1999-2019 witnessed major shifts: from Soviet-era divisions between official, unofficial 41
and émigré literature to a more integrated literary field; from purely offline interaction to 42
the use of internet tools to drive a community; and from chaotic and almost underground 43
cultural movements to an increasingly commercialised literary infrastructure. Since 44
2020, this literary ecosystem has undergone even more dramatic changes — first through 45
the forced digitisation of cultural life by the COVID pandemic, and then through the 46
profound disruption and geographical dispersion of literary networks following the 47
events of 2022. Our analysis thus preserves a detailed record of the last decades of a 48
literary world that has since been fundamentally transformed. 49

2. Network Analysis in Literary Studies 50

The computational analysis of literary networks has evolved through distinct method- 51
ological paradigms, each implementing specific algorithmic approaches to capture 52
different dimensions of literary relationships. Initial frameworks focused on three pri- 53
mary data architectures: the algorithmic extraction of character interaction networks 54
(David Elson 2010), bibliometric analysis of publication and citation patterns (So and 55
Long 2013), and the computational mapping of translation flows (Roig-Sanz and Fólica 56
2021). Moretti’s seminal work (2005) established network visualization as a foundational 57
analytical framework, subsequently expanded through contemporary investigations of 58
digital literary spaces (Basnet and Lee 2021). 59

Traditional bibliometric approaches examine co-authorship patterns and publisher affil- 60

iations to reveal formal literary relationships. Institutional data provide information on organisational memberships and collaborations, while social media analysis enables the mapping of contemporary digital literary communities. Biographical sources — including memoirs, personal documentation and travel records — provide complementary evidence for understanding historical literary networks.

Correspondence network analysis has proved particularly valuable in the study of historical literary figures. Notable projects include the [Republic of Letters](#) and the [correspondence network of early modern merchants](#). While these analyses provide valuable insights into specific literary figures and their immediate connections, there are obvious limitations to their scope.

While these approaches have significantly advanced our understanding of literary networks, we believe that the potential of network analysis extends far beyond texts, quotations, and correspondence. Cultural events — readings, discussions, festivals, and informal gatherings — represent a rich but largely untapped source of data on the formation of literary communities. These events reflect actual patterns of interaction and collaboration that often precede or exist independently of textual production. By treating event records as historical sources, we can examine how literary communities form and evolve through direct participation rather than through textual traces alone.

3. Event-Based Network Analysis

This event-based approach introduces an experimental framework for analysing literary networks, focusing on cultural events as the primary unit of interest. Here we have a possibility to observe direct social interactions as they occur in physical spaces. This direct observation reveals informal relationships and emerging communities that may never be recorded in published works or correspondence. This provides a different picture of how literary networks actually function.

While social media analysis captures casual acquaintances and declared or performative connections, co-participation in events identifies deeper conceptual and aesthetic alignments between participants. Co-participation in poetry readings, book presentations or literary discussions indicates not only physical co-presence, but also meaningful cultural collaboration or artistic affinity. Moreover, event-based analysis describes interactions across generations, including influential figures from older cohorts who have never established a digital presence. This focus on real-world cultural engagement documents both operational and aesthetic relationships, revealing how literary networks function through concrete patterns of artistic collaboration and shared cultural projects.

The event-based methodology captures a broader range of actors than traditional analyses. Beyond examining authors solely through their published works, the data reveals the organisational and curatorial activities performed by poets, writers, and other cultural actors who form literary life through event programming and community building. These figures, often invisible in traditional literary histories focused on textual production, emerge as key nodes in the network of cultural production and transmission through their dual roles as both creative practitioners and cultural mediators. They perform crucial mediating functions of gatekeeping (selecting speakers/themes), con-

necting (bringing together diverse participants), legitimizing (providing platforms for emerging voices), and framing (shaping how literary activities are perceived and categorized) (Janssen and Verboord 2015). This reveals how literary communities are sustained not only through textual creation but through the organizing labor that creates spaces for cultural exchange and collaboration.

3.1 Events as Community-Structuring Mechanisms

Cultural events serve as powerful mechanisms for structuring literary communities, creating patterns of interaction that sculpt the literary landscape. Events are not merely passive reflections of existing networks, but active sites where communities form and evolve. Each event contributes to the establishment of literary connections, while patterns of participation reveal how different groups within the literary world interact.

The spatial dynamics of literary life matter. Venues vary in their centrality to literary life, and their geographical distribution affects patterns of access and participation. Some spaces become cultural hubs through repeated use, while others remain peripheral, creating distinct patterns of literary activity across the urban landscape. For example, some venues become regular meeting places for particular literary communities, while others facilitate interaction between different groups. The cultural geography of St. Petersburg creates hierarchies of venue appeal rooted in both practical accessibility and literary memory. Historically significant venues like the Podval Brodyachey Sobaki (Stray Dog Cellar) or the Pushkin Museum at Moyka 12 carry profound cultural resonance, connecting contemporary literary events to the city's literary past and adding symbolic weight that transcends their immediate practical function. Established institutions in the historic center benefit from this layered cultural prestige alongside mainstream visibility, making them accessible to diverse audiences and facilitating broad community interaction. In contrast, peripheral venues — local district libraries, night clubs, or alternative spaces in city margins — serve as essential spaces for literary communities that exist outside the mainstream cultural hierarchy: alternative groups who deliberately reject heritage culture and institutional legitimacy, and marginalized communities (such as naive poetry groups) who are excluded from prestigious venues. These peripheral spaces provide necessary cultural territory for authentic artistic expression beyond the constraints of official literary culture. This dynamic means that venue selection reflects not just aesthetic preferences but strategic decisions about cultural legitimacy, audience reach, and connection to St. Petersburg's literary tradition.

4. Saint Petersburg's Case

Event-based approach appears particularly promising for analysing the literary scene in St. Petersburg. The city's dense network of cultural institutions, which mix traditional venues (such as the Akhmatova Museum) with alternative spaces (such as the Poryadok Slov bookshop or the city's streets) and informal meeting places (including the apartment concerts, квартирники, that continue the Soviet tradition), provides an ideal setting for studying how physical spaces affect literary life. The spatial concentration of literary activity in the historical centre, particularly along Nevsky Prospekt and in the area between the Fontanka and Moika rivers, maintains historical patterns of cultural

geography, while new literary spaces emerge in peripheral areas. 145

The complex interaction between formal and informal literary circles in St. Petersburg 146
makes it a natural case for the event-based approach. The coexistence of multiple cultural 147
venues — from established academic institutions and state libraries to independent 148
bookstores and experimental poetry bars — creates a rich field for studying how different 149
literary groups interact with the city’s environment. Event data includes large-scale 150
events at major cultural institutions and informal gatherings in alternative spaces, giving 151
a full picture of literary life at various scales and in different settings. 152

5. SPbLitGuide Dataset 153

The primary data corpus for the event-based exploration of the literary network is based 154
on the SPbLitGuide newsletter (1999-2019) announcing upcoming literary events, an 155
information bulletin that provides unprecedented longitudinal coverage of St. Peters- 156
burg’s literary ecosystem. Initiated by the philologist and poet Darya Sukhovey, this 157
chronicle project originated in the circles of experimental poetry and academic philology, 158
although its scope expanded significantly over time. 159

The evolution of the newsletter can be traced through three distinct phases. The first 160
phase established distribution through both email and web platforms (via Moscow 161
poet Alexander Levin’s website), primarily serving experimental and academic literary 162
networks. A significant expansion took place in the second phase (2010-2015) through 163
a collaboration with *DK Krupskoy*, a permanent book fair in St. Petersburg. This part- 164
nership expanded the newsletter’s coverage to include mainstream cultural events and 165
commercial venues, creating a more nuanced representation of the city’s literary life. 166

In the third phase, beginning in 2015, the newsletter’s archives and updates were 167
collected and transferred to the digital platform of the independent publishing house 168
Svoe Izdatelstvo. Over the years, thanks to Darya Sukhovey’s methodical approach, 169
the newsletter maintained weekly periodicity and systematic documentation practices, 170
resulting in a consistent and detailed record of both central and peripheral literary 171
phenomena. 172

The period from 1999 to 2019 came to an end prior to two significant disruptions: the 173
COVID-19 pandemic’s forced digitalisation of literary life and the 2022 war against 174
Ukraine’s fundamental reconfiguration of the cultural field. The latter caused a global 175
dispersal of literary actors and new ideological break-ups within the community. The 176
profound impact of these events is echoed in the newsletter’s publication pattern: after 177
February 2022, there was a one-year hiatus before publication resumed with a much 178
reduced frequency (seven issues in 2023) and a modified scope. 179

The scale of SPbLitGuide becomes clear when compared with similar projects. The 180
Moscow-based *MosLitGuide* project (2016-2020) by Anna Golubkova produced about 100 181
issues before being closed during the pandemic. The “Literary Life of Moscow” section 182
of Dmitry Kuzmin’s *Vavilon.Ru* (1997-2003, also reproduced in print) published 66 183
issues. SPbLitGuide stands out with more than 1,400 issues, consistent documentation 184
methods and wide-ranging coverage of the city’s literary life. 185

The newsletter’s explicit selection principles, as stated by the curator, demonstrate a commitment to broad and unbiased coverage from the very start. It focused on publicly accessible literary events in St. Petersburg, presenting information without aesthetic evaluation to allow readers to make their own choices. The newsletter covered contemporary literary activities, including author readings, book launches, discussions of contemporary literature, and autograph sessions. While it excluded closed writing groups, routine activities of professional unions, and purely theatrical or musical events, it did include academic conferences on contemporary authors and art exhibitions related to the current literary situation. Significantly, with the permission of the organisers, it also documented informal events such as street actions and home readings. This deliberate inclusivity suggests that while the project originated in experimental poetry circles, its documentary approach aimed to capture the full spectrum of the city’s literary landscape.

5.1 Event Entries and Role Identification

Event descriptions in the SPbLitGuide newsletter range from very brief notices to detailed multi-part announcements, but all consistently include the date, time, and place as core attributes. Addresses for all venues are typically listed at the end of each newsletter, which may include anywhere from one to thirty events per issue, depending on the season and level of cultural activity. The source of each entry — be it event organisers, venue owners, presenting authors or the curator herself — is often specified, and this variety of authorship results in significant stylistic diversity: some entries are concise and factual, while others are highly appraising or expressive. Below are two examples:

24.04.06 понедельник 19.00 Платформа
Поэтический вечер. Александр Горнон.

28.04.06 пятница 19.00 Библиотека им. Маяковского
«АЗиЯ-плюс» представляет. Юбилейный вечер к 70-летию Виктора Сосноры.
В программе вечера примут участие: Виктор Соснора, артисты Сергей Дрейден и Лев Елисеев, музыканты Евгения Логвинова и Николай Якимов, а также петербургские литераторы и издатели. Будут представлены аудиокнига с авторским чтением стихов «В. Соснора. Избранное» из серии «Голос поэта» («АЗиЯ-плюс», 2006) и книга «Куда пошёл? И где окно?» (переиздание — СПб., «Пушкинский фонд», 2006) В фойе — выставки книг, архивных фотографий и авторской графики Сосноры.

Almost every event description lists the names of active participants — such as speakers, performers, organizers, or moderators. Sometimes these roles are explicit; in other cases, they are implied by context. Alongside these, event texts may mention other individuals: as part of an organization’s name, as the subject of commemoration, or in promotional contexts highlighting connections with well-known figures. Although references to absent or associated figures can emphasise broader cultural connections, our analysis focuses on actual participation. Hence, we only extract the names of individuals who were directly involved in the events, as these represent veritable social connections within the literary community.

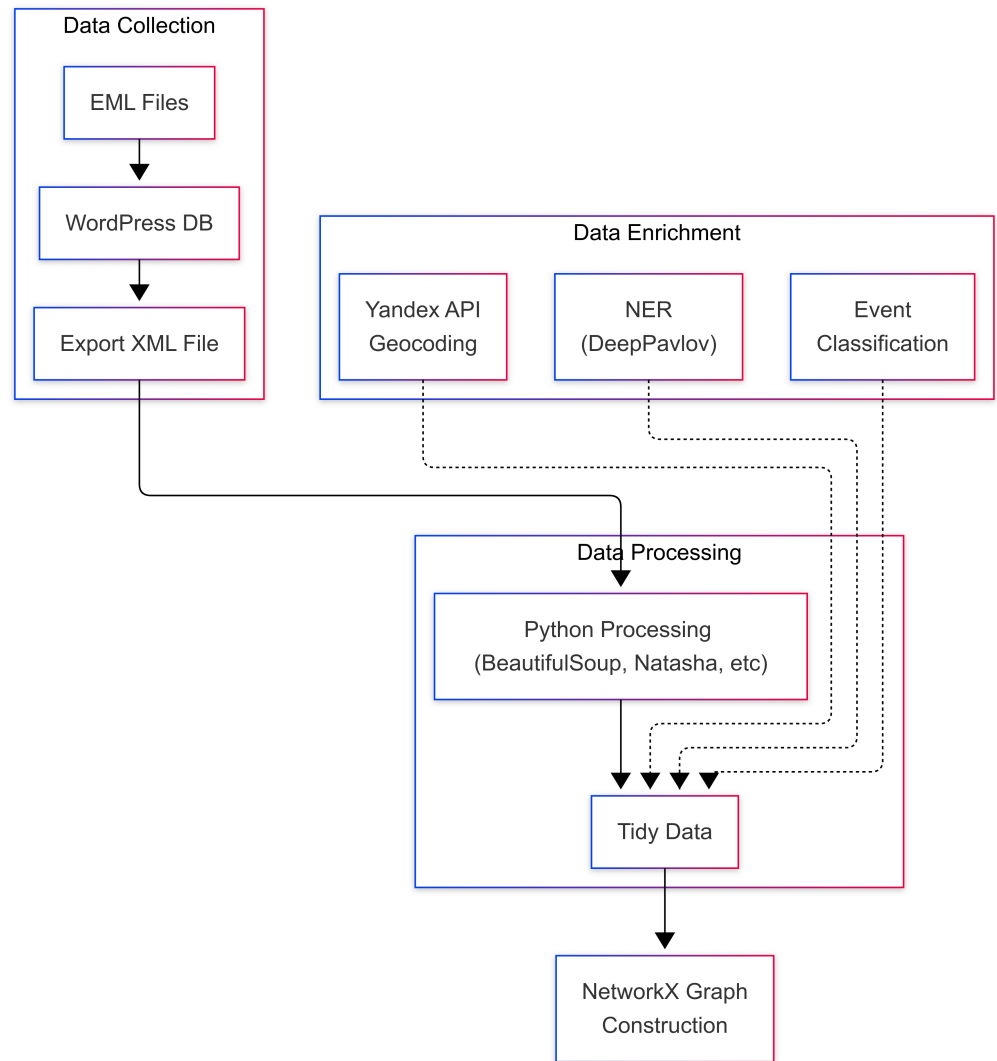


Figure 1: Data Processing Pipeline

6. Data processing pipeline

In 2015, during the migration of the newsletter to the *Svoe Izdatelstvo* platform, the entire archive of previous letters was collected from the mailboxes of the maintainer and her friends, which formed the basis for the creation of the dataset. Since then, all new issues have been published through the same database, providing a secure and complete text corpus. The transformation of raw digital born data into a structured analytical dataset required the design and implementation of a multi-stage processing architecture (shown in Figure 1).

The pipeline begins with source data collection, where primary data is preserved in electronic mail (EML) format, preserving original message structures and metadata integrity. This initial corpus is then systematically converted into a structured database format within the WordPress environment, providing a stable storage layer with XML export functionality for future processing operations.

The primary processing layer uses several Python tools to extract and structure the raw data. BeautifulSoup facilitates HTML parsing, while the Natasha library provides a

specific processing feature for Russian language content. String matching operations are handled by the difflib library, complemented by regular expression processing for content extraction.

After initial processing, data is normalised to achieve consistency and compatibility. This stage standardises the extracted information and implements uniform data structures in preparation for the analysis stage. Geographical enrichment follows, using the Yandex API for coordinate extraction and address standardisation, enabling precise spatial mapping of literary events across St. Petersburg.

The entity recognition layer is a critical component of the processing architecture. Building on the systematic evaluation of NER models for Russian cultural texts (Levchenko 2024a), a multi-stage automated pipeline with final manual validation was implemented. This stage used DeepPavlov's multilingual BERT model for named entity recognition, followed by a post-processing step to handle Russian grammatical forms, different writing styles, patronymics and institution names.

The automated pipeline continued with entity enrichment, where identified entities were automatically mapped to VIAF and Wikidata identifiers using their respective APIs. This automated enrichment process significantly improved the interoperability of the dataset with other cultural heritage resources. The entire dataset was then manually validated as a final quality control step, verifying both the entity recognition results and the automated identifier assignments.

The final stage focuses on network analysis, using NetworkX for graph construction and implementing community detection algorithms. This layer enables the computation of various network metrics, providing the analytical basis for understanding the structure and evolution of the St. Petersburg literary communities.

The execution of this pipeline has produced significant results, successfully processing 15,012 discrete event instances and identifying 11,777 normalised attendee entities. The pipeline has also mapped 862 venue nodes to 817 unique geospatial coordinates and documented over 100,000 attendance records.

Yet, processing the SPbLitGuide dataset presented several significant procedural challenges, particularly in the areas of entity recognition and normalisation. Three main categories of challenges arise during the data processing implementation.

First, the complexity of name variations caused a significant difficulty for entity recognition. The dataset contained multiple representations of the same individual across different events and time periods. For example, a single author could appear as both a patronymic and diminutive full name, or with different combinations of initials and surnames. This complexity was multiplied by the diverse cultural origins of the names in the dataset, ranging from Russian and post-Soviet to European and Asian naming conventions. The literary nature of the dataset also introduced different formatting conventions, including the use of pseudonyms, artistic names and alternative spellings.

Second, contextual ambiguity created significant issues for accurate entity resolution. Names often appeared in multiple roles within event descriptions - as organisers, participants or referenced authors - requiring careful disambiguation. The dataset often contained references to historical figures alongside contemporary participants, requiring

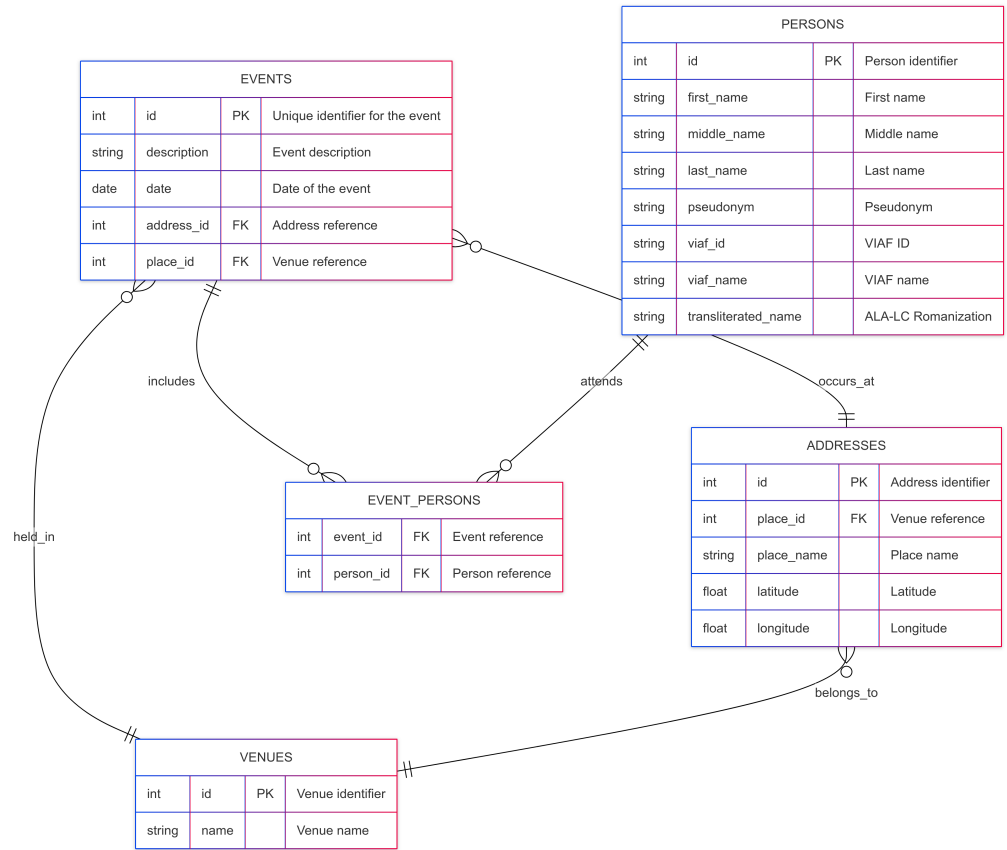


Figure 2: Entity-relationship diagram of the relational structure of the SPbLitGuide dataset

a distinction to be made between actual event participants and mentioned personalities. 286
 This complexity was particularly evident in events such as literary commemorations or 287
 academic conferences, where historical figures were often referenced but not present. 288

Thirdly, the mixed use of formal and informal name presentations required additional 289
 attention. The integration of multilingual content, particularly for international events 290
 or cross-cultural literary gatherings, added another layer of complexity to the processing 291
 pipeline. 292

The processing combined DeepPavlov/Natasha libraries for initial normalization, Lev- 293
 enshtein distance calculations to merge name variants of the same individuals, and 294
 context-based analysis of event descriptions to distinguish different persons with similar 295
 names, with manual validation of all suggestions. The resulting dataset implements a 296
 relational structure optimised for network analysis and spatio-temporal queries (Figure 297
 ??). The data model comprises five core entities: events serve as the central unit linking 298
 persons, venues, addresses, and participation records. This architecture enables diverse 299
 analytical queries: tracking individual activity across communities and venues, map- 300
 ping geographical clustering of communities, analysing temporal patterns in event types 301
 and participation, identifying bridge and key figures, and measuring spatial evolution 302
 of literary activity. The full technical specification and dataset are available via Zenodo 303
 (Levchenko 2024b). 304

7. Network construction methodology

Using the resulting dataset with the list of participants extracted from the event description, we construct an undirected weighted graph based on event co-participation, operating on the premise that shared event attendance indicates social interaction and cultural connection between literary actors. Nodes represent individual participants, while edges represent co-participation in events.

To account for event size differences, we implement a normalisation strategy that reflects the intuition that interactions in smaller gatherings are likely more significant than those in larger events. For each event, if there are n participants, every participant can potentially interact with $(n-1)$ other participants. Therefore, we assign a weight of $1/(n-1)$ to each pair of participants in that event. For example, in a small reading with 3 participants, each pair receives a weight of $1/2$, while in a large festival panel with 10 participants, each pair receives a weight of $1/9$. When participants co-occur in multiple events, their edge weight is the sum of these normalized interaction weights across all shared events.

The complete network consists of 10,656 nodes connected by 106,127 edges, showing a distinct core-periphery structure with 387 separate connected components. The largest connected component contains 9,621 nodes (90% of the participants), representing the core of the active St Petersburg literary community. This main component has a high clustering coefficient (0.753), indicating strong local group formation, with an average shortest path length of 3.702 and a network diameter of 13. The low network density (0.002) and skewed degree distribution (mean: 19.92, median: 8) reveal a selective and hierarchical structure, where a small number of participants maintain extensive connections while most operate in smaller networks.

This network structure exhibits classic “small world” characteristics, combining high local clustering with efficient global connectivity. In particular, the clustering coefficient of our network (0.753) exceeds those found in Broadway musical collaboration networks (0.41, Uzzi and Spiro 2005) and scientific collaboration networks (0.45, M.E.J. Newman 2001), suggesting that literary communities in St. Petersburg form particularly tight local groups. However, this strong local clustering exists alongside multiple unconnected components, reflecting a literary field that combines intense local collaboration with distinct subcommunities.

7.1 Community detection and basic structure

Application of the Louvain community detection algorithm (resolution 1.0) has identified 49 distinct communities within the main component, demonstrating the complex segmentation of the St. Petersburg literary world. These communities show clear differences in size and patterns of activity, with several large groups emerging as particularly significant (see Table 1).

The analysis of the largest detected communities in the St. Petersburg literary network finds remarkably similar structural characteristics despite differences in size. While the communities range from 584 to 1363 members, they maintain comparable internal structural metrics: clustering coefficients fall within a narrow range (0.697-0.781) and



Figure 3: Network visualization of the largest connected component ($N = 9,621$ nodes). Communities identified by modularity optimization are shown in different colors. Edge weights ≥ 13 displayed. Layout: OpenOrd algorithm.

ID	Size	Events/Year	Clustering	Internal Density	External Connections	Key Figures
1	1363	73.55	0.772	0.014	7164	Vladimir Antipenko Maria Agapova Ilya Zhigunov
5	1286	120.98	0.777	0.024	16603	Darya Sukhovey Arsen Mirzaev Dmitry Grigoriev
3	911	85.32	0.729	0.010	5180	Yakov Gordin Andrey Arieu Alexander Kushner
0	866	60.44	0.781	0.014	6633	Alexander Skidan Pavel Arseniev Arkady Dragomoshchenko
4	605	33.83	0.778	0.022	3107	Ivan Pinzhenin Roma Gonza Andrey Nekrasov
7	590	70.27	0.741	0.027	8014	Evgeny Myakishev Evgeny Antipov Galina Ilyukhina
17	584	70.71	0.697	0.014	4244	Pavel Krusanov Sergey Nosov Alexander Sekatsky

Table 1: Major Literary Communities in St. Petersburg (1999-2019): Size, Activity, Network Metrics, and Key Figures (sorted by community size). Key figures identified by highest degree centrality within each community, representing the most connected participants

conference version

internal densities are consistently low (0.010-0.024). 347

The most notable quantitative difference is in external connections, where Community 348
5 has a much higher connectivity (16,603 external connections) than the other commu- 349
nities (ranging from 4,244 to 8,014). However, this difference in external connections 350
does not correspond to substantial differences in internal structure, as evidenced by the 351
similar clustering and density values. 352

The consistency of these network metrics across communities of different sizes sug- 353
gests that literary groups in St. Petersburg tend to develop similar patterns of internal 354
organisation, regardless of their size or central figures. The Louvain algorithm suc- 355
cessfully identified stable groupings, but their structural similarities suggest that these 356
communities, while distinct, follow comparable patterns of connection and interaction. 357

7.2 Aesthetic Validation of Detected Communities 358

The communities identified through event co-participation by the Louvain algorithm 359
could be qualitatively examined to see if they correlate with known aesthetic groupings 360
or stylistic schools within the St. Petersburg literary scene: as the physical manifestations 361
or activations of these latent, often text-centered, communities of interest, interpretation, 362
and affective connection. We have an opportunity to explore whether these structural 363
cleavages correlate with distinct aesthetic schools, ideological stances, or institutional af- 364
filiations that actively maintain boundaries and limit interaction with "outside" groups. 365
For instance, do traditionalist poets, who might cluster in one computationally detected 366
community, consciously avoid (or remain uninvited) to events dominated by experi- 367
mental poets, who cluster in another? Such dynamics would suggest that the network 368
structure reflects not just passive preference but active processes of distinction and 369
boundary maintenance driven by aesthetic or ideological commitments. 370

Event co-participation forms our empirical basis: if two writers frequently appear at 371
the same readings or panels, we infer a latent affinity. Yet an "aesthetic community" 372
implies deeper commonalities — shared poetics, interpretive frameworks, thematic 373
preoccupations — publicly enacted and negotiated at literary gatherings. Because events 374
serve as sites where aesthetics are performed, debated, and transmitted, we can test 375
whether attendance patterns indeed serve as reliable proxies for these richer, affective 376
connections. 377

Below, we demonstrate three communities identified in Table 1 that map convincingly 378
onto established aesthetic schools, institutional affiliations, and critical networks docu- 379
mented in prior scholarship. 380

Community 0 (Experimental/Avant-Garde Poetry). Key figures: Alexander Skidan, 381
Pavel Arseniev, Arkady Dragomoshchenko (also Dmitry Golynko-Volfson, Roman Os- 382
minkin, Galina Rymbu, Natalia Fedorova). 383

This cluster precisely maps onto what Bozović terms the *Translit* avant-garde circle — a 384
cohesive literary formation with explicit institutional structures, shared experimental 385
poetics, and collective political commitments (Bozović 2023). The group centers on 386
the *Translit* almanac, where Arseniev serves as co-editor and Skidan sits on the advi- 387
sory board, creating both institutional coherence and collaborative initiatives like the 388
"Laboratory of Poetic Actionism" (Bozović 2023; Platt 2017). Their aesthetic program 389

unites around experimental strategies that synthesize 1920s avant-garde traditions (LEF, Russian Formalism) with contemporary critical theory. Skidan's collage-based, deconstructive poetics and Dragomoshchenko's "quantum" ideogrammatic experiments represent sophisticated engagements with language poetry and conceptual art practices (Hock 2021; Orlitskiy 2017). Critical recognition confirms their status as a named avant-garde circle with shared poetics, political commitments, and institutional structures (Bozović 2023). Multiple scholars treat them as a cohesive unit rather than loose affiliations, validating the computational detection of their network boundaries (Hock 2021; Platt 2017; Vivaldi 2019).

Community 3 (Literary Traditionalism "Thick Journals"). Key figures: Yakov Gordin, Andrey Arieiev, Alexander Kushner (also Valery Popov, Samuil Lurie, Natalia Sokolovskaya, Daniil Granin).

This cluster corresponds to St. Petersburg's established intelligentsia tradition, epitomized by the "thick journal" model — particularly *Zvezda* and *Neva*, and structuring discourse around continuity with Russia's literary past. Yakov Gordin (historian, writer) and Andrey Arieiev (literary scholar, critic, prose writer) have served as co-editors-in-chief of *Zvezda* since 1992. Within Bourdieu's framework (Bourdieu 1983), they occupy a segment of the field where cultural capital derives from custodianship of tradition rather than avant-garde innovation. The community's defining mindset centers on cultural stewardship and historical consciousness. Rather than pursuing formal experimentation, they embrace what might be termed a "guardianship mentality" — viewing themselves as thoughtful preservers and reinterpreters of Russia's literary inheritance. This orientation manifests in their commitment to neo-classical aesthetics, particularly evident in Alexander Kushner's Neo-Acmeist poetics, which deliberately emphasizes clarity and cultural continuity over radical innovation (Ar'ev 2019).

Community 17 ("New Prose" Petersburg Fundamentalists). Key figures: Pavel Krusanov, Sergey Nosov, Alexander Sekatsky (also Tatiana Moskvina, Viktor Toporov, Andrey Astvatsaturov, Nikolai Yakimchuk, Ilya Boyashov).

This group epitomizes the so-called "new prose" movement, often labeled *the Petersburg Fundamentalists*. Krusanov and Nosov's novels — published by *Amfora* and *Limbus Press* — exemplify an "imperial novel" aesthetic, fusing patriotic or nationalist discourses with mythological motifs and postmodern irony (Fenghi 2023). Sekatsky's philosophical writings (e.g., *The Mogs and Their Might*) provide the group's conservative-esoteric underpinnings (Fenghi 2023). Their work frequently acts as a reaction against 1990s postmodern nihilism, seeking a new cultural myth rooted in neo-Eurasianist and occultist subcultures (Lipovetsky 2008; Noordenbos 2011). Critical recognition confirms their conscious self-definition as a literary circle, with manifestos, public performances, and dedicated institutional support (Fenghi 2023; Noordenbos 2011).

The remarkable alignment between algorithmically detected communities and published accounts of St. Petersburg's literary factions confirms that event co-participation reliably indexes deeper aesthetic affinities and institutional ties.

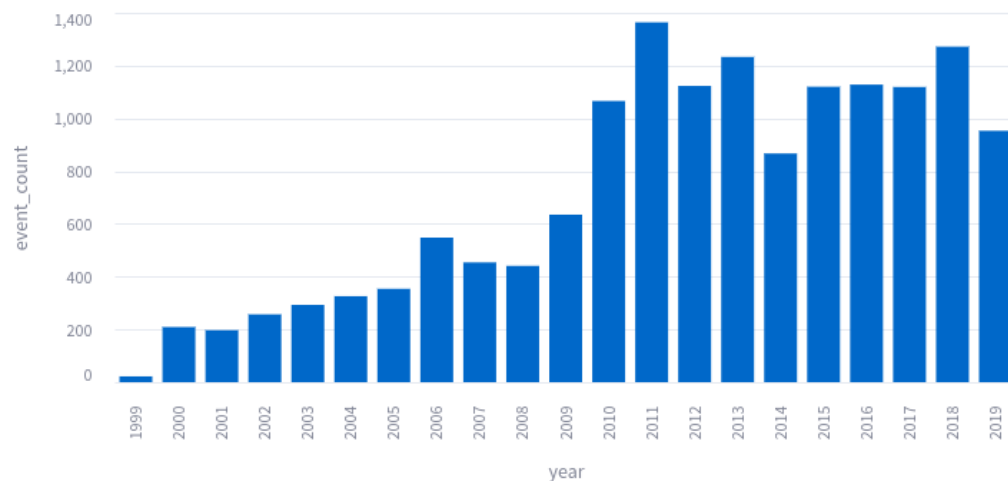


Figure 4: Annual Event Frequency: the total number of events that occurred each year

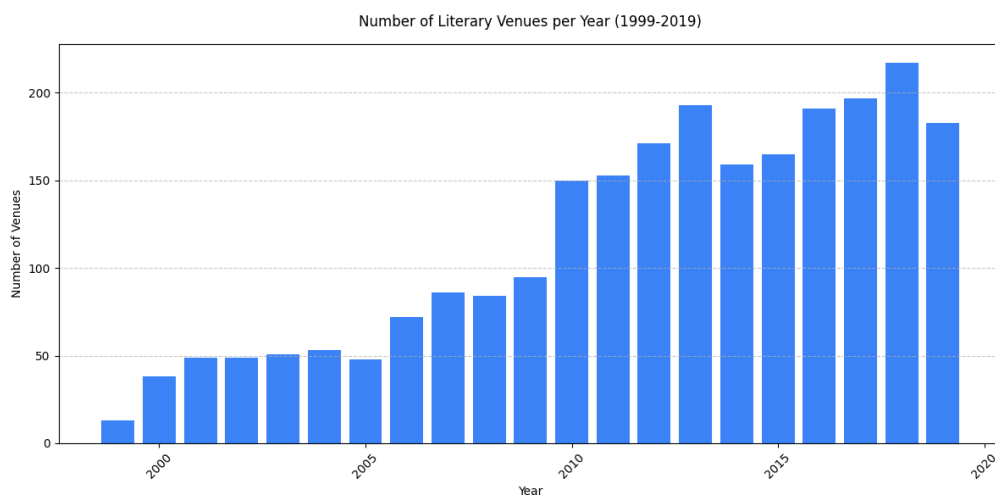


Figure 5: Number of Active Venues

8. Temporal Evolution

431

The evolution of St. Petersburg's literary landscape reflects the wider post-Soviet cultural transformation. The launch of SPbLitGuide in 1999 coincided with - and helped to document — a crucial moment when the city's literary scene was being fundamentally reshaped. This period marked the inclusion of previously unofficial literary trends into public visibility, alongside the rise of new independent venues and voices. The increase in the number of documented venues from 1999 to the following years reflects not only improved documentation, but also the formation of a new literary infrastructure that bridged Soviet underground traditions with post-Soviet cultural energies.

439

The data then show two subsequent major shifts. The first occurred around 2010 and was marked by dramatic growth in both events and venues (Figures 4-5). The number of active venues increased from 95 to 150, reflecting both the increased coverage following SPbLitGuide's collaboration with *DK Krupskoy* and the actual expansion of the literary scene, particularly with the development of commercial venues such as the *Bookvoed* network.

445

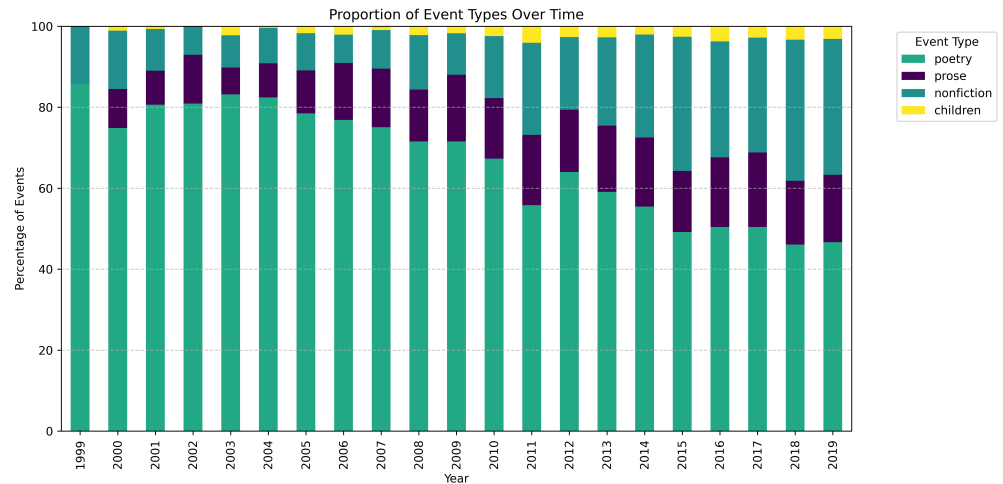


Figure 6: Proportion of Event Types Over Time

A second shift occurred in 2014, when the economic crisis following geopolitical events had a significant impact on the cultural infrastructure. The sharp decline in the number of venues (from 193 in 2013 to 159 in 2014) particularly affected independent spaces, which were more vulnerable to economic pressures.

The post-2014 period shows a pattern of resilience and adaptation. While the number of venues fluctuated between 159 and 217, the literary scene maintained a significantly higher baseline than in the pre-2010 period. This resilience suggests that the diversification of literary spaces achieved in the early 2010s created a more solid cultural ecosystem. Traditional institutions provided stability, while surviving independent venues and commercial spaces continued to support diverse forms of literary activity despite economic challenges.

Another perspective on the evolution of St Petersburg's literary landscape is provided by the AI-based classification of event types. Event descriptions were automatically classified using OpenAI's language model (o3-mini) with a predefined taxonomy of 21 tags covering event formats, genres, and characteristics. Each event was assigned up to 4 relevant tags through structured prompts (classification process used OpenAI's batch API with JSON schema validation to ensure consistent output format). The stacked bar chart (Figure 6) focuses on four primary content categories: poetry, prose, nonfiction, and children's literature events, illustrating the proportional distribution of these core literary content types over time.

While St Petersburg has always been a poetry city, the graph shows that since 2010, poetry's relative share of events has decreased as the literary scene diversified. This shift reflects not a decline in poetry activities, which remained relatively stable in absolute numbers, but rather significant growth in prose and nonfiction events. The increasing prominence of non-fiction events may indicate a move towards analytical, journalistic and educational discourses within the literary community, in line with wider cultural and intellectual developments in Russia during the 2010s.

The monthly distribution of events (Figure 7) shows consistent seasonal rhythms in St. Petersburg's literary life: activity peaks in the spring (March-May) and autumn

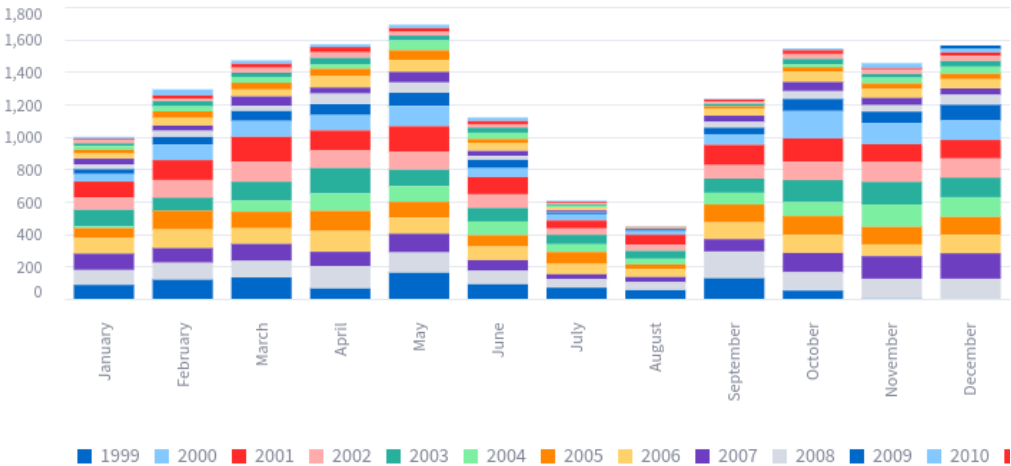


Figure 7: Monthly Event Frequency Over the Years

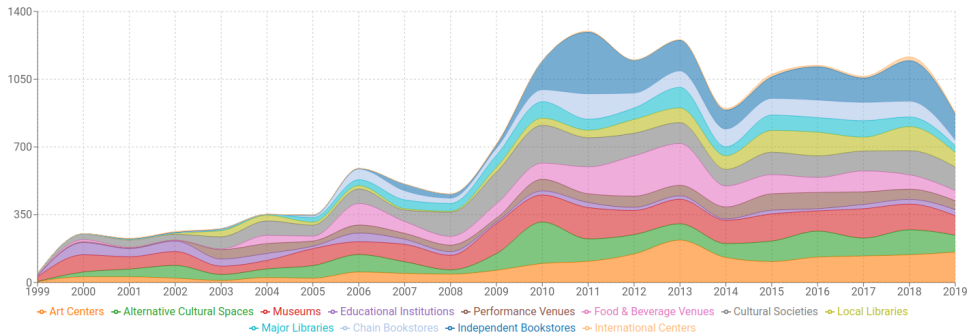


Figure 8: Annual Distribution of Literary Events by Venue Type in Saint Petersburg, 1999-2019 (absolute numbers)

(October-December), with a significant decline in the summer months (July-August). This pattern, which lasted throughout the study period, reflects both institutional calendars and established cultural traditions. Even as the literary scene expanded and diversified after 2010, it maintained these characteristic seasonal fluctuations.

The variation in venue types (Figure 8-9) highlights significant shifts in the spatial organisation of literary life in St. Petersburg from 1999 to 2019. The most striking change occurred around 2010, marked by the dramatic rise of independent bookstores (shown in dark blue) as cultural spaces. This growth coincided with broader changes in the commercial book trade, but represented a distinct phenomenon: indie bookstores weren't just commercial spaces aimed primarily at the reading public, but became active cultural centres, hosting literary events that were important for literary development and bringing together key figures from the city's literary landscape.

Another notable trend is the steady growth of art centers (orange) and alternative cultural spaces (green) throughout the 2000s, which provided flexible venues for literary events outside of traditional institutional frameworks. This diversification of venue types suggests a diversification of literary space away from the Soviet-era model, where literature was primarily housed in official cultural institutions or privately.

The data also show the resilience of traditional venues such as museums (red) and

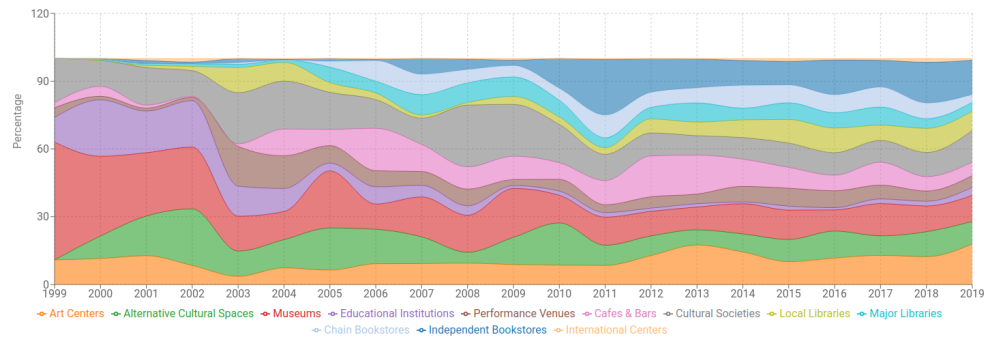


Figure 9: Relative Distribution of Literary Events by Venue Type in Saint Petersburg, 1999-2019 (percentage of total events per year)

educational/academic institutions (purple), which maintained a consistent presence throughout the period. However, their relative share of the overall venue landscape declined as new types of spaces emerged. The growth of cafes and bars (pink) as literary venues, particularly after 2010, indicates another significant shift: the integration of literary events into unconventional settings.

The period after 2014 shows interesting adaptations to economic pressures. While there was some fluctuation in the total number of events, the diversity of venue types remained relatively stable, suggesting that the literary scene had developed sound networks across different types of spaces.

9. Spatial Evolution

The spatial dimension of literary events displays the concentration of literary life across St. Petersburg's urban landscape. As shown in Figure 10, the most intense literary activity is located in the historical centre, particularly in the area bounded by the Fontanka River and Nevsky Prospekt. This core zone has the highest density of events, with notable hotspots around major cultural institutions such as the Akhmatova Museum and the Mayakovsky Library.

However, this aggregate view masks significant venue specialization and community-specific spatial preferences. Literary venues in St. Petersburg operate along a spectrum from generalist to highly specialized spaces. Generalist venues such as major bookstore chains (Bukvoed network) and large cultural institutions (Mayakovsky Library) host diverse events across different literary communities and genres. In contrast, culturally engaged venues develop strong aesthetic affiliations: independent bookshops like Poryadok Slova become closely associated with experimental literature and cultural studies communities, while alternative spaces like Fish Fabrique Nouvelle cater to underground and performance-based literary activities.

Different literary communities exhibit distinct geographical preferences, as illustrated by the comparative analysis of Communities 0 and 4 (Figure 11). Community 0 (experimental poetry, centred around Alexander Skidan and Pavel Arseniev) demonstrates concentrated activity in the historical centre, with strong clustering around the Poryadok Slova and Andrey Belyj centres. It also includes street events on the Neva embankment and post-industrial spaces such as old marine ports, reflecting their preference for es-



Figure 10: Heat Map of Event Frequency at Various Locations in Saint-Petersburg

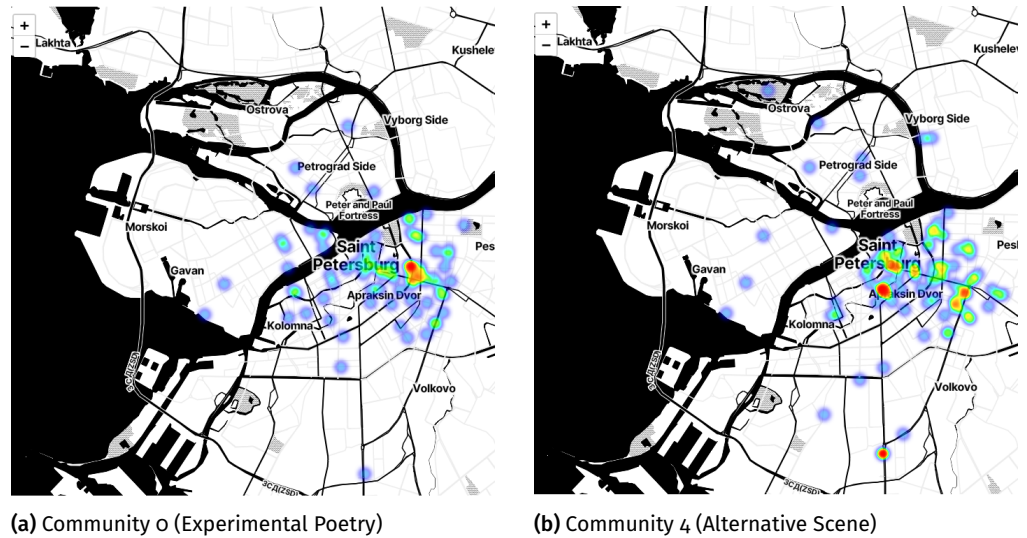


Figure 11: Spatial distribution comparison showing distinct venue preferences and geographical patterns between communities

established alternative cultural spaces combined with experimental urban interventions. 524
In contrast, Community 4 (the younger alternative scene, led by Ivan Pinzhenin and 525
Roma Gonza) exhibits a more dispersed pattern, extending into peripheral areas and 526
utilising unconventional venues such as bars and nightclubs. 527

The spatial data also reveals individual literary careers trajectories through venue tran- 528
sitions. Prose authors like German Sadulaev, Andrey Astvatsaturov, and Ilya Stogoff 529
demonstrate a characteristic migration pattern from independent alternative spaces 530
(Platform, Fish Fabrique Nouvelle) to mainstream commercial venues (Dom Knigi, 531
Bukvoed). This spatial mobility reflects not only literary success and increased reader- 532
ship, but also the evolution of authors' relationships with different literary communities 533
and their integration into broader cultural institutions. 534

This venue-community co-evolution demonstrates how literary groups actively reshape 535
the cultural geography of the city, while individual careers create bridges between 536
different spatial and social literary worlds. 537

10. Conclusion 538

Network structure and community formation. The St Petersburg literary ecosystem 539
is characterised by dense local clusters with strategic connections. The high clustering 540
coefficient (0.753) suggests that literary activity takes place primarily within established 541
communities, while the presence of influential bridging figures enables cross-community 542
exchange. The hierarchical structure of the network is reflected in the skewed degree 543
distribution, with an average of 19.92 connections but a median of only 8. This disparity 544
suggests that while most participants operate in relatively small circles, certain key 545
figures maintain extensive connections across the literary landscape, acting as crucial 546
nodes for information flow and community bridging. Betweenness centrality analysis 547
confirms the strategic importance of these bridge figures: while the network mean 548
is 0.0003, key intermediaries show dramatically higher values, with Арсен Мирзаев 549
(0.0399), Дмитрий Григорьев (0.0386), and Дарья Суховой (0.0365) emerging as the 550

most critical bridges. These figures, concentrated in Community 5, facilitate the strongest inter-community connections in the network, particularly the extensive links between Communities 0, 5, 7, and 11. The existence of 387 separate components in the network depicts a literary world composed of distinct subcommunities with limited interaction, suggesting that despite the presence of bridge figures, significant barriers to cross-community interaction remain.

Spatial and temporal dynamics. The growth from 13 venues in 1999 to 217 in 2019 represents a massive expansion of cultural infrastructure, even if the trajectory was not linear. A significant decline after 2014 particularly affected independent spaces, while the emergence of commercial venues such as the Bookvoed bookshop chain introduced new patterns of literary participation. Geographically, venues remained concentrated in the historical centre of St. Petersburg, maintaining traditional cultural patterns, while, after 2010, new literary spaces emerged in peripheral areas. Throughout these changes, certain venues, such as Poryadok Slova and the Akhmatova Museum, maintained their positions as community anchors, providing stability in the evolving literary landscape.

Historical transitions. The dataset covers three distinct periods in St. Petersburg's literary evolution. The post-Soviet transformation (1999-2009) saw the integration of formerly unofficial literary trends into public visibility, alongside the emergence of new independent venues and the establishment of regular event cycles. This was followed by a period of commercial expansion (2010-2013), marked by dramatic growth in both events and venues, particularly through the entry of commercial bookstore chains and the diversification of event types. The final period (2014-2019) reflects economic adaptation, characterised by a decline in independent venues, while established institutions have shown resilience and literary events have shifted towards more commercially viable formats. Each period represents not just changes in infrastructure, but fundamental shifts in how literary life is organised and sustained. Significantly, the dataset documents the last major phase of predominantly offline literary activity in St. Petersburg before the dramatic disruptions of 2020-2022. This makes the dataset particularly valuable as a record of literary practices and community structures that have since undergone radical transformation.

Methodological Implications and Limitations. The potential of event-based network analysis for understanding literary communities also has important methodological limitations. It can't capture audience information, and we can only analyse the active participants in literary events, not their full social impact. And our method of network construction, which gives equal weight to all instances of co-participation, may oversimplify the complex nature of literary relationships and interactions, whether those interactions take place in formal institutions or informal settings.

The data collection process itself reflects interesting network dynamics. While SPbLitGuide maintainer Darya Sukhovey personally documented many events, her high centrality in our network analysis (0.37) indicates her position as a trusted information hub. Event organisers actively submitted announcements to the newsletter, recognising its role as a key communication channel for the literary community. This organic flow of information suggests that while the dataset may have initially been selection biased due to its origins, it evolved to capture a broader range of literary activities as the newsletter became an established cultural institution.

Future directions. Similar event-based data may exist for other cities and historical periods, from pre-revolutionary literary chronicles to contemporary cultural news sites. In Russian literary studies alone, several publications document early 20th-century literary gatherings in detail comparable to the dataset (Galushkin 2006; Lavrov 2002, 2017). This methodological approach could be applied to the analysis of such historical records, allowing a systematic comparison of literary community structures across periods and locations.

One particularly promising approach is to combine event-based analysis with textual and publication data in order to create comprehensive models of literary community formation. While our event networks capture patterns of social interaction and collaboration, they represent only one dimension of literary relationships. Future research could integrate publication networks (e.g. co-authorship, citation patterns and publisher affiliations), textual influence networks (e.g. intertextuality, stylistic borrowing and translation flows) and institutional networks (e.g. journal editorships, prize committees and academic affiliations) with event participation data. This multi-layered approach would address fundamental questions about how social literary life corresponds to textual production. Do communities that frequently gather together also influence each other's writing? How do patterns of co-participation in events correlate with citation networks, collaborative publications or shared aesthetic preferences? Developing new computational methods to link social and textual data would be required for such integration, but it could further investigate whether the communities we identify through events represent real artistic movements or primarily social phenomena.

A uniquely comprehensive dataset of literary events can illuminate community structures across multiple analytical dimensions. By systematically documenting over 15,000 events between 1999 and 2019, the SPbLitGuide newsletter allows us to combine network, spatial, and temporal approaches to understand literary life in detail. This integrated analysis helps to visualise patterns of community formation and evolution. The dataset's rich documentation of literary life in St. Petersburg before 2019 preserves an original historical record of cultural practices that have since undergone radical change. Combining these different aspects of analysis opens up new possibilities for understanding how cultural communities function and evolve, and provides a framework that could be productively applied to similar historical records from other times and places.

11. Data Availability 628

Data can be found here: <https://zenodo.org/records/13753154> 629

12. Software Availability 630

Software can be found here: https://github.com/mary-lev/literary_communities 631

References 632

Ar'ev, Andrei (2019). "At a Poem's Distance (The Poetry of Aleksandr Kushner)". In: *Russian Studies in Literature* 55.1, 8–50. [10.1080/10611975.2019.1622957](https://doi.org/10.1080/10611975.2019.1622957).



- Basnet, Ankit and James Jaehoon Lee (2021). "A Network Analysis of Postwar American Poetry in the Age of Digital Audio Archives." In: *Journal of Cultural Analytics* 6(2), 180–233. [10.22148/001c.22223](https://doi.org/10.22148/001c.22223).
- Bourdieu, Pierre (1983). "The Field of Cultural Production, or: The Economic World Reversed". In: *Poetics* 12.4–5, 311–356.
- Bozović, Marijeta (2023). *Avant-Garde Post—: Radical Poetics after the Soviet Union*. Cambridge, MA: Harvard University Press.
- David Elson Nicholas Dames, Kathleen McKeown (2010). "Extracting social networks from literary fiction". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 138–147.
- Fenghi, Fabrizio (2023). "The Absolute Elsewhere: Pavel Krusanov and the Countercultural Sources of Russian Imperialism". In: *Ab Imperio* 3, 255–290.
- Galushkin, Alexander Yu. (2006). *Literaturnaya zhizn' Rossii 1920-kh godov. Sobytiya. Otkrytye sovremennikov. Bibliografiya. Moskva i Petrograd. 1917-1920 gg.* Ed. by Alexander Yu. Galushkin. Vol. 1.1. Moscow: IMLI RAN, 768.
- Hock, David (2021). "From a Space out of Time: Russian Poetry and Aesthetic Ideology after the Soviet Union". PhD dissertation. Princeton, NJ: Princeton University.
- Janssen, Susanne and Marc Verboord (2015). "Cultural Mediators and Gatekeepers". In: *International Encyclopedia of the Social & Behavioral Sciences*. Ed. by James D. Wright. 2nd. Vol. 5. Oxford: Elsevier, 440–446. [10.1016/B978-0-08-097086-8.10424-6](https://doi.org/10.1016/B978-0-08-097086-8.10424-6).
- Lavrov, Alexander V. (2002). *Letopis' literaturnykh sobytii v Rossii kontsa XIX - nachala XX v. (1891 - oktyabr' 1917). 1891-1900.* Ed. by Alexander V. Lavrov. Vol. 1. Moscow: IMLI RAN, 526.
- (2017). *Letopis' literaturnykh sobytii v Rossii kontsa XIX - nachala XX v. (1891 - oktyabr' 1917). 1901-1904.* Ed. by Alexander V. Lavrov. Vol. 2.1. Moscow: IMLI RAN, 528.
- Levchenko, Maria (2024a). *Evaluation of Named Entity Recognition Models for Russian News Texts in the Cultural Domain*. <https://github.com/mary-lev/NER>. Accessed: 2024-06-01.
- (2024b). *Literary Events in Saint Petersburg (1999-2019) from SPbLitGuide Newsletters*. Zenodo. [10.5281/zenodo.13753154](https://doi.org/10.5281/zenodo.13753154).
- Lipovetsky, Mark (2008). "Paralogii: Transformatsii (post)modernistskogo diskursa v russkoi kul'ture 1920-kh—2000-kh godov". In: *Novoe Literaturnoe Obozrenie* 140, 98–124.
- M.E.J.Newman (2001). "Collaboration and Creativity: The Small World Problem". In: *Proc. Natl. Acad. Sci. U.S.A.* 98(2), 404–409. [10.1073/pnas.98.2.404](https://doi.org/10.1073/pnas.98.2.404).
- Noordenbos, Boris (2011). "Ironie Imperialism: How Russian Patriots Are Reclaiming Postmodernism". In: *Studies in East European Thought* 63.2, 143–163.
- Orlitskiy, Yury (2017). "The Characteristics of Arkady Dragomoshchenko's "Quantum" Writing". In: *Novoe Literaturnoe Obozrenie* 145.3, 221–245.
- Platt, Kevin M. F. (2017). "Fire in the Head: Pavel Arseniev, Aesthetic Autonomy, and the Laboratory of Poetic Actionism". In: *Novoe Literaturnoe Obozrenie* 145.3, 37–64.
- Roig-Sanz, Diana and Laura Fólica (2021). "Big translation history: Data science applied to translated literature in the Spanish-speaking world, 1898–1945". In: *Translation Spaces* 10, 231–259. [10.1075/ts.21012.roi](https://doi.org/10.1075/ts.21012.roi).
- So, Richard Jean and Hoyt Long (2013). "Network analysis and the sociology of modernism". In: *boundary* 2, 147–182. [10.1215/01903659-2151839](https://doi.org/10.1215/01903659-2151839).

- Uzzi, Brian and Jarrett Spiro (2005). "Collaboration and Creativity: The Small World Problem". In: *American Journal of Sociology* 111(2), 447–504. [10.1086/432782](https://doi.org/10.1086/432782).
- Vivaldi, Giuliano (2019). "You Cannot Even Imagine Us". In: *Tribune Magazine*. May 21, 2019 issue.

From Readers to Data

Uncertainty in Computational Literary Citizen Science

Gilad Aviel Jacobson¹ 
 Itay Marienberg-Milikowsky² 
 Yael Dekel² 

1. Department of Humanistic Studies, Shalem College , Jerusalem, Israel.
2. Department of Hebrew Literature, Ben Gurion University of the Negev , Beer-Sheva, Israel.

Citation

Gilad Aviel Jacobson, Itay Marienberg-Milikowsky, and Yael Dekel (2025). "From Readers to Data. Uncertainty in Computational Literary Citizen Science". In: *CCLS2025 Conference Preprints* 4 (1). [10.26083/tuprints-00030142](https://doi.org/10.26083/tuprints-00030142)

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-02-07

Keywords

Citizen Science, Phenomenology, Reader Response, Uncertainty, Hebrew, Novel

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. We examine uncertainty in computational literary citizen science by analysing The Hebrew Novel Project, a large-scale initiative collecting reader interpretations of Hebrew novels. While citizen science projects typically treat uncertainty as noise, we demonstrate the value of treating it as meaningful data. Through statistical-phenomenological analysis of 1,026 questionnaire responses from 349 readers, we study how readers express uncertainty, from simple question-skipping to explicit rejection of interpretive frameworks. We uncover theoretically meaningful uncertainty patterns - certain literary concepts consistently elicit more uncertainty than others, and individual readers show varying but consistent levels of epistemic humility across different aspects of literary interpretation. We argue that this "productive uncertainty" provides insight into both the nature of literary texts and the process of reading, suggesting new directions for computational literary studies that embrace interpretive ambiguity. By taking uncertainty seriously, citizen science projects can address a wider scope of interpretive phenomena while maintaining methodological rigour.

1. Introduction

Citizen science has been increasingly used in research in recent years, primarily in the natural sciences but also in the social sciences and, more recently, in the humanities (Tauginienė et al. 2020). Despite the fundamental differences between these fields, most studies of citizen science share a common characteristic: they aim to address problems by acquiring data that is difficult to obtain otherwise. This is evident, for example, in projects that involve the public in documenting and counting of species in nature, in characterising galaxies (Galaxy Zoo), in river monitoring initiatives, and similar endeavors (Dickinson et al. 2010, Haklay 2012).

Nonetheless, citizen science is not limited exclusively to crowdsourcing. As researchers have demonstrated, public involvement often includes broad exposure to scientific findings, active contributions to additional stages of the research process, and sometimes even full participation in the entire research process (Wiggins and Crowston 2011). Moreover, even with minimal components of crowdsourcing, citizen science projects often demonstrate additional values that extend beyond the narrow academic scope: It empowers participants, individuals as well as communities, involving them in the

research and thereby serving the community in different ways (Bonney et al. 2016). A community participating in the process of monitoring a river, or in bird counting, may cultivate a closer connection to nature and assume greater responsibility for its interactions with the environment (Brossard et al. 2005). This often serves as one of the primary justifications for choosing this method, despite the various challenges it presents from a narrower traditional academic perspective, be they methodological, conceptual, or ethical. Taking this into account, from a scholar's perspective, citizen science may also foster a sense of contribution to the world beyond the purely scientific achievements.

While this notion appears fairly intuitive in environmental citizen science projects and similar initiatives, it is less apparent in the humanities, where the engagement with data, quantitative methods, as well as collaborative work has traditionally been limited. Digital Humanities - applying digital tools and computational methods to the study of the humanities - is opening new avenues in this regard as well (Tauginienè et al. 2020), especially in cases in which citizen science projects are related to the preservation of cultural heritage. In such cases, citizen science projects provide the public involved in it with symbolic rewards, similar to those associated with nature conservation. People who assist in deciphering medieval manuscripts or in tagging cultural objects are not merely contributing data; they are also helping to preserve culture.¹

Although it is not a cultural heritage preservation project in the traditional sense of the term, this insight informed the development of the Hebrew Novel Project (Dekel and Marienberg-Milikowsky 2021), from which the data presented and analyzed in this article is taken. The Hebrew Novel Project aims to collect comprehensive literary data (poetic, thematic, and bibliographic) on Hebrew novels, from the first published novel in the mid-19th century to the present day. The project employs reader questionnaires relating to a corpus of approximately 8,500 novels spanning various genres, places of publication, and positions with regards to the literary canon. When launching the project in 2020, we sought not only to engage the public in data collection but also to foster a deeper understanding of the novel as a social phenomenon. We aimed at studying this using a methodology that actively engages with some of its social dimensions. This endeavor resonated with the public, as demonstrated by significant media and social media interest, surpassing that of the academic community.

However, we soon encountered a significant challenge, invoked by some of the responses to the questionnaire: we received complex contributions that not only provided valuable answers to our questions, but occasionally - in the designated slots in the questionnaire - also engaged with some of the items in a critical manner. Thus, reader responses reflected their knowledge of the novel they reported on, but also various forms of uncertainty, indeterminacy, epistemic humility and more. This is not surprising: In most citizen science projects, contributor uncertainty is usually reported (e.g. Tikkoun Sofrim project, Wecker et al. 2019), and used for discarding data points, or channeling ambiguous data to experts, with the general aim being to reach a resolution. But the unease that the contributor experiences is informative in and of itself, especially in such a complex questionnaire. Such unease may reflect lack of knowledge about the answer

1. A good example for this is found in the Tikkoun Sofrim Project, which uses crowdsourcing to train algorithms to recognize Hebrew handwriting in medieval manuscripts. See: <https://tikkoun-sofrim.firebaseio.com/en>

or the terminology used in the question; it can also reflect an ambiguity in the novel that does not lend itself easily to the questions asked or the answers provided. In this paper, then, we shift our gaze from the typical questions and answers to the ones usually neglected, and ask: how does one derive meaningful research value from a systematic documentation of uncertainty?

The field of empirical literary studies (hereinafter: ELS) offers a different framework for thinking about our project. Dating back to the 19th century (Salgaro 2021), contemporary research has seen a growing interaction between ELS and various aspects of computational literary studies (hereinafter: CLS).² With the increasing integration of citizen science practices in CLS, this development is intensifying (Salgaro 2021, 538; Herrmann et al. 2021; Rebora et al. 2021): The combination of data-oriented research, quantitative analysis, and real readers' reactions to literature brings the two scholarly domains closer together, even if their objects of study, their methodologies, and their paradigmatic emphases may differ. This process is still in its early stages, and there is much to be done: We believe, for example, that scholars of CLS have a lot to learn from the well-established use of questionnaires in ELS, and the conceptual framework of cognitive poetics. At the same time, we are hesitant of the outright rejection of interpretive subjectivity that is sometimes advocated in ELS.³ In this regard, the affiliation of CLS with non-computational and non-empirical approaches to literary study provides, for us, an important balancing anchor.

Indeed, while ELS engages with real readers and the ways they interact with literature, other approaches—particularly those widespread in the second half of the 20th century—tend to focus on abstract constructions of theoretical readers. For these approaches, uncertainty is viewed as a hypothetical reaction to the object of study. As is well known, some of the most influential schools of literary studies—from reader response criticism to post-structuralism—celebrate interpretative freedom, over-interpretation, ambivalence, and disagreement in different ways. Similar notions had already influenced literary studies earlier, notably in the work of Roman Ingarden and particularly his concept of indeterminacy, which he saw as inherent to literature due to its attempt to represent real objects.

Thus, the scope of our study, its grounding in the community of readers, its ambition of creating a 'democratic' database of novels and the reactions that they evoke, and, lastly, the inevitable computational analysis of the results, point to a complex negotiation between various interpretive traditions and data-driven approaches. On the one hand, it embraces the appeal of a plurality of interpretive voices; on the other hand, it imposes a normalizing framework on them. When it relates to readers as a resource for data collection, deliberately limiting their interpretive freedom by providing a structured mechanism for collecting the data, it faces a clear challenge vis-à-vis some of the traditional intellectual conventions. However, allowing space for uncertainty, and treating indeterminacy as valuable data – and not just as noise, as something to be regulated, validated or simply deleted – can bring the Hebrew Novel Project closer, in some senses,

2. The various activities of the International Society for the Empirical Study of Literature (IGEL) and its journal (Scientific Study of Literature SSOL), are all worthy of consideration, when wishing to integrate citizen science methods and goals within CLS.

3. For instance, Dixon and Bortolussi 2011 assert that "scientific methods require that observations be repeatable, and this requirement rules out subjective analyses that vary across individuals" (p. 65).

to traditional literary studies. 101

There are more difficulties in the implementation of citizen science in literary studies, 102
and specifically in CLS. First, in clear contrast to literary studies as shortly described 103
above, computational research often treats data, at least in its processed form, in a 104
robust manner, as if it were transparent and free of interpretive biases (Piper 2020). 105
Second, in CLS research that relies on annotations (by expert researchers or trained 106
assistants), the norm of an extensive work with annotation guidelines while striving for 107
inter-annotator agreement has been justifiably established (Gius et al. 2021). Thus, in 108
addition to the consideration of uncertainty as data, the introduction of a less-controlled 109
project, driven by amateur contributions, seems to undermine the very foundations of 110
the field's (traditional as well as computational) interpretative concepts; it resonates 111
with past schools of literary theory and criticism (formalism, structuralism) as well as 112
with the concept of indeterminacy as suggested by Ingarden (Ingarden 1973).⁴ 113

But if answers to a highly detailed questionnaire dedicated to the characterization of 114
complicated literary phenomena reflect, to some extent, indeterminacy, what should 115
one do with such data, often considered as noisy or messy? A widespread tendency is 116
to focus on agreed, validated information, to adjust and normalize disagreement, or 117
to ignore uncertainties in different ways (e.g., using reports of uncertainty to redirect 118
data to experts, enlarging the number of reports for those data to allow estimation of 119
some underlying "consensus"). In some of the outcomes of the Hebrew Novel Project, 120
we, too, strive for the agreed. However, in the present article, we choose to celebrate 121
indeterminacy, treating it not as a potential source of noise in the data, but rather as a 122
source of knowledge. Based on this, we seek to conceptualize indeterminacy in a way 123
that will show its benefits to our project as well as other studies. 124

The next part of the article will be devoted to a brief review of the use of citizen science 125
in CLS. We will then provide a detailed description of the approach we developed in 126
The Hebrew Novel Project. Following that, the article will delve into a few specific 127
findings, highlighting indeterminacy in response to a variety of items in the Hebrew 128
Novel project's questionnaire. Lastly, we will turn to discuss the findings, using a 129
statistical-phenomenological approach. 130

2. Computational Literary Studies and Citizen Science 131

The integration of citizen science into the humanities is still in its infancy, and, as noted 132
earlier, is used primarily in digital humanities and more specifically in contexts of 133
cultural digital preservation. Its presence in the subfield of CLS is still scarce, found 134
only in a handful of innovative projects. These projects — some of which we will present 135
here — can be seen as the beginnings of a new scholarly direction, which we propose to 136
call *Computational Literary Citizen Science* (hereinafter: CLCS), linked also to the well- 137
established tradition of ELS. Most of these projects draw on a relatively wide community 138
of non-professional readers, keeping the task simple, sometimes referring to sociological 139
and demographic aspects of the project participants, and usually also combining the 140
crowdsourced findings with various automated techniques. Yet, in many ways, these 141

4. A different issue that will not be discussed here is disagreement between different readers of the same novel. We reserve this discussion for further accounts.

projects also differ from one another, and examining these differences will help us better
situate our own work.

A recent example, “The DisKo project” (Diversitäts-Korpus [diversity corpus]), led by
Marie Flüh, Mareike Schumacher and Peter Leinen, involves the use of citizen science to
collect titles of novels that feature various non-binary gender representations.⁵ This is
achieved through a short questionnaire that includes some demographic questions, a
request to list relevant titles, and an option to provide comments. The goal of this ongoing
project is to compile a sufficiently large list of books – one that could not be compiled
without the assistance of many readers – for future annotation by a professional team
that will explore methods for automatic identification of non-binary gender characters
in literature.

While the DisKo project collects titles, *Project Endings*, led by Helena Michie, Robyn
Warhol and Huw Edwards-Evans, asks readers to delve into books and collect structural
elements.⁶ This recent literary citizen science initiative invites the readers to choose
a serial Victorian novel from a predefined list and mark, using a Google Forms ques-
tionnaire, the narrative’s strategies for the ending and the beginning of each part of the
serial novel. *Project Endings* is rooted in literary studies more than in digital humanities,
and is described by the leading researchers as “a ‘medium data’ study [...] because no
computer application could do the required analysis”.

Focusing on an even smaller literary element, Andrew Piper and colleagues explicitly
integrate citizen science and academic research (Piper et al. 2024). In this computa-
tionally ambitious project, participants are asked to identify predefined types of character
interactions within specific sentences from contemporary literature. This task focuses
on supporting and refining natural language processing (NLP) methodologies and on
validating automated practices. The goal is to acquire accurate and objective informa-
tion, with low-agreement findings used to improve model training. The tagging process
requires minimal interpretation (only one sentence is annotated at a time), and the
emphasis is on achieving high levels of agreement. A similar approach is used in an-
other ongoing project by Piper, which focuses on annotating character emotions.⁷ Both
projects are disseminated through the Zooniverse platform, with the tagline: “Help us
annotate literary characters to build AI that can better understand human storytelling.”
Thus, Piper’s projects clearly demonstrate what appears to be a typical human-machine
interrelationship: the primary goal of the human contribution is to improve the algo-
rithm, and not necessarily explore the different human perspectives. In the end, the
purpose of human annotation is to serve the machine, even if eventually, the compu-
tational results will serve the human. The results are noteworthy: “With respect to
Citizen Science as a mechanism of crowd-sourced text annotation, we find annotation
quality on par with trained student annotators. As prior work has suggested, Citizen
Science projects achieve the same quality standards as other approaches and bring with
them the affordances of a volunteer, community-based approach to scientific discovery”
(Piper et al. 2024, 479). Following this success – in terms of data accuracy – the authors
voice the hope that “more projects in NLP and DH will utilize this significant resource”.

5. <https://msternchenw.de/disko-das-diversitaets-korpus/>

6. This is an ongoing part of a larger project on the Victorian novel, whose details are found here: <https://readinglikeavictorian.osu.edu/>

7. <https://txtlab.org/2024/09/new-citizen-science-project-reading-emotions/>

Although their focus differs, citizen science was employed in the three studies reviewed so far to obtain unambiguous data: to expand the corpus of literature featuring non-binary characters in the first case, to characterise beginning and ending strategies in the second, and to improve the accuracy of automated literature analysis models in the third.

The following study, which is actually the earliest, takes a different direction, one closer to that of the empirical study of literature. Karina van Dalen-Oskam's *The Riddle of Literary Quality* is an extensive two-stage citizen science project (Dalen-Oskam 2023). In the first stage, almost 14,000 readers filled out a survey about the subjective literary quality of contemporary Dutch and translated novels, from a list of best-selling novels. The second stage consisted of computational text analysis of the same novels. The survey (titled The National Reader Survey) was opened for seven months in 2013 and included sixteen questions, both demographic and pertaining to the participants' opinion on the literary quality of the novels they have read (Koolen et al. 2020). Interestingly, *The Riddle* did not use the term Citizen Science or similar terms. Moreover, it dealt with agreement and disagreement (notions that can be seen as related to some extent also to indeterminacy) as part of what can be described as the sociology of literature, actively creating a more diverse profile of respondents based on their gender and geographic location.

The focus of The Hebrew Novel Project is neither the reader, nor sociology of literature. The subjective perspective of its participants (whose demographic and sociological backgrounds are not made explicit in the questionnaire) is apparent in the data through its interpretive literary as well as thematic questions. The data arising from the project suggests a novel question: how does indeterminacy contribute to the research of literature itself?

3. The Hebrew Novel Project

The Hebrew Novel Project was born out of two seemingly contradictory intellectual passions: on the one hand, the urge to organize, to systematically map the entire large-enough yet not-too-large corpus of the Hebrew novel, and on the other, an impulse to disrupt, shown in the enthusiasm for the noise that arises from as many human thorough readings as possible. Interestingly, the tension has been particularly significant in the development of CLS, especially in light of the implicit dialogue between Franco Moretti's "Conjectures on World Literature" (Moretti 2000) and Erich Auerbach's "Philology of World Literature" (Auerbach 2012 [1952]). In short, while Auerbach was criticizing the very idea of a research based on collective work, Moretti proposed a research method based on second-order reading that therefore relies on more than one reader. In the Hebrew Novel Project we took this intention a step further, as both these scholars certainly did not consider literary research based on a *non-scholarly* community, a community of 'ordinary' readers whose variety of *different* readings include uncertainties – rather than a unifying synthesis that adjusts them. Our interest in these different readings is phenomenological, as we want to better understand what can be learned from indeterminacy as such. This phenomenological subjectivity resonates with Wolfgang Iser's understanding of the role of the reader in filling *gaps* in the text: Indeterminacies engage

the readers and require them to participate in the meaning-making of the text, a process that is highly subjective (Iser 1980).

Finally, in order to better understand the essence of the Hebrew Novel Project, we will describe its similarities and differences with other literary projects, traditional as well as computer assisted. First, the Hebrew novel project is not a close reading project. While in traditional literary studies the most widely accepted approach is that of close reading of individual texts, here we tackle a different problem – the Hebrew novel in general – by gathering data on as many texts as possible. Despite this, the Hebrew Novel project is actually based on close readings: the readers who participate in the project fill out an exhaustive questionnaire about a Hebrew novel they have recently read, and are advised to hold the book near them while answering the questionnaire. Most of the questionnaire items require participants to reflect on the novel, delving into some of its stylistic and thematic features. This is a form of second-order distant reading which we named elsewhere *distant public reading* (Dekel and Marienberg-Milikowsky 2021).

Second, as a whole, it is not a typical computational text analysis project. While computing power takes place in different stages of our project – from data gathering (with Google forms) to its statistical analysis (with Excel, R and MatLab) – it has no role in the reading itself. The reading is done by humans, without any algorithmic element, and part of our focus in analysing the reports is to highlight the individual readings that are attested to by the different contributions. It should be noted that while we have digital access to many of the novels, for the current article which focuses on the readers and their uncertainties, we are refraining from processing them with text analysis tools. It should also be noted that some of the other parts of the project rely more than the one presented here on text analysis techniques.

Third, in contrast to another common approach in computational literary studies, the Hebrew Novel Project is also not an annotation project in the usual sense of the term. Typical annotation projects aim both to enable distant reading and to document close reading. We, however, do not use in-line annotations at all, as the comments of those who participate in our project are not attached to specific textual segments; rather, the readers provide their structured feedback at the level of the entire novel (genre, plot, characters, time, space, etc.), and, to some extent, to its external circumstances (e.g., in questions of reception and importance). However, as we have argued elsewhere (Münz-Manor and Marienberg-Milikowsky 2023), the tension between describing a work as a whole and a detailed tagging of its text is a fertile tension for a more sophisticated annotation theory and practice.

As argued by Gius and Jacke, not all disagreements should be processed equally; some can (or should) be resolved but others not: “literary analysis should more often be inspired by the shared effort of agreed disagreement” (Gius and Jacke 2017, 251). The same can be said about uncertainties. Yet, within the framework of our project, we cannot judge the veracity of readers’ claims, except in cases of a clear mistake (about some of the non-interpretative bibliographical data). Since the focus of the current paper is phenomenological, we are not concerned with the veracity of readers’ responses. The question of errors, agreement and consensus may be dealt with in future papers, which will approach the same data through a different prism.

4. Findings

271

Our questionnaire was designed to collect data about several categories (bibliography, 272
narratology, time and space, themes, language) using multiple-choice items, linearly 273
scaled items, and a few short-answer questions that allow for more personal and inter- 274
pretive free text responses. And yet, although the readers mostly choose the best option 275
(or multiple options) out of a few given answers, many of these choices (or, all of them, 276
except the bibliographic ones) depend on interpretation. While most of the questions 277
are required and non-skippable, in a few cases, pertaining to complex literary concepts 278
which nevertheless were explicated in the questionnaire, we allowed the readers to 279
skip a question in cases of uncertainty. Thus, this structured questionnaire calls for 280
interpretation, disagreement, ambivalence and indeterminacy. 281

It is important to note that the Hebrew Novel project was constructed as a Citizen Science 282
project, and our sensitivity to reader uncertainty and ambiguity grew from studying the 283
corpus of filled questionnaires. Therefore, the data analysed is uneven, in the sense that 284
items provided heterogeneous opportunities for expressing uncertainty and ambiguity. 285
The analysis should therefore be assessed for what it is: a demonstration of possible 286
modes of expressing uncertainty and ambiguity, and the kinds of insights we may glean 287
from them, while not providing an exhaustive exploration of all aspects of uncertainty 288
and ambiguity relevant for each item. 289

We first demonstrate the simplest form of reader uncertainty manifested by skipping an 290
item, as items in the questionnaire were occasionally skipped. The questionnaire, which 291
contained 77 items in total of varying types, contained 9 scaled items (see appendix A). 292
Of these 9 items, readers were allowed to skip 4 (see Fig. 1a): 293

- “How would you estimate the typical linguistic register of the novel (1: very low - 294
5: very high)?” (register l-h; n=13/987 skipped). Readers were requested to skip 295
this item if the answer to the previous item (“was Hebrew a spoken language 296
at the time the novel was written?”) was negative. We therefore excluded such 297
skips in our analysis, and only included skips in this item if the previous item was 298
answered in the affirmative (n=987/1026). 299
- “To what extent does the plot leave gaps that the reader must fill using their 300
knowledge or imagination”? (gaps; n=26/1026 skipped). 301
- “Where along the conventional-experimental axis would you locate the novel?” 302
(conv.-exp.; n=53/1026 skipped). 303
- “To what extent, in your opinion, does the novel employ intertextuality?” (inter- 304
text.; n=106/1026 skipped). 305

These items elicited different degrees of skipping (1.3%-10.3%), which we interpret as 306
expressing varying degrees of uncertainty or ambivalence. The uncertainty may result 307
from unfamiliarity with the term (such as intertextuality, that while explained briefly in 308
the questionnaire, is not necessarily familiar to the non-professional reader), a property 309
of the novel, or its perception by the reader, that defies an easy response. In these scaled 310
items, it is impossible to disentangle these disparate explanations, as the readers had no 311
means of providing a more detailed account of the type of difficulty they encountered. 312

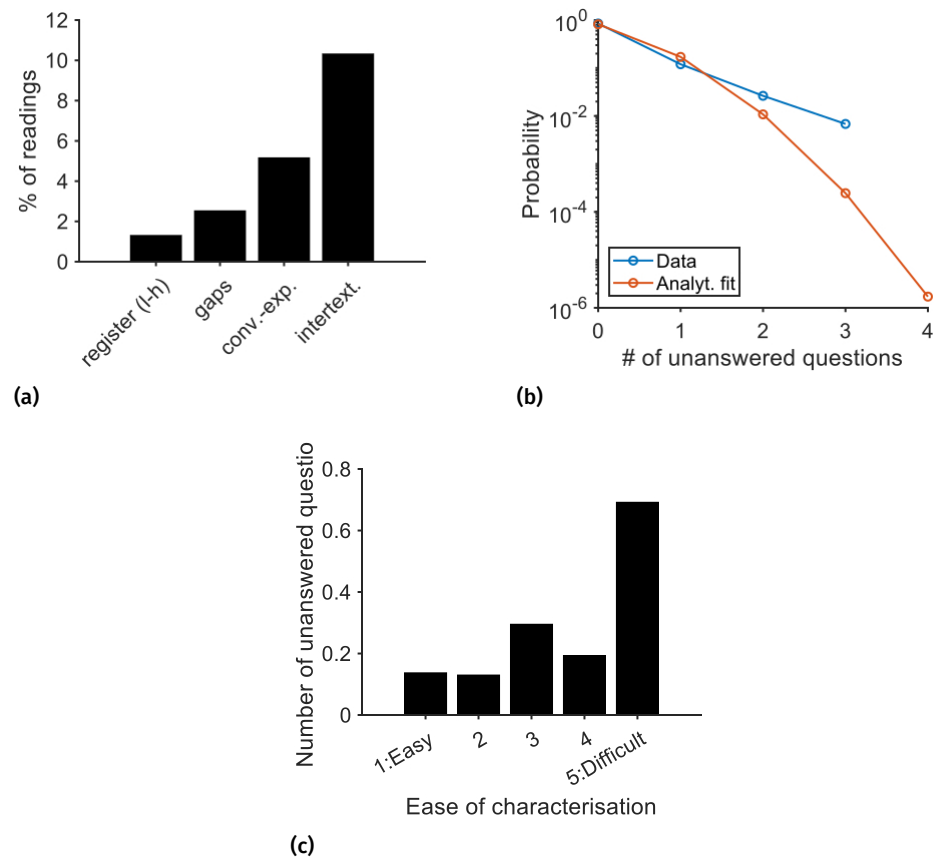


Figure 1: Reader ambivalence in scaled items. Four of the nine scaled items in the questionnaire allowed readers to express ambivalence by skipping the item. **(a)** Percent unscored questions across all questionnaires ($n=1026$), sorted from least to most skipped. **(b)** Probability of observing questionnaires with 0-4 skipped items. Blue line: real data. Red line: analytic fit, using marginals and assuming items are skipped independently. **(c)** Readers who found it most difficult to characterise the book using the questionnaire (5 on x-axis), skipped on average more scaled items (y-axis) (1-way ANOVA: $F=8.91$, $p<10^{-6}$).

Nevertheless, we were able to demonstrate that item skipping tends to cluster in certain questionnaires more than predicted by random distribution. To this end, we calculated the frequency of skipping each item across all questionnaires, and calculated the expected frequency of questionnaires with 0, ..., 4 skips under the assumption that the skips are independent of each other. As shown in Fig. 1b, the data (blue line), when compared to the above calculation (orange line), shows an excess of questionnaires with 2 skips ($\times 2.4$) and 3 skips ($\times 27.8$). This indicates that item skipping within a questionnaire is correlated. Such correlations may arise either from a property of the readers (some readers exhibiting higher ambivalence, epistemic doubt, or lack of acquaintance with terminology, compared to others), or the novels (some novels eliciting more ambivalence in readers across different questions).

Finally, we asked whether item skipping exhibits a relationship to the last, reflective, question in the questionnaire, a scaled item in which readers were asked to report how easy it was for them to characterize the novel using the questionnaire. As seen in Fig. 1c, the mean number of skipped items tends to increase when the reported difficulty in novel characterisation increases. Thus, an explicit report of ambivalence was statistically linked to an implicit one—the number of skipped scaled items, with readers reporting the maximal difficulty (5 vs. 1 – 4, post-hoc contrast after a one-way ANOVA test; $F = 8.91, p = 4.5 \cdot 10^{-7}$).

Next, we extended our characterisation of uncertain responses to a wider range of items, as readers were provided with different means of expressing uncertainty and ambiguity in different items. In some, there was no opportunity provided (e.g. multiple choice questions or scaled items that could not be skipped). In others, one or two of the answers that allowed readers to express their uncertainty or ambiguity (such as “unknown terminology”, “hard to define”) were provided. In items that contained the option for free text, readers could add other categories of uncertainty / ambiguity that were not offered to them.

To demonstrate the different kinds of ambiguity and uncertainty in the questionnaire, we analysed a subset of 23 items that represented the various item types: scaled items, numerical items, and various types of items providing multiple choice, free text, or combinations thereof. For the items with free text answers, we manually tagged all answers that reflected some degree of uncertainty or ambiguity. We then divided uncertain or ambiguous answers into nine categories, according to the common features they share:

1. No answer (the reader skipped answering this item).
2. “Term unknown” (uncertainty regarding question).
3. “It is unknown” (objective uncertainty regarding answer).
4. “Impossible to answer” (a more emphatic form of 3).
5. “Hard to define” (a less emphatic form of 3).
6. “I do not know” (subjective uncertainty regarding answer).
7. “I do not know” + informative answer.

8. "I do not remember accurately". 354
9. Rejection of question. 355

Figure 2a depicts the prevalence of these categories of uncertainty/ambiguity for the 23 selected items. Some categories were infrequent (category 3 (it's unknown): $n = 3$; category 8 (memory): $n = 4$), while others appeared with high frequency (category 5 (hard to define): $n = 550$; category 1 (no answer): $n = 342$). It is clear that expressions of uncertainty/ambiguity that were offered as options in the questionnaire, either implicitly (skipping) or explicitly (choosing an uncertain/ambiguous answer provided in a multiple-choice item) were much more frequent, while those that entailed free text were less frequent. We suspect that this difference is governed both by the additional effort required to conceptualise, phrase and write a free text answer, and by the heterogeneity across items, with many items not providing an option for free text.

Readers varied in the degree of uncertainty/ambiguity they expressed in the questionnaire (see Fig. 2b). Of the 23 items analysed, 340 questionnaires (33%) contained no item with the above indicators of uncertainty/ambiguity, 337 questionnaires (33%) contained a single such item, and the maximal number of uncertain/ambiguous answers was 10, in a single questionnaire. The mean number of uncertainty reports per questionnaire, restricted to the above 23 items, was (mean \pm standard deviation) 1.3 ± 1.5 .

As explained above, the source of reader uncertainty is sometimes itself uncertain, and it is not always possible to determine if it stemmed from a property of the specific novel reported, the specific questionnaire item and the terminology it used, or from a property of the reader, assuming that different readers possess varying degrees of epistemic doubt. It is therefore informative that reports of uncertainty were not independently distributed across questionnaires. Like in Fig. 1b, statistical independence between reports of uncertainty would have resulted in almost no questionnaire with > 5 reports of uncertainty, and in our data, there is an excess of questionnaires with 6 – 10 reports of uncertainty. This excess of uncertainty in some questionnaires may result from properties of the specific reader or the specific novel reported.

Last, we can see that within each item type, different items elicited varying degrees of uncertainty/ambiguity. Figure 2c summarises this data visually. Even within each item category, different items elicited varying degrees of uncertainty. For example, in the multiple choice questions with more than 3 suggested answers (MC (>3)), the item requesting readers to describe the tense of the narration elicited few instances of the uncertain response "hard to define" ($n = 13/1026$), while the item requesting readers to describe the location of the novel's exposition elicited almost a five-fold increase in the same response type ($n = 60/1026$). We must further stress that in the exposition item, we provided readers with yet another answer classified by us as uncertain / ambiguous: "I'm unfamiliar with the term". Thus, we can safely assume that in both these items, the "hard to define" answer reflects a difficulty in assessing the novel itself, and not in understanding the question, and that novels tend to ambiguate the location of the exposition more than ambiguate the grammatical tense.

It is worth highlighting some of the reader contributions to the categories of uncertainty and ambivalence, which were provided in items that allowed free text answers. An interesting example is given in response to the item in which readers were asked to

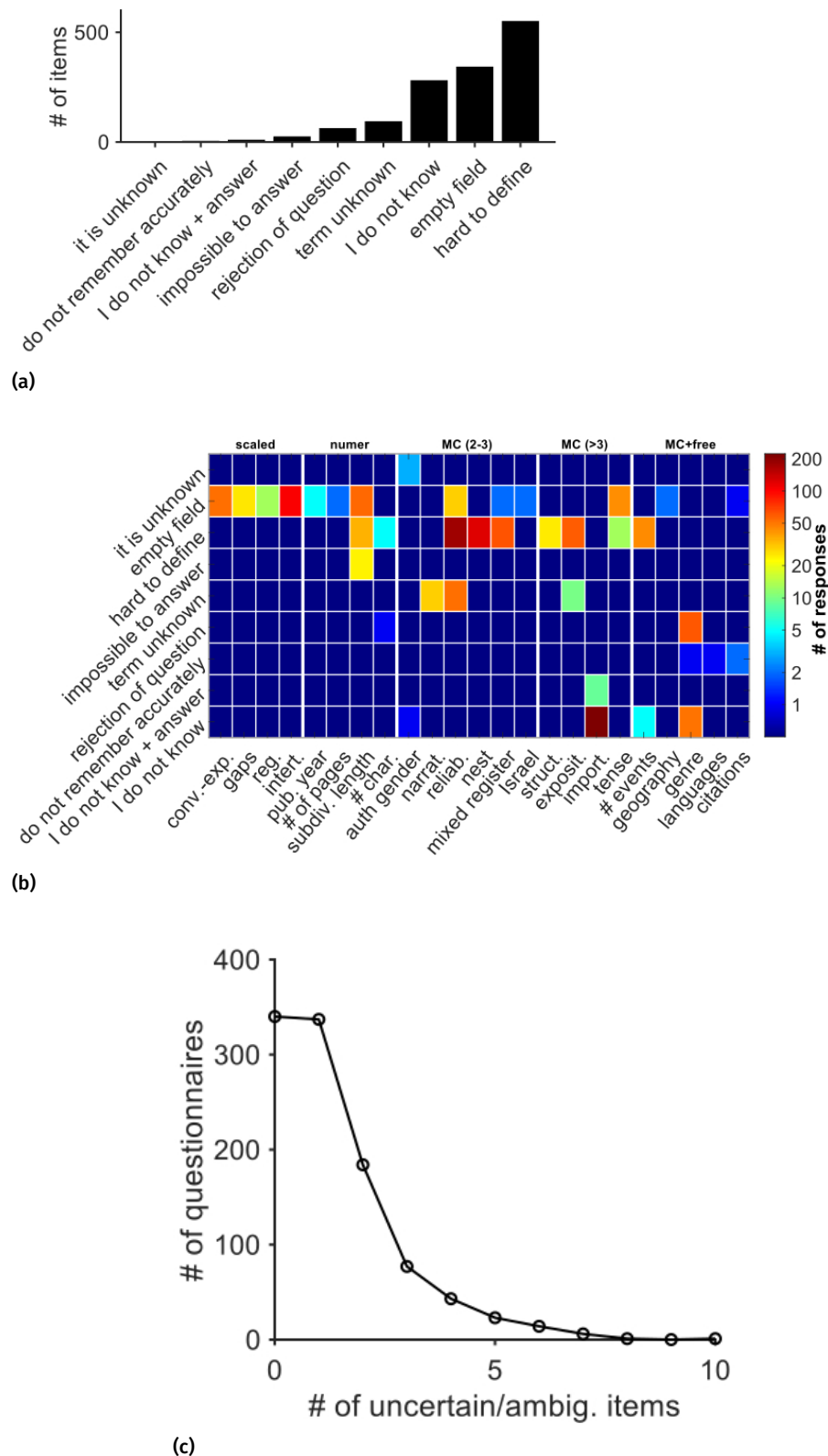


Figure 2: Different modes of expressing uncertainty and ambiguity. Readers express their uncertainty in different ways. (a) Nine categories of uncertainty and their prevalence in the questionnaire corpus. (b) Distribution of the number of uncertain / ambiguous responses across all questionnaires ($n = 1026$; $mean \pm sd = 1.3 \pm 1.5$ per questionnaire, with a maximum of $n = 10/23$). (c) Distribution of different categories of uncertainty (y-axis) across 23 questionnaire items of different types. The types are separated by a thick, white vertical line. *scaled*: items with scale 1-5. *numer*: numerical items. *MC*: multiple choice items, further divided into binary/tertiary items (*MC (2-3)*); items with more than 3 choices (*MC (>3)*); and items that allow both MC and free text (*MC+free*).

report how many main and secondary characters the novel had. One reader wrote: 398
 “I claim that ... the novel is more complex than the framing of some of the questions 399
 aimed at. ... It is indeed possible that there is one main character and several secondary 400
 ones, but the structure of the novel challenges this by having the characters change parts 401
 ...”. They thus questioned, or rejected, the relevance of the question itself, while still 402
 providing a hint to what their answer might have been if the issue was forced. This 403
 response was thus categorised as “rejection of question”. Another item that elicited 404
 answers in the same category requested readers to describe the sub-genre of the novel. 405
 Four different readers replied with answers that rejected the suggested genres, and 406
 one even doubted the book is a novel (e.g. “none of the definitions [suggested] is 407
 accurate”). While a free text option may complicate analysis, and is often avoided in 408
 multiple-choice questions, the examples discussed suggest that they allow contributors 409
 not only to provide what they perceive as accurate answers, but also to comment on 410
 their own unease, uncertainty and ambivalence. 411

5. Discussion: Assessing Uncertainty— 412

A Statistical-Phenomenological Approach 413

Integrating citizen science into projects whose primary objective is to collect data that 414
 cannot be efficiently gathered by other means, seems quite natural. In the so-called 415
 information age, where so many have access to the internet, and scientific endeavors 416
 are more data-driven than ever, it simply makes sense. The challenges arising from this 417
 method, as noted in the introduction, are offset by the advantages of its non-scientific 418
 added value. It is not surprising, then, that when citizen science has been integrated 419
 into CLS, it has primarily been used to collect data, and often in ways that contributed 420
 not only to the scientific work in the narrow sense of the term. It is also not surprising 421
 that in some cases that were described above in detail, it has been done in order to 422
 support computational work, in one way or another. However, we believe that this 423
 new research strategy offers an opportunity not only to collect—and preserve—cultural 424
 data, and not only to build a *useful* datasets that will enhance computational findings, 425
 but also to re-examine the *role* of data in CLS; and, more specifically, to rethink the 426
 place of reading-based data, in relation to prominent currents of literary criticism in 427
 the past century, whether empirical-oriented or theoretical-oriented. This approach 428
 can challenge how we perceive textual content, much like Ingarden’s indeterminacy 429
 theory, Iser’s (and others’) reader-response criticism, and French poststructuralism 430
 have done before. By doing so, we adopt the very idea of operationalization in CLS, as 431
 described more than a decade ago by Moretti 2014: “the process whereby concepts are 432
 transformed into a series of operations—which, in their turn, allow to measure all sorts 433
 of objects. Operationalizing means building a bridge from concepts to measurement, 434
 and then to the world. In our case: from the concepts of literary theory, through some 435
 form of quantification, to literary texts.” 436

Having said that, it is important to note that this method of operationalization might 437
 also challenge well-established CLS practices, such as annotation. While real readings 438
 are usually collected in CLS as in-line text annotation, we suggest comparing them 439
 with readings gathered as structured reflections on the literary text as a whole, as 440

an interpretative perspective that extends beyond mere details (Münz-Manor and Marienberg-Milikowsky 2023). After all, both methods indicate that the text is not just words on a page (or a screen), but a complex communicative act in which the recipient, not just the text itself, plays a part; they just treat this act differently.

It should be emphasized, however, that the use of a structured, research-oriented questionnaire (rather than, for example, collecting reader impressions and reviews from commercial websites or reader communities forums), restricts the respondents' interpretive horizons. Hence, the potential perception of the text in computational literary citizen science, might seem closer (but not at all identical, as Gius and Jacke have shown) to approaches that were dominant around the mid-20th century and onwards, until the rise of post-structuralism (Gius and Jacke 2022).⁸ Under such conditions of a standard questionnaire, the chance of getting a provocative and fruitful overinterpretation (Culler 2007), seems quite low. Yet, our findings suggest that forced, controlled, and data-oriented reading in which interpretive freedom is – at the same time – kept and limited, and restricted to the assessment of the text after its reading, contains valuable information.

Here is where a statistical-phenomenological approach comes into play. Considering different readings as definite data (so-to-speak), and, at the same time, as potentially undecided reactions, allows quantitative-conceptual analysis to better characterize indeterminacy. Indeed, as delineated above, uncertainty can be seen as relating to the complexity of literary characterization in general. This is demonstrated by figure 1a, rating a few literary concepts, in which some are easier to decipher (linguistic register) while others are perceived as more difficult (intertextuality). This is even more evident in the relationship between these specific expressions of uncertainty, and the explicit evaluation of the questionnaire as a suitable means of assessing the novel, as documented in the last, reflexive question of the entire questionnaire (Fig. 1c).⁹

Using the extent of item skipping as a proxy for item difficulty as experienced by readers, helps shed light on uncertainty or ambivalence as being consistent among certain readers and the ways in which the questionnaire resonates their reading experience. Taking this into account, we suggest that ambivalence should be evaluated as such, rather than being normalized for the sake of adjusting the results on the one hand, or validating them on the other. Moreover, the skipping of items may suggest that readers engaged thoughtfully with the challenges posed by the questionnaire. Based on the results, we suggest that skipping may not stem from inexperience in reading literature, but rather could imply a thoughtful and reflective engagement with the text.

We have to address the difficulty in the terminology used in this paper to describe a variety of engagements of readers with the questionnaire. The term uncertainty itself is ambiguous: It may reflect an epistemic uncertainty of the reader, but also

8. We refer here only implicitly to the “digital humanities-as-structuralism” narrative which Gius and Jacke engage with in their article, because, as they demonstrate, the title “structuralism” includes many variants that are productive to literary studies but cannot be described here. Moreover, some of our more explicit sources of inspiration (Ingarden, Iser) might have roots in structuralist thinking, but are not perceived as being under this umbrella. The Hebrew Novel Project, and especially our main concern here – namely, indeterminacy, uncertainty – echoes several (sometimes seen as contradictory) thinkers and approaches.

9. We use a similar method in annotation-based projects in our lab: When the annotation aim is conceptually complicated, we add a question in which annotators have to note if, or to what extent, they are sure about their annotations. The data that such a question provides is not only useful in the process of validation and re-examination of the annotations, but also in and of itself.

an uncertainty about the aptness of the question itself or the answers provided in the questionnaire. It would be useful to consider the variety of terms that may be applicable, to different extents, to the various cases we have presented here: uncertainty, ambivalence, ambiguity, epistemic doubt/humility, rejection. They all share a degree of defiance or an outside view on the question itself, even when not refraining from partially answering the question itself. They all, thus, share a degree of unease towards the question asked. An extreme instance of a combined answer and epistemic doubt can be observed in response to the question about the novel's significance, in which 9 readers chose the answer "I do not know", while marking an additional, informative answer. Future work would have to address and create a taxonomy of the different types of uncertainty and ambiguity, in the vein of Empson's "Seven Types of Ambiguity" (Empson 1973 [1930]).¹⁰

6. Conclusions

We presented in this paper an analysis of uncertainty in reader evaluations of novels, within the framework of The Hebrew Novel Project. While the obvious motivation for CLCS is extensive data collection and annotation, one should not ignore the subjective nature of individual contributions. The study of reader uncertainty and its enrichment of our understanding of reader engagement with literary texts, is not something that we set out to do when starting the project, but was revealed to us serendipitously when examining the resulting corpus of questionnaires. We believe that there is a lot to be learnt from adopting a prism that focuses on the phenomenological, subjective perception of literature by readers, irrespective of the theoretical framework it is cast within. We suggest that CLCS projects may gain something by considering, at the planning stage, providing participants with a variety of means to express their uncertainty, ambivalence, and other facets of their unease with the questions. We also believe that uncertainty and ambiguity can play a much larger role than typically done when collecting data in citizen science projects, in science, social studies and humanities alike. This article provides a step in this direction.

Uncertainty and ambiguity are but one facet of the complex data collected in the Hebrew Novel Project. The same corpus lends itself to multiple analyses and perspectives. One can, for example, focus on disagreements between different readers reporting on the same novel, and return to a close reading of novels that elicit divergent reactions; one can also examine what can be learned from *resolving* disagreements and employing a distant reading approach to the consensus dataset (two directions that we are currently pursuing simultaneously). The use of diverse, and at times conflicting approaches, to the same dataset, ultimately highlights the inherent complexity of literature and its reading, reminding us that, as in the past, nothing should be taken for granted. Data can be interpreted in multiple ways, and our article suggests that ambiguity itself can be treated as an additional dataset — one that is also open to interpretation.

10. Similar to Ingarden and Iser as mentioned above, Empson is another example of a theorist who worked long before post-structuralism, and even structuralism, and yet his theory might be highly relevant for computational literary studies, and used as an inspiration.

7. Methods

518

The questionnaire, its theoretical premises, creation and dissemination was previously explained (Dekel and Marienberg-Milikowsky 2021). The corpus of readings used in the current paper ($n=1026$) was extracted on August 12, 2024 into a spreadsheet format. Number of unique readers in this corpus, $n=349$; Number of unique novels, $n=700$. Data analysis was performed on Matlab, v. R2024b. The analytic fit in Fig. 1b was calculated using the Poisson binomial distribution in the following way. First, our sampling space is $\Omega = \{0, 1\}^4$, whose elements are of the structure $\bar{b} = (b_1, b_2, b_3, b_4)$, $b_i = 1$ implies skipping and $b_i = 0$ implies not skipping, and define the random variable $X(\bar{b}) = \sum_i b_i$. The probability of skipping is different for each item i and denoted by p_i , and these values are estimated from the data. Then the probability of a certain outcome is:

$$P\{\bar{b}\} = \prod_{b_i=1} p_i \cdot \prod_{b_i=0} (1 - p_i)$$

and the probability of a certain number of outcomes k is given by:

$$P(X = k) = \sum_{\bar{b} \in (X=k)} P\{\bar{b}\} = \sum_{\bar{b} \in (X=k)} \prod_{b_i=1} p_i \cdot \prod_{b_i=0} (1 - p_i)$$

8. Appendix A: questionnaire items

519

The Hebrew Novel questionnaire includes the following scaled (1-5) questions, some are skipable (indicated below) and others which are compulsory:

1. Where along the conventional-experimental axis would you locate the novel? If you don't know, please skip the question). [from 1: the most conventional to 5: the most experimental]
2. How would you define the pace of events in the novel's plot? [from 1: very slow plot to 5: very quick plot]
3. To what extent do you think the novel's plot leaves gaps for the reader to fill in using their own knowledge, reasoning, or imagination? This refers to fundamental gaps between events, to unclear causal connections, or to essential gaps in the description of characters, landscapes, and occurrences. If you do not know, please skip the question. [from 1: very little to 5: very much]
4. Try to characterize the key events in the novel's plot. If there are multiple key events, refer only to the central ones. To what extent did they surprise you? [from 1: did not surprise at all to 5: I was really surprised]
5. To what extent, in your opinion, does the novel end in an open-ended way (where it is unclear what happens to the characters, the conflicts remain unresolved, the questions unanswered, etc.) or has a closed ending (such as a marriage, death, or 'and they lived happily ever after')? [from 1: completely open to 5: completely close]
6. If you marked 'yes' in the previous question (was Hebrew a spoken language at the time the novel was written?), how would you assess the typical linguistic

register in the novel in relation to the spoken language of the time when it was written? If you marked 'no' in the previous question, please skip this question. [from 1: very colloquial to 5: very literary]	542 543 544
7. To what extent do you think the novel employs intertextuality? That is, to what extent does the novel maintain a linguistic, formal, or thematic connection — direct or indirect, explicit or implicit — to other texts? If the concept is unclear, please skip this question [from 1: little usage to 5: extensive usage]	545 546 547 548
8. How readable was the novel for you? That is, did you find it easy to read, was the plot easily understood, and was the reading experience not challenging? [from 1: very readable to 5: very unreadable]	549 550 551
9. To what extent was it easy for you to characterize the novel using the questionnaire? [from 1: very easy to 5: very difficult]	552 553
Figure 1 provides an analysis of the skipping of items in the scaled items 1,3,6 and 7 in the above list.	554 555
Figure 2 provides an analysis of 23 items that represent the different types of questions in the questionnaire (scaled; numeric; multiple choice questions with 2-3 options; multiple choice questions with more than 3 options; multiple choice and free text). All 23 items enable the reader to express at least one type of uncertainty.	556 557 558 559
Scaled items:	560
• Items 1,3,6,7 in the above list.	561
Numeric items:	562
• Year of publication	563
• Number of pages	564
• Length of subdivisions	565
• How would you describe the network of characters in the novel? In your answer, please refer only to the main characters and to significant secondary characters, not all the characters appearing in the novel.	566 567 568
Multiple choice questions with 2-3 answers:	569
• Author's gender	570
• Type of narrator (diegetic, non-diegetic, alternating narrators, term unknown)	571
• How would you assess the reliability of the narrator? The reliability of the narrator is usually determined by the degree of alignment between the narrator's value system and knowledge framework and that of the implied author, which is perceived as the value system underlying the text.	572 573 574 575
• Is the novel structured as a nested story?:	576
• Does the novel distinctly mix different registers of the Hebrew language? For example, when a certain character uses a colloquial form of language while the narrator uses a literary form, or vice versa.	577 578 579

- To what extent Israel is central to the novel? 580

Multiple choice questions with more than 3 answers: 581

- How would you describe the division of the novel into units and sub-units? 582
- How can the exposition in the novel be characterized? Exposition is the part of the story that presents the background necessary for understanding the plot. 583 584
- In your opinion, what is the importance of the novel? You may mark more than one option. 585 586
- What is the main grammatical tense in which the story is narrated? The question refers to the primary tense used by the narrator. 587 588

Multiple choice and free text: 589

- What is the nature of the events in the plot? According to the common distinction between 'key events' that are important for advancing the plot and 'filler areas,' which include simple everyday events, descriptions of landscapes and characters, pauses, etc., try to characterize the density of key events in the plot. 590 591 592 593
- Geographically, where does the main plot or the main plots take place? You can mention more than one possibility. 594 595
- Try to define the subgenre of the novel. You may mark more than one option. 596
- Are languages other than Hebrew being used in the novel? If so, which are they? 597
- Does the novel include elements from different artistic genres? The question refers to elements that are distinctly separate from the main plot and/or form of the novel, yet are still an integral part of it. 598 599 600

9. Data Availability 601

Data and software can be found here: <https://github.com/ga-jacobson/JCLS2025a/> 602

10. Acknowledgements 603

This research was generously supported by grant No. 1223 from the Israeli Ministry of Science and Technology. 604 605

11. Author Contributions 606

Gilad Aviel Jacobson: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Resources, Software, Visualization, Writing – original draft, Writing – review & editing 607 608 609

Itay Marienberg-Milikowsky: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing 610 611 612

Yael Dekel: Conceptualization, Data curation, Methodology, Project administration, 613
Resources, Validation, Writing – original draft, Writing – review & editing 614


References 615

- Auerbach, Erich (2012). "Philology and Weltliteratur (1952)". In: *Trans. Maire Said and Edward Said. The Centennial Review* 13, 1–17. 616
- Bonney, Rick, Tina B Phillips, Heidi L Ballard, and Jody W Enck (2016). "Can citizen science enhance public understanding of science?" In: *Public understanding of science* 25.1, 2–16. 618
- Brossard, Dominique, Bruce Lewenstein, and Rick Bonney (2005). "Scientific knowledge and attitude change: The impact of a citizen science project". In: *International Journal of Science Education* 27.9, 1099–1121. 619
- Culler, Jonathan D (2007). *The literary in theory*. Stanford University Press. 620
- Dalen-Oskam, Karina van (2023). *The riddle of literary quality: a computational approach*. Amsterdam University Press. 621
- Dekel, Yael and Itay Marienberg-Milikowsky (2021). "From distant to public reading the (Hebrew) novel in the eyes of many". In: *Magazén* 2.2, 225–252. 622
- Dickinson, Janis L, Benjamin Zuckerberg, and David N Bonter (2010). "Citizen science as an ecological research tool: challenges and benefits". In: *Annual review of ecology, evolution, and systematics* 41.1, 149–172. 623
- Dixon, Peter and Marisa Bortolussi (2011). "The Scientific Study of Literature: What Can, Has, and Should Be Done". In: *Scientific Study of Literature* 1.1, 59–71. 624
- Empson, William (1973). *Seven types of ambiguity*. Chatto & Windus. 625
- Gius, Evelyn and Janina Jacke (2017). "The hermeneutic profit of annotation: On preventing and fostering disagreement in literary analysis". In: *International Journal of Humanities and Arts Computing* 11.2, 233–254. 626
- (2022). "Are Computational Literary Studies Structuralist?" In: *Journal of Cultural Analytics* 7.4. 627
- Gius, Evelyn, Marcus Willand, and Nils Reiter (2021). "On organizing a shared task for the digital humanities—conclusions and future paths". In: *Journal of Cultural Analytics* 6.4. 628
- Haklay, Muki (2012). "Citizen science and volunteered geographic information: Overview and typology of participation". In: *Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice*, 105–122. 629
- Herrmann, J. Berenike, Arthur M. Jacobs, and Andrew Piper (2021). "Computational Stylistics". In: *Handbook of Empirical Literary Studies*. Ed. by Kuiken and Jacobs. Berlin and Boston: De Gruyter, 451–486. 630
- Ingarden, Roman (1973). *The literary work of art: an investigation on the borderlines of ontology, logic, and theory of literature: with an appendix on the functions of language in the theater*. Northwestern University Press. 631
- Iser, Wolfgang (1980). "Texts and readers". In: *Discourse Processes* 3.4, 327–343. 632
- Koolen, Corina, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout (2020). "Literary quality in the eye of the Dutch reader: The National Reader Survey". In: *Poetics* 79, 101439. 633
- Moretti, Franco (2000). "Conjectures on world literature". In: *New left review* 2.1, 54–68. 634

- Moretti, Franco (2014). ““Operationalizing”: or, the Function of Measurement in Modern Literary Theory”. In: *The Journal of English Language and Literature* 60, 3–19. [10.15794/jell.2014.60.1.001](#).
- Münz-Manor, Ophir and Itay Marienberg-Milikowsky (2023). “Visualization of categorization: How to see the wood and the trees”. In: *Digital Humanities Quarterly* 17-3.
- Piper, Andrew (2020). *Can we be wrong? The Problem of Textual evidence in a Time of Data*. Cambridge University Press.
- Piper, Andrew, Michael Xu, and Derek Ruths (2024). “The Social Lives of Literary Characters: Combining citizen science and language models to understand narrative social networks”. In: *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, 472–482.
- Rebora, Simone, Peter Boot, Federico Pianzola, Brigitte Gasser, J. Berenike Herrmann, Maria Kraxenberger, Moniek M. Kuijpers, Gerhard Lauer, Piroska Lendvai, Thomas C. Messerli, and Pasqualina Sorrentino (Oct. 2021). “Digital humanities and digital social reading”. In: *Digital Scholarship in the Humanities* 36.Supplement_2, ii230–ii250. [10.1093/llc/fqab020](#). <https://doi.org/10.1093/llc/fqab020>.
- Salgaro, Massimo (2021). “The History of the Empirical Study of Literature from the Nineteenth to the Twenty-First Century”. In: *Handbook of Empirical Literary Studies*. Ed. by Don Kuiken and Arthur M. Jacobs. Berlin and Boston: De Gruyter, 515–542.
- Tauginienė, Loreta, Eglė Butkevičienė, Katrin Vohland, Barbara Heinisch, Maria Daskolia, Monika Suškevičs, Manuel Portela, Bálint Balázs, and Baiba Prūse (2020). “Citizen science in the social sciences and humanities: The power of interdisciplinarity”. In: *Palgrave Communications* 6.1, 1–11.
- Wecker, Alan J, Uri Schor, Dror Elovits, Daniel Stoekl Ben Ezra, Tsvi Kuflik, Moshe Lavee, Vered Raziel-Kretzmer, Avigail Ohali, and Lily Signoret (2019). “Tikkoun sofrim: A webapp for personalization and adaptation of crowdsourcing transcriptions”. In: *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 109–110.
- Wiggins, Andrea and Kevin Crowston (2011). “From conservation to crowdsourcing: A typology of citizen science”. In: *2011 44th Hawaii international conference on system sciences*. IEEE, 1–10.

A Powerful Hades is an Unpopular Dude

Dynamics of Power and Agency in Hades/Persephone Fanfiction

Julia Neugarten¹ 

1. Department of Arts and Culture Studies, Radboud University , Nijmegen, The Netherlands.

Citation

Julia Neugarten (2025). "A Powerful Hades is an Unpopular Dude. Dynamics of Power and Agency in Hades/Persephone Fanfiction". In: *CCLS2025 Conference Preprints 4* (1). [10.26083/tuprints-00030143](https://doi.org/10.26083/tuprints-00030143)

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-02-06

Keywords

fanfiction, power, gender, NLP, Classical reception

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. This paper employs Riveter (Antoniak et al. 2023) to analyze the dynamics of power and agency between the characters of Persephone and Hades in 482 short works of fanfiction (369,809 words total) about Greek myth. Where a difference was measured, on average Persephone has higher power scores and Hades has higher agency scores. I plot the development of differences in character-power and agency over time and examine how they correlate with several other story features, including popularity metrics and the occurrence of different types of violence in the stories. Hades' power correlates negatively with story-popularity, while Persephone's agency correlates positively with it.

1. Introduction

The academic study of fanfiction – which I defined in previous work as “stories by and for fans, inspired by existing stories and exchanged for free online” (Neugarten 2021, 80) – is motivated by a claim about its transformative potential: by rewriting popular stories, the thinking goes, fans perform a kind of creative criticism, an act of “re-storying” (Thomas and Stornaiuolo 2019). Through rewriting, fanfiction can expose and interrogate the ideological claims on topics like gender (Wills 2013), sexuality (Floegel 2020) or race (Fowler 2019) that underlie popular stories. In doing so it may challenge or change the worldviews of its readers. One such area of potential social change through rewriting concerns gendered power dynamics. By, for example, shifting the narrating perspective from male to female, transforming narrative events or making changes to a storyworld, (fan)fiction may be able to address gender bias and inequality. In this paper, I use Riveter (Antoniak et al. 2023), an NLP-pipeline that detects and evaluates semantic frames, to analyze the gendered power dynamics in a corpus of fanfiction about Greek mythology. The aim of this analysis is to understand this fanfiction’s capacity to transform patriarchal power dynamics.

Scholarship on fanfiction can be divided into several phases or waves. The first of these was characterized by a celebratory attitude towards fan practices. Fanfiction was viewed as a kind of ‘poaching’ (Jenkins 2013), a way for fans to reclaim ownership over folk stories that the financially-motivated copyright system had taken from them. In this context, every fannish act of creation was cast as inherently political. This sense was strengthened because fanfiction communities were and are overwhelmingly non-male (Rouse and Stanfill 2023), and so were understood to some extent to represent the perspective of audiences commonly underrepresented and underserved in mainstream

popular media. In later phases, fan studies scholars began to critique this celebratory view of fan activity, pointing to some of the ways that fan communities perpetuate rather than challenge existing social ills like racial injustice (Lothian and Stanfill 2021; Pande 2018) and online harassment (Stanfill 2024). Scholarship in this wave also called out the “gendered boundary-policing practices within fan communities” (Scott 2019, 223). In a similar vein, media scholars turned attention to the dynamics of anti-fandom and the prevalence of negative affects and dislikes in (online) fan communities (Click 2019).

These different approaches to (the study of) fan culture point to a fundamental division in how fanfiction and other fan practices should be understood. Are they transformative of oppressive cultural and social norms in a way that is politically powerful and therefore laudable? Or do they provide echo-chambers where inequalities that exist in mainstream culture are repeated and perhaps even exacerbated? In other words: what are the limits of fan culture’s transformative potential?

In addition to contributing to the debate on fanfiction’s capacity to transform dominant ideologies, this paper also adds to the cultural sociology of literature, the area of study that explores the ways texts allow audiences to theorize their social world (Váña 2025). Combining analysis of literature and the social draws attention to the ways that narratives can operate “as sites, like social situations, where multiple forms cross and collide, inviting us to think in new ways about power” (Levine 2015, 122). This paper fits within that area of scholarship by examining narrative’s capacity to represent and explore social dynamics between male and female characters. In doing so, fanfiction may be able to deepen readers’ and writers understanding of these dynamics.

2. Research Questions

This paper addresses a set of questions that fit into these larger debates on fanfiction’s possible transformativity and narrative’s capacity to explore social dynamics. I ask:

- In short-form fanfiction about the relationship between the mythological characters of Hades and Persephone, how is their power and agency portrayed?
- Is their power dynamic gendered, i.e. does the gender of these characters impact their level of power or agency in the stories?
- How does this dynamic shift over time?
- How does it compare to existing research (Neugarten 2024; Neugarten and Smeets 2023a,b) on violence and gendered power dynamics in fanfiction about Greco-Roman Antiquity?
- How does it compare to existing research (Yang and Pianzola 2024), on power dynamics in Omegaverse-stories, a popular subgenre of fanfiction that presents a speculative conception of gender with clear power hierarchies?
- Do differing depictions of gendered power relations in this corpus of fanfiction impact the popularity of the stories among readers?

In what follows, I first explain my decision to focus on short-form Hades/Persephone fanfiction as a case study (Section 3). I then outline the method of data collection and

the way the Riveter tool operationalizes and measures dynamics of power and levels of agency (Section 4). I present Riveter-scores of both power- and agency for the fictional characters Hades and Persephone (Section 5.1), examine shifts in these scores over time (Section 5.2), and compare results to some existing computational analysis of similar case studies on violence (Section 5.3) and the Omegaverse (Section 5.4) in fanfiction. I also examine correlations between Riveter-scores and stories' user-generated popularity metrics (Section 5.5). I then reflect on these results and pinpoint some areas for future research (Section 6.1) Finally, I return to the research questions described above and the overarching question of fanfiction's transformative and social potential (Section 6.2).

3. Short-form Hades/Persephone Fanfiction as a Case Study

The dataset used in this paper contains short-form fanfiction rewriting the relationship between the mythological characters Hades and Persephone. I focus on this material for several reasons. Firstly, fanfiction about Greek mythology is a suitable case study for examining whether fanfiction is transformative. This is because the cultural material that this fanfiction is based on has historically overwhelmingly been characterized by gendered inequality, and many translations of this material have also been characterized by what has been called a "patriarchal bias" (McCarter 2022, 148). This material thus offers fans a clearly patriarchal cultural baseline to transform.

Second, it is relevant to study fannish rewritings of Greco-Roman myth because in contemporary online spaces, references to Antiquity are often used to support right-wing ideologies (Hodkinson 2022; Müller 2022; Zuckerberg 2018), so the cultural material in question is already quite heavily politicized online. This makes it a suitable case study to test whether fans are using the material to different political ends, as well. The popularity of referencing Antiquity in right-wing online spaces also shows that the cultural material being received, transformed and evaluated online is often not directly based on mythological sources, but on a highly mediated contemporary understanding of these materials. Instead of adhering to a linear relationship between a culturally dominant source and a subversive or subcultural rewriting, internet users are engaging with a dense web of intertexts in a wide variety of ways. For this reason, the current research also does not compare contemporary fanfiction directly to the ancient cultural materials that fan communities are rewriting and responding to. Instead, I compare works of fanfiction to each other and to insights into these stories taken from previous work, to assess the extent to which fanfiction is transformative of a set of implicit cultural norms that structure the reception of Antiquity today.

Third, within the corpus of fanfiction about Greek myth – which on popular fanfiction-website *Archive of Our Own* (AO3) is called *Ancient Greek Religion and Lore* – it makes sense to focus on Hades and Persephone because the dynamic between these two characters is also characterized by inequality in the culturally dominant myth: in most tellings of this story, such as the ancient "Hymn to Demeter" (Anonymous 1914), Ovid's *Metamorphoses* (Ovid 2010) but also more recent anthologies of myth (Mellenthin and Shapiro 2017) and retellings in Young Adult literature (see for example Bracke 2025; Gloyn 2019), Persephone is abducted by Hades and sexually assaulted. This unequal power dynamic is the culturally dominant representational norm that I read fanfiction

as – often implicitly – responding or writing back to. 108

Fourth, it makes sense to take this relationship as a case study because it is relatively 109
popular in the fanfiction community. In 2022, 844 out of 5,154 stories on AO3, or a little 110
more than 16% of all stories that had been written about Greek myth, were tagged with 111
the Hades/Persephone relationship, making it the most popular ‘ship’ to write fanfiction 112
about in that fandom and thus in some sense indicative of many fans’ preferences. 113

Finally, this case study selection is designed to account for one of the most striking 114
stylistic features of fanfiction texts: their brevity. A short word count is characteristic of a 115
particular kind of fanfiction. As Catherine Tosenberger notes, fanfiction has the “ability 116
to compress a great deal of meaning within a small space” (Tosenberger 2014, 16). In 117
other words, much fanfiction delights in densely packing as much intertextual meaning 118
as can fit into as short a word count as possible. This stylistic property, which has also 119
been called “intimate intertextuality” (Busse 2017) is exemplified by the Drabble-genre; 120
stories of exactly 100 words. Because the stylistic property of brevity is characteristic of 121
much fanfiction, it is important for the analysis of fanfiction to examine short texts. I thus 122
limit the analysis to very short stories of fewer than 10,000 characters (482 stories, or 9.4% 123
of all stories and 57% of Hades/Persephone stories at the time of data collection). This 124
also accounts for the length-limitations imposed by Riveter’s co-reference resolution. 125
This length limitation has been identified as a significant drawback of the Riveter tool 126
(Neugarten 2025, forthcoming). I nonetheless find it defensible here because it turns 127
attention to the short stories that are often so characteristics of fanfiction. As we will see, 128
the remaining dataset is sizable enough to generate interesting insight into the corpus. 129

4. Data and Method 130

4.1 Data 131

A dataset of fanfiction was collected using the AO3-Scraper (Radiolorian 2022). The 132
metadata is described in *MythFic Metadata* (Neugarten and Smeets 2023a,b) The 5,154 133
stories in the dataset all originate from *Archive of Our Own* (AO3), and were published 134
between the platform’s inception in 2008 and the data collection in 2022.¹ All stories 135
were tagged by AO3-users as belonging to the fandom *Ancient Greek Religion and Lore*, 136
although the set also contains some overlap with other fandoms, both those related to 137
Greek mythology (such as the popular young adult book series *Percy Jackson*), and those 138
unrelated to Greek mythology (such as *Sherlock Holmes* and *Harry Potter* fandom). All 139
stories were written in English. From the *MythFic* dataset, I selected stories tagged by 140
their authors with the Hades/Persephone relationship and meeting the length criteria 141
described in Section 3, resulting in a dataset of 482 short stories. Descriptive statistics 142
for the dataset are given in Table 1.² 143

1. Although *Archive of Our Own* went into open beta in 2008, the platform offers users the option to backdate fanworks that have been re-uploaded from other websites or archives.

2. Kudos are best understood as a kind of upvotes or likes, a one-click expression of appreciation.

Metric	Statistic	Value
Number of stories	total	482
Word count	total	369,809
	average per story	767.24
	standard deviation	483.85
	median	674.5
Authors	total	327
	average per author	1.47
	median per author	1
Hits	total	907,996
	median	998.5
Kudos	total	5 2,194
	median	58
Comments	total	2,529
	median	4
Bookmarks	total	5,708
	median	6

Table 1: Descriptive statistics for the corpus

4.2 Method

This paper analyzes connotation frames, a concept first introduced by Richard Fillmore (Fillmore 1976) to describe the conceptual frames that words evoke. For instance, the verb ‘exchange’ implies the existence of a giver, a taker, and a good being exchanged. Connotation frame analysis relies on the assumption that by connecting entities together through a predicate, texts “subtly connote a range of implied sentiments and presupposed facts about the entities” (Rashkin et al. 2016, 311). For example, the sentence “He violates her” casts the entity “he” as a perpetrator and “her” as a victim, and may evoke sympathy or pity. It may even imply that the female entity is valuable or desirable. Connotation frame analysis uses the meanings or connotations implied in agent-verb-theme relationships to assess such linguistic framing of dynamics between textual entities.

In this paper, I use Riveter (Antoniak et al. 2023) to measure the power and agency of two entities, ‘Hades’ and ‘Persephone’, on a large scale. Riveter parses texts, detects and clusters entities, extracts agent-verb-theme triples and matches these against a pre-selected lexicon of connotation frames – in this case, lexicons for power and agency (Sap et al. 2017) – to assign the extracted entities scores in the relevant semantic dimensions. It is worth noting that Riveter’s scoring system can lead entities to be assigned positive or negative scores, because different connotation frames can add or subtract scores. For example, the verb ‘defeat’ increases an entity’s power in relation to a theme, while ‘apologize’ decreases it. Because scores are aggregated at the level of entire stories, it is also possible for these scores to cancel each other out so an entity ends up with a power-score of zero. While power is always calculated through these connotation frames as relational – if I defeat you, my power is increased in relation to yours – agency is not. The agency of an entity can be increased or decreased depending on the verbs they are associated with, independently of other entities in a text. For example, ‘managing’ increases an entity’s agency while ‘waiting’ decreases it.

Advantages of Riveter include its ease of use and the interpretability of its results. Compared to other lexicon-based tools, Riveter also has the benefit that its pipeline takes

grammatical structure into account. One important drawback of Riveter, or any lexicon-based tool, is that the results are only ever as good as the lexicons being applied, and so domain-specific indicators of power or agency may be overlooked. This is perhaps especially troubling since fanfiction presents an online and in some ways subcultural domain of language use. In previous research applying Riveter to fanfiction-texts (Neugarten 2025, forthcoming), a small-scale manual error evaluation conducted by a single annotator found the tool's scoring accurate when detecting power 57% of the time and accurate when detecting agency 89% of the time. Most errors had to do with an inability to detect metaphorical language use and failing to account for the ambiguity of power dynamics in some instances.

On the other hand, the connotation frames for power and agency applied here have been shown to perform well on contemporary movie scripts (Sap et al. 2017), a domain that somewhat similar to language use in contemporary online fanfiction. Following previous research that applied Riveter to fanfiction texts (Yang and Pianzola 2024), I then calculate the power differences between the two relevant entities – Hades and Persephone – for each individual story. I also add a comparison of agency-scores between the two characters.

Another drawback of the Riveter tool is that it reduces gender to a binary variable (male versus female) and assumes that each detected Hades-entity is male while each detected Persephone-entity is female. Fanfiction communities have a relatively large contingent of participants whose gender identity goes beyond the binary (Rouse and Stanfill 2023), and fanfiction correspondingly explores and represents the experiences of fictional characters with nonbinary gender identities more often than mainstream fiction does, even if these explorations are not always explicit (Leetal 2022). However, in *MythFic Metadata*, tags describing gender identities beyond the binary were highly infrequent and not explicitly linked to Hades or Persephone. Unfortunately, distant readings (almost always trade in a measure of granularity or specificity to gain a broader view, and I find the tradeoff acceptable in this case because of the low frequency of tags indicating the presence of nonbinary genders in the stories.

5. Results

5.1 Power and Agency Scores

Table 2 provides descriptive statistics of power- and agency-scores per entity, aggregated over the entire corpus. It also provides descriptive statistics of the power- and agency-differences between Hades and Persephone. For both the power and agency dimensions, these difference-scores were calculated as follows:

$$score^{Hades} - score^{Persephone} = score^{difference}$$

This means that a positive difference indicates that Hades scored higher on power or agency, while a negative difference indicates that Persephone scored higher.

The average scores indicate that both characters have negative power-scores, with Persephone (-0.09) scoring lower than Hades (-0.07). A t-test revealed that this difference

	Power		Agency	
entity	Hades	Persephone	Hades	Persephone
count	141	145	146	146
mean	-0.07	-0.09	0.22	0.21
std	0.36	0.28	0.30	0.27
t-test		0.48		-0.54
p-value		0.63		0.59
	Power Difference		Agency Difference	
count		69		71
mean		-0.03		0.02

Table 2: Descriptive Statistics for Riveter scores

was not statistically significant ($t = 0.48$, $p = 0.63$). In 37 instances, Persephone was assigned more power, while in 29 instances Hades was assigned more power and in 3 instances they were assigned equal amounts of power, canceling each other out and leading to a power difference of 0. Both characters have positive agency scores, with Hades (0.22) scoring slightly higher than Persephone (0.21). Again, the difference was not statistically significant ($t = -0.54$, $p = 0.59$). In 35 instances, Persephone was assigned more agency, while in 33 instances Hades was assigned more agency and again the scores were equal in three stories, cancelling each other out. When looking at the size of the differences, the negative mean power difference (-0.03) indicates that Persephone tends to outdo Hades when it comes to the power dynamics between them, while the positive mean agency difference (0.02) shows that Hades tends to outdo Persephone in terms of agency.³ Keep in mind that the power- and agency difference scores are based only on those stories where both characters were assigned a power- or agency-score so that a difference between the two could be calculated – these difference-scores thus represent a smaller subset of the corpus (only 69 and 71 stories respectively) than the entity-scores.

5.2 Shifts over Time

Existing research on the way literature represents and may impact the social world, and gender dynamics in particular, has hypothesized that the position of female characters in narrative may have shifted over time with regards to empowerment and oppression, because narrative may either reflect or shape the emancipation of women in the real world. However, this emancipatory hypothesis has been rejected, at least for Dutch-language literary novels (Smeets 2024), but also when it comes to increasing the real-world prominence of female authors (Underwood et al. 2018) and their prestige in literary circles (Koolen 2018). Existing scholarship thus seems to indicate that literature is not as progressive in terms of gender politics as academics may like to think.

Fanfiction, however, has a subcultural position outside of the literary establishment, so it is possible that these stories do not adhere to the same patterns when it comes to representing gendered social dynamics as published literature. This raises the question:

3. Another interesting way to compare entities is to examine the distribution of different verbs contributing to the power- and agency-scores for each entity. In previous work applying Riveter to fanfiction about Greek myth (Neugarten 2025, forthcoming) I found no marked gendered differences in verb distribution except with the word 'smile', which contributes positively to an entity's agency and was more often connected to female entities.

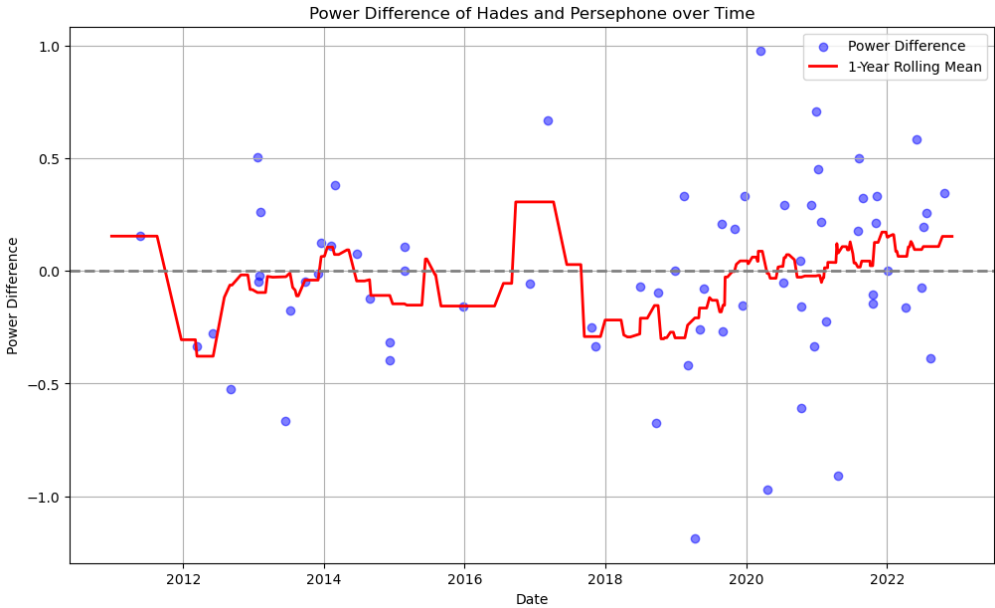


Figure 1: Plot of Power Difference of Hades and Persephone over Time

does the fanfiction case study show a change in gendered dynamics of power and agency over time? On the one hand, it is important to note that the period covered in this fanfiction corpus (2007 – 2022) encompasses a relatively short span of time. On the other hand, during this time a lot has changed when it comes to the reception of the Hades/Persephone relationship in popular culture. At least fifteen retellings of the Hades/Persephone myth aimed at young adults have been published during this period, including the immensely popular web comic *Lore Olympus* (Smythe 2021) which does not portray Hades as a domineering, oppressive or abusive partner, but rather as a soft-spoken, attractive love interest. It is possible that the increasing popularity of these retellings has impacted the dynamics of power and agency portrayed between Hades and Persephone in fanfiction over time.

5.2.1 Power Shifts over Time

Figure 1 presents a plot of all power-differences over time. Each blue dot represents the power difference assigned to a single text, with negative scores indicating that Persephone had more power and positive scores indicating that Hades had more power. The red line indicates the 1-year rolling mean of the power difference. Around 2012, several stories were published with a marked power-difference in favor of Persephone. This is interesting, although it is perhaps too few to speak of a true trend. Around 2019, stories started exhibiting a power difference favoring Hades. This increasing fanfiction production with a power difference favoring Hades indicates that fanfiction is not empowering its central female character, Persephone, more and more over time. In other words, this development is not progressive when it comes to representing gender equality. It will be interesting to see if this trend continues into the future.

5.2.2 Agency Shifts over Time

Figure 2 presents a plot of all agency-differences over time. Each blue dot represents the agency-score difference between Hades and Persephone calculated for a single text,

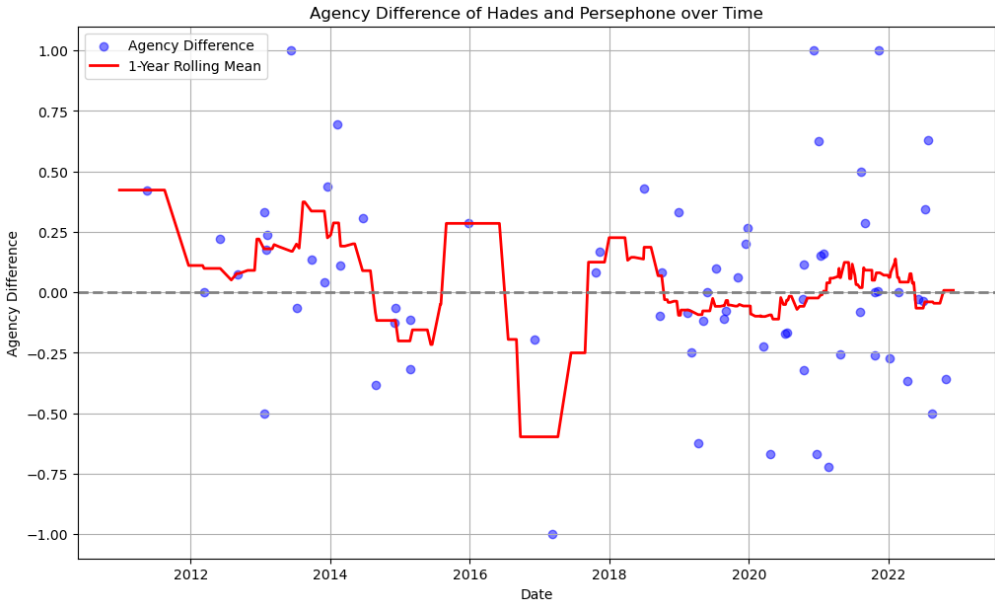


Figure 2: Plot of Agency Difference of Hades and Persephone over Time

with negative scores indicating that Persephone was ascribed a higher agency score by Riveter, and positive scores indicating that Hades was ascribed more agency. The red line indicates the one-year rolling mean agency difference between the two entities. From 2013 to 2015, there was a pattern of decreasing agency difference-scores over time, which is to say that Persephone’s average agency relative to Hades increased over that period, although around 2016 and again around 2018, stories with an agency difference favoring Hades disrupted this trend. In recent years, these mean difference scores have hovered around the zero-line, indicating no strong inequality when it comes to the distribution of agency between Hades and Persephone.

5.3 Comparison to Existing Metadata: Does Power or Agency Correlate with Story Violence?

It is perhaps to be expected that those stories where Riveter detects a large difference in the power or agency between characters are also more likely to contain incidents of violence. After all, violence can be understood as an extreme – usually physical – exertion of power by one entity over another. Violence can also have the effect of limiting the agency of its victim. Fanfiction writers on *Archive of Our Own* tend to attach detailed and accurate content-oriented tags to their stories, making it easy for readers to find stories that fit their tastes and to curate their reading experiences with regards to potentially undesired themes or topics. This tendency to attach detailed metadata (tags) to stories gives scholars in fan studies valuable insight into the story-level content of fanfiction, although these tags are not always very fine-grained. In *MythFic Metadata*, for example, the 5,154 stories for which metadata is provided are accompanied by 1,3936 unique additional tags. A little more than 600 stories in the dataset are not tagged with any additional tags, but for those stories that have been tagged, the tags often provide valuable content-level information about genre classifications, plot elements, characterization or storyworld characteristics. Previous research has used these tags to measure correlations between different types of romantic relationships

	hades_power	persephone_power	power_diff	hades_agency	persephone_agency	agency_diff	physical	noncon	captivity	death
hades_power	nan	0.00	0.40	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00
persephone_power	0.00	nan	-0.30	-0.09	-0.00	-0.21	-0.00	0.00	-0.00	0.11
power_diff	0.40	-0.30	nan	0.00	-0.16	0.26	0.00	0.00	-0.00	0.00
hades_agency	-0.00	-0.09	0.00	nan	0.16	0.28	0.00	0.00	0.00	0.00
persephone_agency	-0.00	-0.00	-0.16	0.16	nan	-0.23	-0.00	0.10	-0.00	0.00
agency_diff	0.00	-0.21	0.26	0.28	-0.23	nan	0.00	-0.00	0.00	0.00
physical	0.00	-0.00	0.00	0.00	-0.00	0.00	nan	0.00	-0.00	0.17
noncon	0.00	0.00	0.00	0.00	0.10	-0.00	0.00	nan	0.28	-0.00
captivity	0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	0.28	nan	-0.00
death	0.00	0.11	0.00	0.00	0.00	0.00	0.17	-0.00	-0.00	nan

Figure 3: Correlations (Spearman’s ρ) between violence and Riveter-scores

(including straight, lesbian and homosexual) and the presence of violence in fanfiction (Neugarten 2024). Using the same tag-based operationalization of violence, I was able to calculate correlations between Riveter’s power- and agency-scores and the presence of tags indicating violence. Because the presence of violence in stories was indicated as a binary variable (1 for violence, 0 for no violence), I calculated Spearman’s ρ correlations to be able to compare this binary data with the Riveter-scores.

I used four of the five types of violence identified in previous work (Neugarten 2024): physical violence, sexual violence, captivity, and death. The specific tags used as a proxy for the presence of these different types of story-violence are listed under supplementary materials (Section 7). I disregard the fifth category of violence – roughness – because evidence suggests that the tags related to this category (rough sex, biting, hair-pulling and spanking) were not reliable indicators of violence that indicate unequal dynamics of power and agency. Instead, these tags were often used to describe consensual sexual acts.

Figure 3 presents Spearman’s ρ correlations (with a significance threshold of 0.05) between the presence of violence as indicated by tags and Riveter-scores.⁴ It is unsurprising that the two characters’ power- and agency-scores correlate with the power- and agency-differences between them,⁵ and that some types of violence correlate positively with each other, such as physical violence with death (0.17) and sexual violence with captivity (0.28). This final correlation is nonetheless interesting, because it points to the existence of stories that reflect the dominant canonical version of the Hades/Persephone myth, in which Persephone is abducted and raped, in the dataset.

Fanfiction featuring death has a weak positive correlation with Persephone’s power scores (0.11) which raises the question of who is dying in these stories. Surprisingly, Persephone’s agency scores are slightly positively correlated with the presence of non-consensual sex acts (0.10). This is counterintuitive for two reasons. Firstly, I expect that Persephone is more often the victim than the perpetrator of sexual violence in the fanfiction corpus because that is also the dynamic most prevalent in the myth’s culturally dominant version. Secondly, I do not associate the narrative or semantic role of being the victim of sexual violence with a high level of agency.

4. Following fan community jargon, sexual violence is labeled ‘noncon’ in Figure 3. This term is short for non-consent or non-consensual sex.
5. Keep in mind that for the difference-scores, a negative score indicates a positive difference in favor of Persephone. This explains the correlation between the agency-difference and Persephone’s agency (-0.23) and also between the power-difference and Persephone’s agency (-0.16).

Power			Agency		
Hades	Persephone	Diff.	Hades	Persephone	Diff.
-0.07	-	-	0.29	-	-
-	0.10	-	-	0.33	-
0.33	-	-	0.33	-	-
-	0.07	-	-	0.31	-
0.17	-0.50	0.67	0	1.0	-1.0
0	-0.11	0.11	0	0.32	-0.32
-0.23	0.09	-0.32	0.14	0.20	-0.06
0.06	-	-	0.44	-	-

Table 3: Power and Agency Scores by Entity for Stories Tagged Non-Con

Closer inspection of the data shows that thirteen stories in the corpus were tagged with sexual violence. Of those, four were not assigned any Riveter scores. The Riveter scores for the remaining nine stories are listed in Table 3. It then becomes evident that Persephone is assigned positive agency scores in six of the stories and positive power scores in three. Some of these stories call into question the assumptions underlying the metadata analysis. Firstly, in some instances, tags indicating sexual violence that is ‘implied’ or ‘past’ – but not literally described in the narrative text – account for the correlations. In other instances, the *direction* of sexual violence is not aligned with what one would expect to see in a patriarchal story world, for example when Persephone is perpetrating sexual violence on Hades instead of vice versa. These stories, in which the expectations of gendered inequality are reversed or the past occurrence of sexual violence is used as a narrative basis for a story of empowerment, may be interesting candidates for close reading in future research.

5.4 Comparison to Existing Omegaverse Analysis: Does Gendered Power Behave Differently than A/B/O Power?

Previous research (Yang and Pianzola 2024) examined the gendered power dynamics of the Omegaverse, a popular subgenre of fanfiction. At the time of writing, *Archive of Our Own* hosts over 240,000 stories tagged with this trope. In the Omegaverse, the culturally dominant division of gender as a male/female binary is expanded because characters have a secondary gender. For this secondary gender, three options exist: alpha (the dominant gender), omega (the submissive gender) and beta (the neutral gender). For this reason, the Omegaverse is also often referred to as A/B/O, an abbreviation of these three genders. In most Omegaverse-stories, A/B/O hierarchy dictates how people interact in erotic and romantic situations, but in some versions their entire social world is structured around this hierarchy. As noted by Milena Popova, “an alternate universe where gender and sexual scripts work radically differently to ours, such as the Omegaverse, is the perfect tool to explore the effect of scripts and dominant ways of thinking on our actions and our ability to meaningfully negotiate consent” (Popova 2021, 58). In other words, the imagined society of the Omegaverse can be a way for fanfiction to engage transformatively with dominant cultural ideas structuring gendered power relations. This also makes it interesting to compare gendered power dynamics between fanfiction that takes place in the Omegaverse and fanfiction in storyworlds that are more similar to real-world societies in their conceptions of gender.

Because Omegaverse-stories imagine gender and its associated power dynamics in a way that is radically different from real-world societies, Yang & Pianzola decided to map the gendered power dynamics between selected Alpha- and Omega-characters in a dataset of Omegaverse fanfiction. They found that gender power difference between Alphas and Omegas can be more-or-less stably detected over time within particular fandoms, although “most fandoms exhibit more within-group consensus when more fans start writing” (Yang and Pianzola 2024, 914). Although the Hades/Persephone dataset used here is smaller than the multi-fandom dataset used by Yang and Pianzola, Figures 4 and 5 suggest a similar trend in this use case, with both power- and agency-differences fluctuating less as more and more fanfiction is written in the fandom year by year.

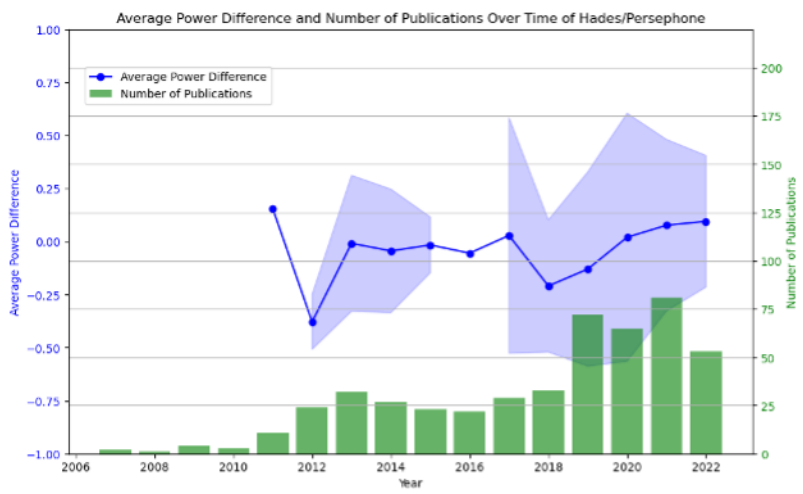


Figure 4: Average Power Difference and Number of Publications over Time

Comparing gender power difference between the Yang and Pianzola study and the case study presented in this paper presents one methodological difficulty: because Persephone and Hades always have the same gender in the corpus used here, I have simplified the calculation of gender power difference to account for this, so that negative power differences indicate a positive power difference for Persephone. Conversely, in the dataset used by Yang & Pianzola, characters can have different A/B/O genders in

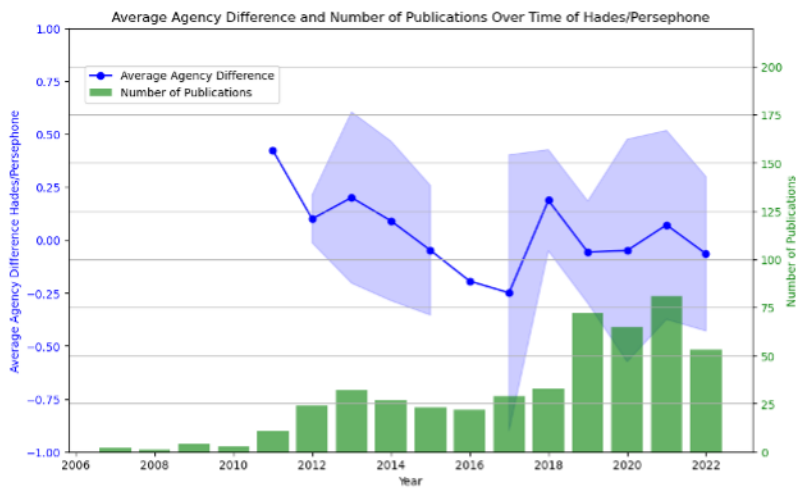


Figure 5: Average Agency Difference and Number of Publications over Time

different stories. In their research, a positive gender power difference always indicates more power for the Alpha (the traditionally dominant partner), while a negative gender power difference means more power for the Omega (the traditionally submissive partner). In terms of the real-world construct of the gender binary, all characters included in their dataset are male.

Table 4 provides descriptive statistics of the gender power differences from the Omegaverse (Yang and Pianzola 2024), with the same statistics for the Hades/Persephone case study for comparison. Because Yang & Pianzola did not measure agency, no comparison can be made on that variable.

Fandom	Relationship	Avg. Diff.	std
My Hero Academia (anime)	Bakugou Katsuki /Midoriya Izuku	0.08	0.23
My Hero Academia (anime)	Katsuki Yuuri / Victor Nikiforov	0.05	0.22
BTS (K-pop)	Jeon Jungkook /Park Jimin	0.07	0.19
BTS (K-pop)	Jeon Jungkook /Kim Taehyung	0.08	0.21
Supernatural (TV)	Castiel/Dean Winchester	0.09	0.25
Hannibal (TV)	Will Graham/Hannibal Lecter	0.29	0.38
Marvel Cinematic Universe	James ‘Bucky’ Barnes/Steve Rogers	0.08	0.22
Marvel Cinematic Universe	Steve Rogers/Tony Stark	0.05	0.25
Greek Myth	Hades/Persephone	-0.03	0.39

Table 4: Average Power Difference Scores for Hades/Persephone Compared to Omegaverse Study

Compared to the A/B/O relationships analyzed by Yang & Pianzola, then, we can conclude that the average power difference between Hades and Persephone is small, and slightly in favor of what may perhaps be called the traditionally less powerful partner: the young girl Persephone. In contrast, all average differences calculated by Yang & Pianzola result in positive scores, pointing to a power difference in favor of the traditionally dominant alpha. It is worth noting that all relationships analyzed by Yang & Pianzola are slash – romantic or sexual relationships between men. In comparison to their dataset, then, it seems that the Hades/Persephone fanfiction analyzed in this paper is slightly less unequal, if we take equality to mean that women are portrayed as powerful main characters in relation to their (male) love interests. On the other hand, the Hades/Persephone relationship has a higher standard deviation than most relationships examined by Yang & Pianzola, suggesting that the gender power difference between them is more variable than for other relationships.

One possible explanation for these observed differences is that the Omegaverse-fanfiction studied by Yang & Pianzola relies for many of its genre conventions on a rigid hierarchy of genders, with alphas (almost) always portrayed as dominating over omegas in every aspect of their interpersonal, sexual, and social interactions. In these genre conventions, the Omegaverse differs from the corpus of Hades/Persephone stories. If interpersonal dynamics between Hades and Persephone are more flexible in relation to genre conventions than between alphas and omegas, this could also explain the higher standard deviation in their power difference-scores.

5.5 Does Power or Agency Correlate with Common Popularity Metrics on AO3?

As described in Section 4.1, *Archive of Our Own* also provides statistics on various popularity metrics for fanfiction, and these metrics are included in *MythFic Metadata*. The different kinds of popularity-related metrics provided on AO3 are listed and described in Table 5.

metric	description
comments	number of times a reader has left a comment after reading a story.
kudos	number of readers who have left Kudos for a story.
bookmarks	number of readers who have bookmarked the story to find it later.
hits	number of times a story has been viewed.

Table 5: Description of popularity metrics

Each of these popularity metrics indicates a different kind and level of engagement. A hit does not necessarily mean that a story has actually been read. Kudos are quickest and easiest to give, and while they communicate a positive evaluation or encouragement to the author, this may not be as engaged as bookmarking a story to revisit it later, which has been described as a “stronger and stickier form of approval than a simple ‘kudos’” (Vadde and So 2024, 24). Finally, a typed-up comment may be considered the strongest indicator of approval. In a previous study on fanfiction comments, I observed that “commenters can be characterized as above-averagely engaged or committed readers, since they invest the time and effort to comment” (Neugarten et al. 2024, 2020). A 2013 census of *Archive of Our Own* also found that only 43.6% of the platform’s users regularly leave comments (centrefortheselights 2013). In this sense, commenters are not necessarily representative of fanfiction’s readership as a whole. Finally, hits are perhaps best understood as indicators of what readers find appealing at first glance or based on a brief description, while kudos, comments and bookmarks are indicative of what readers appreciate after (mostly) reading the full text of a story.

Figure 6 presents Spearman’s ρ correlations (with a significance threshold of 0.05) between popularity metrics and Riveter scores. It is unsurprising that popularity metrics correlate strongly with each other. However, it is notable that Hades’ power scores have a mild negative correlation with three of the popularity metrics: -0.14 for kudos, -0.11 for bookmarks and -0.12 for hits. In other words, stories that represent Hades as more powerful are slightly less popular. The power difference between Hades and Persephone also has a mild negative correlation (-0.10) with both kudos and hits, indicating that greater inequality in the power relation between the characters is less popular among readers. Persephone’s agency correlates positively with many measures of popularity: 0.16 for kudos, 0.15 for bookmarks and 0.11 for hits, so stories that portray her as having more agency tend to be more popular. Hades’ agency also correlates positively with kudos (0.11), though not with other popularity metrics.

These numbers also generate some insight into how power and agency circulate in these stories. For both characters, their own levels of agency and power do not correlate with each other, although Persephone’s agency has a weak positive correlation (0.16) with Hades’ agency, suggesting that some stories ascribe more agency to both characters, and the agency of one need not come at the expense of the agency of the other. Interestingly,

	comments	kudos	bookmarks	hits	hades_power	persephone_power	power_diff	hades_agency	persephone_agency	agency_diff
comments	nan	0.53	0.53	0.40	-0.00	-0.00	-0.00	-0.00	0.00	-0.00
kudos	0.53	nan	0.87	0.87	-0.14	-0.00	-0.10	0.11	0.16	0.00
bookmarks	0.53	0.87	nan	0.78	-0.11	-0.00	-0.00	0.00	0.15	0.00
hits	0.40	0.87	0.78	nan	-0.12	-0.00	-0.10	0.00	0.11	0.00
hades_power	-0.00	-0.14	-0.11	-0.12	nan	0.00	0.40	-0.00	-0.00	0.00
persephone_power	-0.00	-0.00	-0.00	-0.00	0.00	nan	-0.30	-0.09	-0.00	-0.21
power_diff	-0.00	-0.10	-0.00	-0.10	0.40	-0.30	nan	0.00	-0.16	0.26
hades_agency	-0.00	0.11	0.00	0.00	-0.00	-0.09	0.00	nan	0.16	0.28
persephone_agency	0.00	0.16	0.15	0.11	-0.00	-0.00	-0.16	0.16	nan	-0.23
agency_diff	-0.00	0.00	0.00	0.00	0.00	-0.21	0.26	0.28	-0.23	nan

Figure 6: Correlations (Spearman's ρ) between popularity metrics and Riveter-scores

there is no correlation between the power scores for the two characters, suggesting that power dynamics are not actually a zero-sum game. Persephone's power does have a weak negative correlation with Hades' agency (0.09), suggesting that his ability to take actions in a particular storyworld sometimes limits her power in that universe. Furthermore, it stands out that the power difference between the two characters correlates positively with the agency difference (0.26) – when the dynamics in a story are more unequal in terms of power, this correlates with more unequal dynamics in terms of agency.

6. Discussion and Conclusion

6.1 Discussion

Two main areas present themselves as fruitful for future research. Firstly, it is of course possible that gendered power dynamics operate differently in longer stories, especially because their buildup of narrative tension may operate in a different way than a short or very short story does. Future research may therefore want to turn attention to those longer stories about Hades and Persephone – or other characters from mythological narratives or popular fiction – that have not been covered in this paper.

Secondly, as noted in previous research (Neugarten 2025, forthcoming) it is difficult to draw fine-grained conclusions based on Riveter scores without contextualizing and evaluating these scores through close reading. Although detailed close readings fall outside the scope of the current paper, this also presents an interesting avenue for future research.

6.2 Conclusion

To conclude, let me address the research questions raised in Section 2 one by one.

- In short-form fanfiction about the relationship between the mythological characters of Hades and Persephone, how is their power and agency portrayed?

Both Hades and Persephone are portrayed in fanfiction about their relationship as relatively disempowered, because Riveter assigns both entities negative power scores on average (-0.07 for Hades and -0.09 for Persephone). The average difference between the two characters is not large (-0.03). For agency, both characters are assigned low but positive scores (0.22 for Hades and 0.21 for Persephone), and the average difference is even smaller (0.02).

- Is their power dynamic gendered, i.e. does the gender of these characters impact their level of power or agency in the stories? 470 471

In stories where both characters are assigned scores in a given category (meaning that a difference could be calculated) Persephone tends to score higher than Hades on power while Hades tends to score higher than Persephone on agency. This indicates that something transformative is going on in this corpus of fanfiction: the character with many identity-characteristics traditionally associated with weakness or disenfranchisement – youth and femininity – tends to have the power difference in her favor in this small subset of the corpus ($n = 69$). 472 473 474 475 476 477 478

- How does this dynamic shift over time? 479

Around 2019, stories became marked by a mean power difference favoring Hades. This roughly coincides with increased fanfiction-production focusing on the Hades/Persephone relationship. Between 2013 and 2015, there was a pattern of decreasing mean agency differences over time. In other words, Persephone's average agency relative to Hades increased over that period. After that time, this pattern was no longer discernible. It is an interesting focus for future research to look at how these patterns have developed since 2022, the year of data collection for *MythFic Metadata*. Since then, the total number of works of fanfiction in the *Ancient Greek Religion and Lore* fandom on *Archive of Our Own* has more than doubled, from 5,154 to 13,246 at the time of writing, and the total number of stories about the Hades/Persephone relationship has increased from 844 to 1,073. 480 481 482 483 484 485 486 487 488 489 490

- How do power and agency scores compare to existing research (Neugarten 2024; Neugarten and Smeets 2023a,b) on violence and gendered power dynamics in fanfiction about Greco-Roman Antiquity? 491 492 493

It was not surprising to find positive correlations between the different types of violence under analysis. It was surprising, however, that stories tagged with 'death' showed a weak but positive correlation (0.11) with Persephone's power scores, suggesting either that the death in many cases may not have been her own or that death in the context of becoming Hades' partner and the queen of the Underworld can be an empowering experience. It was also surprising to find that stories tagged with non-consensual sex showed a weak positive correlation (0.10) with Persephone's agency scores, prompting a closer examination of what was going on in those stories. 494 495 496 497 498 499 500 501 502

- How do these scores compare to existing research (Yang and Pianzola 2024), on power dynamics in Omegaverse-stories, a popular subgenre of fanfiction that presents a speculative conception of gender with clear power hierarchies? 503 504 505

Compared to the power dynamics present in the Omegaverse-fanfiction studied by Yang & Pianzola, the power difference between Hades and Persephone was relatively small. The average power difference between Hades and Persephone was -0.03 while the scores reported by Yang & Pianzola – which reflect the gender power difference between the fictional genders of alpha and omega rather than those between men and women – ranged from 0.05 to 0.29. It thus seems that the fanfiction analysed here was less unequal than the stories about the various 506 507 508 509 510 511 512

alpha/omega relationships studied by Yang & Pianzola, provided we take Riveter scores as indicative of gender power difference and consider a smaller gender power difference to be less unequal than a larger difference.

- Do differing depictions of gendered power relations in this corpus of fanfiction impact the popularity of the stories among their readers?

What stands out in this regard is that a higher power score for Hades is negatively correlated with a number of popularity metrics. These stories receive fewer hits (-0.12), suggesting that fanfiction-readers are less likely to click on them. They also receive fewer kudos than other stories (-0.14), suggesting that readers are less likely to compliment these stories' authors and express their enjoyment, and are bookmarked less often, suggesting readers are less likely to want to revisit these stories. Conversely, stories' popularity correlates positively with Persephone's agency score. This applies to hits (0.11) – suggesting these stories are more often clicked on – kudos (0.16) – suggesting these stories are more often liked – and bookmarks (0.15) – suggesting these stories are more often revisited.

What, then, can we conclude about fanfiction's capacity to transform the culturally dominant gendered power dynamics of the Hades/Persephone myth, and perhaps more broadly its capacity to transform culturally dominant discourses – related to gender, but also other topics – through rewriting?

On average, we see power differences favoring Persephone and agency differences favoring Hades. This suggests that there is a pattern of difference structuring the distribution of power and agency between these two characters, although more research would be needed to determine whether these patterns of difference can be considered representative of each character's gender. In comparison to the (fictional) genders 'alpha' and 'omega', power differences between these two (male and female) characters are small. Trends in the dynamics of power and agency between Hades and Persephone became visible over time, and point to some interesting avenues for future research.

The most interesting finding, perhaps, is that stories with more power assigned to Hades are significantly less popular, while stories with more agency assigned to Persephone are significantly more popular. In the end, fanfiction is a reflection of the kinds of stories fans are most interested and invested in, and in these correlations we can see a desire from the fanfiction-reading audience for gendered power relations to be less unequal. It is clear that fanfiction has the capacity to be transformative of unequal power dynamics between male and female characters like Hades and Persephone, though not all stories are. The stories that are transformative, however, are rewarded with more appreciation and engagement from their readership.

7. Supplementary Materials: Description of Metadata Tags Used to Operationalize Violence

The metadata tags used to operationalize violence in this paper were taken from existing research (Neugarten 2024), which identified five categories of violence based on the most frequently-occurring additional tags in *MythFic Metadata*: physical violence, sexual

violence, roughness, captivity, and death. Four of those categories were used in this paper, and the tags used to operationalize them are as follows:

Physical Violence: Canon-Typical Violence, Violence, Blood, Blood and Violence, Non-Graphic Violence, Minor Violence, Torture, Cannibalism, Pain, Implied/Referenced Torture, Past Abuse.

Sexual Violence: Implied/Referenced Rape/Non-con, Incest, Dubious Consent, Sibling Incest, Rape/Non-con Elements, Past Rape/Non-con, Rape, Bestiality, Gang Rape, Mildly Dubious Consent, Implied/Referenced Incest.

Captivity: Kidnapping, Abduction, Captivity, Imprisonment.

Death: Death, Implied/Referenced Character Death, Minor Character Death, Murder, Temporary Character Death, Past Character Death.

8. Data Availability

To protect the privacy and copyright of authors in the fanfiction community, full-text works of fanfiction are not made available for reuse. However, the metadata set collected by Neugarten and Smeets (2023a,b) is available here: https://data.ru.nl/collecties/ru/rich/mythfic_metadata_dsc_550.

9. Software Availability

Code and derived data are available at: <https://github.com/julianeugarten/CCLS2025>.

References



- Anonymous (1914). "Homeric Hymn to Demeter". Trans. by Hugh G. Evelyn-White. In: *The Homeric Hymns and Homerica with an English Translation by Hugh G. Evelyn-White*. <https://people.uncw.edu/deagona/women%20F12/HH%20Dem%20w%20DQ.pdf>.
- Antoniak, Maria, Anjalie Field, Jimin Mun, Melanie Walsh, Lauren Klein, and Maarten Sap (July 2023). "Riveter: Measuring Power and Social Dynamics Between Entities". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Toronto, Canada: Association for Computational Linguistics, 377–388. 10.18653/v1/2023.acl-demo.36. <https://aclanthology.org/2023.acl-demo.36>.
- Bracke, Evelien (2025). *Persephone and the Beast: The construction of female power in Girl Fantasy literature*. preprint.
- Busse, Kristina (2017). "Intimate intertextuality and performative fragments in media fanfiction". In: *Fandom: Identities and communities in a mediated world*. Ed. by Jonathan Gray, Cornel Sandvoss, and C. Lee Harrington. New York University Press, 45–59.
- centreoftheselights (2013). *AO3 Census: Site Use*. <https://archiveofourown.org/works/16988355/chapters/39932727>.
- Click, Melissa A., ed. (2019). *Anti-fandom: dislike and hate in the digital age*. Postmillennial Pop. New York: New York University Press.




- Fillmore, Charles J (1976). "Frame semantics and the nature of language". In: *Annals of the New York Academy of Sciences* 280.1, 20–32. 592 593
- Floegel, Diana (2020). "'Write the story you want to read': world-queering through slash fanfiction creation". In: *Journal of documentation* 76.4, 785–805. 10.1108/JD-11-2019-0217. 594 595 596
- Fowler, Megan Justine (2019). "Rewriting the school story through racebending in the Harry Potter and Raven Cycle fandoms". In: *Transformative Works and Cultures* 29. 597 598 599 10.3983/twc.2019.1492.
- Gloyn, Liz (2019). *Tracking classical monsters in popular culture*. London: Bloomsbury Academic. 600 601
- Hodkinson, Stephen (2022). "Spartans on the Capitol: recent far-right appropriations of Spartan militarism in the USA and their historical roots". In: *Classical Controversies: Reception of Graeco-Roman Antiquity in the Twenty-First Century*. Ed. by Kim Beerden and Timo Epping. Leiden: Sidestone Press, 59–84. 602 603 604 605
- Jenkins, Henry (2013). *Textual poachers: television fans and participatory culture*. Updated 20th anniversary ed. New York: Routledge. 606 607
- Koolen, Cornelia Wilhelmina (2018). *Reading beyond the female: the relationship between perception of author gender and literary quality*. ILLC dissertation series. Amsterdam: Institute for Logic, Language and Computation, University of Amsterdam. 608 609 610
- Leetal, Dean (Sept. 2022). "Revisiting gender theory in fan fiction: Bringing nonbinary genders into the world". In: *Transformative Works and Cultures* 38. 10.3983/twc.2022.2081. <https://journal.transformativeworks.org/index.php/twc/article/view/2081>. 611 612 613 614
- Levine, Caroline (2015). *Forms: Whole, rhythm, hierarchy, network*. Princeton: Princeton University Press. 615 616
- Lothian, Alexis and Mel Stanfill (Sept. 2021). "An archive of whose own? White feminism and racial justice in fan fiction's digital infrastructure". In: *Transformative Works and Cultures* 36. 10.3983/twc.2021.2119. <https://journal.transformativeworks.org/index.php/twc/article/view/2119>. 617 618 619 620
- McCarter, Stephanie (2022). "Breasts, Shame, and Disgust in English Translations of Ovid's *Metamorphoses*". In: *The Classical Outlook* 97.4, 148–158. <https://www.jstor.org/stable/27191941>. 621 622 623
- Mellenthin, Jessica and Susan O. Shapiro (2017). *Mythology Unbound: An Online Textbook for Classical Mythology*. <https://uen.pressbooks.pub/mythologyunbound/>. 624 625
- Müller, Julia (2022). "Pop Culture against Modernity: New Right-Wing Movements and the Reception of Sparta". In: *Classical Controversies: Reception of Graeco-Roman Antiquity in the Twenty-First Century*. Ed. by Kim Beerden and Timo Epping. Leiden: Sidestone Press, 103–22. 626 627 628 629
- Neugarten, Julia (2021). "Brittle: Re-thinking Narratives of Disordered Eating through Fanfiction". In: *Frame: Journal of Literary Studies* 34.2. <https://www.frameliteraryjournal.com/34-2-writing-the-mind/34-2-julia-neugarten/>. 630 631 632
- (2024). "MythFic Metadata: Gendered Power Dynamics in Fanfiction about Greek Myth". In: *Digital Humanities Benelux Journal* 6. 10.5281/zenodo.14169050. https://journal.dhbenelux.org/wp-content/uploads/2024/11/8_Neugarten_individual.pdf. 633 634 635 636
- (2025). "Using Riveter to Map Gendered Power Dynamics in Hades/Persephone Fanfiction". In: *Transformative Works and Cultures*, forthcoming. 637 638

- Neugarten, Julia, Tess Dejaeghere, Pranaydeep Singh, Amanda Robin Hemmons, and Julie M. Birkholz (2024). "Catching Feelings: Aspect-Based Sentiment Analysis for Fanfiction Comments about Greek Myth". In: *CHR 2024: Computational Humanities Research Conference*. Aarhus, Denmark, 217–231. <https://ceur-ws.org/Vol-3834/paper23.pdf>.
- Neugarten, Julia and Roel Smeets (Apr. 2023a). *MythFic Metadata: Exploring Gendered Violence in Fanfiction about Greek Mythology*. 10.34973/2MYE-8468. https://data.ru.nl/collections/ru/rich/mythfic_metadata_dsc_550.
- (May 2023b). "MythFic Metadata: Exploring Gendered Violence in Fanfiction about Greek Mythology". In: *DH Benelux 2023*. Brussels, Belgium. 10.5281/ZENODO.7941533. <https://zenodo.org/record/7941533>.
- Ovid (2010). *Ovid Metamorphoses*. Trans. by Charles Martin. Norton Critical Edition. New York: WW Norton Company.
- Pande, Rukmini (2018). *Squeezed from the margins: fandom and race*. Fandom & culture. Iowa City: University of Iowa Press.
- Popova, Milena (2021). *Dubcon: fanfiction, power, and sexual consent*. Cambridge, Massachusetts: The MIT Press.
- Radiolorian (2022). *AO3Scraper*. <https://github.com/radiolorian/AO3Scraper>.
- Rashkin, Hannah, Sameer Singh, and Yejin Choi (Aug. 2016). "Connotation Frames: A Data-Driven Investigation". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, 311–321. 10.18653/v1/P16-1030. <https://aclanthology.org/P16-1030/>.
- Rouse, Lauren and Mel Stanfill (Feb. 2023). *Over*Flow: Fan Demographics on Archive of Our Own – Flow*. <https://www.flowjournal.org/2023/02/fan-demographics-on-aos/>.
- Sap, Maarten, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi (2017). "Connotation Frames of Power and Agency in Modern Films". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2329–2334. 10.18653/v1/D17-1247. <http://aclweb.org/anthology/D17-1247>.
- Scott, Suzanne (2019). *Fake Geek Girls: Fandom, Gender, and the Convergence Culture Industry*. New York: New York University Press.
- Smeets, Roel (Dec. 2024). "Emancipatie en de roman. Vrouwelijke personages in Nederlandstalige romans tussen 1960 en 2010". In: *Nederlandse Letterkunde* 29.2, 180–210. 10.5117/NEDLET.2024.2.003.SMEE. <https://www.aup-online.com/content/journals/10.5117/NEDLET.2024.2.003.SMEE>.
- Smythe, Rachel (2021). *Lore Olympus*. Vol. 1. London: Del Rey.
- Stanfill, Mel (2024). *Fandom is ugly: networked harassment in participatory culture*. Critical cultural communication. New York: New York University Press.
- Thomas, Ebony Elizabeth and Amy Stornaiuolo (2019). "Race, storying, and restorying: What can we learn from Black fans?" In: *Transformative Works and Cultures* 29. 10.3983/twc.2019.1562.
- Tosenberger, Catherine (2014). "Mature Poets Steal: Children's Literature and the Unpublishability of Fanfiction". In: *Children's Literature Association Quarterly* 39.1, 4–27. 10.1353/chq.2014.0010. http://muse.jhu.edu/content/crossref/journals/childrens_literature_association_quarterly/v039/39.1.tosenberger.html.

- Underwood, Ted, David Bamman, and Sabrina Lee (2018). "The Transformation of Gender in English-Language Fiction". In: *Journal of Cultural Analytics* 3.2. 10.22148/16.019. <https://culturalanalytics.org/article/11035>.
- Vadde, Aarthi and Richard Jean So (Mar. 2024). "Fandom and Fictionality after the Social Web: A Computational Study of AO3". In: *MFS Modern Fiction Studies* 70.1, 1–29. 10.1353/mfs.2024.a921546. <https://muse.jhu.edu/article/921546>.
- Váña, Jan (2025). "Sally Rooney and Cultural Sociology of Literature: Towards Epistemological Symmetry within Literature-and-Society Research". In: *Journal of Literary Theory* 19.1, 28–55.
- Wills, Emily Regan (2013). "Fannish discourse communities and the construction of gender in The X-Files". In: *Transformative Works and Cultures* 14. 10.3983/twc.2013.0410.
- Yang, Xiaoyan and Federico Pianzola (2024). "Exploring the Evolution of Gender Power Difference through the Omegaverse Trope on AO3 Fanfiction". In: *CHR 2024: Computational Humanities Research Conference*. Aarhus, Denmark, 906–916. <https://ceur-ws.org/Vol-3834/paper27.pdf>.
- Zuckerberg, Donna (2018). *Not all dead white men: classics and misogyny in the digital age*. Cambridge, Massachusetts: Harvard University Press.

The Outward Turn: Geocoding the Expansion of Fictional Space in Russian 19th Century Literature

Daniil Skorinkin ¹ 
Boris Orekhov ^{2, 3} 

1. Digital Humanities Network, University of Potsdam , Potsdam, Germany.
2. Laboratory for Digital Research of Literature and Folklore, Institute of Russian Literature , Saint Petersburg, Russia.
3. School of Linguistics, HSE University , Moscow, Russia.

Citation

Daniil Skorinkin and Boris Orekhov (2025). "The Outward Turn: Geocoding the Expansion of Fictional Space in Russian 19th Century Literature". In: *CCLS2025 Conference Preprints 4* (1). [10.26083/tuprints-00030144](https://doi.org/10.26083/tuprints-00030144)

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-02-10

Keywords

geocoding, maps, Russia, 19th century, prose, novel

License

CC BY 4.0 

Reviewers

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract.

We examine the large-scale geospatial dynamics of Russian prose literature in the 19th century. Specifically, we analyze how the distribution of location mentions shifts from the early 19th-century romantic era to the late 19th-century realist period. We demonstrate how realist literature, with its emphasis on portraying 'typical characters in typical settings', moves away from the historical (and often heavily mythologized) landscapes of Russia, Poland, Ukraine, and the Baltics. Instead, it increasingly focuses on the then-new capital, Saint Petersburg, as well as Western Europe and the expanding eastern and southern peripheries of Russia, reflecting the country's ongoing military and economic expansion.

1. Introduction

Of all 'distant reading' methods, geocoding is the one that most tangibly embodies the 'distance' metaphor. With maps, one can literally zoom in and out of vast research material, possibly consisting of thousands of texts, all laid out on a geographic projection of the Earth, and produce conclusions, generalizations, and interpretations on a grand scale.

This does not mean that every geocoding of literature is always meaningful - as Döring (Döring 2013) put it, 'the benefit of any map of literature has to be that it visualizes things that would otherwise remain invisible' and for some literary maps, "[t]here seems to be hardly any analytical value in" them. Literature reduced to dots, lines, and polygons (the basic units of any map) loses most of its inner complexity, and there is always the danger of throwing the baby out with the bathwater. But at the same time, reduction is exactly what gives strength to any modelling attempt in research: only by reducing the complexity and detail, we can see the large drifts of literary movements and the long dynamics of cultural development that is not inferable from close reading of a selection of 'significant' texts.

In our work, we apply mapping and geocoding to study the large-scale geospatial dynamics of Russian prosaic literature over the course of the 19th century, a time when a Russian novel became a global cultural phenomenon through the works of Gogol, Tolstoy,

Turgenev, Dostoevsky and other authors. We analyse the changes in the distribution of mentions of geotaggable toponyms between the two extremely important periods of Russian literature: the early 19th-century romantic era and the late 19th-century realist period. We show how realist literature with its tendency to depict 'typical' characters in 'typical' settings (Fridlender 1971, 105) and not shy away from 'ordinary' and 'average', turns from the mythologized landscapes of historical Russia, Ukraine, Poland and the Baltics, to Western Europe, the then-new northern capital and trading outpost of Saint-Petersburg, and the 'new' eastern and southern peripheries of Russia as the country continues its military, cultural and economic expansion in all directions.

2. Corpus and research design

As members of the PyZeta team put it in the description of their project, "[t]he methodological and epistemological paradigm of comparison is deeply rooted in the Humanities" (Du et al. 2025). In our experience, a research endeavour in computational literary studies typically benefits from having a clear two-sided comparison. Even if such comparison comes at a price of some simplification. Therefore we chose to structure our research around the comparison between the prosaic works of Russian 19th-century realism and the Russian romantic prose that preceded it.

The problem of defining realism in literature is a long-standing one. As Fanger (Fanger 1998, 3) put it, "few literary terms have suggested more and signaled less than 'realism'". Realism often seems too broad a term, combining too many things that lack a common denominator. To quote Molly Brunson (Brunson 2016, 2), "this monolithic presence of realism more often than not splinters into equivocation or endless classification. It is little wonder, given the dizzying array of objects that must crowd beneath this singular term". And for scholars of Russian literature, this was additionally complicated by the ideologically charged understanding of realism in the Soviet era, which led to a strong aversion to the term in the post-Soviet times (see e.g. Vdovin et al. 2020). However, Vdovin et al. (Vdovin et al. 2020) also show that removing the term completely, while continuing to talk about romanticism, classicism and other traditionally labeled literary movements, does not seem feasible either. It is therefore reasonable to keep using it, acknowledging the ambiguity and inner contradictions of the term.

Luckily, in this particular academic endeavour we neither intended nor needed to answer the 'what is realism' question. For us, it was enough to adopt a functional definition that would allow us to make a split in a collection of Russian prose (without any prior genre or literary movement markup) and obtain a 'realistic enough' corpus for computational analysis. We therefore followed the chronological approach. In many cases Russian literary realism is defined as something that started in the 1840es with the projects of the so-called Natural School (Brunson 2016) or more specifically mid-1840es (Bowers 2022, 2). 1845 was the year of the publication of the 'The Physiology of Saint-Petersburg' (Физиология Петербурга), the first artistic manifesto of the Natural School, compiled by Nikolai Nekrasov. In 1846 the second one, 'The Petersburg Collection' (Петербургский сборник) was published by Nekrasov. The first one was a compendium of short 'physiological sketches' by Vissarion Belinskiy, Nikolai Nekrasov, Dmitry Grigorovich, Vladimir Dal, and the Ukrainian writer Yevhen Hrebinka. The

second, bigger volume contained the first large novel by Fyodor Dostoevsky ('The Poor Folk'), as well as texts by Ivan Turgenev, Alexander Herzen, Ivan Panayev, Apollon Maikov and Vladimir Odoevsky. Also in 1846 Belinsky, the most prominent Russian critic of the era, called for new literature that "dealt with life and reality in their true light". We chose to adopt the year 1845 as the 'starting date' of the realist period in our corpus.

As the end of the clearly realist period we selected 1890. This year is frequently named as the starting point of modernism in Russian literature (Douglas Clayton 2016, Ioffe 2009). Of course, there were many realistic works created after that date too (realism never really 'stopped' the same way e.g. classicism did), but without any reliable metadata we had to rely on temporal borders and chose to stop at 1890 to keep modernism out of our corpora.

In the end, having consulted with a number of specialists in Russian 19th-century prose, we received their blessing to consider for the purposes of our quantitative investigation everything written between the years 1845–1890 to belong to realism and everything that fell into the period between 1800 and 1840 to belong to romanticism. We are aware, of course, of how imprecise this division is. However, as Algee-Hewitt et al. (Algee-Hewitt et al. 2018) put it, "[d]irty hands are better than empty," so we continued our research, hoping that the size of the corpus would rectify the lack of quality in our crude criterion for the split.

To compose a corpus of Russian 19th-century prose for our study, we used two main sources of texts. One of them was the 'Corpus of Russian narrative prose' by Oleg Sobchuk, published in the Open Repository on Russian Literature and Folklore (Sobchuk and Lekarevich 2025). Another source, also published in the same repository, was the corpus of the 'Forgotten novels of Russian writers from the collections of the Pushkin House (1857–1917)' by Elena Kazakova (Kazakova 2024). We then filtered out everything that was written outside of the periods we were interested in (1800–1841 and 1845–1890). Our resulting corpus consists of 506 texts between the years 1800 and 1890, of which 96 belong to the romantic subcorpus and 421 – to the realist subcorpus. The list of all texts and their dates of publication is available here.¹

While this corpus is far from being a comprehensive source of Russian 19th-century literary heritage, it contains prosaic work by all the well-known authors of the period (Pushkin, Lermontov, Gogol, Turgenev, Goncharov, Dostoevsky, Tolstoy), as well as a large number of lesser known writers. The total number of word tokens in the corpus is 46.4 mln.

3. Methods

Geocoding literary texts to explore the relationship between literature and geography has a long tradition that spans more than a century. As early as 1910s (Bartholomew 1914) one can find numerous literary maps based on the works of Balzac, Dickens, Dumas, George Eliot, and other authors. In the field of Digital Humanities, the application of geocoding in literary studies has been notably championed by Franco Moretti in his

1. https://github.com/DanilSko/mapping_russian_prose

book ‘The Atlas of the European Novel’ (Moretti 1999). He stated that “geography is not an inert container, is not a box where cultural history ‘happens’, but an active force that pervades the literary field and shapes it in depth.” Mapping literature, according to Moretti, makes visible “the connection between geography and literature” and reveals “significant relationships that have so far escaped” scholarly attention. Through a series of case studies, he examined the geographical dimensions of 19th-century European literature, highlighting the prominence of Paris in French novels, contrasting depictions of urban and rural environments in English literature, and analyzing representations of the Russian landscape in the works of authors such as Dostoevsky and Tolstoy. Since Moretti’s work, there has been, as Döring described it, “a small boom in maps of literature” (Döring 2013). In Bodenhamer et al. 2010 the emergence of the Spatial Humanities was proclaimed, stating that by 2010, there had been “wide application” of Geographic Information Systems (GIS) to “historical and cultural questions.” Multiple scholars have contributed to this growing field. To cite just a few examples, Döring (Döring 2013) examined the toponymy of Berlin in German literature after 1989; Kuzmenko and Orekhov (Kuzmenko and Orekhov 2016) geocoded the Russian national poetic corpus and analyzed the frequency of references to countries and cities; and Barbaresi (Barbaresi 2018) mapped the satirical literary magazine *Die Fackel*. More recent example is the paper by Wilkins et al. (Wilkins et al. 2024), who mapped the geographies of American fictional books and compared them to those found in non-fiction texts.

In all those recent cases, NER tools were used to extract toponyms. We followed these footsteps and for the initial location extraction we utilized Natasha (natasha/natasha 2024), a natural language processing library for the Russian language with NER toolkit. We extracted approximately 12,000 unique locations from our corpus, which were then manually filtered to eliminate evident homonyms. Specifically, we excluded toponyms frequently used as surnames in our corpus (e.g., ‘Rostov’, which in 90% of the cases was the surname of one of the member of the Rostov family in *War and peace*) and those typically employed metaphorically (e.g., ‘Babylonian’).

The filtered toponyms were subsequently geocoded using the ‘wikipedia’ Python package. Geocoding helped us remove duplicates: different spellings and word forms of the same city (e.g. Saint-Petersburg can be spelled in at least 6 different variants in our corpus), country, or river were merged based on the matching coordinates. Thus, the pair of latitude and longitude became the primary ID of each location that we analysed. Extracted and geocoded locations are available here.²

We then produced symbol maps that overlay frequencies in the texts onto geographical locations. We analysed the raw frequencies of locations and their relative increase or decrease in frequency between the periods of romanticism and realism.

It is important to note that our analysis of geographical material focused not on where events take place but on all mentions of place names. This highlights the writers’ and Russian society’s attention to these parts of the world.

Additionally, to compare contexts of the same toponyms in the two periods we used word2vec (Mikolov et al. 2013). With this we attempted to detect the contextual change for some of the most frequent locations that were found in both corpora.

2. https://github.com/DanilSko/mapping_russian_prose

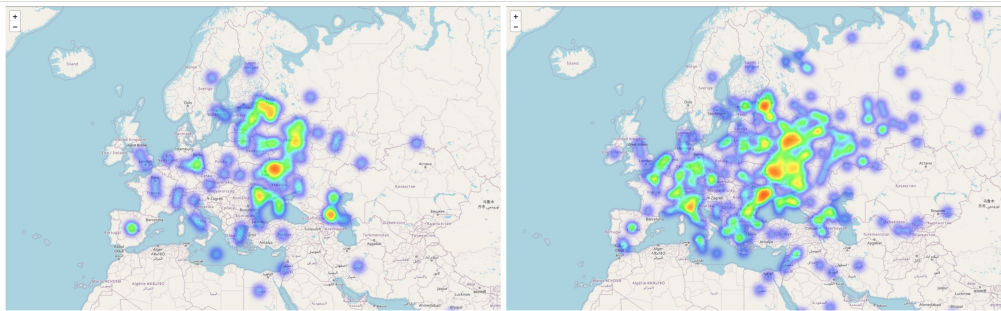


Figure 1: A heatmap depicting location frequencies in romantic (left) and realist (right) texts, visualized through surface occupied and color intensity. Focused on Eurasia.

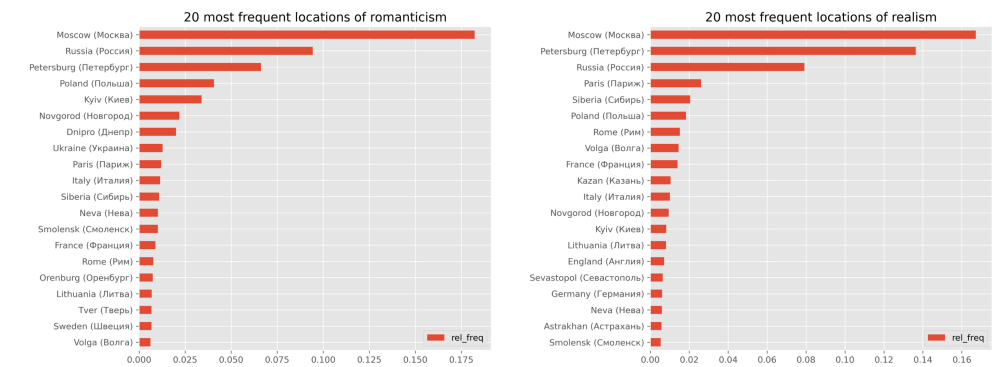


Figure 2: Top 20 locations by relative frequency in romantic (left) and realist (right) texts.

4. Results

A comparison of the geographical distribution of locations in the romanticist and realist corpora reveals a discernible shift. Figure 1 shows two heatmaps reflecting the frequencies of geotagged locations in both corpora.

This visualisation already demonstrates certain key differences, such as relatively more attention to Western Europe in the realist period, as well as a bigger relative presence of Saint Petersburg. However, it is hard to analyze such heatmaps in detail. Figure 2 contains a more traditional bar plot diagram with the top 20 locations by relative frequency in each of the two subcorpora, providing a more detailed zoom into their differences.

Although Moscow is the most frequently mentioned location in both corpora, its dominance significantly diminishes in the realist texts. In the romanticist corpus, Moscow's mentions surpass those of Saint Petersburg by a factor of 2.5, whereas in the realist corpus, Saint Petersburg's mentions are only 20% fewer than Moscow's.

The emergence of Saint Petersburg as a prominent location is unsurprising; it serves as a primary setting for many significant Russian novels of the realist period. Dostoevsky's 'Crime and Punishment' and other works, Goncharov's 'An Ordinary Story' and 'Obломov,' as well as Tolstoy's 'Anna Karenina' and 'War and Peace', are set in Saint Petersburg. Additionally, many lesser-known works of Russian realism are set there. Russian literary tradition often attributes to realism a focus on depicting 'typical' characters in 'typical' settings, and these characters were frequently situated in the then-capital and administrative hub of Saint Petersburg.

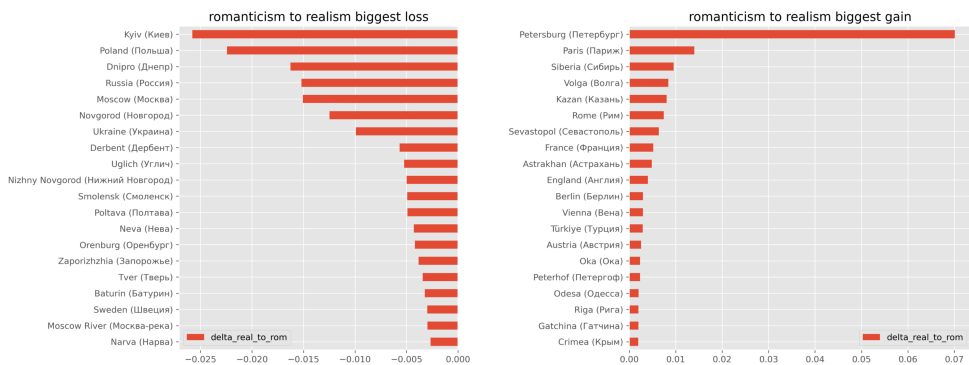


Figure 3: Top 20 biggest relative loss (left) and relative gain (right) from romanticism to realism.

A second significant shift in the realist corpus is the diminished prominence of Kyiv and other Ukrainian locations. While Kyiv ranked fifth in the romantic corpus, it declines to twelfth place in the realist texts, being surpassed not only by Western capitals such as Paris and Rome but also by peripheral Russian locations, including Kazan, the Volga River, and Siberia. Similarly, other Ukrainian locations, such as the Dnipro River and Poltava, exhibit a noticeable decrease in relative frequency.

To systematically capture these changes and emphasize the locations that underwent the most substantial shifts, we calculated the relative overall change in location frequencies and ranked them accordingly. This approach enables a clear visualization of the locations that experienced the most pronounced relative increase or decrease in the realist subcorpus compared to the romantic corpus. The corresponding ranking is presented in Figure 3.

Among the locations that experienced the most significant decline in frequency during the transition from the romantic to the realist period (Figure 3, right), a distinct group comprises Ukrainian toponyms, including Kyiv, Dnipro, Poltava, Baturin, and Zaporizhzhia. This category can be further extended to include neighboring Polish and Baltic locations (Poland, Narva). These Ukrainian, Polish, and Baltic territories — historically contested regions of Eastern Europe — played a crucial role in the literary landscape of Russian historical fiction.

Key historical events, such as the Time of Troubles (Smuta), the Polish–Russian War of 1605–1618, the Cossack uprisings against Polish and Russian rule, and the Great Northern War of 1700–1721 (which accounts for the inclusion of Poltava and Narva), unfolded largely within the territories of present-day Ukraine and the Baltic states. These events provided a rich source of inspiration for numerous Russian-language authors of the romantic era, including Alexander Pushkin, Nikolai Polevoy, Mikhail Zagoskin, Faddey Bulgarin, and Fyodor Glinka. Within their works, these contested lands of Eastern Europe function similarly to Scotland in the novels of Walter Scott, serving as a backdrop for narratives of conflict, heroism, and national identity.

Another significant group of locations prominent during the romantic era but less favored during the realist period includes historical cities of Central Russia, such as Novgorod (the capital of the Novgorod Republic and a popular ‘unrealized alternative’ to monarchical centralized Moscow), Uglich (known for the death of Tsarevich Dmitry,

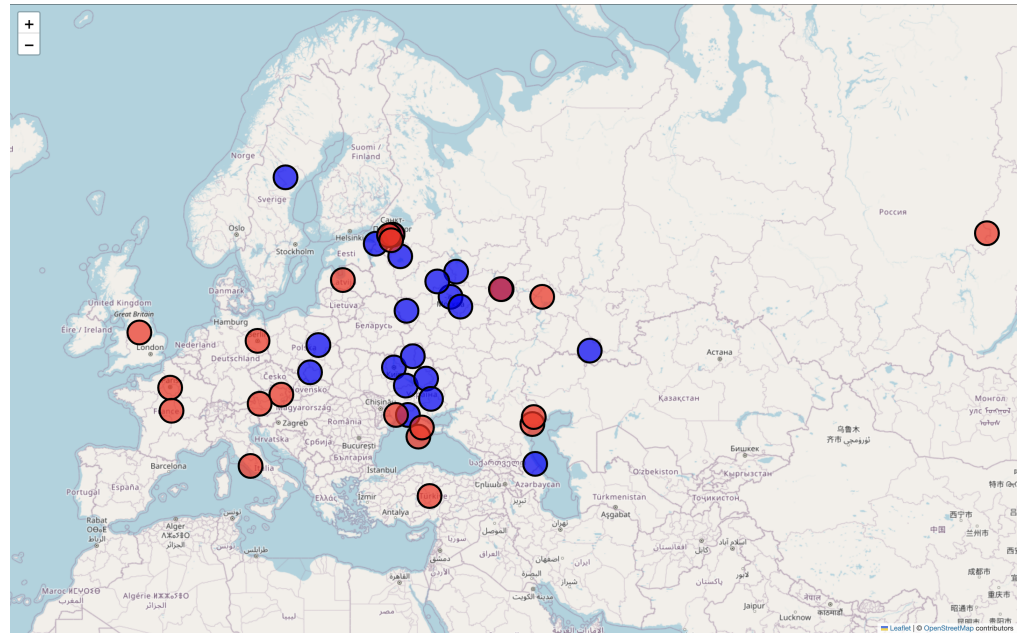


Figure 4: Top-20 locations with the biggest loss (blue) and the top 20 locations with the biggest gain (red) in the realist subcorpus as compared to the romanticist one.

a pivotal event in Russian history), and Moscow itself. Both Moscow and Kyiv, which were among the most frequently depicted locations during the romantic period, lost their literary prominence to Saint Petersburg as Russian literature shifted its focus from a romanticized past to the contemporary present.

Realist literature, oriented towards the present, also shifted its geographical focus westwards — away from the historically contested lands of Eastern Europe and the Baltics, towards the Western European capitals (Paris, Rome, Berlin, Vienna) and countries (France, England, Austria, Switzerland). The characters of realist novels no longer engage in battles in Poland, Lithuania, or Ukraine; instead, they travel to and from France, Italy, or Switzerland, often by train, much like the protagonist of Dostoyevsky's 'The Idiot' or characters of Tolstoy's 'Anna Karenina'.

Concurrently with the westward shift, there was an eastward expansion in literature. In the 19th-century, Russia was actively colonising territories in the Volga region, the Urals, and beyond into Siberia. Notably, "Siberia" exhibits the third largest relative increase in frequency within the realist subcorpus, following only Paris and Saint Petersburg. Other prominent locations in this context include the Volga, the Urals, Kazan, Ufa, and Saratov. While some of these locations possess historical significance, during the realist period they were primarily associated with new economic development. At the same time, these remote places play a bigger role in the ever-growing wave of literature dealing with the topics of prison, penal labour system (katorga) and penal exile of political prisoners, typically members of revolutionary movements.

In Figure 4 we mapped the 20 locations that saw the biggest loss (blue) and biggest gain (red) in their relative frequency in transition from romanticism to realism.

Figure 4 demonstrates that the overall picture is clearly that of an expansion. With the advent of realism, Russian literature transitions from its historical roots in East

	Romanticism	Realism
Saint Petersburg	Moscow (Москва) a village (деревня) a city/town (город) Simbirsk (Симбирск) Kabarda (Кабарда) service (as in army service, government service) (служба) a capital (столица) Kursk (курск) Siberia (Сибирь) to practice (упражняться)	Moscow (Москва) a university (университет) Paris (Париж) a province (провинция) a village (деревня) a grammar school (гимназия) a city/town (город) a capital (столица) Germany (Германия) Kyiv (Киев)
Moscow	Petersburg (Петербург) a city/town (город) a village (деревня) Simbirsk (Симбирск) empty (obsolete) (порожний) an army (армия) kursk (Курск) Kabarda (Кабарда) to practice (упражняться) a tavern (трактир)	Petersburg (Петербург) a city/town (город) a village (деревня) Paris (Париж) Kyiv (Киев) a capital (столица) Petersburg (colloquial) (питер) a monastery (монастырь) Russia (Россия) a university (университет)
Kyiv	an army (армия) Paris (Париж) Smolensk (Смоленск) a fortress (крепость) a monastery (монастырь) a neighborhood (соседство) resurrection (воскресение) a tavern (корчма) a gang (шайка) a province (губерния)	Astrakhan (Астрахань) Kazan (Казань) Berlin (Берлин) Paris (Париж) Vienna (Вена) a horde (typically the Golden Horde) (орда) Germany (Германия) Petersburg (colloquial) (Питер) Ryazan (Рязань) Siberia (Сибирь)

Figure 5: Top 10 contextual neighbours for Saint Petersburg, Moscow, and Kyiv in romanticist (left) and the realist (right) corpora.

Slavic civilization (Novgorod, Kyiv, Moscow) to a focus on contemporary life in Saint Petersburg. This shift also facilitates connections with Western Europe (England, France, Italy, Germany) and provides insights into developments at new trading outposts and ports of the Empire, such as Astrakhan, Kazan, Crimea, and Siberia. As the nation undergoes economic and military expansion, new territories are also being explored by its literature.

Of course, there are limitations to what one can find out looking at frequency changes only. Not only do frequencies of toponyms change, but also the contexts in which they are used. To look into the differences we trained two word2vec models on our corpora. We then compared the contextual semantic neighbours (i.e. words with the closest vectors in the model) for the three most prominent capital cities in our corpus: Kyiv, Moscow, and Saint Petersburg. Figure 5 lists the top 10 most similar words for each of the three cities in both the romanticist and the realist corpora.

Comparing the sets of contextual neighbors for these three cities in two models, we can see that in the case of Saint Petersburg there is a very obvious modernisation of contexts.

While the closest word is Moscow in both models (which is totally understandable given the nature of word2vec mechanics: the two capitals appear in very similar functional positions in texts), the second is a village in the romanticist corpus and a university in the realist corpus. Other realist connotation in the list of contextual neighbors is a grammar school (гимназия) — none of those modern education-related words are present in the romanticist contexts. Notable is also the generally more ‘western’ selection of locations that appear most similar: Germany, Paris.

In the case of Moscow, such modernisation of contexts is much less visible. A village remains as the third closest contextual neighbour, while a university is only on the 10th position, below both the word monastery as well. This highlights Moscow’s more traditional and non-modern connotations, which likely contribute to its relative decline in frequency that we reported in Figure 3.

As for Kyiv, we likely see the total change of its function in the texts. Its contextual neighbors in the romanticist corpus suggest Kyiv being the centre of historical action: the mentions of armies, taverns, gangs... In the realist corpus (where, as we remember, there is much less Kyiv, so this should be taken with a grain of salt), on the other hand, Kyiv becomes just one item in the list of many locations, which in our view is an indicator of the city losing its function as the setting of literary plots.

5. Conclusion and discussion

Our research is an early attempt at modeling the spatial component of Russian 19th-century prose through geocoding and mapping. Our approach obviously lacks many important nuances. For one, we do not differentiate between different functions of toponyms inside the texts, be it a random mention of a place or a location important for the development of the plot. But what we were interested in was primarily the expansion of ‘mental’ geography of Russian writers and readers. Regardless of whether a certain city or country was just ‘mentioned’ or actually was part of the plot, its appearance in the text is a clear sign that it entered the mental map of Russian literate society. Secondly, and maybe more importantly, we did not normalize locations in proportion to the size of work. A lengthy novel set in Moscow can contain hundreds of mentions of the city and will inevitably skew the whole map towards it. We intend to handle this issue in the next iterations of this work.

Despite these and other caveats, we believe that our results demonstrate the utility of the method as a tool to track large scale literary changes on relatively big corpora in the paradigm of distant reading. Most of the novels we worked with belong to the ‘great unread’ of Russian 19th-century literature. The ability to derive conclusions regarding the evolution of literature in relation to the economic, political, and cultural developments of the Russian Empire — without the necessity of reading hundreds of individual texts — presents a promising research perspective. The detection of an expansion in geographic boundaries during the second half of the 19th century through quantitative analysis further demonstrates that methods of distant reading can yield meaningful insights into literary corpora.

6. Data Availability 281

Data can be found here: https://github.com/DanilSko/mapping_russian_prose/tree/main/geodata 282
283

7. Software Availability 284

Software can be found here: https://github.com/DanilSko/mapping_russian_prose/tree/main/code 285
286

8. Author Contributions 287

Daniil Skorinkin: Project administration, Writing – original draft, Writing – review editing, Visualization 288
289

Boris Orekhov: Conceptualization, Formal Analysis, Data curation, Investigation 290



References 291



- Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser (2018). *Canon/archive : large-scale dynamics in the literary field*. ISSN: 2164-1757 Issue: 11 Pages: 14 Publication Title: Stanford Literary Lab: Pamphlets ; 11 Series: Pamphlets of the Stanford Literary Lab Type: workingpaper Volume: Stanford Literary Lab. <https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf>. 292
293
294
295
296
- Barbaresi, Adrien (Mar. 15, 2018). "Toponyms as Entry Points into a Digital Edition: Mapping Die Fackel". In: *Open Information Science* 2, 23–33. 10.1515/opis-2018-0002. 297
298
299
- Bartholomew, J. G. (John George) (1914). *A literary & historical atlas of Europe*. In collab. with University of California Libraries. London ; Toronto : Dent ; New York : Dutton. 272 pp. <http://archive.org/details/literaryhistatlas00bartrich> (visited on 07/15/2024). 300
301
302
303
- Bodenhamer, David J., John Corrigan, and Trevor M. Harris, eds. (June 28, 2010). *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Edition Unstated. Bloomington, Ind.: Indiana University Press. 203 pp. ISBN: 978-0-253-22217-6. 304
305
306
- Bowers, Katherine (2022). *Writing Fear: Russian Realism and the Gothic*. University of Toronto Press. ISBN: 978-1-4875-2692-4. <https://www.jstor.org/stable/10.3138/j.ctv2hvfjgr> (visited on 02/09/2025). 307
308
309
- Brunson, Molly (Sept. 10, 2016). *Russian Realisms: Literature and Painting, 1840–1890*. Cornell University Press. 283 pp. ISBN: 978-1-5017-5753-2. 310
311
- Döring, Jörg (2013). "How Useful Is Thematic Cartography of Literature?" In: <https://www.semanticscholar.org/paper/How-Useful-Is-Thematic-Cartography-of-Literature-D%C3%B6ring/27e15ff9349db2aeeddfb73e59e853068633c88d> (visited on 07/12/2024). 312
313
314
315

- Douglas Clayton, J. (2016). "Russian Modernism (1890–1934)". In: *Routledge Encyclopedia of Modernism*. 1st ed. London: Routledge. ISBN: 978-1-135-00035-6. [10.4324/0123456789-REM1879-1. https://www.rem.routledge.com/articles/russian-modernism-1890-1934](https://www.rem.routledge.com/articles/russian-modernism-1890-1934) (visited on 02/07/2025).
- Du, Keli, Julia Dudar, Cora Rok, and Christof Schöch (2025). *Project – Zeta and Company*. <https://zeta-project.eu/de/project/> (visited on 02/08/2025).
- Fanger, Donald (1998). *Dostoevsky and Romantic Realism: A Study of Dostoevsky in Relation to Balzac, Dickens, and Gogol*. Northwestern University Press. 332 pp. ISBN: 978-0-8101-1593-4.
- Fridlender, G. (1971). *Pojetika russkogo realizma: Ocherki o russkoj literature 19 veka*. "Nauka", Leningr. otd-nie.
- Ioffe, Dennis (2009). "The Poetics of Personal Behaviour: The Interaction of Life and Art in Russian Modernism (1890-1920)." PhD thesis. Universiteit van Amsterdam [Host].
- Kazakova, Elena (Jan. 18, 2024). *Zabytye romany russkih pisatelej iz fondov Pushkinskogo Doma (1857–1917)*. Version 2. [10.31860/openlit-2023.12-C007. https://dataverse.pushdom.ru/dataset.xhtml?persistentId=doi:10.31860/openlit-2023.12-C007](https://dataverse.pushdom.ru/dataset.xhtml?persistentId=doi:10.31860/openlit-2023.12-C007) (visited on 02/08/2025).
- Kuzmenko, E. and Boris V. Orekhov (2016). "Geography Of Russian Poetry: Countries And Cities Inside The Poetic World". In: Digital Humanities Conference. <https://www.semanticscholar.org/paper/Geography-Of-Russian-Poetry%3A-Countries-And-Cities-Kuzmenko-Orekhov/4e506e7c533f219c5ade38e6193162f6d35d2cf5> (visited on 06/19/2024).
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (Sept. 7, 2013). *Efficient Estimation of Word Representations in Vector Space*. [10.48550/arXiv.1301.3781. arXiv: 1301.3781\[cs\]. http://arxiv.org/abs/1301.3781](http://arxiv.org/abs/1301.3781) (visited on 02/08/2025).
- Moretti, Franco (Sept. 17, 1999). *Atlas of the European Novel: 1800-1900*. Google-Books-ID: ja2MUXS_YQUC. Verso. 232 pp. ISBN: 978-1-85984-224-9.
- natasha/natasha (July 14, 2024). original-date: 2016-08-03T09:49:51Z. <https://github.com/natasha/natasha> (visited on 07/16/2024).
- Sobchuk, Oleg and Evgenija Lekarevich (Jan. 17, 2025). *Korpus narrativnoj prozy 19 veka*. Version 4. [10.31860/openlit-2020.10-C004. https://dataverse.pushdom.ru/dataset.xhtml?persistentId=doi:10.31860/openlit-2020.10-C004](https://dataverse.pushdom.ru/dataset.xhtml?persistentId=doi:10.31860/openlit-2020.10-C004) (visited on 02/08/2025).
- Vdovin, Alexey, Margarita Vaysman, Ilya Kliger, and Kirill Ospovat (2020). "«Realizm» i russkaja literatura XIX veka". In: *Russkii realizm XIX veka: Obshchestvo, Znanie, Povestvovanie*, edited by M. Vaisman, A. Vdovin, I. Kliger, and K. Ospovat. Moscow: Novoe literaturnoe obozrenie, 431–451.
- Wilkens, Matthew, Elizabeth F. Evans, Sandeep Soni, David Bamman, and Andrew Piper (Sept. 26, 2024). "Small Worlds: Measuring the Mobility of Characters in English-Language Fiction". In: *Journal of Computational Literary Studies* 3.1. Number: 1 Publisher: Universitäts- und Landesbibliothek Darmstadt. ISSN: 2940-1348. [10.48694/jcls.3917. https://jcls.io/article/id/3917/](https://jcls.io/article/id/3917/) (visited on 02/08/2025).

Making BERT Feel at Home

Modelling Domestic Space in 19th-Century British and Irish Fiction

Svenja Guhr¹ 
 Jessica Monaco² 
 Alexander J. Sherman² 
 Matt Warner² 
 Mark Algee-Hewitt² 

1. fortext lab, Technical University of Darmstadt , Darmstadt, Germany.
2. Literary Lab, Stanford University , Palo Alto, USA.

Citation

Svenja Guhr, Jessica Monaco, Alexander J. Sherman, Matt Warner, and Mark Algee-Hewitt (2025). "Making BERT Feel at Home. Modelling Domestic Space in 19th-Century British and Irish Fiction". In: *CCLS2025 Conference Preprints* 4 (1). 10.26083/tuprints-00030145

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-02-07

Keywords

British and Irish fiction, 19th Century, Victorian, domesticity, space, classification

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. We introduce a novel approach to detecting domestic space in literary texts beyond explicit spatial markers like "home" or "house." Using a pre-trained English BERT model fine-tuned on manually annotated passages from a corpus of 19th-century British and Irish novels, we develop a method to operationalize and quantify domesticity in fiction. Our model captures the nuances of domestic space by analyzing contextual and relational cues rather than relying solely on toponymic and other explicit references. This approach offers new insights into the representation of space in literature, revealing the fluid and dynamic nature of domesticity in 19th-century British and Irish fiction.

1. Introduction

She went upstairs, emerging all at once into the full morning sunshine in the hall, which dazzled and appalled her. [...] She went into Clara's room first. [...] Clara's maid was seated, fast asleep, before a table on which a candle was burning pitifully in the full daylight. The room looked trim and still as a room does which has not been occupied in that early brightness. The maid woke with a shiver as Mrs. Burton entered. "Oh, Miss Clara, I beg your pardon," she said. "It is no matter. My daughter will not want you tonight. Go to bed, Jane," said Mrs. Burton. (*At His Gates*, Margaret Oliphant)

What makes space domestic in fiction? Is it the mention of keywords like "home" or "room"? Is it the presence of characters discussing private matters? Or is space domestic when characters are engaged in private or intimate interactions, as in *At His Gates* when Mrs. Burton checks on her daughter's room while talking to the housemaid? Space occupies an important place in literary theory, and domestic space in particular gains importance in 19th-century fiction. As, for example, Davidoff and Hall (1987) detail, the Victorian period is marked by "separate spheres," in which men participated in public and professional life while women were responsible for the home as the central organizers of domestic life, combining both physical space and domestic ideology. In fiction, as Cohen (2017) explains, portrayals of domesticity both criticized and upheld domestic ideologies as novels populated their homes with characters ranging from the angelic

Agnes Wickfield in Dickens' *David Copperfield* to the villainous figures of sensation fiction. Such attention has resulted in a significant body of criticism on domesticity in Victorian fiction, including but not limited to Armstrong (1987)'s additional focus on class in *Desire and Domestic Fiction*, Freedgood (2006)'s emphasis on empire and materiality in *The Ideas in Things*, and Marcus (2007)'s work on friendship and sexuality in *Between Women*. However, these studies and others tend to prioritize addressing the concept of domesticity over a strict account of domestic space. In both fiction and literary criticism, spatial information offers a concrete link between domestic settings and ideologies and allows readers to orient themselves as characters move through the fictional worlds they inhabit. Such settings are also implicated in themes of gender, class, and colonialism. Our project therefore sought to operationalize space in fiction, (especially domestic space) in order to trace the patterns of domesticity and its associated cultural constructs through the British and Irish 19th-century novel.

The operationalization of space has a long history in the context of computational literary studies. Moretti (1999) and Piatti (2016) concentrate on the importance of geographic plotting in the construction of narrative meaning, while Ryan et al. (2016) and Wilkens (2013) have applied computational methods to map fictional settings onto real-world entities. Other examples include Bamman et al. (2019), who annotated and automated the recognition of named spatial entities in BookNLP (Bamman 2021), as well as Bologna (2020) and Schumacher (2023), who similarly operationalize space by identifying Bamman et al. (2019)'s sets of spatial keyword classes (e.g. GPE, LOC, FAC, etc.) using machine learning techniques. These approaches rely on explicit spatial references, such as named entities like toponyms or spatial entities such as "marketplace" or "sitting-room," which are the focus of the most recent work by Kababgi et al. (2024), who fine-tune a BERT language model to automatically detect and recognize non-named spatial entities (NNSEs) from manually annotated training data. Using sentence-based annotations, they first identify sentences containing NNSEs and then classify them as "rural," "urban," "natural," or "interior." While these methods have proven adept at detecting explicitly spatialized passages, passages without these entities are often difficult to spatially identify. This problem underscores the challenge of identifying implicit space.

In this paper, we introduce a new method for the automated detection of both explicit and implicit domestic space in English-language fiction based on the probability of a passage being set in domestic space. Our approach offers a departure from the implicit ideological or ontological framework of previous approaches – where domestic space is predefined as a static concept – by adopting a phenomenological one. Instead of asking "Is this space domestic?" we switch to the question "How likely is it that the passage is set in domestic space?" This change allows us to explore how domesticity manifests in ways that challenge traditional assumptions as we identify the domestic qualities of unexpected or liminal spaces like gardens, carriages, or even ships. To that end, we propose the calculation of a "domesticity score": a score based on the probability assigned to a passage by a fine-tuned English BERT classifier (trained on manually annotated data) of a passage being set in "domestic space." This modeling approach offers new possibilities of the analysis of fictional spaces that are not explicitly described but are discursively constructed through dialogue, context, and emotional tone.

In our paper, we first describe our operationalization of “domestic space” and then detail the annotation process that we used to operationalize our corpus of 19th-century British and Irish fiction. Second, we introduce a multilingual transformer model fine-tuned to compute the probability of a passage being set in “domestic” space through a two-step classification task performed on six-sentence passages. Using our model, we calculate the “domesticity score” for each passage in our corpus, which we can then summarize across each novel. We then provide an analysis of chosen texts by canonical authors to offer a new perspective on implicit domestic space. This intervention opens new opportunities for analyzing space, character, and plot in fiction.

2. Operationalizing Domestic Space

Our project to identify domestic space in 19th-century British and Irish fiction began with a derivative approach to annotation-based concept operationalization recommended by Pichler and Reiter (2022). We similarly do not start from a specific working definition of explicitly and implicitly represented “domestic space” in fiction. Rather, we approach the concept through approximation, using exploratory annotation, inter-annotator agreement calculation, and discussions resulting in the iterative development of a decision tree for the annotation task. Our rationale is that, while theoretical frameworks in narratology operationalize space via narrated action involving characters or descriptions of physical environments, textual clues to setting are often absent from narrative discourse (see Fludernik and Keen 2014; Ryan 2014). For instance, the spatiality of an event may be inherited from descriptions in previous scenes (frequent in novels with long dialogues), remain implicit in character interactions, or be altogether absent in reflective passages that are narrated non-spatially. Focusing on examples of domestic space, we recognized that domestic spatiality is not clearly bound to entities, but is a fluid literary concept that varies contextually. For instance, a garden may sometimes function as a domestic space within a narrative about children playing or adults discussing romantic entanglements during a stroll, becoming an extension of the private sphere of the home. In other contexts, however, gardens can be part of publicly accessible parks, whether or not they are adjacent to the homes of the wealthy. This example only emphasizes the difficulty of setting the boundaries for a clear definition of domestic space in fiction in a way that captures its full ideological, historical, and cultural dimensions.

In contrast to existing definitions that risk excluding the ambiguities that make domestic space so central to fiction, we adopt an inductive approach to operationalizing domestic space. Rather than imposing a fixed definition, we fine-tune a language model on agreed-upon examples of domestic space, allowing the model to infer patterns and associations that characterize settings. By approaching the classification of space with machine learning methods based on contextual embeddings, we offer a fluid definition of domestic space through a “domesticity score” that measures the probability of a passage being set in domestic space against it being set in another type of space or being non-spatial. In our manual annotation process, we include passages set in living rooms, kitchens, bedrooms, etc. that provide strong indicators of what constitutes a domestic space beyond named or non-named entities, i.e. through resolved coreferences, deictics, or other explicit and implicit clues detectable by human annotators. In the same way, we also include passages with explicit settings that are not domestic (for example battles,

ships at sea, or carriage rides). By using these clear examples as part of a training sample, we enable the model to detect domesticity even in passages that lack overt spatial markers. In this way, we can extrapolate from explicit examples of domestic space to implicit examples recognized by the model as sharing all of the same features except the explicit references to domestic space. The model's ability to generalize from training data allows it to classify all of the passages in our corpus and reveal patterns of domesticity.

We aimed to categorize passages into two primary classes: "domestic space" and "other." The choice to limit the classification to these two classes was driven by the nature of our future research interest: By focusing on domesticity, we aimed to isolate passages of interest for broader inquiries into themes of gender, colonialism, and social hierarchies in Victorian fiction. Attempts to differentiate the class "other" into subcategories (e.g., public spaces, natural landscapes, or non-spatial passage) proved impractical for several reasons. On the one hand, annotators often struggled to achieve consensus on subcategories, given the inherent fluidity and overlapping boundaries of non-domestic spaces, as well as the limited context given in the annotation passages. For example, from a six-sentence passage, it was often impossible to tell if the passage was spatialized non-domestically or just non-spatialized. On the other hand, exhaustively classifying different types of spaces was not our primary goal of accurately recognizing domestic spaces.

To transform the abstract concept of domestic space into measurable units, we defined these units as fixed-length six-sentence text segments. This segmentation allows us to systematically apply annotations and later model predictions across the corpus. We relied on intersubjective interpretation during the annotation process. This system involved iteratively creating a set of guidelines that balanced theoretical rigor with practical applicability. Annotators were tasked with identifying passages that unambiguously depicted either domestic settings, such as interiors of homes, or non-domestic settings such as workplaces or otherwise public settings. Ambiguous or marginal cases were excluded. This operationalization ensured that the training data for our model represented the clearest possible examples of both domesticity and non-domesticity to minimize uncertainty in the machine learning process.

3. Data and Method

3.1 Data Preprocessing

We used a corpus of 19th-century British and Irish novels (see Table 1), sourced from the University of Illinois libraries and *Chadwyck-Healey Nineteenth-Century Fiction* collection (Chadwyck-Healey Literature Collections and ProQuest 2024). The corpus represents a curated selection of literary texts, including canonical and lesser-known literary prose. Additionally, the collection offers detailed metadata, such as publication dates and author information which enables diachronic and comparative analyses. Although not the largest corpus available and not strictly representative of 19th-century novelistic prose, our corpus offers relatively clean OCR (many of the texts were hand-keyed) and a sample of both canonical and non-canonical texts.

Texts	2,865
Words in total	557,097,804
Individual authors (+ 126 “unknown”)	1,250
British authors	1,226
Irish authors	24
Texts by Irish authors	118
Time period	1748–1899
Six-sentence passages	3,684,727
Manually annotated passages	1,227
“domestic space” passages	521
“other” passages	678
“trash” passages	28

Table 1: Research corpus metadata summary.

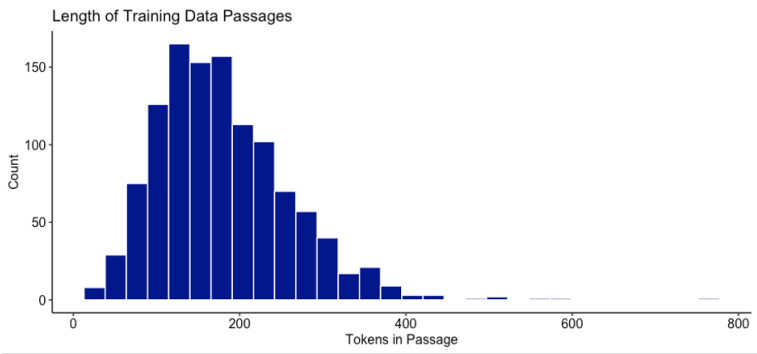


Figure 1: Varying passage lengths in the training corpus. Passage length did not correlate with the class choice of the model.

One of our first decisions for the project lay in our chosen resolution for the passages we 152
wanted to classify. Chapters made up of multiple scenes would be too long to classify 153
as domestic or not (the action of a chapter might move from a bedroom to a garden 154
to a carriage), while sentences would be too short (within a given chapter, only a few 155
sentences actually contain information on the setting). Paragraphs, although closest to 156
our desired resolution, are too inconsistent in length (particularly when representing 157
dialogue) for reliable classification with our transformer model. In our previous close- 158
reading approaches, six-sentence passages proved to be the Goldilocks zone: long 159
enough to get enough spatial information, short enough to be mostly one space and 160
to be read and classified quickly enough by human readers while manually tagging 161
the passages (see Figure 1). Furthermore, the six sentences strike a balance between 162
granularity and context. They capture enough of the narrative to identify domestic 163
space without introducing excessive noise. During the annotation process, six-sentence 164
passages provided sufficient context for human annotators to make informed decisions 165
about spatial settings that aligned with the model’s training needs. 166

3.2 Manual Annotation 167

Following the recommended workflow for annotation guideline creation by Reiter 168
(2020), we defined annotation classes and developed a decision tree (see Figure 2) 169
giving annotators an ordered set of decisions to follow before declaring a passage to be 170
set in “domestic space” or “other.” As a third class for manual annotation, we defined 171
“trash” for passages that either contained paratextual material (such as bibliographic 172

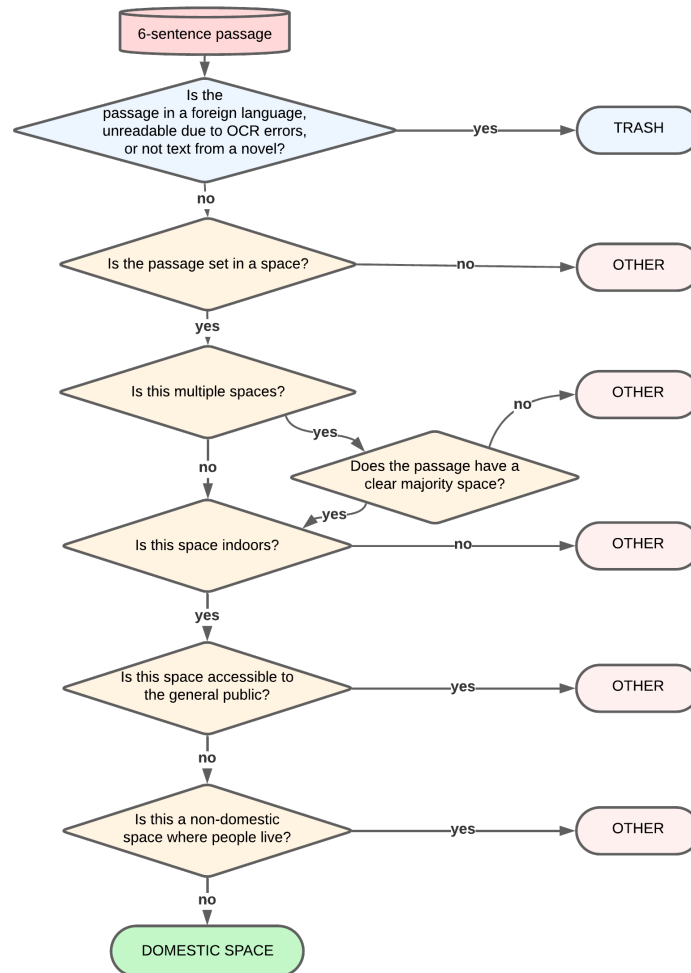


Figure 2: Decision tree for the manual annotation of a passage as “domestic space,” “other,” or “trash.”

references or advertisements) or were unreadable to human annotators due to excessive 173
 OCR errors or foreign language. The annotation guidelines were iteratively developed 174
 through their application, annotation evaluation, discussion, and guideline refinement 175
 to ensure clarity, consistency, and alignment with the conceptual framework. We define 176
 the classes as follows: 177
 178

- **“domestic space”**: passages set in clear, unambiguous domestic settings, such as 179
 interiors of homes, 180
- **“other”**: passages set in non-domestic or ambiguous spaces, including public 181
 places, natural landscapes, or spaces where the setting was unclear or non-spatial 182
 as in reflective passages or summaries, 183
- **“trash”**: passages with poor OCR quality, foreign language, or extra-textual or 184
 paratextual elements. 185

Five experts trained in literary studies and two student assistants manually annotated 186
 1,375 passages. The passages were selected partly because they contained a domestic 187

seed term, such as “kitchen”¹, and partly at random from the corpus. Each passage was annotated by at least two independent annotators, resulting in a total of 3,657 individual annotations. Following the decision tree for manual annotation, each passage was tagged with one of the three classes: “domestic,” “other,” or “trash.” In an earlier version of the annotation guidelines we added the category “I don’t know” (“IDK”) to the decision tree to distinguish between passages that were unambiguously non-domestic or non-spatial from passages that gave no information on their spatiality at all.

As part of the iterative development of the annotation decision tree, we added a majority decision step when encountering “IDK” passages containing information on more than one space or non-spatial elements mixed with some spatial information. Namely, for passages containing more than one space, annotators were told to classify the passage based on the location of the majority of its sentences. For instance, in the case of a six-sentence passage in example 3.1, the first four sentences cover a setting in domestic space as the characters prepare to go outside. Their exit is narrated at the end of the passage in the last two sentences as they walk “towards the gardens.” The agreed-upon decision for this passage by all annotators was the class “domestic space.” At a later stage, we incorporated “IDK” into the *ex-negativo* class “other” to focus our annotation on the detection of domestic space. Example 3.2 contains a passage that has been manually annotated as “trash.” Out of all 1,375 passages annotated, about 30% of the passages were classified as “domestic space,” 67% as “other,” and 3% as “trash” giving us an initial benchmark against which to measure the automated performance of the model.

Ex. 3.1 It is warm and mild now, and we shall be back in time for luncheon, I will just get my hat.” He went into his bedroom as he spoke, and after a moment came back with his hat in his hand. John had left the room and was standing just outside the door. As Sir Lionel came through the sitting-room, he watched him furtively, but closely; and as soon as he was fairly in the corridor, John shut the door, and, forgetting his usual deference, led the way briskly through the porch. They walked towards the gardens; but presently John said: “I fear you will have some further trouble with James, I hope he will go this afternoon.” “I hope so, these scenes of howling and supplicating are very tiresome.”
(passage from *Riding out the Gale* by Annette Lyster labeled as “domestic”)

Ex. 3.2 THE LAWS OF WAR AFFECTING COMMERCE AND SHIPPING. By H. BYERLEY THOMSOX, of the Inner Temple. Second Edition, greatly enlarged. 8vo, price 45. %d. boards. LECTURES ON the ENGLISH HUMOURISTS OF THE 18th CENTURY. By W. M. (metatextual element labeled as “trash”)

3.3 Validation of the Annotations through Ground Truth

To assess the reliability of our manual annotations, we calculated the inter-annotator agreement (IAA) using Krippendorff’s Alpha, a statistical measure for categorical data annotated by more than two annotators (Krippendorff 2018). The overall Krippendorff’s Alpha for our annotations was 0.58 across five annotators, which is below the standard threshold of 0.8, but consistent with the inherent subjectivity and ambiguity observed

1. The list with the used annotated seed words can be found in the GitHub repository.

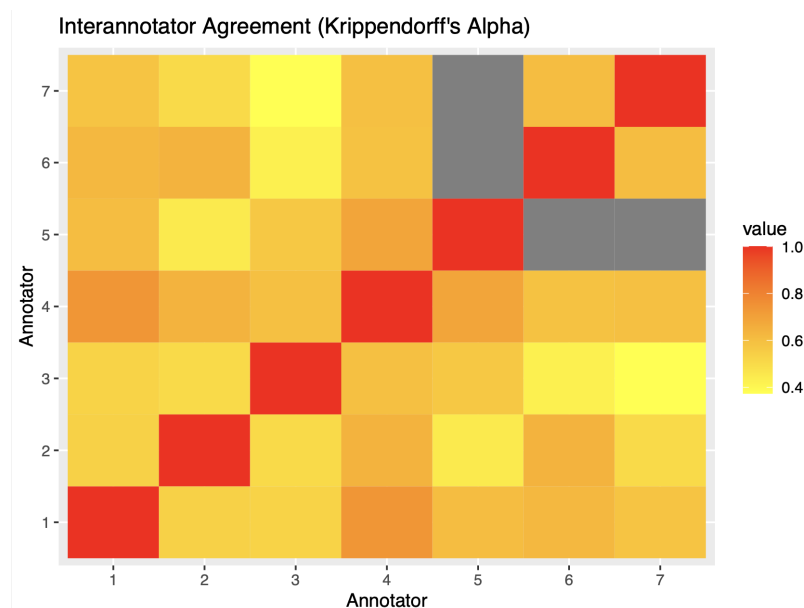


Figure 3: Heat map showing the (dis)agreement between annotators calculated with Krippendorff's Alpha.

in similar literary annotation tasks (see Figure 3).

Despite the relatively low Alpha, the qualitative comparison of the manual annotations did not reveal any systematic deviations or rogue annotators. Instead, disagreement was evenly distributed, reflecting the underlying complexity of identifying fictional space. Assessing annotation quality, however, extends beyond inter-annotator agreement. While Baledent et al. (2022) question whether high agreement necessarily ensures accuracy, a key challenge remains: Annotators may converge on errors, making intersubjective consensus yield a lower quality annotation than ground truth. To evaluate the validity of our annotated data, we established a set of ground truth annotations for passages where the consensus converged on one annotation rather than the other, despite the absence of explicit spatial markers in the extracted text segment.

As Pichler and Reiter (2022, 14) explain, validity serves as the critical “link between theory and measurement,” allowing researchers to evaluate whether their methods genuinely align with their conceptual objective. Similarly, Krippendorff (2018, 361) emphasizes that a measurement instrument is valid “if it measures what it claims it measures.”

In literary studies, intersubjectively recognized annotations – those agreed upon by multiple annotators – are considered a robust measure of validity. As gold annotations, they are used as the basis for text analysis and interpretation as well as for training models for automation. However, during our annotation process, we observed a key limitation: While high inter-annotator agreement confirmed the reliability of our classifications, the annotations themselves did not always capture the true spatial context of a passage. On the contrary, given a six-sentence passage without an explicit lexical marker for spatial information, the annotators had to decide whether the passage was set in domestic space based on the given information, such as private dialogues or intimate actions, which are more likely to be set in domestic space than in public space and have to be spatial by default since characters are present. Nevertheless, in the discussion

rounds, the annotators often could not justify their annotation decision by referring to elements on the textual surface, even when an intersubjective annotation decision was given. Consider the example 3.3, where the annotators initially only saw six sentences of the dialogue, which they agreed contained little spatial information and suggested an “other” classification. The passages in set brackets (presented here in abbreviated form), however, show the surroundings of the dialogue, taken from the novel, which clarify that it actually takes place in a domestic space.

Ex. 3.3 {But he had not had time to finish his sentence before the door of the house was thrown open, and Stephanie Harcourt appeared upon the threshold.
 “Bella” she cried to her friend hysterically, “it is all over. I am dismissed without salary, and I can’t even pay you my share of the week’s rent ! The sooner I go to the Tombs with that scoundrel the better!”
 “Hush, hush, dear ! there is a stranger present,” said Miss Vavasour compassionately. [...] “My poor child, how came you to marry him?”
 “I can’t tell you that. I was frightened into it in a way that you would hardly understand. Only, thank heaven, I am now delivered from him.”}
 “But after his two years’ incarceration are over, he will come out again and claim you.”
 “I will have broken the chain by that time. I will have gone far away where he shall never find me.”
 “And you met Cortes in San Francisco?”
 “Yes, sir.”
 “And that scoundrel Sandie Macpherson had some hand in your marrying him?” {
 The girl’s cheek became as white as ashes. “Who has told you that?”
 “No one. I guessed it”}
 (*Phyllida. A Life Drama* by Lean Florence, 1882)

This is the key difference between gold annotations and ground truth. While we achieved a high inter-annotator agreement in manually classifying six-sentence passages as “domestic” or “other,” we wondered whether our intersubjective class choices actually represented valid annotation choices for the passages when we took the greater context of the passage into account (context that was unavailable to our annotators and which would be unavailable to our model). Accordingly, we decided to go beyond our gold annotations and manually verify the spatial setting of a given passage by looking at where each passage fit within the novel, and by searching outside of the passage (before or after) for contextual information about the actual space in which the passage is set.

We conducted this contextual validation on a subset of 15 passages, with additional annotations informed by the surrounding text. This process revealed some new findings: Many passages that were initially labeled as “other” in the gold annotations were reclassified as “domestic space.” For instance, dialogues that appeared spatially ambiguous within the passage, i.e., due to the lack of any spatial marker in the dialogue itself, were often revealed to occur in domestic settings when viewed in context. Going back and forth several pages before and after the passage (sometimes up to 30 pages needed for long dialogue passages and on average ten minutes needed for the classification of one passage²) allowed us to find spatial referents for our target passages, and thus

2. In comparison, the preparation of gold annotations took approximately 30 seconds for reading and deciding a class for a six-sentence passage.

enabled a ground truth classification for the six-sentence passage. Passages containing dialogues or transitional scenes (e.g., characters moving between spaces) were the most likely to be reclassified. These results highlight the challenges of detecting implicit domestic space based on limited textual context alone and underscore the importance of ground truth annotations for classification tasks beyond the gold annotations that annotators agree on. While gold annotations provide a standardized and efficient means of generating training data, ground truth annotations offer more fidelity to the actual text being annotated.

However, creating ground truth annotations is even more expensive than creating gold annotations because of the extra labor involved in tracking down the contextualizing information. Furthermore, for the purpose of automating the classification task, we had to consider that the state-of-the-art transformer models we use are also constrained to a limited context. Therefore, the decision to use six-sentence passages proved to be an appropriate heuristic: While the passages are short enough for manual examination, they provide a relatively high level of contextual information for the classification task. Since we could not provide a large number of ground truth classifications for training, we kept the ground truth annotations out of the training set entirely and limited their use to an additional evaluation step with an extended test set.

4. Automation: Make BERT Feel at Home

Transformer-based architectures have emerged as a preferred approach for classification tasks in computational literary studies (CLS), offering greater transparency than large language models (LLMs), which are often optimized for language generation rather than classification (see e.g. Bamman et al. (2024)). Pre-trained models from the BERT family (Devlin et al. 2019) have been successfully applied in various literary and linguistic classification tasks, including genre attribution (Zundert et al. 2022), character gender identification (Schumacher et al. 2022), emotion classification in plays (Dennerlein et al. 2023), and the detection of dubitative passages (Parigini and Kestemont 2022). For automated space recognition, recent studies have demonstrated the superior performance of fine tuned BERT-based models over LLMs such as GPT-3.5 and GPT-4 (Kababgi et al. 2024; Soni et al. 2023). Given these findings, we selected a BERT-family model for our sequence classification task, specifically the TensorFlow Universal Sentence Encoder (USE) model (Yang et al. 2021).

As we describe above, unlike prior work on spatial classification that relies on entity detection (Kababgi et al. 2024; Soni et al. 2023), our study shifts the focus from explicit spatial markers to the implicit discursive construction of domestic spaces. To implement our approach, we fine-tuned a pre-trained English BERT model from TensorFlow Hub on our manually annotated training data. Initially, we used TensorFlow's BERT_en_uncased preprocessor with an English BERT model pre-trained on Wikipedia and BooksCorpus and fine-tuned on the Multi-Genre Natural Language Inference (MNLI) dataset (Devlin et al. 2019; Google 2023a). While BERT_en_uncased is widely used for NLP task and designed for token-level tasks like question answering and named entity recognition, capturing bidirectional word context, the Universal Sentence Encoder (USE) generates fixed-size sentence embeddings, making it more effective for semantic similarity and

sentence classification. Consequently, the USE model offers superior performance in complex, higher-order tasks (such as classifying space). It is also multilingual, offering an additional advantage for passages containing foreign language words (a semi-regular occurrence in nineteenth-century novels) and outperformed the BERT model for our classification task. For these reasons, we ultimately selected USE due to its strong performance in sentence-level embeddings and its effectiveness in transfer learning, particularly in low-data settings. The model employs a Transformer-based sentence encoding architecture that computes context-aware representations of words while preserving both word order and surrounding context (Cer et al. 2018; Google 2023b). This enables effective sentence-level transfer learning for our six-sentence segments, providing higher classification performance with a small set of training data.

To develop a classifier for detecting domestic space in British and Irish fiction, we fine-tuned an up-to-date (2023) USE model using TensorFlow and Keras. The training process followed a two-step classification approach. First, we trained a binary classifier to filter out “trash” passages with the understanding that these would not be relevant for further classification. This first model was trained on manually labeled data, where passages were categorized as either “trash” or “not_trash.” The data was preprocessed using the USE multilingual preprocessor, tokenized, and passed through the USE encoder. The model was trained with categorical cross-entropy loss and optimized using the Adam optimizer³, incorporating early stopping to prevent overfitting. Once trained, this model was used to filter out irrelevant passages from the dataset, ensuring that only meaningful textual segments were passed to the second classification step.

The second model classified the remaining passages into “domestic space” or “other” categories. This model was trained in a similar manner, using a labeled dataset where passages were tagged accordingly. Again, we used the USE preprocessor and encoder to generate sentence-level embeddings, which were then fed into a neural network with a dropout layer to mitigate overfitting. The trained model was saved for reuse, allowing for batch classification of unseen textual data.

4.1 Prediction

After training, the models were deployed to classify new texts. Raw passages from unseen datasets were first processed through the trash detection model, filtering out irrelevant segments. The remaining passages were then analyzed by the domestic space classifier, which assigned probabilities to each passage being “domestic space” or “other.” The classification results were compiled into structured tables for further analysis. This two-step approach proved more successful than a three-way classification task (“domestic space” vs “other” vs “trash”) as the dual binary classifications allowed for the development of separate specialized models for recognizing trash and identifying domestic passages respectively. This enabled us to ensure high-quality predictions while leveraging the strengths of USE’s sentence-level embeddings for transfer learning in a low-data setting.

We predicted the domesticity score for each six-sentence passage in the corpus using

3. Adam is an algorithm that combines the advantages of the Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp) by adjusting learning rates for each parameter based on estimates of first and second moments of the gradients (Kingma and Ba 2017).

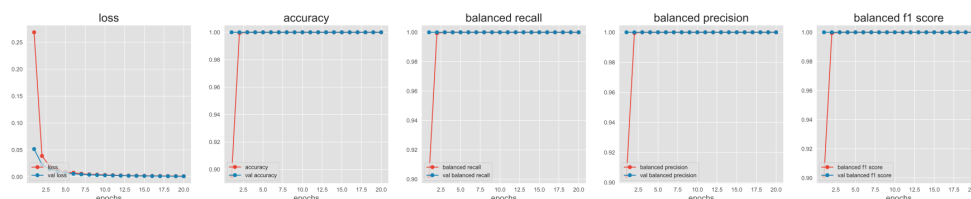


Figure 4: Performance of the “Trash Detector” model.

a rolling-window approach. To facilitate prediction on unseen data, we employed a sequential two-model pipeline. The input consists of an Excel or CSV file containing segmented literary plain texts, where each cell contains a six-sentence passage generated through prior segmentation using the spaCy sentence splitter. The first stage of the prediction pipeline is trash detection, where the input text is processed by a trash detection model. This model assigns a probability score indicating whether the passage is classified as “trash” or not. Segments with a high probability of being non-trash are retained for further analysis. The filtered output consists solely of text segments deemed relevant for domesticity classification.

In the second stage, the domesticity prediction model processes the filtered text. The model reads the cleaned dataframe and predicts a domesticity score for each six-sentence segment, determining the likelihood of its setting being domestic. The prediction operates independently for each segment, meaning that the surrounding textual context – both preceding and following passages – is not considered. The classification is based exclusively on the content within each individual cell.

The dataset, available in our GitHub repository, includes an extraction of the 1,000 passages with the highest and lowest domesticity scores, offering insight into the model’s classification of domestic settings. Model evaluation was conducted using a held-out test set alongside ground-truthed annotations. To avoid sampling bias introduced by initial keyword-based selection, the held-out test set was randomly sampled from the full corpus, ensuring a more representative and independent evaluation. Our assessment of results at both the novel and passage levels suggests alignment with established critical expectations.

Finally, we acknowledge that the model is highly overspecialized to detect 19th-century domesticity, as it has been trained specifically for this purpose. For example, if applied to texts from Latin American Boom fiction in translation, it would still attempt to assign domesticity scores using the criteria it has learned from 19th-century novels despite contextual differences. However, this historical specificity aligns with the goals of our project, which aims to capture and analyze domesticity as it was conceptualized in the 19th century.

4.2 Evaluation of the Model Performance

The model was evaluated using key metrics, including categorical accuracy, recall, precision, and F1-score. Training performance was visualized over multiple epochs to monitor improvements (see Figure 4 and Figure 5), and early stopping was applied to optimize performance.

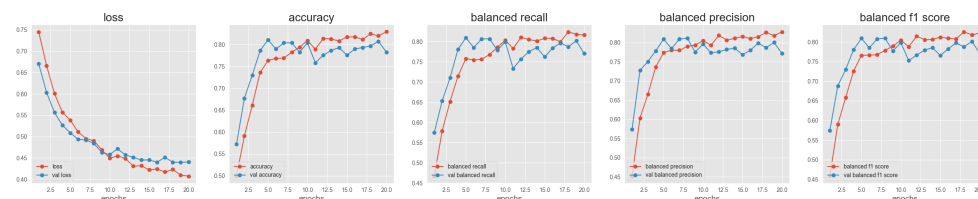


Figure 5: Performance of the “Domestic Space Detector” model.

	Trash Detector	Domestic Space Prediction
Accuracy	1.00	0.8159
Recall	1.00	0.8105
Precision	1.00	0.8092
F1-score	1.00	0.8097
Loss	0.0012	0.4288

Table 2: Model evaluation results for the “Trash Detector” and the domestic space prediction.

Trash Detector In the first step of our pipeline, the most frequent misclassification of trash occurs when the trash detector fails to filter out foreign-language passages (French in particular) that were manually labeled as “trash” in the training data. This indicates that the model was not explicitly trained to use language as a distinguishing criterion for “trash.” For example, passages with non-English dialogue, but also segments of foreign language texts (see example 4.2), which were missed during the manual cleaning of the data set, remain for the second prediction step. However, with the transition from BERT uncased, pre-trained on English texts, to the multilingual Universal Sentence Encoder, the model retains the ability to predict whether a passage is set in a domestic setting. Another common misclassification occurs with segments of low OCR quality, which manual annotators labeled as “trash,” but which are included in the later stage of analysis based on the criteria employed in the automated approach (see example 4.3). Given that foreign-language passages are relatively rare in the dataset (but still present despite manual checks for foreign-language texts, duplicates, and English translations of non-English texts, which are subject to human error), and that the model’s ability to accurately classify passages of low OCR quality based on their setting is an advantage rather than a disadvantage, this limitation does not affect the overall effectiveness of the pipeline. On the contrary, the trash detector still performs well on these segments, outperforming⁴ human annotators in these cases.

Ex. 4.1 FALSE STEPS 1 64 XIII. WANT OF MONEY 179 XIV. IN THE GLOAMING 1 97 CHATTER PAGE XV. [...] (True positive: index manually labeled as “trash,” automatically predicted as “trash” with a probability of 0.15 (“not-trash”) to 0.85 (“trash”))

Ex. 4.2 Enfin, ils se sont tous ruinée, et un M. Stanlej a acheté le bien. Si je ne me trorape, il était le premier mari de Ladj Clarancourt et il lui a laissé le Manoir, mais seulement en usufruit. [...] (False negative: manually labeled as “trash,” automatically predicted as rather “not-trash” with a probability of 0.57 (“not-trash”) to 0.43 (“trash”))

4. While we do not suggest that the model outperforms human annotators in theory-driven classification tasks, in the specific case of the “trash” category, characterized primarily by textual noise rather than interpretive ambiguity, the model shows greater consistency, particularly in detecting low-quality OCR passages that annotators often disagreed on.

Ex. 4.3 [...] He addressed a most affectionate letter to ttubert, informing him of the death of Mrs. 445
 Sedley, and the total change which had ken, place; adding, that in consequence towliieb be added, 446
 lfe- fird\$ uite î6M- šAedi; and oa b& irettrm šhi)d((îlb plearore yfeld bis wife up t64Â^ fie tberr 447
 added, that dthe i fidnlfiicflîlîSft 'wcM not allow her to WrilbifB' 'dlf, she liad requested himÛ 448
 petfonsli tIMIt office for her. 449
 (False negative: manually labeled as “trash,” automatically predicted as rather “not-trash” with a 450
 probability of 0.54 (“not-trash”) to 0.46 (“trash”)) 451

Prediction of Domesticity As already reported in (anonymous), we validated the 452
 domesticity prediction model by selecting an additional random sample of 120 passages 453
 from the corpus, manually annotating them, and comparing the results with the model’s 454
 classifications. The model and annotators aligned in 71% of cases (85 out of 120), 455
 surpassing the initial inter-annotator agreement (IAA). Further analysis of the model’s 456
 probability scores reinforces these findings. In 84 instances, the model assigned a high- 457
 confidence probability (either above 70% or below 30%) for a passage being categorized 458
 as “domestic,” with annotators agreeing 82% of the time. For passages where the model 459
 showed greater uncertainty (probabilities between 40% and 60%), agreement dropped 460
 to 44%. These results indicate that most discrepancies arose in passages the model itself 461
 recognized as ambiguous. 462

In a further validation step, we did an error analysis of the predicted domesticity scores of 463
 the segments that were included as part of the ground truth data set (see [subsection 3.3](#)). 464

The predicted domesticity scores for passages labeled as “domestic” in the ground truth 465
 data reveal intriguing patterns. Among the 19 passages identified as pure dialogue 466
 without explicit spatial markers, the model assigned an average domesticity score of 0.45 467
 with a standard deviation of 0.2. Notably, 15 of these 19 passages received a score below 468
 60%, suggesting that the model frequently registered uncertainty when encountering 469
 dialogue without explicit spatial cues. 470

Conversely, the seven passages categorized as pure dialogue in “other” settings showed 471
 the model’s tendency to correctly assign them to non-domestic spaces. These passages 472
 had an average domesticity score of 0.26, corresponding to a 0.74 probability of being 473
 “other,” with a standard deviation of 0.18. Moreover, five of the seven passages received 474
 a low domesticity score (<40%, i.e., >60% as “other”), indicating a clearer classification. 475

These findings raise interesting questions about the role of dialogue in spatial classifica- 476
 tion. While dialogue alone does not strongly signal domesticity, it appears that the model 477
 struggles more with assigning high domesticity scores to dialogue-heavy segments with- 478
 out explicit spatial markers. This suggests that contextual cues beyond six-sentence 479
 windows, such as speaker identity, dialogue patterns, or adjacent descriptions, may play 480
 a critical role in determining domesticity. Further investigation of dialogue structure as 481
 a latent feature in domesticity classification will be discussed in [subsection 5.3](#). 482

4.3 Domesticity Score

483

Since the output of our classification tasks consists of numerical values between 0 and 1, the received numbers provide a way to identify passages with a high probability of being set in “domestic space” or “other” (or of being “trash” for the first classification task respectively) and can be taken directly as a score indicating the relative domesticity of the passage. With this approach, we are able to provide information about the likelihood of a passage being set in “domestic space” or “other” rather than providing forced binary decisions for one class. As a result, passages of ambiguous spatial nature are present (and identified as such), as well as passages that tend toward one of the two classes. Based on this, each passage considered in the second classification task was assigned a domesticity score between 0 and 1. The output of the classification task is a dataframe in which each classified passage is identified by a distinguishing passage ID and the classification value for being set in “domestic space” or “other,” enriched with metadata about the title of the text from which the passage was taken, the author’s name, and the publication date.

The analysis of domesticity scores highlights key patterns in how the model interprets domestic space in fiction. Passages with the highest domesticity scores, such as those from *The Ill-tempered Cousin* by Elliot Frances (see example 4.4) and *Ombra* by Margaret Oliphant (see example 4.5), exhibit rich domestic imagery, explicit spatial markers, and detailed descriptions of household activities. For example, in *The Ill-tempered Cousin*, the passage’s focus on household disorder, personal belongings, and family interactions contributed to its nearly perfect domesticity score of 0.996. Similarly, the passage from *Ombra*, with a score of 0.978, features a cozy, well-defined domestic setting, emphasizing warmth, comfort, and familial intimacy. In contrast, passages with low domesticity scores often lacked clear spatial markers or were dominated by dialogue without explicit references to domestic settings. The model showed greater uncertainty when processing such ambiguous segments, particularly in cases where dialogue occurred without contextual grounding. This suggests that while the model effectively identifies overtly domestic scenes, it – like many readers – struggles with less explicitly defined spaces, reinforcing the need for further analysis of latent features such as dialogue patterns and indirect spatial cues.

Ex. 4.4 Everything in the house that morning was in confusion. The housemaid had put coarse sheets on Lady Danvers’ bed, and forgotten the muslin curtains to the window. [...] A letter, too, had come from John Bauer (how many hours the excellent John had spent over its composition in the solitude of Wood’s Green, who can say?) telling of the deep impression Miss Escott had made on him, and requesting his aunt’s permission to return, “Only to be allowed to look at her,” wrote honest John, in a strictly business hand, with dots on all the i’s, and the t’s crossed to such a nicety, it would have been a pleasure to look at them, to anyone less worried than Aunt Amelia. [...] (Passage from *The Ill-tempered Cousin* by Elliot Frances, automatically predicted as “domestic” with a probability of 0.996)

Ex. 4.5 Mrs. Anderson’s room was a large one; opening into that of *Ombra* on the one side, and into an ante-room, which they could sit in, or dress in, or read and write in, for it was furnished for all uses. It was a petit appartement, charmingly shut in and cosy, one of the best set of rooms in the house, which Kate had specially chosen for her aunt. Here the mother and daughter met

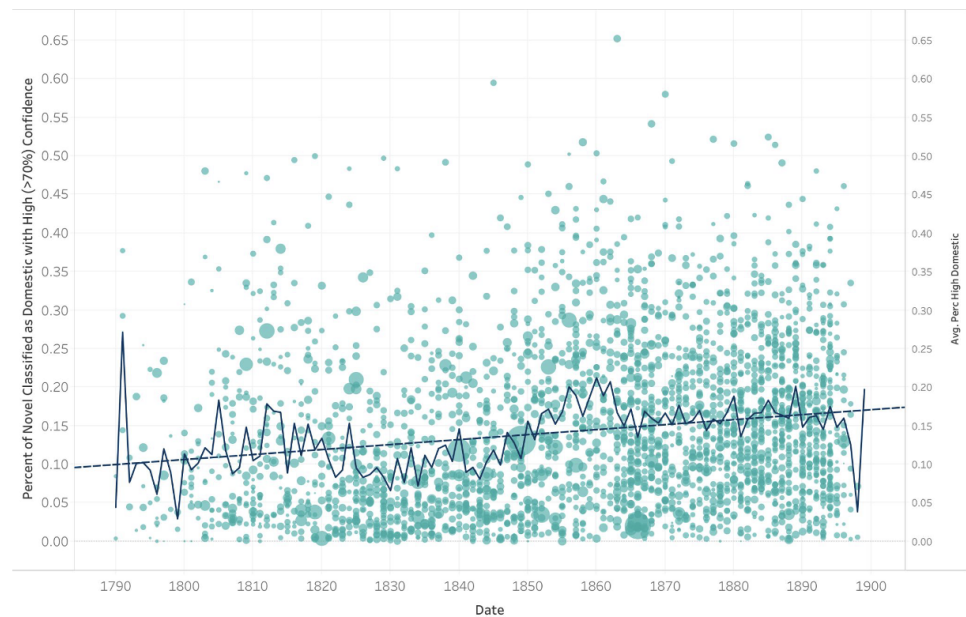


Figure 6: Domesticity score trendline of the 19th-century novel corpus. Due to the limited data points provided for the respective years, the beginning and end of the line plot are not representative.

one night after a very tranquil day, over the fire in the central room. [...] Ombra came in from her 527
own room in her dressing-gown with her dusky hair over her shoulders. Dusky were her looks 528
altogether, like evening in a Winter's twilight. 529
(Passage from *Ombra* by Margaret Oliphant, automatically predicted as “domestic” with a proba- 530
bility of 0.978) 531

5. Analysis and Results 532

In this section, we compile the predicted domesticity scores across the texts in our corpus 533
and visualize them diachronically to get a new perspective on domesticity within British 534
and Irish literary history over time (see subsection 5.1). We then focus on authors 535
(see subsection 5.2). We also addressed the challenges posed by dialogic passages 536
(see subsection 5.3) to detect domestic spaces. As the proportion of “domestic” to 537
“other” classifications in our automated classification echoes the percentages found 538
by our annotators (described above), we take this as an additional validation for our 539
model. The strong performance of our model in detecting the specific space class 540
“domestic” based on manually labeled data highlights the potential of our classification 541
approach and suggests that similar techniques could be successfully applied to other 542
space classification tasks, such as identifying urban settings in detective fiction or 543
automobile scenes in American short stories. 544

5.1 Domesticity and Literary History 545

To gain new insights into the diachronic development of domesticity across the corpus, 546
we visualized the predicted domesticity scores for each novel by calculating the percent- 547
age of passages it contained that were classified with a greater than 70% probability 548

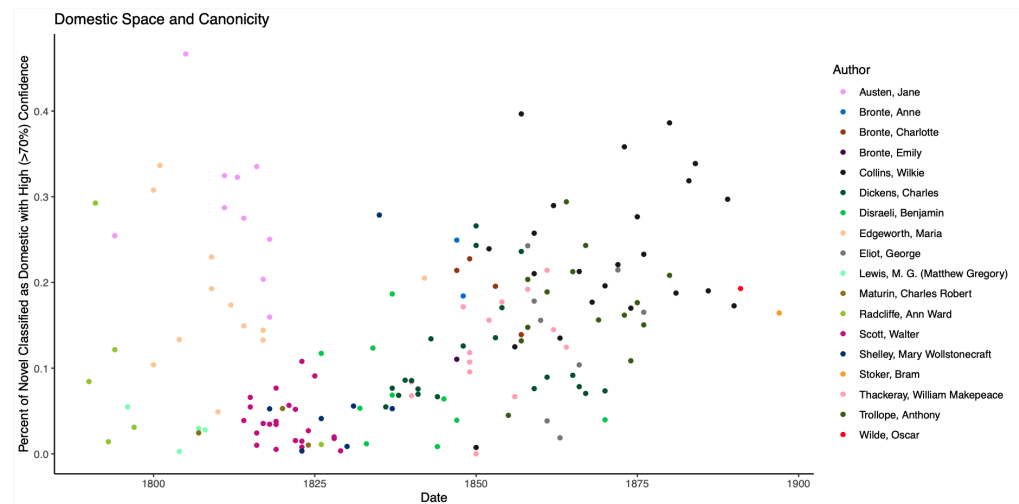


Figure 7: Visualization of a set of canonized authors’ texts and the percentage of passages with a likelihood of being domestic above 70%.

of being domestic. Each data point in Figure 6 represents the percentage of highly domestic passages within an individual novel, while the variable trendline reflects a rolling average of high domesticity throughout the long 19th century. The canonical novel with the highest percentage of passages predicted as “domestic” (highest dot in Figure 6 at 0.65) is Julia Kavanagh’s *Queen Mab* (1863) – an Irish author known for her “fashionably domestic [...] style” and writing for young women readers (Sutherland 1989, 343). The next highest point is at 0.60, which is British author Elizabeth Missing Sewell’s novel *Gertrude* (1845), primarily set in the home of the female protagonist and stressing the importance of familial responsibilities (Frerichs 1974). The third and fourth highest dots are again written by Julia Kavanagh, namely *Silvia* (1870) with 0.58 and *Dora* (1868) with 0.54.

The trendline provides a lens to examine the shifting prevalence of the domestic in different novelistic genres over time. For instance, the late 18th century, characterized by slightly lower domesticity levels, coincides with the popularity of Gothic romances and travel narratives set abroad. In the 1810s, Jane Austen’s domestic novels emerge, followed by the rise of historical and Newgate fiction in the 1820s and 1830s. From 1850 to 1870, there is a noticeable increase in domesticity, likely linked to the prominence of domestic spaces in both realist novels and sensation fiction. Toward the end of the 19th century, the growing popularity of adventure fiction, which by default does not represent domesticity, reshapes the Victorian novel, with the trendline reflecting this shift.

5.2 Domesticity and Canonicity

For the authors writing in the British Romantic period, from the last decades of the 18th century through the earliest decades of the 19th, the points representing their novels tend to form distinct clusters. These clusters also tend to correspond to particular novelistic genres. Ann Radcliffe and Matthew Lewis, whose points group together in the bottom left-hand corner, are both writers of Gothic fiction. Gothic novels in the Romantic period often take place in castles (which could be tagged as domestic spaces according to our annotation guidelines referring to the public or private access to the room in question) or

convents (which, despite that people live in them, were always tagged as non-domestic within our annotation guidelines). Ann Radcliffe in particular is known for her long, descriptive scenes of sublime landscapes and outdoor travel. To the right of their clusters, the points representing novels by Walter Scott also form a distinct group. Walter Scott's historical novels tend to focus on public spaces and represent the characters' experiences within large historical events (see also Lukács (1983)). A slight exception to this pattern of highly-clustered authors is Jane Austen, whose marriage plots spend so much time in houses that two of her novels – *Mansfield Park* (1814) at 0.28 and *Northanger Abbey* (1817) at 0.16, which is an old abbey converted into a domestic space – are named after them⁵. The location of the biggest outlier among her works, *The Watsons* (1805) at 0.46, seems to be, in part, a factor of length, since it was never published and exists only as novel fragment of 21,505 words.

In the Victorian period (1837–1901), realism and sensation fiction dominate the graph. Although their plotting differs – realism prioritizes everyday life, whereas sensation fiction foregrounds exceptional crimes and secrets – both genres often take place in homes. That being said, unlike the canonical authors represented in the earlier part of the century, authors like Charles Dickens, George Eliot, and Wilkie Collins are often spread out across a range of percentages for passages classified as highly domestic. For Dickens, for example, the most “highly domestic” novel is *David Copperfield* (1850) at 0.27, the one that, fittingly, has Angel-in-the-House Agnes Wickfield. However, most of Dickens's novels hover around 0.05 to 0.15 and show investment in representing both work and home environments. Even *Bleak House*, a novel named directly after two houses with that exact name and, arguably, after many other bleak homes represented alongside them, is only slightly more “highly domestic” at about 0.14 than the other Dickens novels represented by the points on either side of it. Given Dickens' interest in representing the courts and the slums of London in *Bleak House*, this does not come as a surprise. George Eliot's novels hover mostly around 0.2, with some above and some below; *Middlemarch* (1872) at 0.21, known for being a canonical example of Victorian realism, includes several marriage plots and their respective domestic spaces, but it is also steeped in the politics and labor of the town of Middlemarch and the surrounding countryside. Of Wilkie Collins's sensation novels, *The Dead Secret* (1857) at 0.4, is the most “highly domestic” according to the model's classifications; like many works of sensation fiction, this novel centers an inheritance plot and themes of family and illegitimacy.

The placement of some points on the visualization may be surprising. Oscar Wilde's *The Picture of Dorian Gray* (1890) and Bram Stoker's *Dracula* (1897) could be identified as about 0.19 and 0.16 “highly domestic,” respectively. Although early iterations of the Gothic novel as practiced by Radcliffe and Lewis rarely take place in domestic spaces, in the more urban Gothic of Wilde and Stoker, these Gothic plot lines more often do; take, for example, the location of Dorian's portrait in his own home.

5. While *Northanger Abbey* is indeed titled after a domestic site, much of the novel's action actually unfolds in public and quasi-public settings like Bath, with the abbey serving more as a site of symbolic and imagined significance than as the primary narrative location. However, despite this detail, the Austen texts provide a very high number of domestic space passages on average in relation to the other authors' texts, underscoring her sustained focus on the interior and private spheres.

	Dialogue	No Dialogue	Totals
Domestic Space	411,718	104,462	516,180
Ambiguous	1,035,524	347,836	1,383,360
Other	1,062,109	520,714	1,582,823
Totals	2,509,351	973,012	3,482,363

Table 3: Number of passages in domestic space, as classified by the model with probability > 0.7, compared to the number of passages containing at least some dialogue, as estimated by the presence of quotation marks.

5.3 Domesticity and Dialogue

618

In our annotation process, we noticed that dialogue was a common source of difficulty. Passages containing dialogue often seemed to be set in domestic spaces, but they lacked any explicit signs of their location, and we thus often could not definitively tag them for inclusion in our training data when restricted to the six-sentence passages. Our intuitions aligned with literary critical arguments about the correlation between household interiors and dialogue in domestic fiction. We were vindicated when, during our ground-truthing process (see subsection 3.3), we found passages consisting wholly of dialogue that our model correctly identified as set in domestic space.

619
620
621
622
623
624
625
626

To investigate this relationship between dialogue and domestic space further, we conducted a short exploratory study, where we found that passages containing dialogue were more likely to be set in domestic space and vice versa, a strong signal for their connection. As a proxy for the presence of dialogue, we found all the passages that contained single or double quotation marks, excluding those used as apostrophes. This method is somewhat imperfect: It misses passages from the middle of monologues, while catching those that might contain only a short portion dialogue at the end or beginning. It also encounters some problems due to OCR, dialogue without quotations, and quotation marks at the end of passages. In a sample of 100 passages, the method’s recall for finding passages with dialogue was 0.92, the precision was 0.9, and the F-score was 0.91. However, we judged these results sufficient for an exploratory study of the correlation. Our results (see Figure 8 and Table 3) show that dialogue is present, even predominant, across all spatial categories. However, passages with dialogue are 53% more likely to be in domestic spaces than those without dialogue, and passages in domestic space are 19% more likely to include dialogue compared to passages set in unambiguously non-domestic spaces. As we move from non-domestic to ambiguous and finally domestic spaces, there is more and more dialogue. The bidirectional relationship, contrast with non-domestic spaces, and high number of observations across categories imply a strong connection between domestic space and dialogue. Future work might explore the underlying factors in this relationship; we hypothesize, based on an analysis of the words distinctive of domestic spaces, that the prevalence of names, personal address, and family titles in dialogue plays a role. But our brief analysis here underlines our larger methodological arguments. Literary texts represent space much more complexly than just through mentions of place names and spatial terms, including in dialogues between characters that do not include any explicit spatial information yet still signal a domestic setting. Our method is able to detect these pervasive, nuanced, and fundamental aspects of literary space.

627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653

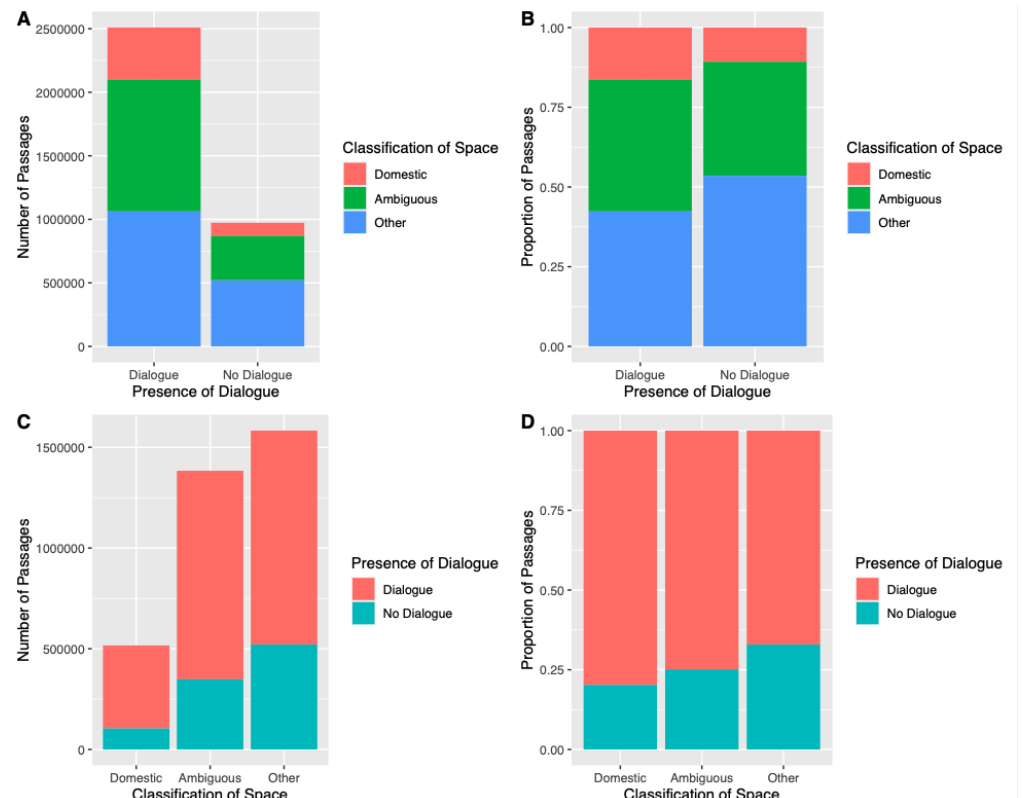


Figure 8: Visualization of the absolute numbers and proportions of dialogue in domestic space.

6. Discussion

654

Our analysis highlights the complex interplay between domestic space and domesticity, emphasizing that expected domesticity does not always align with physical domestic spaces. While houses frequently serve as markers of domestic settings, domesticity is not solely confined to them. The model successfully identified high domesticity scores in traditionally domestic environments, yet it also revealed instances of unexpected domesticity in unconventional locations such as gardens, carriages, and even ships. These findings suggest that domesticity extends beyond physical structures, emerging instead through relational and behavioral cues, such as familial interactions, caregiving, or moments of emotional intimacy.

The classification further underscores the gendered and classed nature of domesticity. Passages featuring female characters engaged in household affairs or emotional reflection were more likely to receive high domesticity scores, reinforcing historical associations between women and domestic spaces. Meanwhile, lower-class settings often exhibited a more ambiguous domesticity score, particularly in spaces where work and home life intersected. This suggests that domesticity is not merely a spatial designation but also a socio-cultural construct shaped by class and gender expectations.

Finally, instances of unexpected domesticity, such as domestic-like interactions occurring on ships or characters finding moments of intimacy in liminal spaces, challenge rigid binaries between public and private spheres. These cases highlight how domesticity can emerge in transient, mobile, or even hostile environments, as seen in characters engag-

ing in intimate conversations in carriages or tending to one another in non-traditional 675
 settings. The model's handling of these cases suggests that while domesticity is often an- 676
 ticipated in certain spaces, its presence can also surface in other places where characters 677
 engage in acts of care, reflection, or emotional connection. 678

7. Conclusion 679

Our approach to modeling domestic space in 19th-century English and Irish fiction pro- 680
 vides new insights into both the concept of domesticity and computational approaches 681
 to analyzing literary settings. Our findings challenge conventional narratives that rigidly 682
 define domesticity by location, instead emphasizing the importance of activities and 683
 interactions that create domesticity in a variety of spaces within the novel. By mov- 684
 ing beyond toponymic markers and incorporating non-traditional spaces, our model 685
 demonstrates the fluidity of domesticity and its dependence on relational and narrative 686
 cues. 687

The validation of our model against ground truth data reinforces its reliability while 688
 also highlighting areas of ambiguity, particularly in dialogue-heavy passages. This 689
 methodological approach addresses a critical gap in digital humanities research, offering 690
 a scalable way to analyze non-toponymic spaces computationally. In doing so, our 691
 study contributes to a new quantitative history of domestic space, revealing unexpected 692
 patterns in where and how domesticity is represented across 19th-century novels. 693

Ultimately, our results reveal the 19th-century novel not as a monolithic expression of 694
 gendered and classed domesticity, but as an evolving exploration of what domestic space 695
 could be. The strong language of domesticity captured by our model suggests that these 696
 novels were not merely reinforcing hegemonic ideals but experimenting with different 697
 forms of domestic representation. By rethinking domesticity through a computational 698
 lens, we uncover a more nuanced and dynamic portrayal of space, identity, and social 699
 structure in the literary imagination of the period. 700

8. Data Availability 701

Data and Code can be found here: [https://github.com/literarylabb/jcls_domestic](https://github.com/literarylabb/jcls_domestic_space) 702
[_space](https://github.com/literarylabb/jcls_domestic_space). 703

9. Acknowledgements 704

We thank Annie Lamar and our student assistants Sophie Schwarzhappel and Julia 705
 Gershon, for the support in the annotation work. We also thank Kent Chang for his idea 706
 of splitting the recognition approach into two steps, which increased the performance 707
 of our model. 708

10. Author Contributions 709

Svenja Guhr: Project administration, Conceptualization, Methodology, Data curation, 710
 Formal analysis, Writing – original draft 711

Jessica Monaco: Conceptualization, Methodology, Data curation, Formal analysis,	712
Writing – original draft	713
Alexander J. Sherman: Conceptualization, Methodology, Data curation, Formal analysis,	714
Writing – original draft	715
Matt Warner: Conceptualization, Methodology, Data curation	716
Mark Algee-Hewitt: Project administration, Conceptualization, Methodology, Writing	717
– review & editing	718

References


719


- Armstrong, Nancy (1987). *Desire and Domestic Fiction: A Political History of the Novel*. Oxford University Press. 720
- Baledent, Anaëlle, Yann Mathet, Antoine Widlöcher, Christophe Couronne, and Jean-Luc Manguin (2022). “Validity, Agreement, Consensuality and Annotated Data Quality”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, 2940–2948. <https://aclanthology.org/2022.lrec-1.315> (visited on 01/29/2025). 722
- Bamman, David (2021). *BookNLP*. <https://github.com/booknlp/booknlp> (visited on 11/04/2024). 723
- Bamman, David, Kent K. Chang, Lucy Li, and Naitian Zhou (2024). “On Classification with Large Language Models in Cultural Analytics”. In: *Proceedings of the Computational Humanities Research Conference 2024*. CEUR Workshop Proceedings 3834. CEUR, 494–527. <https://ceur-ws.org/Vol-3834/paper119.pdf> (visited on 01/29/2025). 724
- 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)*. Association for Computational Linguistics, 2138–2144. [10.18653/v1/N19-1220](https://doi.org/10.18653/v1/N19-1220). 725
- Bologna, Federica (2020). “A Computational Approach to Urban Space in Science Fiction”. In: *Journal of Cultural Analytics* 5.2. [10.22148/001c.18120](https://doi.org/10.22148/001c.18120). 726
- Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil (2018). “Universal Sentence Encoder for English”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 169–174. [10.18653/v1/D18-2029](https://doi.org/10.18653/v1/D18-2029). 727
- Chadwyck-Healey Literature Collections and ProQuest (2024). *Nineteenth-Century British Fiction*. https://collections.chadwyck.com/marketing/products/about_iloc.js?collection=nfc (visited on 10/02/2024). 728
- Cohen, Monica F. (2017). “Domesticity in Victorian Literature”. In: *Oxford Research Encyclopedia of Literature*. Oxford University Press. [10.1093/acrefore/9780190201098.013.252](https://doi.org/10.1093/acrefore/9780190201098.013.252). 729


- Davidoff, Leonore and Catherine Hall (1987). *Family Fortunes: Men and Women of the English Middle Class, 1780-1850*. Women in Culture and Society. University of Chicago Press. 753 754 755
- Dennerlein, Katrin, Thomas Schmidt, and Christian Wolff (2023). "Computational Emotion Classification for Genre Corpora of German Tragedies and Comedies from 17th to Early 19th Century". In: *Digital Scholarship in the Humanities*, fqado46. 10.1093/llc/fqad046. 756 757 758 759
- 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*, 4171–4186. 10.18653/v1/N19-1423. 760 761 762 763 764
- Fludernik, Monika and Suzanne Keen (2014). "Introduction: Narrative Perspectives and Interior Spaces in Literature Before 1850". In: *Style* 48 (4), 453–460. 10.5325/style.48.4.453. 765 766 767
- Freedgood, Elaine (2006). *The Ideas in Things: Fugitive Meaning in the Victorian Novel*. University of Chicago Press. 768 769
- Frerichs, Sarah Cutts (1974). "Elizabeth Missing Sewell: A Minor Novelist's Search for the Via Media in the Education of Women in the Victorian Era". PhD thesis. Brown University. 770 771 772
- Google (2023a). *experts/bert*. Kaggle. <https://www.kaggle.com/models/google/experts-bert> (visited on 01/29/2025). 773 774
- (2023b). *universal-sentence-encoder*. Kaggle. <https://www.kaggle.com/models/google/universal-sentence-encoder> (visited on 01/29/2025). 775 776
- Kababgi, Daniel, Giulia Grisot, Federico Pennino, and Berenike Herrmann (2024). "Recognising Non-named Spatial Entities in Literary Texts: A Novel Spatial Entities Classifier". In: *Proceedings of the Computational Humanities Research Conference 2024*. Vol. 3834, 472–481. <https://ceur-ws.org/Vol-3834/paper59.pdf> (visited on 12/10/2024). 777 778 779 780 781
- Kingma, Diederik P. and Jimmy Ba (2017). *Adam: A Method for Stochastic Optimization*. 10.48550/arXiv.1412.6980. 782 783
- Krippendorff, Klaus (2018). *Content Analysis: An Introduction to Its Methodology*. 4th ed. SAGE. 784 785
- Lukács, Georg (1983). *The Historical Novel*. Bison Books. University of Nebraska Press. 786
- Marcus, Sharon, ed. (2007). *Between Women: Friendship, Desire, and Marriage in Victorian England*. Princeton University Press. 787 788
- Moretti, Franco (1999). *Atlas of the European Novel, 1800-1900*. 789
- Parigini, Margherita and Mike Kestemont (2022). "The Roots of Doubt. Fine-tuning a BERT Model to Explore a Stylistic Phenomenon". In: *Proceedings of the Computational Humanities Research Conference 2022*. CEUR Workshop, 72–91. 790 791 792
- Piatti, Barbara (2016). "Mapping Fiction: The Theories, Tools and Potentials of Literary Cartography". In: *Literary Mapping in the Digital Age*. Ed. by David Cooper, Christopher Donaldson, and Patricia Murrieta-Flores. Routledge, 88–101. 793 794 795
- Pichler, Axel and Nils Reiter (2022). "From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities". In: *Journal of Cultural Analytics* 7 (4). 10.22148/001c.57195. 796 797 798

- Reiter, Nils (2020). "Anleitung zur Erstellung von Annotationsrichtlinien". In: *Reflektierte algorithmische Textanalyse*. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. De Gruyter, 193–202. [10.1515/9783110693973-009](https://doi.org/10.1515/9783110693973-009). 799–800
- Ryan, Marie-Laure (2014). "Space". In: *The Living Handbook of Narratology*. <https://www-archiv.fdm.uni-hamburg.de/lhn/node/55.html> (visited on 01/23/2025). 802–803
- Ryan, Marie-Laure, Kenneth E. Foote, and Ma'oz 'Azaryahu (2016). *Narrating Space/Spatializing Narrative: Where Narrative Theory and Geography Meet*. Theory and interpretation of narrative. The Ohio State University Press. 804–806
- Schumacher, Mareike (2023). *Orte und Räume im Roman: Ein Beitrag zur digitalen Literaturwissenschaft*. Digitale Literaturwissenschaft. Berlin, Heidelberg: Springer Berlin Heidelberg. [10.1007/978-3-662-66035-5](https://doi.org/10.1007/978-3-662-66035-5). 807–809
- Schumacher, Mareike, Marie Flüh, and Marc Lemke (2022). "The Model of Choice. Using Pure CRF- and BERT-based Classifiers for Gender Annotation in German Fantasy Fiction". In: *Book of Abstracts*. ADHO 2022 - Tokyo. <https://dh2022.dhii.asia/dh2022bookofabsts.pdf> (visited on 01/23/2025). 811–813
- Soni, Sandeep, Amanpreet Sihra, Elizabeth Evans, Matthew Wilkens, and David Baman (2023). "Grounding Characters and Places in Narrative Text". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, 11723–11736. [10.18653/v1/2023.acl-long.655](https://doi.org/10.18653/v1/2023.acl-long.655). 814–818
- Sutherland, John (1989). *The Stanford Companion to Victorian Fiction*. Stanford University Press. 819–820
- Wilkens, Matthew (2013). "The Geographic Imagination of Civil War-Era American Fiction". In: *American Literary History* 25.4, 803–840. 821–822
- Yang, Ziyi, Yinfei Yang, Daniel M Cer, Jax Law, and Eric Darve (2021). *Universal Sentence Representations Learning with Conditional Masked Language Model*. <https://openreview.net/forum?id=WDVD4lUCTzU> (visited on 01/23/2025). 823–825
- Zundert, Joris J. van, Marijn Koolen, Julia Neugarten, Peter Boot, Willem van Hage, and Ole Musmann (2022). "What do we talk about when we talk about topic?" In: *Proceedings of the Computational Humanities Research Conference 2022*. CEUR Workshop, 398–410. 826–829


Urban Transportation in Latvian Novels or Why do you use a 19th-century horse-drawn cab when you have a 20th-century taxi?

Eva Kristsons-Eglāja¹ 

Anda Baklāne² 

Valdis Saulespurēns² 

1. Institute of Literature Folklore and Art, University of Latvia , Riga, Latvia.

2. Department of Digital Development, National Library of Latvia , Riga, Latvia.

Citation

Eva Eglāja-Kristsons, Anda Baklāne, and Valdis Saulespurēns (2025). "Urban Transportation in the Latvian Early Novels or Why do you use a 19th-century horse-drawn cab when you have a 20th-century taxi?" In: *CCLS2025 Conference Preprints* 4 (1). [10.26083/tuprints-00030146](https://doi.org/10.26083/tuprints-00030146)

Date published 2025-06-17

Date accepted 2025-04-18

Date received 2025-02-07

Keywords

Latvian novels, urban transportation, Word2Vec, large language models, horse-drawn cab, taxi

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check <https://jcls.io> for the final journal version.

Abstract. The article explores the depiction of urban transportation in Latvian novels from the late 19th century to the mid-20th century, aiming to understand how these images reflect broader societal modernization and technological and social changes in an increasingly urbanized environment. The first part of the study explores the frequency and dissemination of mentions of urban vehicles in the novels. Two methods were used to identify relevant transport terms – Word2Vec and the Gemini 1.5 language model –, comparing the results of both approaches. In the second part of the study, particular attention is given to the horse-drawn cab and the taxi, which illustrate economic development, societal modernization, and the growing disparity between different social strata. The study emphasizes that transportation is a practical means of mobility and a significant cultural and social symbol. The study uses the data set *Corpus of Latvian Early Novels*, which includes novels published between 1879 and 1940.

1. Introduction

Literature as a source for the history of everyday life has become a vast field of research, offering profound insights into people's life experiences, emotions, and social processes, which are often absent from traditional historical sources. The link between everyday life and modernization processes is particularly noteworthy in prose fiction: technological progress, urbanization, and changes in social structure are documented in the literature both directly and through symbols, revealing not only the external world but also changes in the psychology and culture of society.

Urban transportation is not only a functional infrastructure; it is also a powerful cultural symbol. As cities modernize, modes of transport shape and reflect changing experiences of space, time, identity, and technology. In the literature, vehicles such as horse-drawn cabs¹ (*ormanis*, *fūrmanis*, *važonis*) and taxis (*taksis*, *taksītis*, *taksometrs*) often serve as more than a means of travel; they become narrative devices that express social aspiration, class tension, emotional crisis, or ideological transition. This article explores how the evolution of urban transportation – from horse-drawn carriages to motorized taxis – is

1. In this study, the term "horse-drawn cab" (in Latvian *ormanis*) will be used to refer to carriages commonly employed during the 19th century for passenger transportation. This term refers to vehicles such as the "hansom cab" and "hackney carriage," which were widely used in urban areas of the period.

reflected in Latvian fiction across the late 19th and early 20th centuries. The core research question is: How does the transition from horse-drawn cab to taxi reflect shifting urban subjectivities and sociotechnical imaginaries in Latvian literature?

To approach this question, we adopt a hybrid methodology that combines computational analysis of literary corpora with close reading of selected narrative episodes. Our approach begins with the analysis of transport-related vocabulary using word embeddings and the large language model for keyword extraction across the *LatSenRom* corpus, which comprises Latvian novels published between 1876 and 1940. To contextualise the prevalence of the concepts of the horse-drawn cab and taxi among other vehicles, we also examine the frequency of mentions of other land vehicles, including both mechanised vehicles modern for their time (car (*auto*), automobile (*automobilis*), railway (*dzelzceļš*), tram (*tramvajs*), etc.) and horse-drawn transport (carriage (*pajūgs*), wagon (*vezums*), farm wagon (*ore*), etc.). We then turn to interpretative analysis, examining how various modes of transport function symbolically and narratively within individual texts, particularly in relation to representations of modernity, class, gender, and psychological experience (Kohlrausch and Behrends 2014).

This dual approach is situated within the broader field of computational literary studies, which has increasingly emphasized the value of combining distant reading with interpretive frameworks. The methods employed in the article engage and contribute to the vast tradition of computational and digital literary studies, as well as digital history studies in particular sharing the ambition to analyze bodies of text that span long periods to capture changes in language, style, genre, as well as culture and society that occur over time (Moretti 2005, Moretti 2013, Jockers 2013, Underwood 2019, Piper 2018, Graham et al. 2016, Fridlund et al. 2023). In approaching the study of cities, urbanization, and modernization, it is notable that the representation of urban and rural spaces has long occupied a central place in scholarly research. Franco Moretti, in his pioneering work *Atlas of the European Novel 1800–1900*, examined not only the geographical distribution of literary forms but also the spatial dynamics depicted within the novels themselves (Moretti 1998). Building on this foundation, other scholars have expanded the exploration of how rural and urban spaces are modeled in literature. Dennis Yi Tenen, for example, integrated narratological concepts such as diegetic density and clutter distance to capture the complexity of spatial representation in literary texts (Tenen 2018). Federica Bologna, meanwhile, investigated the lexical presence of urban-related terms in twentieth-century English science fiction (Bologna 2020).² Moving beyond the mere cataloging of place names, current research on urban spaces increasingly foregrounds the vocabulary of urban material culture, highlighting its significance for modeling and identifying various types of spaces. Our study builds on these foundations by applying such methods to a small-language literary tradition, where digital tools remain underutilized. This paper, however, does not yet attempt to model broader semantic domains, opting instead to concentrate on the more narrowly scoped domain of vehicles.

While 21st-century scholars have increasingly embraced complex methodologies, particularly those leveraging machine learning, to move beyond the traditional corpus analysis

2. In the context of Latvian studies of literary geography, Zita Kārkla and Eva Eglāja Kristsone have studied the geographical places in women's prose fiction; see Kārkla and Eglāja-Kristsone 2022.

paradigm of word search and frequency analysis, the rise of new-generation language models has sparked a renewed interest in earlier approaches. Word embeddings and new-generation language models offer unprecedented opportunities to identify terms that align with the specific semantic domains a researcher aims to explore.

In this study, we explore a mixed-methods approach, integrating digital tools, simple document frequency and word frequency counts, and qualitative interpretative methods. We begin by employing the Word2Vec machine learning algorithm (Mikolov et al. 2013) alongside with Gemini 1.5 language model (Georgiev 2024) to identify transport-related concepts. This is followed by computer-assisted frequency counts to trace the occurrence of these concepts across a corpus of novels. Next, we use the concordance and word frequency list features of a corpus analysis platform ³ to investigate linguistic patterns in greater depth. Finally, we conduct a close interpretative analysis of representations of the horse-drawn cab and taxi in the LatSenRom corpus, drawing on perspectives from modern and urban material culture studies within Latvian as well as broader literary and cultural histories. An earlier stage of this research that did not include the usage of the Gemini 1.5 model was documented in a publication in Latvian (Eglāja et al. 2024).

The most technologically complex aspect of the study lies in the methodology used for identifying vehicle-related terms within the text. In the past, researchers have relied on manual or semiautomated methods to discover and annotate concepts of interest. In recent years, efforts have increasingly scaled up with the adoption of word embedding-based techniques (Mikolov et al. 2013), followed more recently by zero-shot large language models (LLMs) (Karjus 2023, Fan et al. 2023, Törnberg 2024, Ziems et al. 2023). Methods based on embeddings have proven particularly effective in automating the identification of semantic similarity and difference in terms or larger discourse segments (Rodman 2020, Rodriguez et al. 2023).

The dataset used in the study is the *Corpus of Latvian Early Novels (1879–1940)* ("Latviešu senāko romānu korpuss"; hereafter LatSenRom), which includes novels written in Latvian and published in book form between 1879 and 1940. As a corpus, LatSenRom exemplifies historical digitized datasets that span several decades and reflect a paradigm shift in typeface usage and orthographic norms. It highlights persistent challenges that humanities researchers face, such as optical recognition errors, evolving writing conventions, and polysemy. Moreover, there is a significant disparity in the number of works published before and after the 1920s. Working with such data underscores the importance of addressing data heterogeneity and quality issues. It advocates for digital analysis not as a "quick fix" but as a relatively slow, iterative process allowing backtracking to correct errors or refine approaches. Without the ability to contextualize the analysis results within the text, or when datasets are too large for manual verification, the reliability of those results may be compromised. These considerations emphasize the need for meticulous methodologies and critical engagement when working with digitized historical data.

The case study was inspired by the question posed in Kārlis Lapiņš's novel *Students in the Farm* ("Studenti fermā"): "Why do you drive a 19th-century carriage when you have a 20th-century taxi at hand?" (see Lapiņš 1934b, 129). The question succinctly describes

3. The corpus analysis platform of the National Library of Latvia: <https://korpuss.lnb.lv>.

the tension between tradition and modernity, symbolically embodied by the transition from the carriage to the taxi, which examines the impact of this transition in the broader context of urban transport development in Latvian literature.

The introduction of motorized taxis was a significant shift in urban mobility, offering greater efficiency and a new form of urban anonymity that contrasted sharply with the more personalized and slower journeys provided by horse-drawn carriages. Viewed through this lens, the transition from horse-drawn carriages to motorized taxis not only reshaped urban transport but also left a lasting mark on the cultural and literary landscape of the era. From a narrative analysis perspective, the depiction of urban transport reflects the characters' emotional state, social status, and the broader cultural environment. The increasing presence of urban transportation – horse-drawn carriages, trains, trams, buses (or omnibuses), and cars – within the literature appears as a significant element that serves both as a backdrop for the narrative's action and as a symbol of the broader social changes of the era.

The insights gained from the LatSenRom research into the prevalence of vehicle concepts in late 19th- and early 20th-century Latvian long prose are based on an examination of the largest Latvian fiction dataset to date. The results provide information on the mention of the horse-drawn cab, the taxi, and other vehicles in the data studied and, thanks to the juxtaposition with the facts of transport history and the insights of transport history researchers, can also contribute to studying material history and cultural history in general.

2. Dataset and the methodology of term extraction

2.1 Dataset

The LatSenRom dataset was developed at the National Library of Latvia in consultation with researchers at the Institute of Literature, Folklore and Art at the University of Latvia. The corpus version used for this article contains 463 works (novels and parts of books, trilogies, tetralogies) by 190 authors. The dataset contains about 36.3 million tokens, including about 28.8 million words.⁴

The creation of the LatSenRom began as part of an international project – the COST action "Distant Reading for European Literary History" (CA16204), which took place from 2017 to 2022 (Schöch et al. 2021). The project's main objective was to establish the *European Literary Text Collection* (ELTeC). ELTeC is a collection of datasets comprising corpora of novels in various languages, compiled according to unified principles. Each language corpus was designed to include 100 works published in European countries between 1840 and 1920. These works were carefully selected from the broader range of publications based on balancing criteria: a balanced number of works from each decade, author gender categories, work length categories, and the canonical status of the works. Additionally, in line with the selection algorithm, no more than three works by any single author were included in the corpus.

The strict selection criteria served well in creating a representative corpus for major

4. For a more detailed account of creation of the corpus, see Baklāne et al. 2024.

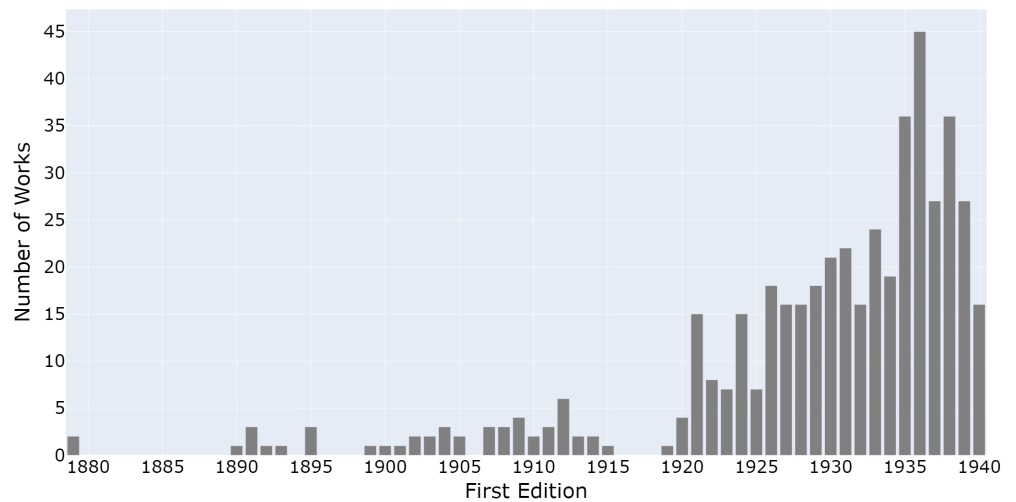


Figure 1: LatSenRom first editions from 1879 to 1940

European languages, where it was possible to choose from a vast selection of thousands of works. At the same time, the distinctiveness of smaller languages became strikingly evident – in several countries, the novelistic literary tradition developed later, involved fewer women, and had a relatively small number of works considered canonical. This experience prompted the creators of the Latvian novel corpus to adopt a different approach for its further development: instead of forming a representative selection, a decision was made to include in the dataset all original novels published in book form in the Latvian language in the present-day territory of Latvia or abroad.

Figure 1 illustrates the frequency of new novel publishing over time. It is important to note that the dataset excludes novels published solely in periodicals during the period under consideration. Based on book publishing data, the novel as a genre remained relatively unfamiliar in Latvia at the end of the 19th and the beginning of the 20th centuries. Until 1899, only 12 novels were published, with an additional 38 appearing between 1900 and 1918 (year of the founding of the independent Republic of Latvia). However, following Latvia's independence, publishing of books surged, with more than 400 new novels released between 1919 and 1940.

Data processing and analysis were carried out using a normalized, morphologically tagged version of the LatSenRom corpus. The normalization process involved the automatic conversion of texts originally printed in an older orthographic tradition that employed the Fraktur script into the contemporary orthography using the Antiqua-based alphabet. These texts comprise approximately 16 percent of the corpus. Morphological tagging allows all word forms to be linked to their base forms (lemmas), enabling comprehensive retrieval across the corpus. Working with lemmas also improves processing efficiency and supports the generation of Word2Vec word embeddings.

2.2 Methodology

The first phase of the study focused on identifying vehicle terms that appear in LatSenRom. Various methodologies could be applied to detect transport-related concepts, such as subjectively hypothesizing vehicle names, compiling vocabularies from dictionaries or research sources dedicated to a particular topic (see the dictionary-based selection

method used in Bologna 2020) or manually annotating them within the texts. However, compiling vocabularies subjectively or from external sources risks incompleteness, as it lacks systematic verification, while the manual annotation is highly labor-intensive. To maximize the identification of transport-related terms, the study initially employed the natural language processing technique based on Word2Vec to analyze the texts. These results were later cross-referenced with those generated by the Gemini 1.5 language model, ensuring a more comprehensive and robust analysis.

Although large language models (LLMs) in all likelihood outperform Word2Vec, this is not yet consistently the case for the Latvian language at the time of writing this paper. The performance of LLMs remains notably weaker for low-resource languages and the methodologies employed still require further validation at this stage of model development. A comprehensive comparison of multiple LLMs was beyond the scope of this study, as the primary objective was not the evaluation of methods but the accurate identification of as many transportation-related terms as possible. Preliminary testing, however, identified several state-of-the-art LLMs that demonstrated acceptable performance for Latvian; among them, Gemini 1.5 was selected for the exploration of LatSenRom.⁵ While non-commercial models would be preferable for research purposes, fully open-source alternatives currently fail to provide satisfactory results for texts in Latvian. Additionally, LLMs in general remain prone to various bias, anomalies, and hallucinations (Törnberg 2024). To mitigate errors, a hybrid approach should be implemented for cross-verifying results. In this study, Word2Vec outputs and human verification were used to cross-examine the findings generated by Gemini 1.5.

The Word2Vec method, widely known since a 2013 study by Google scientists (Mikolov et al. 2013), uses neural networks to create word embeddings. Compared to other early language model-based methods such as GPT and BERT, Word2Vec is directly optimized for finding similarities between concepts. Although Word2Vec's capabilities are limited compared even to BERT and early GPT, Word2Vec is more accessible for everyday use in research practice as a LatSenRom-sized model using a pre-trained list of lemmas can be trained in a day on a standard personal computer, without the need for a supercomputer or cloud computing solutions. The Python programming language external Gensim library (Řehůřek and Sojka 2010) with built-in Word2Vec support was used in the study.

Two datasets were processed to generate Word2Vec embeddings. The first model was based on the LatSenRom corpus, consisting of 463 documents. After text cleaning and optimization, embeddings for 105,677 lemmas were derived from the raw dataset of over 36 million tokens. However, empirical analysis of the model's performance indicated that its size and lexical coverage were insufficient to identify a comprehensive list of transport-related concepts.⁶ To achieve a broader representation of transport terminology, a second, larger model was developed using Latvian periodicals published between 1920 and 1940. This periodical-based model utilized 172,240 documents (articles) as the primary data source. Unlike the LatSenRom model, the training corpus for this model was constructed by selecting only documents containing the verb "to drive" (*braukt*) in

5. Comparative analysis of GPT-4o and Gemini 1.5 results was presented at the DHNB 2025 conference; see: Baklāne and Saulespurēns 2025.

6. No definite guidelines exist to indicate what size of the corpus is sufficient for acquiring satisfactory results for various tasks performed based on Word2Vec embeddings, however, the larger size of the training data is known to increase the performance (Rodman 2020).

various conjugate forms, rather than incorporating all available articles from the period. 212
 In terms of temporal coverage, the dataset included materials that overlapped with the 213
 publication years of LatSenRom, ensuring consistency. The raw data volume of the 214
 periodical corpus amounted to approximately 140 million text units, yielding vectors 215
 for 565,623 lemmas (the amount of lemmas is exaggerated due to optical recognition 216
 errors). 217

To obtain a comprehensive inventory of vehicle terms, an initial subjective list was 218
 compiled,⁷ and queries containing these keywords were used to extract broader sets 219
 of related concepts from the Word2Vec LatSenRom and Periodicals models. For each 220
 queried term, a list of the most similar words was generated, applying a similarity score 221
 threshold of 0.6. These lists were then manually evaluated. This process identified 222
 dozens of land vehicles, including various horse-drawn carriages and mechanized 223
 vehicles. Different spelling variants were also discovered, which is particularly important 224
 when working with changing orthography and noisy data. In the subsequent phase of 225
 the study, the terms identified in the periodicals were employed to locate references to 226
 vehicles in LatSenRom. For example, only Word2Vec Periodicals model yielded names 227
 of specific car brands, some of which were later found in LatSenRom as well. 228

To identify vehicles in LatSenRom using Gemini 1.5, only the LatSenRom corpus was 229
 utilized. Several prompts were tested to inquire the corpus with Gemini 1.5, including 230
 instructions that included the list of keywords that were used for querying Word2Vec 231
 models as few-shot examples. The selected prompt prioritized precision and produced 232
 the highest number of valid results following data cleaning. The prompt aimed to give 233
 clear, structured instructions; it defines the form of input and output data, stipulates the 234
 expertise in the Latvian language and transportation and emphasizes that no interpreta- 235
 tion and transformation of results is expected; no examples of vehicles are provided: 236
 8 237

*You are an expert on Latvian language and transportation. Please extract a comprehensive list 238
 of all mentioned land transportation vehicles from the given texts. Show all specific land trans- 239
 portation types found in the text as a list. No other information is needed, just the transportation 240
 terms as they appear in the text. Provide the terms only in Latvian, preserving their original 241
 transcription variants as they appear in the text. Latvian Text follows: 242*

The prompt proved to be highly effective for identifying vehicles; it could be further 243
 adjusted to diminish the number of false positives or terms that do not correspond with 244
 the researcher's hypothetical definition of vehicles. 245

The initial extraction of terms from the Word2Vec LatSenRom model after cleaning 246
 yielded only 31 valid ⁹ unique terms (in addition to initial keywords); Word2Vec results 247
 from the Periodicals supplied 127 unique terms (76 later found also in LatSenRom); 248

7. Initial list: horse (*zirgs*), railway (*dzelzceļš*), horse-drawn cab (*ormanis*), rig (*pajūgs*), coach (*kariete*), sledge (*kamanas*), bicycle or velocipede (*velosipēds*), machine (*mašīna*), motorcar (*automašīna*) train (*vilciens*), locomotive (*lokomotīve*), tram (*tramvajs*), dinky line (*bānītis*).

8. The best practices of formulating instructions for language models are discussed in Törnberg 2024.

9. Only tentatively, can we call all non-valid terms false positives. The range of terms considered relevant in this study is highly dependent on the arbitrary boundaries established for its scope. For instance, it remains debatable whether various types of horses should be included within the semantic domain of transportation studies. In addition to the terms considered valid in this study, the results from all models included other transportation-related terms, such as those referring to air and water transport, vehicle parts, infrastructure elements, drivers, passengers, and others.

Gemini 1.5 model results after cleaning yielded 159 unique terms (see Table 1). Notably, the Word2Vec training data due to the nature of the pre-processing generated only single-word terms and compound words. In contrast, the Gemini model identified multi-word expressions that incorporate a base term already present in the vocabulary, typically representing a specific subtype of a vehicle (e.g., car – sports car; railway – electric railway). Therefore, when considering only single-word terms, the difference between combined Word2Vec and Gemini 1.5 outputs is less pronounced. The data cleaning process proved to be similar across both approaches, each exhibiting comparable levels of redundancy. Only 13.8 percent of the Word2Vec LatSenRom results within the similarity threshold (score ≥ 0.6) represented valid terms, compared to 48.2 percent in the Word2Vec Periodicals model, and 11.4 percent in the Gemini 1.5 results.¹⁰ While Gemini 1.5 generated a higher number of useful terms, the prevalence of false positives rendered this approach unsuitable for automated annotation without significant methodological improvements. In contrast, the comparatively larger Word2Vec Periodicals model produced the fewest false positives within the given similarity score range.

Measure	Word2Vec LatSenRom	Word2Vec Periodicals	Gemini 1.5
Number of valid terms	31	127 / 76	159
Percentage of valid terms	13.8	48.2	11.4

Table 1: Number and percentage of valid terms per method.

3. Analysis of the mentions of vehicles in LatSenRom

3.1 Overview of quantitative findings

As shown in the previous section, approximately 160 unique terms and multi-word expressions referring to vehicles were identified in the general scan of the corpus. Regarding the proportion of motorized and horse-drawn transport, approximately 50 percent of the combined Word2Vec findings in LatSenRom referred to horse-drawn vehicles and bicycles, with the remainder being motorized vehicles. In the Gemini results, close to 49 percent of the identified terms referred to motorized vehicles. This suggests that, by the end of the 1930s, the vocabulary related to modern vehicles was nearly as extensive as that associated with traditional horse-drawn vehicles, despite modern transportation still being a relatively recent phenomenon. Nevertheless, references to modern modes of transportation remain comparatively less frequent than those to horse-drawn vehicles in the novels (see Figure 2).

Mapping mentions of individual concepts is challenging due to the wide variety of forms used to designate the same vehicle: this includes OCR errors, spelling variations, longer and abbreviated forms, as well as literary and colloquial expressions. A number of terms have several meanings, prohibiting simple counting operations. Although the vehicle lists derived from the models were thoroughly reviewed and validated, it is still likely that some terms and variations remain unaccounted for.

10. The percentages reported for the Word2Vec and Gemini 1.5 results are not directly comparable, as the lists were generated using different methodologies; however, these figures serve a descriptive purpose.

When assessing the increase in the frequency of references to specific concepts, it is worth noting that the lack of chronological balance in the corpus results in volatile scores for the relative frequencies in the early years, followed by a substantial increase in absolute counts of mentions in the later years. The absolute document frequency increases significantly over time because of the considerably larger number of editions published: the more novels, the more times all types of vehicles are mentioned (see Figure 3). However, when analyzing a comprehensive corpus that contains all published novels, the apparent increase in mentions should not be dismissed merely as a distortion caused by chronological imbalance. Instead, the absolute document frequency may partly serve as an indicator of the extent to which a concept entered the cultural mainstream, assuming the novel, as a genre, functioned as one of the key vehicles shaping public imagination. Speculatively, it could be hypothesized that, in contrast to lyric poetry and drama, the novel is a medium that captures and documents the everyday life and material culture of an era with particular sensitivity.

A detailed analysis of references to urban transport concepts reveals that the automobile and the bus appear in novels around ten years after their introduction to Riga. It remains to be discovered whether these innovations were documented earlier in prose works published in periodicals. At the same time, the corpus of novels reveals that depictions of the material culture in futuristic novels sometimes precedes real-world developments – for example, electric cars have been running on the streets of Riga since 1930s (Paulockis 1938a, Ģirupnieks 1939).

Figure 2 shows aggregated (average) relative document frequencies of terms pertaining to horse-drawn and motorized means of transportation. While horse-drawn vehicles remain dominant in novels even into the 1930s, it is noteworthy that references to both types of transportation at times tend to follow similar patterns of increase and decrease over time. This parallel suggests that some works may simply feature more travel overall. Supporting this hypothesis, examples from the corpus show that horse-drawn cabs and motorized taxis are sometimes mentioned in the same context – typically when characters are deciding or suggesting which mode of transport to use.

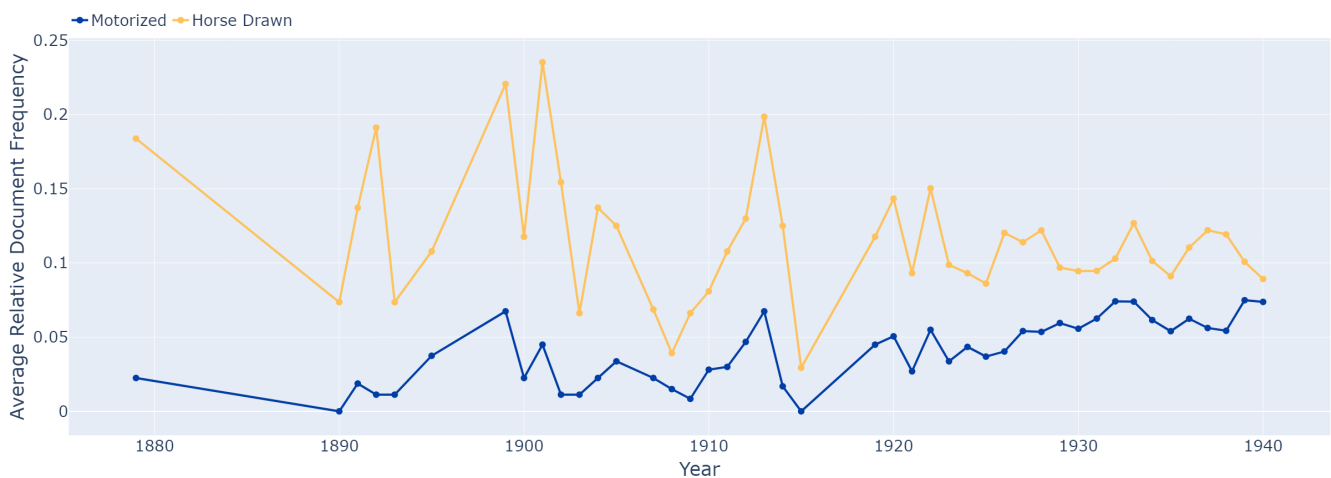


Figure 2: Average relative document frequency of terms referring to motorized and horse-drawn vehicles (all terms)

Figure 3 shows the absolute document frequency of selected prominent vehicles. It supports the observation that even though motorized vehicles were increasingly appearing in the literary vocabulary in the 1920s and 1930s the horse-drawn vehicles were still mentioned in a large number of works. The increase of references to motorized vehicles starting from the 1920s highlights both the growing prominence of the novel as a literary genre and the emergence of modern life in the newly established Republic of Latvia. Among the most frequently used terms are horse-drawn vehicles such as rig (pajūgs), single-horse carriage (vienjūgs), pair carriage (divjūgs), three-horse carriage (trijjūgs), four-horse carriage (četrjūgs), droshky (droška), line-droshky (līnijdroška), farm wagon (ore), horse-drawn cab (ormanis), wagon (vezums), covered wagon (kulba), spring wagon (federrati), coach (kariete), sleigh (kamanas), sled (ragavas), wain (vāģi). Among motorized vehicles frequently mentioned are railway (dzelzceļš) and train (vilciens), car (auto), automobile (automobile), motorcar (automašīna), machine (mašīna, used colloquially for "car"), limousine (limuzīns), bus (autobuss), tram (tramvajs), trolleybus (trolejbuss), taxi or teximeter cab (taksis, taksītis, taksometrs, taksomotors), motorcycle and motorbike (motocikls, motociklets). The modern lifestyle also introduces the use of bicycles (velosipēds, divritenis).

conference version

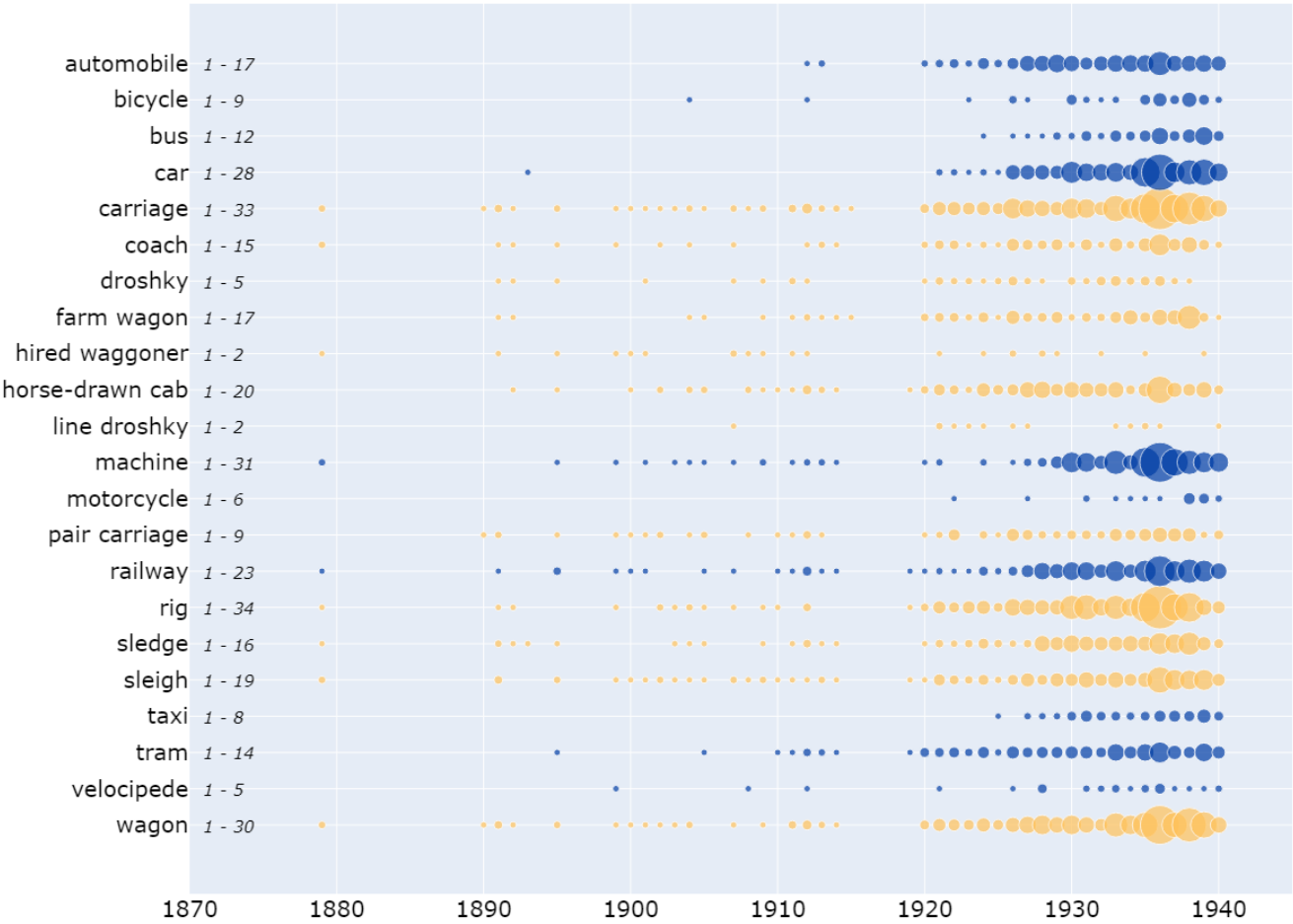


Figure 3: Distribution of selected land vehicles across documents. The range of number of novels for each vehicle is provided next to each term.

The Latvian word *mašīna* (machine), now used colloquially to refer to a car, appeared as early as in the first novels, but it was not initially used to mean "automobile". Machines

first entered rural households as agricultural machinery, e.g., threshing machines. Since the early 20th century, in the works of many authors, machines referred to factory equipment; as early as in 1899 machines also appear as household appliances – coffee machines and sewing machines (Deglavs 1897–1899). Already in one of the earliest novels the concept of “machine” appears as an abstract notion or metaphor: the novel’s protagonist utters that other people were machines released into the world for the sake of the nobles (Māters 1879).

The first mentions of the word *automobilis* (automobile) can be found in novels published in 1912 and 1913 (Skuja 1912, Kaija 1913, Upītis 1913). In total, there are 818 mentions of the word *automobilis* across 157 works by 96 authors. Starting from the 1920s, the shortened version of the word – *auto* (car) – became increasingly popular. The first mentions of this variant appear in 1921 (Akuraters 1921, Upītis 1921). By 1940, there were 1,637 mentions across 35 novels by 111 authors. The word *automašīna* (motorcar) was used comparatively less frequently, appearing only from the 1930s onwards, with a total of 40 mentions across 16 works by 11 authors.

When focusing on modern public and hired land transport vehicles, the railway and train undoubtedly play a central role. The railway (*dzelzceļš*) is mentioned as early as in the first Latvian novel (M. Kaudzīte and R. Kaudzīte 1879), while the first mention of a train (*vilciens*) can be found in 1891 (Deglavs 1891).¹¹

Mentions of trams (*tramvajs*) appear in novels starting from 1895 (Poruks 1895); in total, the LatSenRom corpus contains 611 mentions of the word *tramvajs* across 155 works by 89 authors. The word *autobuss* (bus) appears from 1924 (Skuja 1924). In total, there are 250 mentions across 79 works by 49 authors.

Terms related to horse-drawn cab (*ormanis*) remained relevant up until 1940. Mentions of it can be found in novels starting from 1892 (Purapuķe 1892). In total, there are 1,000 mentions across 183 works by 95 authors.

The search for mentions of taxi is complicated by the various forms in which the word is written. The earliest form, chronologically, might be *taksomotrs* (Skuja 1924). This term appears a total of 19 times across nine works by five authors. Alternative term *taksometrs* (also spelled *taksametr*) has been mentioned since 1925 (Gulbis 1925). Overall, there are 98 mentions across 47 works by 35 authors. Forms like *taksis* and *taksītis* are also encountered. Similar to the case with *vilciens* (train), tracking mentions of these terms is more challenging due to polysemy. The first mention of *taksis* appears in an anonymous author’s work (Anonymous 1926); the word is used a total of 66 times across 33 works by 27 authors. The term *taksītis* first appears in 1927 (Erss 1927); it is used a total of 57 times across 25 works by 19 authors (excluding uses with non-relevant meanings).

The bicycle is also increasingly mentioned as cycling was a cheaper than the horse or the car mean of personal transport and a popular leisure activity. In the journalism and fiction of the period, it was also given a role as a symbol of the New Woman and the

11. A quantitative analysis of the concept “train” is complicated by the fact that the Latvian word *vilciens* primarily means “train” (a railway vehicle), though it also has a secondary, metaphorical meaning derived from the root “vilkt” (to pull or draw). This figurative usage refers to a motion, stroke, or sweeping action and appears in expressions like “elpas vilciens” (the flow or rhythm of breath) and “zīmuļa vilciens” (the stroke of a pencil). These uses evoke the idea of a continuous or pulling motion, adding depth and poetic nuance to the term.

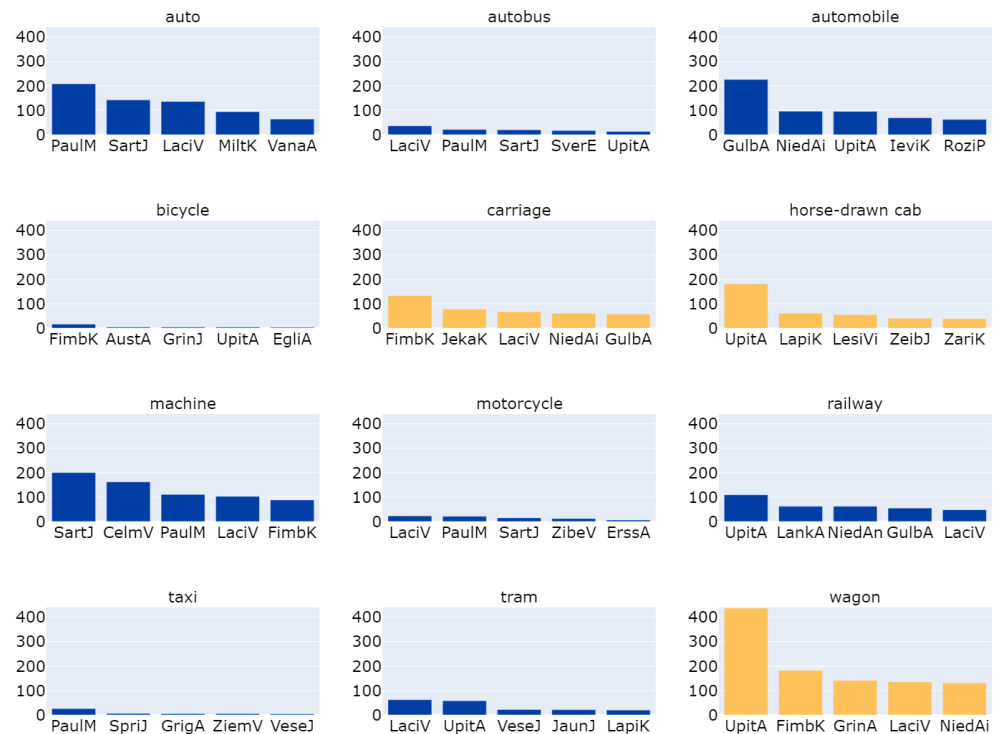


Figure 4: Word frequency of selected vehicles: top five authors with the highest mentions for each term

abolition of *fin-de-siècle* gender restrictions. References to the bicycle date back to 1899, in August Deglavs's novel *The New World* (Deglavs 1897–1899).

Combined Word2Vec and Gemini 1.5 results revealed also less frequently used words, including specific car brands and types: Ford, Mercedes, Chevrolet, Rolls-Royce, Oldsmobile, Alfa Romeo, Buick Essex, sports car, roadster etc. which indicates the consolidation of the role of modern mechanized transport in culture in the late 1920s and 1930s.

An analysis of references to vehicles reveals two key insights. Firstly, as the 1940s approach, the vocabulary of transportation extracted from Latvian novels is nearly evenly divided between terms for horse-drawn and motorized vehicles. Secondly, the ranking of the frequency of mentions of transport includes novels and novelists whose names are less well known in the history of Latvian literature or who have been marginal in the genre of the novel, providing an opportunity to get to know several of them anew, such as Miķelis Paulockis, Kārlis Lapiņš or Ansis Gulbis. This analysis of the large corpus of novels and full-text data is thus not based on canonical texts; on the contrary, it offers a democratic and unencumbered-by artistic quality criteria approach to analyzing a particular phenomenon in a large corpus of texts.

Figure 4 shows the authors whose works mention popular vehicles the most (the height of the bars represents the total number of mentions of a given name, summed over all the authors' works).¹² Although dozens of authors have mentioned various vehicles,

12. Although relative frequencies are more appropriate for studying the importance of a particular term in an author's oeuvre, the absolute frequencies used here offer insight into the extent to which an author may have hypothetically exposed their audience to specific vocabulary, whether in the course of a single work or across several. With relative frequencies, this signal would be subdued for authors with several comparatively long works.

most references are brief, and only for a few does a vehicle become an essential part 390
of the message. For example, "car" plays a prominent role in the works of Miķelis 391
Paulockis, Jānis Sārts, Vilis Lācis, Kārlis Miltiņš, and Arvīds Vanags, while "automobile" 392
is most often found in the novels of Ansis Gulbis, Aīda Niedra, Andrejs Upītis, and 393
Kārlis Ieviņš. Modern vehicles such as the bus and the motorcycle have not become the 394
"main characters" in any of the works of the LatSenRom corpus, i.e. they are represented 395
with a much smaller number of mentions, but it is interesting to observe that they have 396
been important for several authors who have also written about cars and automobiles, 397
such as Vili Lācis, Miķelis Paulockis, Jānis Sārts, Andrejs Upītis. 398

3.2 Towards mechanized transport: Riga as a Baltic metropolis 399

Observations on vehicles entering literature at the end of the nineteenth century point to 400
phenomena that each warrants further study and analysis that exceed the capacity of an 401
article – only the taxi and horse-drawn cab were selected to illustrate the rich semantic 402
load embedded in the literary representation of transportation. 403

As Steven A. Mansbach has observed, Riga was "the only true metropolis in the Baltic 404
region," where "the forces of history and modernism reached an accommodation for 405
northeastern Europe" (Mansbach 2014, p. 261). This metropolitan status was not only 406
symbolic or architectural, but also rooted in Riga's unique role as the industrial and 407
technological hub of the region. At the beginning of the 20th century, Riga was home to 408
several key players in the automotive industry, including *Alexander Leutner & Co.*, the 409
Rossiya factory, and the *Russo-Baltic Wagon Factory*, which produced the first automobiles 410
in the Russian Empire. During the interwar period, Riga further solidified its position 411
as a regional automotive center through the *Ford-Vairogs* plant, which operated under 412
license from the Ford Motor Company. The presence of these factories meant that 413
automobiles – particularly taxis – were more readily available in Riga than elsewhere in 414
the Baltics, which explains their early literary visibility as markers of urban modernity 415
and social mobility. 416

In parallel with this industrial infrastructure, the establishment of the *Supreme Board* 417
of Roads and Structures of Latvia in 1919 introduced a regulatory framework that funda- 418
mentally altered the conditions of mobility. This institution oversaw road development, 419
enforced traffic regulations, and issued licenses, thereby transforming the automobile 420
from a luxury item into a managed element of everyday urban life. The increasing 421
presence of legal norms, signage, and speed control in the public space paralleled the 422
narrative tension visible in Latvian fiction, where taxis often appear as symbols of not 423
only acceleration and freedom, but also of risk, collision, and bureaucratic control. 424

With the development of modern urban spaces, in Riga and other European and Russian 425
metropolises, the landscapes, sounds, scents, and transport flow became an integral part 426
of life, even for those who did not use this transport daily. This experience was present 427
both day and night. It became a significant element in literature that has increasingly 428
turned its focus to the urban environment, encompassing and reflecting the impact 429
of transport on the urban landscape and its inhabitants. Old habits persisted into the 430
transitional decades around the turn of the century, just as horses continued to trot 431

through the streets. Yet, new mobility habits and transportation systems also emerged.¹³ Street traffic is precisely depicted in Jānis Veselis' novel *Dienas krusts* (1931) through the character Mežaks riding a bicycle through Riga:

The evening twilight, bustling with shop assistants coming home, shimmered and sparkled in the glow of illuminated windows in the rushing lights of trams and buses. Where shadows darted, met, and parted at every street corner, Mežaks rode his bicycle to the workers' sports ball. He deftly weaved between the large buses, speedy taxis, and slow pedestrians, sometimes whistling, ringing a bell, braking, and others speeding up. With varying speeds, resembling a perilous collection of beasts, he maintained harmony. He felt like a bird in the air: the true residents of the street were his friends, and the high curve of the street was all alike, from the ten house owners, deputies, and ministers to the poor courier, shoemaker, seamstress, and dockworker. Only those who referred to a bicycle or car as their own or those using it had special rights: they could. (Veselis 1931)

Anchoring the literary data in this industrial and administrative context allows for a richer interpretation of how novels mediate technological change as both experience and imagination. The material culture depicted in novels partly reflects developments in the author's immediate surroundings, but it can also represent advancements that have not yet reached their homeland or even prefigure future changes. The first automobile appeared on the streets of Riga in 1904; in the novel, it was first mentioned in 1912, in an episode set in Moscow (Skuja 1912); further references to automobiles from books published in 1913 are situated in Riga. In Paris, automobiles appeared only slightly earlier – in 1896 – underscoring that Riga has always been a modern city and everything new in Europe arrived in Riga (then under the rule of the Russian Empire) with only a slight delay (Stirna 2024).

Omnibuses, trams, and buses represent large-scale public transportation, typically carrying sizable crowds of passengers, which was a new development in the context of inner-city mobility (Biedriņš 2021). The predecessors of trams and buses – horse-drawn omnibuses – first appeared in the streets of Riga in 1852 (Budiloviča 2024) and were also mentioned in the first Latvian novel (M. Kaudzīte and R. Kaudzīte 1879). In the following decades, the term appeared in only seven works, with its last mention in Pāvils Rozītis's novel in 1936, where it is portrayed as a crude, dirty, and slow vehicle – so much so that a girl steps out of it to avoid staining her dress (Rozītis 1936). When considering the modes of transport that carried citizens through the streets of Riga, much more attention in novels is given to the tram. The inception of tram services dates back to 1882, when the first horse-drawn trams appeared on the streets of Riga. The first electric tram line was launched in 1899 in Liepāja; in 1900, the Riga Trams Joint Stock Company was established, and in 1901, Riga's electric tram services were launched (Budiloviča 2024). The tram has been mentioned in Latvian novels since 1895, for the first time in the narration that takes place in Dresden (Poruks 1895). Further references, starting from 1900, are likely situated in Liepāja and Riga. In addition to trams, it is

13. The most comprehensive source for understanding Riga's historical transport system is Andris Caune, *Riga's Transport 100 Years Ago* (Caune 2020).

worth noting the arrival of the bus in Riga in 1913¹⁴ – though it is mentioned in novels only after World War I.

In rural settings, the dominant modes of transport until the First World War were on foot, horseback, or carriage, and these modes are often used to emphasize the deep connection between the characters and the landscape and to open up space for personal reflection and a deeper connection with the environment. Rural modes of transport emphasize a closer connection to the land and a slower pace of life. Horse-drawn carriages are mainly used for agricultural purposes and transporting goods, showing the hard work and traditional way of life of rural life, which is increasingly threatened by the encroachment of modernization; for example, trains often play a different role in rural settings than in urban narratives. While in cities, they symbolize connectivity between distant places and even countries and progress, in rural areas, trains can symbolize the encroachment of the industrial world. They can create both opportunities and significant disruptions to traditional rural communities. The arrival of railways in remote regions is a technological achievement and a catalyst for social change, affecting the social fabric and physical landscape of the rural areas where these stories are set.

Just to briefly touch upon other means of transportation, water transport on rivers and the sea in novels has always symbolized discovery, travel, personal transformation, and confrontation with the forces of nature.¹⁵ The depiction of air transport in literature appeared later and rarely (until the turn of the 20th century). After the First World War, airplanes symbolized the new technological challenge. It is only natural that most novels written in Latvian refer to airplanes and aeroplanes as military means of warfare. Still, also small sports planes are mentioned in *The Silver Sun Leaps* by Skuju Frīdis (Skuja 1924), and one striking mention of the airplane as a private vehicle is in a feminist utopian novel by Amanda Klot's *The Victory of Woman* (Klot 1927), where all the women of the world fly to an extraordinary women's congress in airplanes or, as the novel has it, aero-planes. Another example, notable for its time, is Miķelis Paulockis's science-fiction novel *The Saviour of the World* (Paulockis 1938b). It anticipated technological developments, including constructing a subway in Riga in 1986, which never happened.

3.3 Symbolism and social context of transport

Transport in European and world literature is an essential narrative element and a capacious symbol of the age's social, cultural, and technological changes. Transport scholars have used literary and historical methodologies to explore horse-drawn carriage travel, the waiting associated with rail travel, and the impact of the train window on passengers' perception of the landscape as 'another world' (Kellerman 2019, Livesey 2016).

In addition, works on transport history and literary mobility assess the role of the horse in the city in the 19th century and the changes in perception brought about by rail travel (Gavin 2015, McShane and Tarr 2003).

14. The news that a bus route was opened between Sarkandaugava and Jaunmīlgrāvis can be found in the 1913 issue of the newspaper *Dzimtenes Vēstnesis* (Dzimtenes Vēstnesis 1913).

15. The Gemini 1.5 inquiries identified 46 maritime and 30 air transportation terms. These domains were not researched further in this study.

Horses were vital to industry, commerce, agriculture, and employment, serving other transport systems, including rail and shipping. In addition to the omnibus, taxi, carriage, cart, and coal wagon functions, horses were used for deliveries to stores, postal services, public holidays, refuse removal, emergency services, wedding parties, and funeral ceremonies (Gavin 2015). With trams, which were initially horse-drawn and then electricity from the 1880s onwards, and rails appeared on city streets, trams were a space where people from different social classes could physically meet and appreciate each other (Finch 2023).

Barbara Schmucki has shown that urban dwellers in Britain and Germany had different views and relationships with the horse tram and its electrified descendants, sometimes viewing newcomers with distaste for the associated rails and electric wires. However, "by the 1950s, trams were fully integrated into everyday life" (Schmucki 2012).

The appearance of trains and railways in the literature often marks the development and mobility of society and symbolizes progress. Trams and buses capture the rhythm of everyday urban life, highlighting the interaction between different social classes and the everyday life of urban dwellers. This sense of mobility was particularly characteristic of the turn-of-the-century youth who came to the metropolises: "They are brought to and transported around the capital by the new mobility offered by the rail network and trams, and it is their dreams of upward social mobility that drive them forward." (Ameel 2014) In his study of urban experiences in Helsinki (1890–1940), Finnish comparative literature scholar Lieven Ameel points out, "In many important Finnish literary texts dealing with urban experiences in this period, modes of transport and images of mobility acquire more than a symbolic status. They are central to the development of provincial characters in the city." (Ameel 2014)

An analysis of specific period novels, such as those by Virginia Woolf, James Joyce, and F. Scott Fitzgerald. For example, the role of the omnibus in Virginia Woolf's novels is explored on several levels: as a realistic indicator of technological progress, as a cultural commentary on the British class system, and as a device of Woolf's narrative technique. Having frequently traveled in both horse-drawn and later motorized omnibuses, Woolf knew the colors, routes, and fares of the omnibuses that made up London's route network (McNees 2010). The different modes of transport offer glimpses into a public sphere where anonymity and observation intertwine, as in Virginia Woolf's depiction of London, where buses allow characters to traverse and observe the city's diverse social landscapes.

The period under review witnessed unprecedented advances in various forms of mechanical transport and related technologies, each of which introduced new dimensions to the plot and thematic setting of the novels, which can also be observed in the analysis of the LatSenRom corpus, for example, the technological development is accurately described in Jānis Plaudis' novel *Gymnasium students*:

If a hundred years ago, one was still dreaming of the revolving balloon and relieving pain through surgical operations, now one flies like a bird in an aeroplane and begins to think about the questions of eternal youth. (Plaudis 1935)

As the hero of the novel by Vilis Aizstrauts predicts:

Technology is increasingly making its mark in life. The time is not far off when humankind will do almost all its work with electricity. Electric plows will drill in the soil, and electric airplanes will roar in the air; ships and trains will be powered by electric batteries instead of coal. In short, electricity will rule the world. The question is how to get the power we need most cheaply and conveniently. Along these shores, millions of electrical energy go unused. (Aizstrauts 1933)

Alongside the excitement about the new possibilities, there was also criticism and nostalgia, for example, a quote from Augusts Mežsēts' novel *The Enchanted City* (1929), which reflects a critique of technological progress and a fear of being overwhelmed by the possibilities of the new technologies:

It is true that in our days when people are jolting around the earth in cars and aeroplanes are lifting them into the air, human thought doesn't need to fly. It is enough for the human body to fly. Young lovers wander in armchairs on boulevards, gaze into electric light bulbs, or sit in cinemas and marvel at the enchanting love that fades on the screen in an hour. (Mežsēts 1929)

4. The public transport system in Riga. Horse-drawn cab vs. taxi

As highlighted in introduction, exploring urban transport in the literature provides a fruitful perspective for understanding sociocultural dynamics. This aspect allows for the revelation of not only the technological achievements of its time and the transformations of the urban environment but also a deeper insight into the daily lives of individuals and the social and psychological landscape that emerged during these changes, transitioning from the 19th to the 20th century.

Compared to trams and buses, horse-drawn cabs and taxis represent more private forms of public transport, typically serving just one or a few individuals at a time. An important aspect is that taxis gradually replaced the legendary horse-drawn cabs, but a particularly notable period was the 1920s when both types of vehicles coexisted. Figure 5 presents the relative document frequency of references to horse-drawn and motorized cabs, with the various terms used for each type consolidated into two respective lines. While the yellow line prevails, it is gradually declining, whereas the blue line rises steadily, signaling the arrival of a new era. Therefore, examining the transition from the once-popular horse-drawn cab to the mechanized taxi in early 20th-century Latvian novels is particularly fascinating.

"Taximeter" is the key term for this transition, as its original meaning is 'mechanical meter,' which was attached to the horse-drawn carriage to count the distance traveled and thus calculate the fare.¹⁶ It was introduced due to detected fraudulent practices by drivers to obtain larger payments. The introduction of the taximeter as a fare-measuring device at the end of the 19th century brought a new element to private and public

16. "The city council has just ordered six taximeters from abroad for horse-drawn cab drivers, to whom they will sell them. A taximeter is a device attached in a certain way to the carriage, showing how far has been traveled." See: Baltijas Vēstnesis 1902.

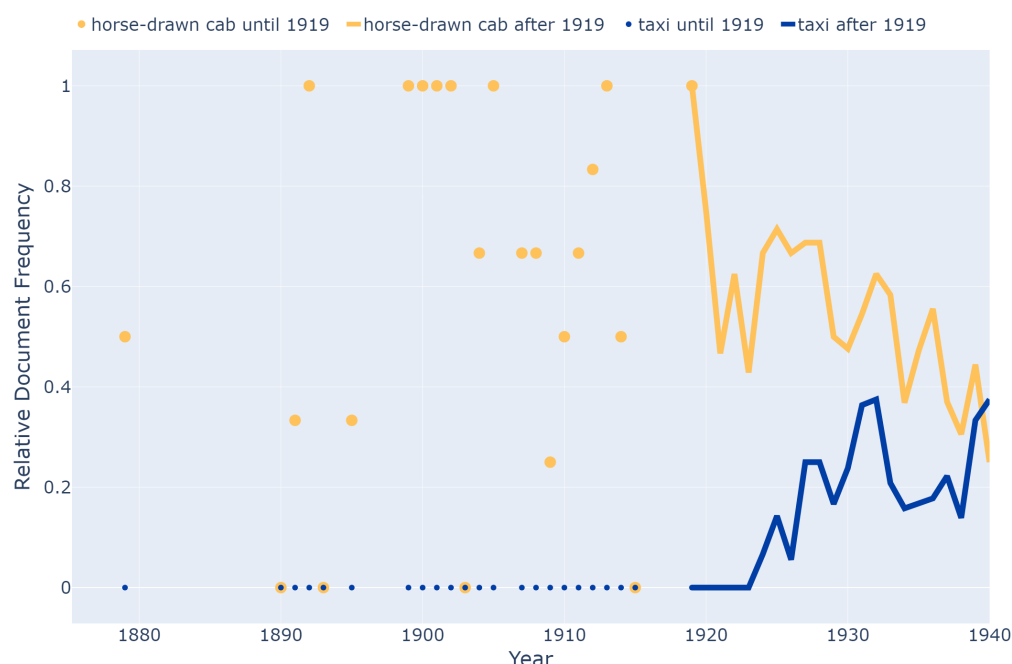


Figure 5: The proportion of novels in which the terms "horse-drawn cab" (including terms *ormanis*, *fūrmanis*, *važonis*) and "taxi" (including terms *taksis*, *taksītis*, *taksometrs*, *taksometris*, *taksomotors*) appear in different years.

transport, ensuring fair pricing based on distance traveled or time. This innovation was crucial for building trust and standardizing tariffs in growing urban centers. The first taxi cars appeared in Riga in 1913; however, their numbers surged after the First World War, particularly during Latvia's period of independence, when they also began to appear frequently in novels (for counts, see subsection 3.1 in this paper).

As mentioned in the introduction, the case study was inspired by the question asked in Kārlis Lapiņš's novel *Students in the Farm*: "Why do you drive a 19th-century carriage when you have a 20th-century taxi at hand?" As cities expanded and modernized, the transition from horse-drawn carriages to motorized taxis marked a significant development in public transport and a new relationship with speed,¹⁷ reflecting changes in urban lifestyles and technological advancement. Therefore, a brief insight into how this is revealed in the corpus of Latvian novels.

4.1 The legacy of the horse-drawn cab in Riga

The horse-drawn cab (*ormanis*), as historian Andris Caune notes, "is the oldest form of public transport in Riga," (Caune 2020) is mentioned in the historical sources of Riga since the 13th century. The horse-drawn cab as a romanticized symbol of Riga is vividly portrayed in Latvian poetry, notably by Aleksandrs Čaks with several poems as *Pathetic Quartets* and *Poem about the Horse-drawn Cab* (Čaks 1930). And also in novels, for example, in Jānis Jaunsudrabiņš's novel *Aija*, the character Jānis, while in the countryside, occasionally feels a sentiment for the urban identity and recalls "the beautiful streets of Riga on winter evenings. The bells of the horse-drawn cabs tinkled. Illuminated by electric lights, people came and went." (Jaunsudrabiņš 1911) Thus the horse-drawn cab

17. For the concept of speed in interwar modernist poetry in connection with European metropolises (Ostups 2024).

is a striking figure associated with the city (see [Figure 6](#)).

Often, the most significant role in novels belongs not to the transport itself but to the driver of the horse-drawn carriage, as seen in Andrejs Upīts's novel *The Last Latvian* (Upīts 1913), which highlights the human characteristics of the carriage driver through various aspects. In Upīts's novel, the carriage driver is more than just a driver; he observes and reacts to Kalve's emotional state, becoming a silent witness to his passenger's inner turmoil. The episode of the novel vividly depicts the bustling and somewhat chaotic urban environment of that time, highlighting different societal attitudes and behaviors through the characters' interrelations, particularly around the figures of the carriage driver and Alberts Kalve. The novel's action takes place in a traffic jam caused by a tram and two timber-carrying carts, a common occurrence on the busy streets of a growing city. This obstacle provokes a series of reactions from Kalve, revealing not only his impatience but also a more profound commentary on the rhythm and priorities of urban life. Kalve's impatience and the subsequent interaction with the tram driver and timber carriers reflect the differences and expectations of social classes regarding service. In this scenario, the "carriage driver" acts not only as a service provider but also as a participant in the broader dynamics of society. His initial inaction and then hasty reaction after Kalve's prompting with a cane emphasize the power dynamics and the expectations placed on individuals based on their roles in society. However, Kalve also perceives something unacceptable in the general attitude towards carriage drivers, namely that a carriage driver must wear a number on his back at the collar, even though the number is visible on the back of the cart, at the front by the pole, and on the sides. This numbering system is part of a strict control and identification mechanism. Kalve, describing this situation with its mechanical and depersonalized approach, ironically suggests that a person should wear a number plate of their home address on their back to conform to this absurd system fully.

Conversely, the episode in which the carriage driver is urged to pick up the pace with a hit of the cane (or another object) is quite typical in accounts of carriage drivers, for example, in Andrejs Upīts's novel *Gold*, where the young woman Made is particularly impressed by the journey with the carriage driver alongside her brother, who has come into wealth. Both how Sveilis, with his cane, touches the carriage driver to halt the carriage and how he exclaims: "You, man, wait: I don't have small change!" (Upīts 1921)

This expression seems astonishing to Made, as it sounds so gentlemanly, contrasting sharply with her previous poverty and experiences in impoverished conditions. Such moments of superiority are particularly striking in the portrayal of the carriage driver, which is not typical in accounts involving taxi drivers. It is noteworthy that carriage drivers are also much more affected by weather conditions, as, for the most part, the driver's seat is not equipped with a roof or any other form of protection against rain, cold, or heat, which impacts the comfort of the journey. For instance, in Ivande Kaija's novel *In the Yoke*, the carriage driver humbly asks a passenger: "Madam, are we going to roam around the city for long? I am starting to feel cold; the horse also needs a rest." (Kaija 1919)

Modern vehicles, which replaced horse-drawn transport, were considered by some as peculiar and indifferent machines that threatened to change human character as well.

A significant example of generational clashes is in Kārlis Lapiņš’s novel *The Degenerate* (Lapiņš 1934a). One morning on Barona Street, a distressing event occurs. The father, Pumpītis Senior, moves slowly along the street with a horse-drawn cab when suddenly a taxi appears from behind and swiftly overtakes them. Pumpītis looks suspiciously at the taxi, in which a familiar wide-brimmed hat seems to flash. He instructs the horse driver to urge the horse to try to catch up with the taxi, but it proves unsuccessful, and the car disappears around the corner, continuously emitting loud signals. Pumpītis hisses, dissatisfied with what happened, as he suspects the taxi passenger could be his son. When they finally arrive home, the driver stops beside the taxi. Seeing his father, the son greets him with a casual “Good morning, father,” but Pumpītis does not respond. He feels uncomfortable and angry, realizing that this is neither the place nor the time to discuss what has transpired as both father and son return from a night out.

4.2 The rise of the taxi and urban modernization

The appearance of taxi in literature is negligible before 1925, which aligns with the historical introduction of automobiles in Riga. From the mid-1920s onward, however, its presence grows steadily. By the early 1930s, taxi achieves narrative parity with ormanis, and in the years leading up to the Second World War, it slightly surpasses it. This pattern corresponds to the accelerating pace of modern urban life and the shifting aesthetic of literature toward speed, anonymity, and urban dynamism (see Figure 6).

Feature	Ormanis (Horse-drawn Cab)	Taxi
Narrative Function	Human-centered, emotional depth	Dynamic, anonymous, fast-paced scenes
Symbolism	Tradition, nostalgia, rootedness	Speed, modernity, societal transition
Social Class Reflection	Often tied to working or poorer class	Emerging middle class, upward mobility
Temporal Feel	Slow, contemplative, historic	Fast, fragmented, urban modern
Driver Role	Named, empathetic characters	Anonymous, functional presence
Gendered Dynamics	Personal, respectful, often passive	Can involve erotic tension, autonomy
Weather Vulnerability	Open to cold, wet, discomfort	Protected, modern convenience

Figure 6: Comparison of Narrative and Symbolic Functions of Ormanis and Taxi

In Pāvils Rozītis’ novel *Ceplis* (Rozītis 1928), the life of Riga in the spring of the 1920s is vividly depicted, highlighting how the changing weather conditions influenced the use of transport. Snow was often followed by windy rain, making the streets muddy. The drivers, who tied their horses to sleighs in the mornings, were forced to switch to carts by noon, only to replace them with sleighs again in the evening. This uncertainty and frequent transition from one mode of transport to another caused inconvenience for both the drivers and their clients. The horses appeared dirty and drenched, reflecting the prevailing gloomy atmosphere of the city. Meanwhile, the few taxis moved with a particular enthusiasm, splashing pedestrians, building walls, and shop windows.

Riga’s residents mainly used taxis when intoxicated, when money was no object, and life was wanted to be experienced at a faster pace than the cab’s leisurely journey. Only the wealthiest or those poisoned by the desire for life’s accelerating tempo would ride taxis sober, relishing the carefree speed that allowed them to race past the sleepy gray façades of buildings and dash across the entire city in mere minutes. However, there were not many of these people in Riga, just as the wealthy were still in the process.

The description cited emphasizes the taxi as a modern, dynamic, and somewhat ex-

travagant choice, contrasting with traditional and slow horse-drawn transport, offering the reader insight into urban life and the impact of technological progress on societal habits and lifestyles. The taxi symbolizes the changing pace of urban environments and the restructuring of society. "Limousines and taxis race along Freedom Boulevard; it is the aorta of the city" (Prūsa 1936) – this is how Emily Prūsa describes the new urban landscape in her novel *The Temptation of Distance* (1936). The taxi provides a notion of the car for those who have not yet traveled in one; this joy of the ride is described in Jānis Veselis's novel *The Uprising of People*, where Meiklis invites guests for a ride in his new car:

The car purred charmingly and pleasantly; the wheels began to roll across the paved yard and rolled out onto the street. Meiklis, who had rarely driven in a taxi, and Lija, who had rarely enjoyed the delights of such a journey, both felt a rocking pleasure as the rubbery wheels swiftly whisked them through the already cobbled streets, racing dreamily past houses, lampposts, people, and autumnal smoky flames, as if all of that were a life to be left behind irrevocably. (Veselis 1934)

The city description seen through the window of a moving car tends to emphasize the experience of speed and the fragmentary and ephemeral nature of observation. Meanwhile, the taxi driver's interaction with clients can reveal urban dynamics and themes of anonymity. The cultural significance of the taxi extends beyond literary representation and is embedded within the broader mediascape of the early twentieth century. In addition to novels, the motorcar emerged as a prominent visual and narrative motif in early cinema. It appeared frequently in dramatic scenes, including chases, romantic encounters, and depictions of urban glamour or danger. These cinematic representations amplified the symbolic associations of the taxi with speed, risk, and emotional intensity.

Like horse-drawn carriages, novels mention taxis in scenes involving personal reflections, character interactions, and dynamic movement. The taxi serves not only as a means of transport but also as a place for character and plot development. Taxis appear most frequently in Miķelis Paulocki's novels. In Paulocki's works, taxis fulfill various narrative functions, from the movement of characters to significant interactions or places for contemplation; they symbolize aspects of modernity, urban life, or transitions, reflecting the characters' experiences and broader societal changes of that era. In Paulocki's novel *The Secret of the Old Lighthouse*, it is notably revealed that the new speed provided by mechanized transport, such as the taxi, does not always fulfill the characters' wishes. The protagonist, Krauze, needs to reach Zaslauks pier; therefore, he begins his journey with a horse-drawn cab. However, although he asks to go faster, the horse can only tread slowly in the small cart, causing Krauze's dissatisfaction. Feeling the pressure of time, Krauze decides to switch to a taxi to speed up his journey. This transition symbolizes not only a physical movement from one vehicle to another but also a shift from the slow pace of the old world to the new – faster – tempo of urban life. Krauze, who despised taxis, was nonetheless compelled to use this new mode of transport. The taxi ride at the corner of Kalnciems Street suddenly turns into a catastrophe as a car collision occurs. Krauze is severely injured and understands that his hope to escape his previous life and the problems related to drinking and loss of property has been shattered. His attempt to

leave the city and find a new beginning in the countryside has stalled because the city's asphalt, symbolically speaking, "held him tightly in its firm grip." (Paulockis 1936)

As part of the new urban aesthetics expressed in the literature of this period, a new central urban experience appears – the night metropolis, seen through the window of a moving car. In several novels and stories, driving at night is depicted as a liminal space that is partially private and partially public. It serves as a space where boundaries are transgressed with a clear sensual and sexual undertone. In such scenes, the enclosed car, much like the night city as a whole, acquires a sexual ambiance; for instance, in Ansis Gulbis' novel *The Legacy of the Fantasts*, a taxi ride is used as a place for expressing intimacy and passion. Romina and Jeremejevs, sitting in the taxi, are physically close to one another, creating an intimate atmosphere. Jeremejevs touches Romina's leg, generating a subtle erotic tension (Gulbis 1925). This moment, occurring within the confines of the automobile contrasts with the anonymity and openness of the outside world, where they are not hidden from view, highlighting the car as a location where private and public boundaries blur, becoming a place for sensual and emotional expressions.

Meanwhile, in an episode from Miķelis Paulockis' novel *Professor Sūna's Wonderful Elixir*, the city and taxi serve as a backdrop and symbol for Herta's inner tension and yearning for life. The refreshing breeze of a late autumn day contrasts with Herta's feverishly hot face, revealing her restlessness and desire to enjoy life. Herta's internal dialogue, where she acknowledges her "mad cravings" and her wish to live and revel in her beauty, directly references her freedom and self-confidence. The taxi she hails becomes a symbol of her hurried pace of life and inner anxiety. The driver, who looks at Herta with surprise and desire, offers her a self-affirming acknowledgment of her beauty and attractiveness. Herta's rush and desire to escape the routine of everyday life and return to the enjoyment of life are reflected in the motion of the taxi:

The taxi sped madly down the boulevard, crossed paths with horse riders in front of its snouts, and splashed pedestrians with mud from puddles still collected from the recently fallen rain. It seemed as though Herta's anxiety had overtaken the lifeless machine, rumbling in its cylinders, hissing as the tires rubbed against the asphalt. Faster! Faster! (Paulockis 1938b)

The episode reveals the intensity of urban life and Herta's crisis, using the taxi ride as a symbol of her inner turmoil and yearning for life.

As seen from the example of the novel, while the horse-drawn carriages evoke nostalgia for a slower, gentler era, taxis symbolize the dynamic, constantly changing nature of the modern city at that moment. These vehicles and their drivers can be either creators or resolvers of conflict, as their interactions with other characters in the novel and cars can lead to significant plot twists, revealing the characters' traits and motivations. Furthermore, they serve as a metaphor for the irreversible transition from tradition to progress, accurately reflecting the complexity of the relationship between nature and humanity in modernization. The shift from the horse as a living being to the engine as a lifeless technology illustrates technological progress and more profound changes in human relationships with nature and technology.

5. Conclusion

786

The LatSenRom corpus expands opportunities for literature researchers by providing access to vast datasets and facilitating more effective text exploration. This resource supports both quantitative and qualitative research, enabling comparisons across authors and tracing the evolution of literary motifs and themes over time. Additionally, digitized texts make it possible to uncover lesser-known authors and works while fostering interdisciplinary research that blends literary studies with history and sociology. This integrated approach enriches our understanding of cultural and societal dynamics.

The use of Word2Vec and Gemini 1.5 language models led to the discovery of over 160 unique terms related to land vehicles, with approximately half referring to modern transportation and the other half to horse-drawn transport. The Gemini 1.5 model proved to be the most effective in identifying terms within texts compared to other techniques. However, even the results from the most productive prompt contained approximately 88 percent false positives. Although approximately half of the transportation-related vocabulary refers to motorized vehicles, the average relative document frequency remains higher for horse-drawn vehicles throughout the entire period studied. An examination of terms referring to horse-drawn cabs and taxis reveals a clear trend: mentions of horse-drawn transportation steadily declined throughout the 1920s and 1930s, while references to motorized taxis increased. The analysis of vehicle term frequencies provides a valuable evidential basis for conjectures about the popularity of various modes of transportation and their distribution across authors.

The portrayal of horse-drawn cabs and taxis in literature during the first decades of the 20th century offers a nuanced insight into the evolution of urban public transport and its impact on social interaction and urban experience. These vehicles facilitate physical movement and promote a deeper exploration of themes such as anonymity, modernity, and the individual's place within the urban landscape. The cabs symbolize mobility, bridging distances and seemingly connecting parts of the narrative, often serving as physical and metaphorical means of transition that facilitate characters' personal transformation and development. During the interwar years, both transport forms coexist in literature, mirroring urban reality. This overlap period is particularly rich for symbolic readings: it is the moment when characters must choose between the familiar ormanis and the modern taxi, often with ideological undertones.

In several interwar novels, similar to the modern poetry of that time, the theme of speed and technology stands out, serving as a metaphor for the challenges of human existence and the changes in the pace of life during modernity. This observation indicates the role of technology as a response to existential questions, making it an important aspect of analysis that provides additional interpretive possibilities in literature. In Latvian literature, like in Europe, transport is closely associated with mobility, freedom, and change themes.

In sum, the quantitative patterns of term usage in Latvian fiction correlate closely with the qualitative shifts in narrative representation. The ormanis and the taxi serve not only as transport modalities but as literary devices charged with ideological and emotional meaning. Their respective presences in novels mark the changing pace of life, class dynamics, gender roles, and psychological atmospheres of the city. The transition from

horse-drawn carriages to taxis thus reflects a deeper literary modernization, in which vehicles become symbols of cultural acceleration, disruption, and the reconfiguration of the self in the urban milieu.

Furthermore, urban studies provide a perspective for identifying how urban transformations, interactions among different social groups, and the impact of technological progress on city life are reflected in literature. Themes of alienation, progress, and public and private space are frequently explored in the novels of this era, particularly through the prism of urban transport. The expansion of the range of urban vehicles is depicted not only as a technological triumph but also as a source of existential anxiety and social fragmentation. The historical context of rapid urbanization and events leading up to World War II imparts a sense of urgency and transformation to these themes.

6. Data Availability

Data can be found here: <https://dom.lndb.lv/data/obj/1554666>.

7. Software Availability

Software can be found here: https://github.com/ValRCS/lmb_transports.

8. Author Contributions

Eva Kristšone-Eglāja: Conceptualization, Writing

Anda Baklāne: Conceptualization, Writing

Valdis Saulespurēns: Language model deployment, Visualization

References

- Aizstrauts, Vilis (1933). *Mīlas un ārpriekšējā varā*. Latgrāmata.
- Akuraters, Jānis (1921). *Pēteris Danga*. A. Gulbja apgādībā.
- Ameel, Lieven (2014). *Helsinki in early twentieth-century literature: urban experiences in Finnish prose fiction 1890–1940*. Studia Fennica Litteraria, 8. Finnish Literature Society/SKS.
- Anonymous (1926). *Kapteinis Tālvāldis un viņa brīnišķīgie piedzīvojumi jeb No kuģapuiķas par kapteini*. Viļņi.
- Baklāne, Anda, Artis Ozols, and Eduards Skvireckis (2024). “Latviešu senāko romānu atkārtotie izdevumi – datu kopas izveide un analīze”. In: *Proceedings of the National Library of Latvia XXX11.12*, 113–141. [10.52197/AQKR3401](https://doi.org/10.52197/AQKR3401).
- Baklāne, Anda and Valdis Saulespurēns (2025). “Extracting Semantically Related Concepts from the Corpus of Latvian Novels: A Comparative Analysis of Word2Vec, GPT-4o, and Gemini-1.5 Results”. In: *DHNB 2025. Programme. Book of Abstracts*. Presented at DHNB2025 ‘Digital Dreams and Practices’. Estonian Literary Museum. <https://dhn.b.eu/conferences/dhnb2025/>.
- Baltijas Vēstnesis (1902). “Rīga”. In: *Baltijas Vēstnesis* 244, 3.

- Biedriņš, Andris (2021). "Rīgas elektriskais tramvajs". In: *Transports. Eiropas kultūras mantojuma dienas*. Mantojums, 56–57. 866
867
- Bologna, Federica (2020). "A computational approach to urban space in science fiction". 868
In: *Journal of Cultural Analytics* 5.2, 37–60. 10.7910/DVN/VXEK7A. 869
- Budiloviča, Evelīna (2024). *Pilsētas transports Latvijā*. Nacionālā enciklopēdija. <https://enciklopedija.lv/skirklis/89597-pils%C4%93tas-transports-Latvij%C4%81> 870
871
(visited on 05/30/2025). 872
- Čaks, Aleksandrs (1930). *Poēma par ormani*. Zelta grauds. 873
- Caune, Andris (2020). *Rīgas satiksme pirms 100 gadiem*. Zinātne. 874
- Degļavs, Augusts (1891). *Starp divām ugunīm*. F. Gēliņš. 875
- (1897–1899). *Jaunā pasaule*. Ernsts Plāte. 876
- Dzimtenes Vēstnesis (1913). "Vietējās ziņas". In: *Dzimtenes Vēstnesis* 123, 2. 877
- Eglāja, Eva, Anda Baklāne, and Valdis Saulespurēns (2024). "Pilsētas transportlīdzekļi latviešu senākajos romānos". In: *Proceedings of the National Library of Latvia XXX* 11.12, 65–879
96. 10.52197/DIOW5621. 880
- Erss, Ādolfs (1927). *Aglonas dievmātes atgriešanās*. Grāmatu Draugs. 881
- Fan, Yaxin et al. (2023). "Uncovering the Potential of ChatGPT for Discourse Analysis in Dialogue: An Empirical Study". In: *arXiv preprint*. arXiv: 2305.08391. 882
883
- Finch, Jason (2023). "Powered modernity, contested space: literary modernism and the London tram". In: *European Journal of English Studies* 27.2, 288–308. 10.1080/13825577.2024.2307035. 884
885
886
- Fridlund, Mats, Michael Azar, Daniel Brodén, and Michael McGuire (2023). "The Cultural Imaginary of Terrorism: Close and Distant Readings of Political Terror in Swedish News and Fiction During the Cold War". In: *DHNB2023 Conference Proceedings* 5.1, 90–104. 887
888
889
890
- Gavin, Adrienne E. (2015). "'I Saw a Great Deal of Trouble Amongst the Horses in London': Anna Sewells' 'Black Beauty' and the Victorian Cab Horse". In: *Transport in British Fiction: Technologies of Movement, 1840–1940*. Ed. by Adrienne E. Gavin and Humphries. Andrew F. Palgrave Macmillan, 101–122. 891
892
893
894
- Georgiev, Petko et al. (2024). "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context". In: *arXiv preprint arXiv:2403.05530*. <https://arxiv.org/abs/2403.05530>. 895
896
897
- Ģirupnieks, Jānis (1939). *Ekspedīcija zemes dziļumos*. Valters un Rapa. 898
- Graham, Shawn, Milligan Ian, and Scott Weingart (2016). *Exploring Big Historical Data: The Historian's Macroscopic*. Imperial College Press. 899
900
- Gulbis, Ansis (1925). *Fantastu mantojums*. A. Gulbis. 901
- Jaunsudrabiņš, Jānis (1911). *Aija*. J. Brigadera apgādība. 902
- Jockers, Matthew (2013). *Macroanalysis: Digital Methods in Literary History*. University of Illinois Press. 903
904
- Kaija, Ivande (1913). *Iedzimtais grēks*. Valters un Rapa. 905
- (1919). *Jūgā*. A. Valtera, J. Rapas un biedru komisijā. 906
- Karjus, Andres (2023). *Machine-assisted Mixed Methods: Augmenting Humanities and Social Sciences with Artificial Intelligence*. arXiv: 2309.14379 [cs.CL]. <https://doi.org/10.48550/arXiv.2309.14379>. 907
908
909
- Kārkla, Zita and Eva Eglāja-Kristone (2022). "Liriskās ģeogrāfijas: literārās telpas kartēšana latviešu rakstnieču romānos un īsprozā". In: *Letonica* 47, 104–127. 910
911
- Kaudzīte, Matīss and Reinis Kaudzīte (1879). *Mērnīeku laiki*. H. Allunans. 912


- Kellerman, Robin (2019). "Waiting for Railways (1830–1914)". In: *Timescapes of Waiting: Spaces of Stasis, Delay and Deferral*. Ed. by Christoph Singer, Robert Wirth, and Olaf Berwald. Brill Rodopi, 35–57. 913–915
- Kloto, Amanda (1927). *Sievietes uzvara*. Nākotnes Sieviete. 916
- Kohlrausch, Martin and J. C. Behrends (2014). *Races to Modernity: Metropolitan Aspirations in Eastern Europe, 1890–1940*. Central European University Press. 917–918
- Lapiņš, Kārlis (1934a). *Pagrimušie*. Grāmatu Draugs. 919
- (1934b). *Studenti fermā*. Autora izdevums. 920
- Livesey, Ruth (2016). *Writing the Stage Coach Nation: Locality on the Move in Nineteenth-Century British Literature*. Oxford University Press. 921–922
- Mansbach, Steven A. (2014). "Capital Modernism in the Baltic Republics: Kaunas, Tallinn, and Riga". In: *Races to Modernity: Metropolitan Aspirations in Eastern Europe, 1890–1940*. Ed. by Jan C. Behrends and Martin Kohlrausch. Central European University Press, 309–328. 923–926
- Māters, Juris (1879). *Sadzīves vilņi*. M. Māters. 927
- McNees, Eleanor (2010). "Public Transport in Woolf's City Novels: The London Omnibus". In: *Woolf and the City: Selected Papers from the Nineteenth Annual Conference on Virginia Woolf*. Ed. by Elizabeth F. Evans and Sarah E. Cornish. Clemson University Press, 31–39. 928–931
- McShane, Clyde and Joel Tarr (2003). "The Decline of the Urban Horse in American Cities". In: *The Journal of Transport History* 24.2, 177–198. 10.7227/TJTH.24.2.4. 932–933
- Mežsēts, Augusts (1929). *Apburtā pilsēta*. Grāmatu Draugs. 934
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv preprint*. 10.48550/arXiv.1301.3781. 935–937
- Moretti, Franco (1998). *Atlas of the European Novel: 1800–1900*. Verso. 938
- (2005). *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso. 939
- (2013). *Distant Reading*. Verso. 940
- Ostups, Artis (2024). *Laicena kunga ātruma izjūta: par dzejnieka Linarda Laicena (1883–1938) Berlīnes dzeju*. <https://www.punctummagazine.lv/2024/03/27/laicena-kunga-atr-uma-izjuta/> (visited on 08/26/2024). 941–943
- Paulockis, Miķelis (1936). *Vecās bākas noslēpums*. Laikmets. 944
- (1938a). *Pasaules glābējs*. Senatne. 945
- (1938b). *Profesora Sūnas brīnišķīgais eliksīrs*. Senatne. 946
- Piper, Andrew (2018). *Enumerations*. The University of Chicago Press. 947
- Plaudis, Jānis (1935). *Gimnāzisti*. A. Gulbis. 948
- Poruks, Jānis (1895). *Pērļu zvejnieks*. E. Plāte. 949
- Prūsa, Emīlija (1936). *Tāle vilina*. A. Gulbis. 950
- Purapuķe, Jānis (1892). *Pasaules lāpītājs*. M. Jakobsons. 951
- Řehůřek, Radim and Petr Sojka (2010). "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, 45–50. 952–954
- Rodman, Emma (2020). "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors". In: *Political Analysis* 28.1, 87–111. 955–956
- Rodriguez, Pedro L., Arthur Spirling, and Brandon M. Stewart (2023). "Embedding Regression: Models for Context-Specific Description and Inference". In: *American Political Science Review* 117.4, 1255–1274. 10.1017/S0003055422001228. 957–959

Rozītis, Pāvils (1928). <i>Ceplis</i> . Valters un Rapa.	960
— (1936). <i>Valmieras puikas</i> . Valters un Rapa.	961
Schmucki, Barbara (2012). “The Machine in the City: Public Appropriation of the Tramway in Britain and Germany, 1870–1915”. In: <i>Journal of Urban History</i> 38.6, 1060–1093.	962 963 964
Schöch, Christof, Roxana Patras, Tomaž Erjavec, and Diana Santos (2021). “Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives”. In: <i>Modern Languages Open</i> 2021.1. 10.3828/mlo.v0i0.364.	965 966 967
Skuja, Frīdis (1912). <i>Zem saules</i> . K. Keirans.	968
— (1924). <i>Sidrābota saule lec</i> . Latvijas Aizsardzības Biedrība.	969
Stirna, Līga (2024). <i>Ātruma ierobežojums – 12 kilometri stundā. Kā pa Rīgu sāka joņot pirmie auto</i> . https://www.delfi.lv/a/120011444 (visited on 07/25/2024).	970 971
Tenen, Dennis Yi (2018). “Toward a Computational Archaeology of Fictional Space”. In: <i>New Literary History</i> 49.1, 119–147. 10.1353/nlh.2018.0005.	972 973
Törnberg, Petter (2024). “Best Practices for Text Annotation with Large Language Models”. In: <i>arXiv preprint</i> . arXiv: 2402.05129.	974 975
Underwood, Ted (2019). <i>Distant Horizons</i> . The University of Chicago Press.	976
Upītis, Andrejs (1913). <i>Pēdējais latvietis</i> . A. Golta apgādībā.	977
Upīts, Andrejs (1921). <i>Zelts</i> . D. Zeltiņa un A. Golta apgāds.	978
Veselis, Jānis (1931). <i>Dienas krusts</i> . Valters un Rapa.	979
— (1934). <i>Cilvēku sacelšanās</i> . Valters un Rapa.	980
Ziems, Calet, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang (2023). “Can Large Language Models Transform Computational Social Science?” In: <i>Computational Linguistics</i> , 1–53. 10.1162/coli_a_00502.	981 982 983




Verse within Prose

Annotating and Classifying Narrative Functions of Embedded Poems in Chinese Qing (1644-1912) Vernacular Fiction

Rongqian Ma¹ 

Keli Du² 

Yiwen Zheng³

1. Luddy School of Informatics, Computing, and Information, Indiana University Bloomington , Bloomington, IN, USA.
2. Trier Center for Digital Humanities, University of Trier , Trier, Germany.
3. Department of East Asian Languages and Cultures, Indiana University Bloomington , Bloomington, IN, USA.

Citation

Rongqian Ma, Keli Du, and Yiwen Zheng (2025). "Verse within Prose. Annotating and Classifying Narrative Functions of Embedded Poems in Chinese Qing (1644-1912) Vernacular Fiction". In: *CCLS2025 Conference Preprints* 4 (1). 10.26983/tuprints-00030147

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-02-06

Keywords

Chinese poetry, Qing fiction, Narrative function, Large language models, Annotation

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. What narrative functions do poems serve when interwoven with vernacular prose? This article takes what has often been labeled as “embedded poems” or “parasitic poems” in late imperial Chinese fiction as the primary subject of study. We examine the narrative roles of these poems within a selected corpus of Qing dynasty fiction, specifically investigating if an approach that combines human annotation with large language models can aptly capture and automatically classify their narrative functions. Through two rounds of iterative annotation and large language model testing, we demonstrate both the potential and limitations of this approach. As one of the few studies that applies large language models to Chinese literary research, our work lays the groundwork for future large-scale investigations into the dynamics between verse and prose in classical Chinese literature, incorporating both canonical works and beyond.

1. Introduction

This article takes what has often been labeled as “embedded (chanru 孳入)” poems or “parasitic (jisheng 寄生)” poems in late imperial Chinese fiction as primary subjects of study. The hybrid genre of prose incorporating verse has been widely practiced ever since the medieval period, especially during the late imperial times—the Ming (1368-1644) and Qing (1644-1912) Dynasties—an age where classical and vernacular Chinese fiction reached its peak. Scholarly perspectives on the function and significance of embedded verse in prose have long been divided—a debate that continues to this day. For example, as early as the Song Dynasty (960-1279), the literatus Luo Ye 羅燁, in his *The Drunken Man’s Talk* (*Zuiweng Tanlu* 醉翁談錄), demonstrated how poetry was used by professional storytellers as a teaser or a form of enticement—an embellishment designed to attract the audience. This practice,

however, was often regarded as lowbrow or associated with petty entertainment (see Luo 1965). In comparison, Zhao Yanwei 趙彥衛, in *Notes from the Cloud-Covered Hill* (*Yunlu Manchao* 雲麓漫鈔), emphasized how great works of the Tang Dynasty were great because they “incorporated diverse genres, including history, poetry, and argumentative essays” 蓋此等文備眾體，可以見史才、詩筆、議論 (see Yanwei Zhao 1966).

In modern times, one major camp readapted some earlier critiques and created the concept of “parasitic verse” as one major approach to studying the phenomenon (see Yishan Zhao 2014). They argue that verse can change or even impede the narrative flows of the text, a phenomenon that they think took stage in literary history momentarily and ultimately disappeared by giving way to the narrative. The other camp, as represented by the scholar Rao Longsun 饒龍隼, opposed the labeling of “parasitic verse” by arguing that the verse and narrative sections of the late imperial Chinese fiction are supplementary to each other in forming a holistic organism and thus should not be taken apart in analysis (see Rao 2023). Another group of scholars focuses on the cause and function of the verse-within-prose phenomenon, pivoting toward intellectual history and without taking a stance. For example, Zhang Zhejun 張哲俊 argues that poetry in verse provides authority to the text, a function that can be traced back to the classics such as *Book of Songs* (*Shijing* 詩經) (see Z. Zhang 2015). Similarly, Guo Jie 郭杰 suggests that the rise of poetry in verse in China originates from and mimics ways of history writing ever since pre-Qin times (see Guo 1995).

Elder generations of Western scholars of Chinese Studies tend to side with opinions on the lack or the diminishing narrative functions of verse in late imperial fiction. John Bishop stated that “originally such verses may have had an integral function in the story; later they served as a commentary, a verification, a means of delaying a climax, or merely as an embellishment” (see Bishop 1965, p. 241). Robert Hegel argued that the use of verse in fiction was mainly due to the writer’s scholarly identity and desire for articulation (Hegel 1985). He claimed that “the pace of action is deliberately slowed in mature novels for the elite by frequent insertion of verse, usually attributable to the narrator or ‘quoted’ by him from earlier, characteristically anonymous, sources...Literati novelists utilized the novel form to meet specific intellectual needs: social and political commentary, philosophical exploration, self-expression, and even their own and their friends’ enjoyment” (see Hegel 1985, p. 126).

However, during recent years, the narrative function of poetry in late imperial Chinese fiction has been increasingly noted and studied more in depth, especially in scholarly works that narrow in on a specific fiction work. For example, scholars have taken an interest in analyzing the narrative functions of poetry in one particular work—the Qing Dynasty fiction *Dream of the Red Chamber* (*Honglou Meng* 紅樓夢, also translated as *The Story of the Stone*). This makes sense because the work is often

considered the epitome of this hybrid genre as its narrative incorporates more than
a number of subgenres of verses. Cai Yijiang 蔡义江 sees the hybrid style in *Dream of
the Red Chamber* as a composite of genres that generate positive meaning, describing
it as “prose equipped with a variety of genres” (wen bei zhong ti 文備众體). Cai
specifically lists out five narrative functions of poems in *Dream of the Red Chamber*:
1) Social critique (借题发挥, 伤时骂世), 2) Part of plot (小说的有机组成部分), 3)
Reflecting social reality (时代文化精神生活的反映), 4) Character portraiture (按头
制帽, 诗即其人), and 5) Prediction of later plot (谶语式的表现方法) (see Cai 2007,
pp. 27-39). Another literary scholar Chia-ying Yeh 葉嘉瑩 divides the use of verse
in *Dream of the Red Chamber* into three types: 1) Pre-introducing characters through
homophonic puns or combination of radicals in poetry; 2) Modeling of characters
through articulation of their imagined voices; and 3) Conveying authorial intention
and emotion through prediction of later plot (see Yeh 2004, p.58). Compared to
Cai’s more comprehensive list that attempts to exhaust possible scenarios from the
perspective of the author, Yeh’s shortlist takes into account both characters and the
narrator and views the verse as conveying emotion and facilitating the plot and
narrative (see Yeh 2004).

Despite extensive interpretive efforts by literary scholars, few studies have examined
embedded poems collectively within the broader context of Ming-Qing fiction,
especially in lesser-known or non-canonical works. This gap may, in part, stem
from the methodological challenges of analyzing large-scale text corpora. In our
work, we engage in the literary debate over embedded poems’ functions, yet aim to
do so by leveraging state-of-the-art computational methodologies.

Recent advancements in natural language processing and computational literary
studies provide a promising means to revisit and analyze the “verse within prose”
literary phenomenon at an unprecedented scale. For instance, a Bayesian hierarchi-
cal generalized linear model can be used to track the relations between emotions
in poems and factors such as period, author profession, and rhyme. This analysis
demonstrates that the connection of emotion with rhyme is as strong as that with
thematic genre, while the connection with profession is as strong as that with gender
(see Konle et al. 2023). Moreover, the rapid development of LLMs (both prompting
and fine-tuning) has proven especially valuable for literary and poetry analysis. In
some cases, LLMs can match or even surpass supervised machine learning models
in distinguishing broadly recognized concepts such as science fiction, westerns,
or the emotional states of characters (see Bamman et al. 2024). Additionally, re-
searchers have tested zero-shot prompts with varying levels of information across
six state-of-the-art LLMs (including GPT-4 and LLaMA 3) to classify poetic forms
and their structural elements (see Walsh et al. 2024).

In the domain of Chinese literary studies, however, the application of computational
approaches as well as emerging LLMs remains relatively under-explored, with a few
pioneering efforts from scholars in East Asian studies. For example, Paul Vierthaler

investigated the “stylistic taxonomy” of the subtle and mixed genre of the late imperial unofficial historical narrative or quasi-history texts, by applying “statistical and linear algebraic analysis of the term frequency lists calculated from digitized transcripts” of these texts (Vierthaler 2016). Liu (Liu et al. 2018) demonstrated the potential of using digital tools to explore Chinese poetry from different aspects such as aesthetic expressions, and personal styles. Additionally, LLMs have shown remarkable capabilities in generating ancient Chinese poetry (see Huang and Shen 2025) and detecting and correcting errors in classical Chinese verse (see Yu et al. 2024).

Extending from the previous works, our article seeks to apply emerging LLMs to Chinese literary analysis, particularly exploring their potential in identifying and classifying the narrative functions of embedded poems in Ming-Qing vernacular fiction. Our approach combines human annotation with an iterative, trial-and-error method for automatic classification. Researchers have argued that literary scholars are best equipped to “explore, define, and exemplify narratological concepts” and encourage “a corpus with annotated concepts” to be created, before “any computer scientist and/or machine learning expert can work on the automatic detection of the concepts” (see Reiter et al. 2019). Following this principle, we selected a sample of Qing fiction, extracting embedded poems along with their immediate narrative contexts to construct a pilot dataset. Using this pilot dataset, we developed an iterative, collaborative annotation process to classify the narrative functions of these poems. The annotated dataset was then employed to test whether LLMs, such as ChatGPT, can automatically classify poems by their function on a larger scale. In this paper, we will describe the creation of the pilot dataset (Chapter 2), our iterative annotation process, as well as two rounds of preliminary testing and their results (Chapters 3 & 4). Our findings underscore both the potential and challenges of applying LLMs to the study of Ming-Qing fiction. Ultimately, this work opens new possibilities for computational approaches in Chinese literary scholarship.

2. Creating pilot dataset

From the Chinese Text Project¹, we obtained approximately 900 titles of Chinese Ming-Qing fiction in plain text. From this complete corpus, we randomly selected 18 novels and extracted embedded poems in them. The extraction was completed in two steps. Generally, the lines in Chinese poetry have a fixed length, typically five, seven, or eight characters per line. However, in some cases, a poem with four lines of eight characters each is written as two lines of sixteen characters. In the first step, we automatically extracted all lines of up to 20 characters, along with two sentences before and after each extracted line. This approach identified many poems but also included unrelated content such as Chinese couplets (duilian 對聯), chapter titles, or text with improperly formatted paragraph breaks. To refine

1. <https://ctext.org/>

the results, we processed the extracted text using ChatGPT (GPT-3.5, free version) 134
 with the following prompt: “Analyze the text below to see if it contains poems, and 135
 if so, retell all the poems you find.” While most of the data used to train ChatGPT 136
 is likely from modern Chinese texts, it successfully identified embedded poems 137
 written in vernacular and classical Chinese dating back to the Qing Dynasty. This 138
 was achieved using semantic cues like “詩曰” (The poem says) and “有詩為證云” 139
 (There is a poem to prove it), as well as formal elements of poetry, such as line 140
 length and number of characters per line. As a result, ChatGPT was able to extract 141
 360 embedded poems as our pilot dataset without requiring fine-tuning on classical 142
 Chinese texts specifically for this task. This dataset is available in the project’s 143
 GitHub page: <https://github.com/dkltimon/EmbeddedPoems>. 144

3. Initial Annotation Framework and Testing 145

3.1 Annotation Framework and Results 146

To build a foundation for the analysis of poems’ narrative functions, we employed 147
 a data-driven, bottom-up approach, in which we established an initial annotation 148
 framework by synthesizing insights from existing research literature with an ex- 149
 ploratory analysis of the 360 poems in our pilot dataset (Cai 2007; Chun 2009). In 150
 this initial framework, we identified five key narrative functions: opening teaser, 151
 character portraiture, commentary, integration within a scene or plot, and conclud- 152
 ing remarks (Table 1). We did not directly adopt the categories used in existing 153
 literary scholarship for two main reasons. First, prior discussions of the narrative 154
 functions of poetry in fiction have largely been based on close reading analyses 155
 of a few canonical works, such as *Dream of the Red Chamber*. These studies focus 156
 on a narrow group of literati authors and reflect poetry’s function within specific, 157
 limited historical contexts. Second, existing scholarship often differs significantly 158
 in both terminology and interpretation, resulting in fragmented and unsystematic 159
 categorizations. 160

Using this initial framework, one author with academic and research backgrounds 161
 in classical Chinese literature annotated the poems, assigning each to one of these 162
 narrative functions. Figure 1 demonstrates the annotation results. As shown in 163
 the graph, “part of a scene or a plot” is the most common category in our dataset, 164
 with approximately 140 poems. This is followed by “commentary,” which includes 165
 around 100 poems. “Character portraiture type” has a moderate presence with 166
 about 60 poems, while “opening teaser” and “concluding remarks” are less frequent, 167
 with roughly 40 and 20 poems, respectively. This distribution graph shows that 168
 poems serving as part of a scene or plot and those providing commentary are the 169
 most prevalent, whereas opening teasers and concluding remarks are comparatively 170
 rare. 171

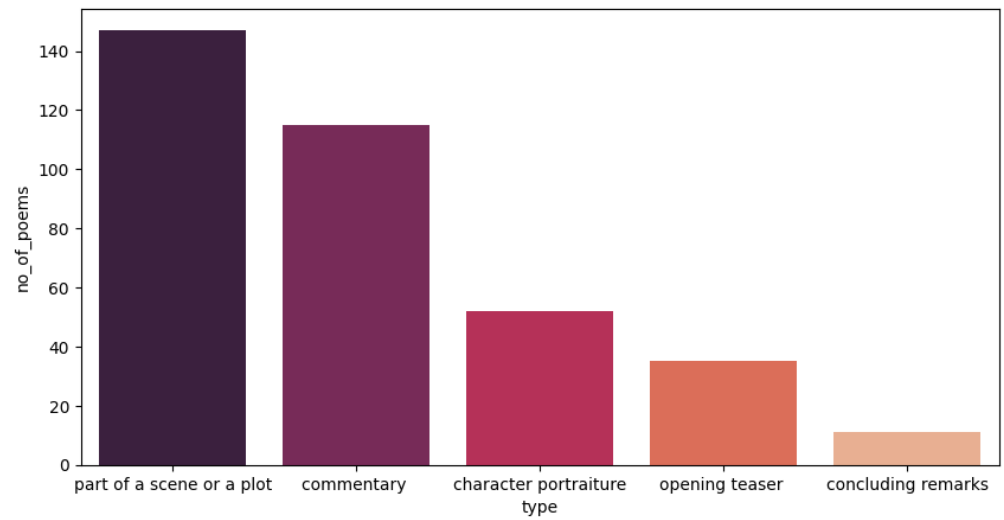


Figure 1: Distribution of the five identified narrative functions of embedded poems in Qing vernacular fiction.

Table 1: Initial annotation framework for the narrative functions of the embedded poems in Qing vernacular fiction.

Function	Description
Opening teaser	The poem is placed at the beginning of a chapter or the entire fiction. The poem may summarize the main message or idea of the fiction and serves as a teaser that grabs the readers' attention. A poem that serves this narrative function is usually composed from a third-person perspective (e.g., the narrator of the fiction).
Character Portraiture	The poem of this function describes and expresses the inner world, such as the emotions, thoughts, and feelings, of characters in the fiction, and is an important literary vehicle that shapes the personality of a character. The poem can be composed from both the character's first-person perspective or a third-person perspective (e.g., the narrator or other characters in the fiction). The poem's position is flexible and often interwoven with the vernacular narrative in the fiction.
Commentary	The poem amplifies and develops a plot from a third-person commentator's perspective. The commentator is usually the narrator and occasionally can be other characters in the fiction. Content-wise, such poems make a comment or critique of a story, a character, or a scene in the fiction. It also serves the purpose of educating the readers on moral lessons in Qing vernacular fiction.
Part of a scene or a plot	Poems of this function are a closely integrated component of a plot in fiction. The poems are usually composed by the characters involved in the plot. Although the specific content of the poems varies drastically depending on the given plot where they appear, overall, these poems facilitate the development of the plot and enrich the plot.
Concluding Remarks	The poem that serves this function is placed towards the end of a chapter or the entire fiction. Such poems usually summarize the entire story in the fiction or reiterate, emphasize, and amplify the main messages of the work. Poems of this function are usually composed from a third-person perspective (e.g., the narrator of the fiction).

3.2 First-round Classification Testing

172

After annotating the poems, we investigated whether LLMs can automatically distinguish their narrative functions. Specifically, we examined whether these

173

174

functions can be identified solely based on the poems' content, independent of their broader context. While this assumption is likely to be incorrect, our goal was to empirically test and potentially falsify it. To this end, we conducted an automatic classification of the poems without incorporating contextual information. Given the limited dataset of 360 poems and the imbalance across narrative function classes, the dataset was insufficient for fine-tuning and evaluating pre-trained models. Instead, we employed Zero-Shot classification and tested the following four models:

- **Erlangshen-RoBERTa-110M-NLI**: A fine-tuned version of the Chinese RoBERTa on several NLI datasets (J. Zhang et al. 2022).²
- **Bart-large-mnli**: The BART-large model (Lewis et al. 2019) after being trained on the MultiNLI dataset (Williams et al. 2018).³
- **XLM-ROBERTA-BASE-XNLI-ZH**: A fine-tuned version of the XLM-RoBERTa-base model (Conneau et al. 2020) using data in Chinese.⁴
- **ChatGPT-3.5 (free version)**: The English translation of the prompt used: "Please tell me which of the following categories this poem can be classified in. Please note that you can only assign the poem to one category. The five categories are: opening teaser, character portraiture, commentary, integration within a scene or plot, and concluding remarks."

As shown in Table 2, the classification performance was generally low, with most results aligning closely with the randomized baseline of 0.2. The best results were obtained using the XLM-Roberta-based Model, yielding an accuracy of 0.38 and an F1-score of 0.20. ChatGPT achieved the same F1-score, but lower accuracy.

Table 2: Zero-Shot classification results using four different models.

	Erlangshen-Roberta-110M-NLI	bart-large-mnli	XLM-ROBERTA-BASE-XNLI-ZH	ChatGPT-3.5
Accuracy / F1-score	0.16 / 0.08	0.21 / 0.13	0.38 / 0.20	0.27 / 0.20

To gain a deeper understanding of the classification results, we analyzed the confusion matrices for XLM-ROBERTA-BASE-XNLI-ZH and ChatGPT, as these two models achieved the highest F1-scores. As shown in Figure 2, the XLM-RoBERTa-based model exhibited a strong bias toward classifying poems as "part of a scene or plot," with approximately 60% of the poems assigned to this category. Since this category contains the largest number of poems in our dataset, the model's overall accuracy was the highest. However, this pattern suggests that the model

2. <https://huggingface.co/IDEA-CCNL/Erlangshen-Roberta-110M-NLI>

3. <https://huggingface.co/facebook/bart-large-mnli>

4. https://huggingface.co/morit/chinese_xlm_xnli

is overfitting to the dominant class rather than effectively distinguishing between 204
narrative functions. In comparison, ChatGPT is less overfitted and classified more 205
poems as “commentary” and “character portraiture” (see Figure 3). 206

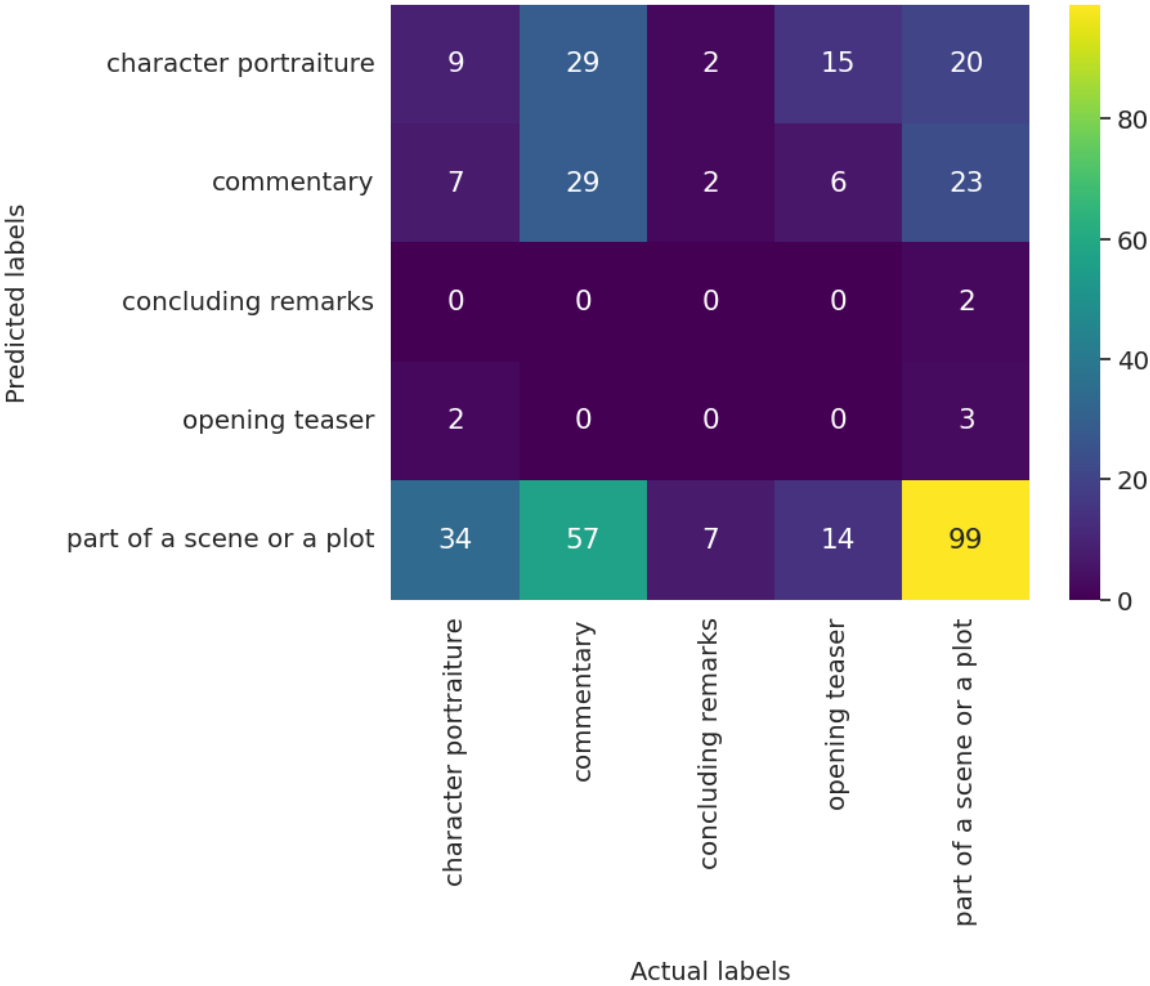


Figure 2: Confusion matrix for Zero-Shot classification using XLM-ROBERTA-BASE-XNLI-ZH.

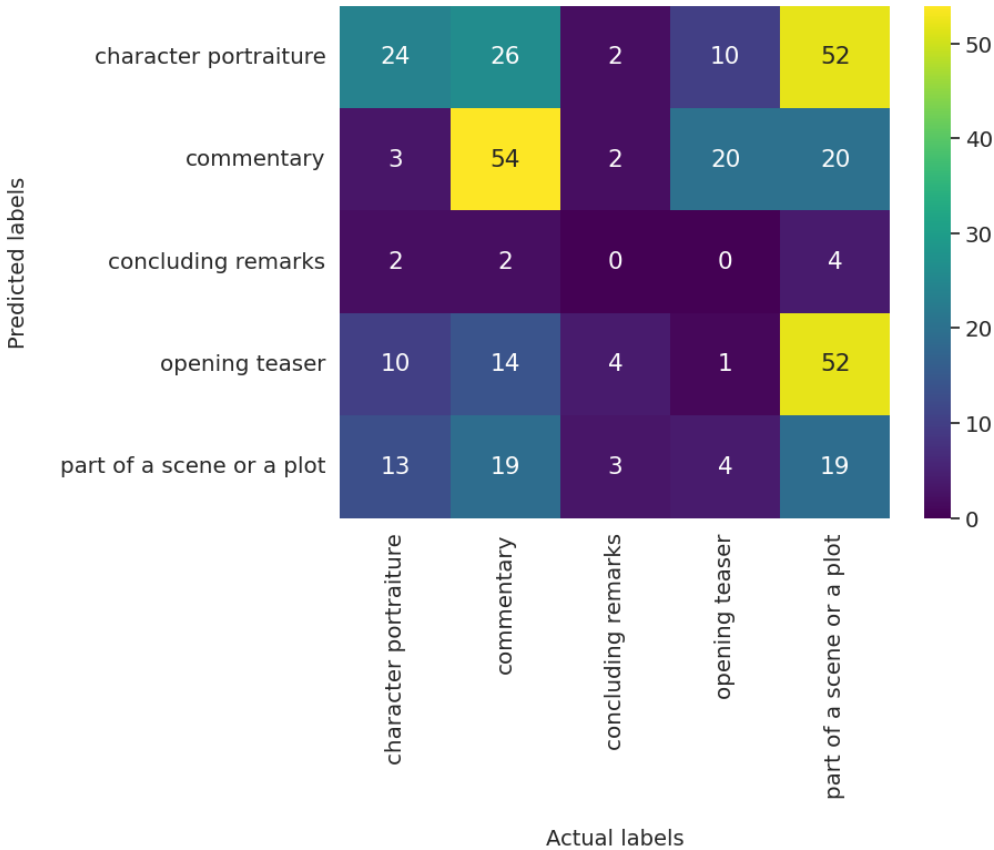


Figure 3: Confusion matrix for Zero-Shot classification using ChatGPT.

3.3 Reflections on the Framework and Results207

The initial annotation and model testing revealed several limitations in using LLMs208
to analyze embedded poems in Qing vernacular fiction. First, regarding the annota-209
tion framework, one poem has only one assigned label in the initial framework that210
encompasses all dimensions of information. For instance, labels such as “opening211
teaser” and “concluding remarks” are primarily defined by their structural roles212
and positions within the text, but they may also reflect thematic or content-based213
features. A poem placed at the beginning or the end of a chapter, for example,214
may simultaneously capture readers’ attention (structural role) and function as a215
narrative summary or commentary on a character (thematic role). Such examples216
demonstrate the need for a more multifaceted framework that captures the com-217
plexity of the narrative functions. Additionally, since the annotation process was218
conducted by one person, the subjectivity of individual interpretations may also be219
a factor that contributed to the models’ inconsistent performance.220

From a technical perspective, the models selected for the first round exhibited221
limitations in their linguistic and cultural adaptability when processing Chinese222
texts, making them less than ideal for this task. Furthermore, the methods used223
for prompt design and implementation may have also influenced the classification224
results.225

Based on these observations, we refined our approach and conducted a second round of annotation and testing. In this subsequent phase, we developed a new annotation framework, revised our annotation process, and adjusted both the model selection and testing approaches to achieve better classification results.

4. Revised Framework and Second-Round Testing

4.1 Revised Annotation Framework and Results

In the revised annotation framework, each poem was analyzed across three dimensions—position, perspective, and content—to better capture the complexity of its narrative functions. The “position” dimension indicates whether a poem appears in the opening section of a chapter, within the middle of the narrative, or towards the end of a chapter. “Perspective” denotes whether a poem is written from a character’s first-person or a narrator’s third-person viewpoints. Finally, the “content” dimension categorizes each poem into one of four types: “character portraiture,” “scene,” “commentary,” and “plot.” Definitions for each category are included in Table 3.

Table 3: Definitions of each category for the content aspect.

Character portraiture	Describe a character, e.g., their appearance, inner feelings, emotions, and personality.
Scene	Describe natural scenery, objects, and nature.
Commentary	Offers comments and critiques of events, society, morality, characters, etc.
Plot	Conveys, narrates, and sometimes summarizes a sequence of events.

Using the revised framework, three annotators independently labeled each poem in the dataset from the three dimensions, assigning one label for each dimension⁵. The inter-annotator agreement, measured using Fleiss’ kappa (Fleiss and Cohen 1973), yielded scores of 0.87 for position, 0.89 for perspective, and 0.66 for content, respectively. To address annotation discrepancies, the three annotators convened to reconcile differences, and this process resulted in a finalized ground-truth dataset for classification experiments. As illustrated in Figure 4, the dataset shows an even distribution of poems between the “narrator’s” and “character’s” categories for “perspective”. Most poems are categorized as “middle” for “position.” For content, only one poem is classified as “plot,” while 36 poems are labeled as “character portraiture.” The remaining two content categories are evenly represented across the dataset.

5. In the second round of annotation, the annotators identified and removed several couplets and duplicate poems from the pilot dataset. As a result, the data size has been reduced from 360 in 18 novels to 339 poems in 15 novels. The new dataset is available in the GitHub page: <https://github.com/dkltimon/EmbeddedPoems>

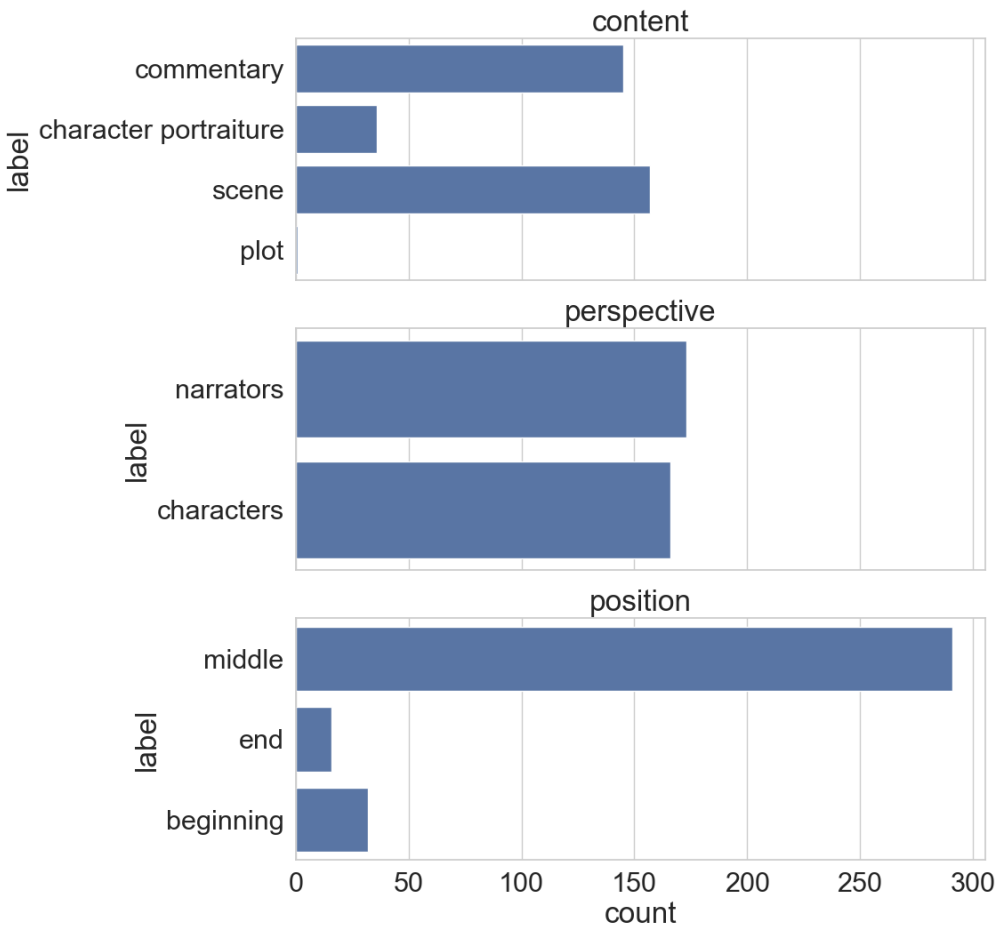


Figure 4: Annotation results for the poems using the revised framework.

4.2 Second-round Testing with the Revised Framework253

In the second round of testing, we improved our approach to classifying poems254
in several ways. First, the results from the first round indicated that the narrative255
function of poems cannot be identified without considering their surrounding256
narrative contexts very well. To address this, we included two sentences before and257
after each poem as contexts for classification in this round.258

Second, compared to multilingual language models fine-tuned on Chinese texts259
such as RoBERTa, the free version of ChatGPT 3.5 demonstrated better performance260
in the classification task. Also, generative models are easier to use. Therefore, we261
used three different generative models for this test: the paid version of ChatGPT262
(GPT-4) and two open-source models—Llama 3.3 (the new state-of-the-art 70B263
model)⁶ and a Chinese Llama model, Llama-3-Chinese-8B-Instruct-v3 (Cui et al.264
2023)⁷.265

Third, we refined the prompting methods and tested the poem content classification266
using three different prompts. The first was a brief prompt asking the model to267

6. See: <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>
7. See: <https://huggingface.co/hfl/llama-3-chinese-8b-instruct-v3-gguf>

select a category without further information. The second was a longer prompt that provided detailed definitions for each category before requesting a selection. The third also explained the categories but instructed the model to use a binary classification approach.⁸ All three models mentioned above were tested using these prompts as follows:

- Short prompt:** “The following Chinese text contains a poem. The beginning and end of the poem are marked with ‘p_s’ and ‘p_e’ respectively. Which of the categories “commentary,” “character portraiture,” “scene” and “plot” does it belong to? You do not have to explain your answer, just output your answer using the given categories. Here is the text with the poem:”
- Long prompt:** “The following Chinese text contains a poem. The beginning and end of the poem are marked with ‘p_s’ and ‘p_e’ respectively. You have three tasks. First task, determine the narrative function of the poem. There are four options: 1. ‘commentary’, which offers comments and critiques of events, society, morality, characters, etc. 2. ‘character portraiture’, which describes a character, e.g., their appearance, inner feelings, emotions, and personality. 3. ‘scene’, which describes natural scenery, objects, and nature. 4. ‘plot’, which conveys, narrates, and sometimes summarizes a sequence of events. Second task, determine the position of a poem in a novel. The position indicates the structural role that the poem plays in the narrative of the fiction. There are three options: 1. ‘beginning’, means that the poem is an opening poem for a chapter. 2. ‘middle’, means that the poem is in the middle of a plot. 3. ‘end’, means the poem comes at the end of the chapter and concludes the storyline. Third task, determine whether the poem is composed or recounted from the first-person perspective of a character in the story or a third-person perspective of the author or a storyteller. If the former, please answer with ‘character’, if the latter, please answer with ‘narrator’. For each task, you must choose one and only one option as your answer. You do not have to explain your answer, just output your answer in this format answer to the first task, answer to the second task, answer to the third task using the given option labels. Here is the text with the poem:”
- Long prompt (binary approach):** “The following Chinese text contains a poem. The beginning and end of the poem are marked with ‘p_s’ and ‘p_e’ respectively. Your task is to determine the narrative function of the poem, considering both the content and the context of the poem. First, determine if the poem offers comments and critiques of events, society, morality, characters, etc. If yes, answer ‘commentary’ . If no, determine if the poem describes natural scenery, objects, and nature. If yes, answer ‘scene.’ If no, determine if the poem describes characters, e.g., their appearance, inner feelings, emotions,

8. The second prompt also asked the LLMs to classify the position and perspective of the poem. This will be addressed later.

and personality. If yes, answer ‘character portraiture’. If no, determine if the poem conveys, narrates, and sometimes summarizes a sequence of events. If yes, answer ‘plot’. If still no, read the text again and choose one from the above-mentioned three options (‘commentary’, ‘character portraiture’, ‘scene’). You must choose one and only one option as your answer. You do not have to explain your answer, just output your answer using the given option labels. Here is the text with the poem:”

As shown in Table 4, the best classification results were achieved by ChatGPT using the long prompt, with an accuracy of 0.55 and an F1-score of 0.43, which is much better than the classification results obtained in the first-round testing (accuracy 0.38, F1-score 0.20). This suggests that providing detailed explanations can help models better understand both the input text and the task. However, the binary approach did not lead to further improvements but slightly reduced both accuracy and the F1-score. In comparison, the Llama models performed worse than ChatGPT. While Llama 3.3 achieved similar accuracy, its F1-scores were much lower. A detailed examination of the results revealed that Llama 3.3 classified 90.6% of the poems as “commentary,” indicating a strong overfitting issue. Surprisingly, the Chinese Llama model performed the worst, with better classification results when the short prompt was used. This model also displayed overfitting, classifying 45% and 53% of the poems as “commentary” and “plot,” respectively. By contrast, only five poems were identified as “character portraiture,” and none were classified as “scene.” To obtain more details on the classification results, we checked the confusion matrix for our best classification (Figure 5). As can be seen, ChatGPT correctly identified 81% of “character portraiture” poems and 57% of “commentary” poems. However, 54% of “scene” poems were misclassified into other categories, highlighting areas that require further refinement.

Table 4: Classification results of poems for the content aspect.

	Short prompt		Long prompt		Long prompt binary	
Model	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
ChatGPT	0.41	0.33	0.55	0.43	0.53	0.39
Llama 3.3	0.40	0.15	0.51	0.33	0.53	0.29
llama3-zh-inst	0.37	0.29	0.21	0.14	0.44	0.17

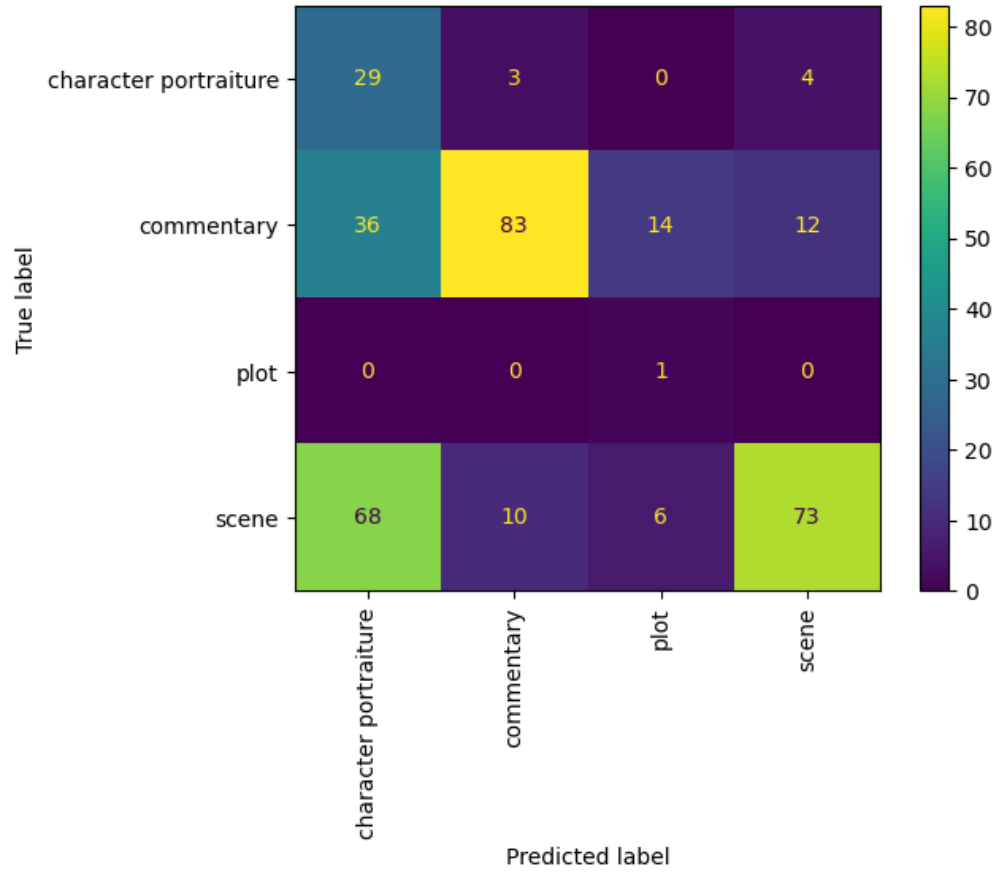


Figure 5: Confusion matrix of content classification using ChatGPT and the long prompt.

Using the long prompt described earlier, we also tested whether the models could identify the position and perspective of the poems. When classifying the position of poems, Llama 3.3 achieved a much higher F1-score than ChatGPT, with only a 0.02 decrease in accuracy. In terms of perspective classification, both Llama models outperformed ChatGPT. Taken together, the classification results across all three dimensions show that ChatGPT and Llama 3.3 have their own strengths. It is worth noting that Llama 3.3 is an open-source, free model and that we can fine-tune using texts from the Ming and Qing periods, which makes it even more appealing for our research in the future.

Table 5: Comparison of classification results using the long prompt across three aspects in the second-round testing.

	Content		Position		Perspective	
Model	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
ChatGPT	0.55	0.43	0.84	0.34	0.63	0.60
Llama 3.3	0.51	0.33	0.82	0.55	0.75	0.73
llama3-zh-inst	0.21	0.14	0.48	0.26	0.66	0.66

4.3 Analysis of Misclassified Examples

342

What caught our attention the most among the results in our second-round testing 343
 was that ChatGPT incorrectly classified 109 poems into the “character portraiture” 344
 category for the content aspect. This represents nearly one-third of the entire dataset. 345
 To better understand the reasons behind these misclassifications, we conducted a 346
 detailed investigation of the mislabeled examples. This analysis uncovered several 347
 distinct patterns, demonstrating the challenges of aligning automated classification 348
 with human expertise in Chinese literary traditions. 349

4.3.1 Scene misclassified as character portraiture

350

One prominent type of misclassification involves 68 poems categorized by three 351
 human annotators as “scene” but classified by ChatGPT as “character portraiture.” 352
 The following example from our data sample aptly demonstrates this argument: 353

“彼時，众人都挪到當中桌子旁邊來，等可人月旦。獨爐湘妃折下一枝菊 354
 花，插在瓶中，放在面前，寫「供菊」一題，見了他二人眼睛，看著福 355
 壽笑了一笑。只見可人前，擺著紅筆朱硯，先看璞玉的詩： 356

懷菊 潤翰公子 357
 獨倚東籬思故友，哀吟淒涼增新愁。 358
 此心郁郁無人問，斜生彎枝知也無？ 359
 涼秋已臨我何急，盛時既去汝太羞。 360
 艷色秀容今何在？曼立香迹猶楚楚。 361

可人看罷，笑道：「璞公此詩，可謂懷之入骨髓矣，真古今之絕唱也。」 362

[Translation]⁹: “At that moment, everyone moved to the table in the 363
 center, waiting for Keren’s monthly commentary. Lu Xiangfei alone 364
 broke off a chrysanthemum branch, placed it in a vase in front of her, 365
 and wrote the title “Offering Chrysanthemums.” When she met the eyes 366
 of the two, she smiled at Fushou. In front of Keren was placed a red 367
 brush and a vermillion inkstone. First, she read Puyu’s poem: 368

‘Cherishing Chrysanthemums’ by Scholar Runhan 369
 Leaning alone on the eastern fence, thinking of an old friend, 370
 Sorrowful chants add new grief. 371
 This heart, melancholic, remains uncared for, 372
 Do the slanted, bending branches know or not? 373
 The cool autumn has arrived, why must I hurry? 374
 The peak of your time has passed, causing you boundless shame. 375
 Where now is your bright, elegant beauty? 376
 Still upright, your fragrant traces are clear and pure. 377

After reading, Keren smiled and said: ‘Scholar Pu’s poem can truly be 378

9. The translations presented in this article were created with the help of ChatGPT (GPT-4o).

said to express longing down to the marrow. It is indeed an unparalleled masterpiece of ancient and modern times.”

This example is quoted from the Qing fiction *One-Story Pavillion* (*Yiceng Lou* 一層樓).¹⁰ This example captures a literati gathering engaging in a chrysanthemum-themed poetry contest set against an idyllic autumn backdrop. In such poetry contests, the participants often wove emotional elements into their poems to demonstrate their creativity and artistic finesse. Therefore, these emotions were not necessarily a reflection of the participants’ or the novelist’s genuine feelings, but rather artistic constructs strategically crafted to enhance the poem’s appeal and increase the participants’ chances of winning literary competitions. Annotators familiar with this cultural and literary tradition interpreted the poem’s primary function as depicting a poetic exchange scene, rather than expressing individual characters’ emotions. ChatGPT appears to have a limited understanding of Chinese literary conventions or did not consider the tradition of poetic exchanges when classifying the poems, which might have led to this misclassification.

Another instance of the “scene-to-character portraiture” misclassification involves poems inscribed on objects as decorative elements. The primary function of these poems is to complement the aesthetic or symbolic value of the objects. For instance, in one example from *Shadows of Dream of the Red Chamber* (*Hong Lou Meng Ying* 紅樓夢影),¹¹ a poem inscribed on a fan was presented and appreciated:

“賈蘭说：「我的扇子也是他送的，姑姑看見沒有？」二人齊说：「沒有，你取去。」賈蘭忙忙下樓，不一刻取來。探春接來一看，也是檀香股、絹面，小楷寫的「擬閨詞」七律四首。探春念道：

東風影里罷梳頭，窗外呢喃聽不休。
藻井待栖雙玉剪，筠簾初上小銀鈎。
疑將軟語商量定，似有柔情宛轉留。
銜得新泥重補葺，余香猶記舊妝樓。”

[Translation]: “Jia Lan said, ‘My fan was also a gift from him [Xue Pan 薛蟠]. Aunt, have you seen it?’ The two replied in unison, ‘No, go fetch it.’ Jia Lan hurried downstairs and returned shortly after with the fan. Tanchun took it and examined it closely. It was also made with sandalwood ribs and a silk surface, inscribed in fine script with four seven-character quatrains titled ‘Imitations of Boudoir Verses’. Tanchun began to read aloud:

Amid the shadows of the spring breeze, she sets her comb aside,
Listening to the incessant chirps outside the window.
Beneath the carved ceiling, the scissor-shaped swallows await their perch,

10. <https://ctext.org/wiki.pl?if=en&chapter=406211&remap=gb#p7>

11. <https://ctext.org/wiki.pl?if=gb&chapter=576940&remap=gb#p16>

The bamboo blinds newly adorned with a silver sickle moon. 417
 Perhaps, tender words have reached an accord, 418
 As if gentle sentiments linger in their winding flow. 419
 Carrying fresh mud, they rebuild what was once their home, 420
 The lingering fragrance marks the old makeup chamber.” 421

Poems as such were crafted to complement and emphasize the value of the objects 422
 that the characters in the story adorned. Therefore, the primary function of these 423
 poems was to vividly describe the objects, bringing them to life, while simultane- 424
 ously showcasing the poetic and literary talents of the fictional writer. The poem 425
 under examination in this case was titled “Imitation of Boudoir Verses” (ni guici 426
 擬閨詞), which represents a typical convention of classical Chinese poetry. Such 427
 poems are typically written by male poets who adopt a female persona to articu- 428
 late women’s emotions, inner worlds, and everyday experiences—often exploring 429
 themes of longing, separation, loneliness, and the transience of youth. Given this 430
 literary tradition, the references to a woman (e.g., the use of “she”) in the poem 431
 should not be interpreted as literal depictions of a fictional character. As such, the 432
 poem would not be classified as “character portraiture” in this context. While it is 433
 possible the poem may indirectly reflect the thoughts, psychology, or personality of 434
 its author (see Rouzer 2001)—and, if the author were indeed a fictional character, 435
 potentially serve a characterizing function—this does not apply to the example 436
 at hand. The poem appears in a scene where the character Jia Lan displays a fan 437
 gifted by Xue Pan. However, close reading of the surrounding narrative reveals no 438
 clear indication of who authored the poem inscribed on the fan. This misclassified 439
 example further underscores LLMs’ struggle to interpret Chinese poetry, especially 440
 when meaning emerges from the nuanced interplay between form, tradition, and 441
 narrative function. 442

4.3.2 Commentary misclassified as character portraiture 443

The second major case of misclassification includes 36 poems that were identified 444
 as “commentary” by human annotators while classified as “character portraiture” 445
 by ChatGPT. The following case from *A Pillow of Wonders* (Yizhen Qi 一枕奇)¹² 446
 demonstrates this point: 447

“莫说丁協公是個富貴公子，他日日要見教的；就是徐鵬子一個窮公孫， 448
 他看他考得利肚里又通，也時常虛賣弄，三兩日來鬼混一場去。總不如 449
 那丁公子與他貼心貼意，分外相投，一刻也離他不得的。這正是： 450
 嫖賭場中箋片，文章社內法喜。 451
 雖然牌挂假斯文，不如尊綽白日鬼。 452
 却说丁協公看了那條字兒，委決不下，躊躇了一夜，次日侵早，著人去 453
 請了白日鬼來。周白日道：「昨日有些小事，不曾會你，場期已迫，看你 454

12. <https://ctext.org/wiki.pl?if=gb6chapter=523546remap=gb>

的氣色好的緊，今科定要高發的。請問呼喚何事見教？」 ” 455

[Translation]: “Not to mention Lord Ding Xie, a wealthy nobleman he 456
receives lessons from every day; even Xu Pengzi, a poor descendant of a 457
noble family, who he finds capable in exams and well-read, is someone he 458
frequently interacts with and flatters. He often engages in lighthearted 459
mischief with Xu every two or three days. However, none of this com- 460
pares to Lord Ding Xie’s deep and heartfelt connection with him. The 461
two are exceptionally close, so much so that Ding cannot bear to part 462
from him for even a moment. This is precisely: 463

A sidekick in gambling dens, an entertainer in scholarly circles. 464
Though wrapped in pretended refinement, 465
Better call him ‘Daylight Swindler’. 466

Now, regarding Lord Ding Xie, he was deeply conflicted after reading the 467
message and deliberated over it for an entire night. Next morning early, 468
he sent someone to invite the ‘Daylight Swindler’ over. Zhou Daylight 469
said, ‘Yesterday, I had some small matters to attend to, so I couldn’t meet 470
with you. With the exam approaching and your energy looking excellent, 471
I’m certain this will be a successful year for you. May I ask, what is the 472
purpose of summoning me today?’” 473

When examined solely within the extracted sample, the poem proved difficult to 474
understand. This confusion may likely come from the absence of context for terms 475
such as “Daylight Swindler” (白日鬼) or the reference to “he” in the excerpt. To 476
resolve this ambiguity, the annotators revisited the original fiction. They discovered 477
that a character named Zhou De, with the nickname “Daylight Swindler,” was 478
introduced in the preceding paragraphs of the same chapter where this extracted 479
sample is located. This contextual information clarified the meaning of the poem: 480
rather than focusing on interpersonal relationships, it delivers a sarcastic critique of 481
Zhou De’s idle and opportunistic nature. Based on this understanding, annotators 482
classified the poem as a commentary on the character’s personality, which falls 483
under the “commentary” category. By contrast, ChatGPT labeled this poem as 484
“character portraiture.” We asked ChatGPT to explain the rationale behind the 485
classification and were given this answer: “The poem describes the relationship 486
between two characters, showing the preference of one character towards another. 487
It highlights a bond and understanding between the two, giving us insight into 488
their personalities and conduct.” This misclassification may stem from the limited 489
context (i.e., only the two-sentence excerpts immediately before and after the poem) 490
provided to ChatGPT during the classification process, where the key contextual 491
details regarding the character Zhou De were absent. This misclassified example 492
has inspired us to reassess whether two sentences before and after the poem provide 493
sufficient context for LLMs to make informed decisions. 494

Another scenario of the “commentary-to-character portraiture” misclassification commonly happens among poems following typical semantic cues, such as “there is a poem that proves it” (有詩為證) and “this is precisely” (這正是). The following case from the data illustrates this scenario:

“鵬子道：「勸你放心。這科包管決中，賠也賠得你一個舉人。若還不中，不但無顏見你，也無面目再見那些親族朋友了。」王氏道：「但願如是，就當拜謝天地。」這正是：

只謂才不如己，爭道巧不猶人。
指望一朝騰霄漢，誰知窮鬼不離身。

却说同學內有一個秀才，姓丁名全，字協公，其人也是世家。乃父累官至工部侍郎，宦途頗順，廣積官資。”

[Translation]: “Pengzi said, ‘I urge you to rest assured. I guarantee that I will pass this examination, and the worst case scenario would be a “recommended man”. If I don’t succeed, not only would I have no face to see you, but I would also have no face to see our relatives and friends.’ Madam Wang replied, ‘I only hope it will be so; if it happens, we will surely give thanks to Heaven and Earth.’ This is precisely:

Some claim others lack their talent,
Never admits their smartness does not surpass others’ .
Hoping one day to soar to the skies,
Yet who knew misfortune clings like a shadow.

Now, among the students, there was a scholar named Ding Quan, with the courtesy name Lord Xie. He came from an established family; his father had steadily advanced in his official career, rising to the rank of Vice Minister of the Ministry of Works. His bureaucratic path had been smooth, allowing him to amass significant capital of officialdom.”

Similarly from *A Pillow of Wonders*,¹³ this poem is composed immediately after the narrator describes the financial and career struggles faced by Pengzi and his family. The poem creates a contrast between Pengzi’s ambition and his lack of fortune and luck, reinforcing the central themes established in the preceding context. The phrase “this is precisely” preceding the poem also signals that the narrator of the story is about to reiterate the content presented previously.

To some extent, this poem was indeed about Penzi, a character in the fiction. So, this misclassification may be due to the inherent ambiguity of the poem. However, human annotators were able to distinguish between a “description of a character” and a “comment on a character’s situation,” the latter being the correct classification for this poem. This may be rooted in their understanding of the structural cues provided by phrases such as “this is precisely,” which emphasize or reiterate the messages

13. <https://ctext.org/wiki.pl?if=gb6chapter=523546remap=gb>

stated previously. These observations suggest that introducing the roles of such structural phrases to LLMs, or providing explicit examples for these ambiguous, boundary cases, may help enhance their classification accuracy.

5. Discussion

The two rounds of iterative annotation and testing show that while LLMs hold significant potential for identifying and classifying poetry’s narrative functions, limitations remain that must be addressed for the approach to be applicable to the task. To address these limitations, our analysis suggests possible approaches such as enhancing the annotation framework, refining prompts, and incorporating technical considerations.

5.1 Annotation Framework

Reflecting on the annotation process, we recognize that the inherent ambiguity and fluid boundaries between categories may have contributed to the challenges of automatic classification. In particular, the “commentary” category of the content aspect in the revised framework still encompasses a wide range of content, including reflections on characters’ personalities, moral lessons, scenes, and even broader societal themes. The lack of clear distinctions between “commentary” on a character or scene and descriptions of a character (“character portraiture”) or a scene (“scene”) may have led to confusion for LLMs. To address these issues in the next round of annotation and testing, we will focus on refining the “commentary” category by dividing it into more specific and well-defined subcategories.

The single-label annotation system with mutually exclusive narrative-function categories may also partly explain why the models appear to misclassify some poems’ narrative functions. To address this potential issue, we plan to experiment with a multi-label annotation framework in place of the current single-label approach, to better capture the complexity and richness of the narrative functions that a poem often serves within fictional storytelling. In addition to revising the annotation framework itself, we will develop a more detailed annotation guideline to support future experiments with few-shot learning. This guideline will incorporate examples for each label, as well as discussions of misclassified cases to clarify the rationale behind function assignments, particularly in boundary or ambiguous instances.

5.2 Prompts Development

Based on our analysis of misclassified cases, we propose the following strategies to enhance prompt design and further improve LLMs classification results: First, incorporate sufficient information and knowledge about classical Chinese literary traditions in the prompts to communicate with LLMs. This addition may help

LLMs develop a more culturally oriented interpretation of the poems in late imperial Chinese novels, such as those in poetic exchange scenes discussed above. Second, expand the accompanying prose contexts for the poems, so that LLMs can look for additional, more accurate cues to understand and interpret the poems. Third, during the second round of testing, we got the best results using the long prompt. We speculate that this may be because the long prompt pushed ChatGPT to develop a more comprehensive understanding of the poems by requiring classification across all three aspects (content, position, and perspective). We plan to test this hypothesis in future work. For example, we will incorporate instructions for classifying position and perspective into the long prompt (binary approach) to see if the overall classification results improve. Finally, our analysis indicates that the semantic diversity of texts may have also confused LLMs. Some of the misclassifications discussed above—particularly those involving the distinctions between a “commentary of a character” and a “description of a character,” as well as those triggered by structural cues like “this is precisely”—could be mitigated by using few-shot classification and providing LLMs with a few examples.

5.3 Technical Challenges

Additionally, we also face technical challenges, the most pressing of which is the difficulty in fully understanding how generative LLMs operate. For example, the Llama home page states the following:

“Llama 3.3 supports 7 languages in addition to English: French, German, Hindi, Italian, Portuguese, Spanish, and Thai. Llama may be able to output text in other languages than those that meet performance thresholds for safety and helpfulness. We strongly discourage developers from using this model to converse in non-supported languages without implementing fine-tuning and system controls in alignment with their policies and the best practices shared in the Responsible Use Guide.”¹⁴

Given this limitation, we tested not only Llama 3.3 but also a fine-tuned version of Llama 3 trained on Chinese text, anticipating that the latter would yield better classification results for Chinese-language data. The results of the test showed that the classification of the majority of the poems was successfully completed by Llama 3.3, except for a few that were answered “*I can’t fulfill this request*” (twice) and “*I don’t have the capability to view or analyze the Chinese text you provided. Could you please copy and paste the text here, and I’ll be happy to help you determine which category it belongs to?*” (three times). What surprises us even more is that Llama 3.3 outperforms the Chinese Llama model across all three aspects of classification and even surpasses ChatGPT in identifying the position and perspective of poems. We do not understand how a model that does not support Chinese can accomplish the task, and it is unclear whether the results of the classification were based on the

14. <https://ollama.com/library/llama3.3>

model’s understanding of the text or it was just a shot in the dark. This suggests that the opacity of LLMs’ training data can significantly limit our understanding of why a model produces certain outputs, particularly for culturally dependent tasks. This limitation applies regardless of whether the model is classified as “open” (e.g., Llama 3) or closed-source (e.g., ChatGPT). In future research, we plan to experiment with models trained on larger Chinese corpora, such as DeepSeek, as well as next-generation LLMs as they become available—particularly those with improved capabilities for processing classical Chinese texts.

6. Conclusion

In this article, we explored the use of LLMs to examine the narrative functions of “embedded poems” in Chinese Qing fiction. Specifically, we presented two rounds of iterative annotation processes and LLMs testing. Our analysis revealed the diverse roles that poetry plays in Qing novels and highlighted both the potential and inherent limitations of LLMs for identifying and classifying these functions. Moreover, we found that an integrated refinement approach that encompasses adjustments in annotation, model selection, and testing methodologies can enhance the performance of LLMs for our classification task. After two rounds of refinements, our findings showed that ChatGPT and Llama 3.3 outperformed the other models in our dataset, each demonstrating unique strengths. Moving forward, we will continue to refine our approach to further improve the robustness and accuracy of the classification results. The ultimate goal of our work is to develop a computational approach that analyzes the narrative function of poetry in late imperial Chinese vernacular writings on a large scale, extending beyond the limited corpus of canonical works.

Our work contributes to both Chinese literary scholarship and research on LLMs. Harnessing the power of LLMs to revisit the storytelling dynamics of this rich literary tradition, we can assess and offer insights into the narrative roles of poetry in vernacular novels on a large scale. From the perspective of LLM research, this study highlighted a key limitation of current LLMs: their difficulty in processing culturally distinct corpora. This underscores the need for more rigorous evaluation and experimentation before LLMs can be applied effectively and responsibly in computational literary analysis. We believe that through careful testing, evaluation, and fine-tuning, LLMs can be developed into powerful tools for analyzing multilingual and linguistically complex text—domains that remain underrepresented in mainstream training data, which is predominantly derived from Western, contemporary, and commercially available Internet sources.

7. Data Availability

Data can be found here: <https://github.com/dkltimon/EmbeddedPoems>

8. Software Availability 647

Software can be found here: <https://github.com/dkltimon/EmbeddedPoems> 648

9. Author Contributions 649

Rongqian Ma: Conceptualization, Data Curation, Methodology, Investigation, 650
Writing – original draft, Writing – review & editing, Funding Acquisition 651

Keli Du: Conceptualization, Data Curation, Methodology, Software, Formal Analy- 652
sis, Visualization, Writing – original draft, Writing – review & editing 653

Yiwen Zheng: Investigation, Writing – original draft, Writing – review & editing 654

References 655

- Bamman, David, Kent K. Chang, Li Lucy, and Naitian Zhou (2024). “On Classifi- 656
cation with Large Language Models in Cultural Analytics”. In: *Computational 657*
Humanities Research Conference (CHR 2024), 494–507. <https://ceur-ws.org/Vol-658-3834/paper119.pdf>. 659
- Bishop, John L (1965). “Some limitations of Chinese fiction.” In: *The Far Eastern 660*
Quarterly 15.2, 239–247. [10.1002/andp.19053221004](https://doi.org/10.1002/andp.19053221004). 661
- Cai, Yijiang (2007). *Hongloumeng shici qufu quanjie* 紅樓夢詩詞曲賦全解. Shanghai: 662
Fudan daxue chubanshe. ISBN: 9780198520115. 663
- Chun, Mei (2009). “Garlic and Vinegar: The Narrative Significance of Verse in 664
‘The Pearl Shirt Reencountered’ ”. In: *Chinese Literature: Essays, Articles, Reviews 665*
31, 23–43. <https://www.jstor.org/stable/20799715>. 666
- Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guil- 667
laume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, 668
and Veselin Stoyanov (2020). *Unsupervised Cross-lingual Representation Learning 669*
at Scale. arXiv: [1911.02116 \[cs.CL\]](https://arxiv.org/abs/1911.02116). <https://arxiv.org/abs/1911.02116>. 670
- Cui, Yiming, Ziqing Yang, and Xin Yao (2023). “Efficient and Effective Text Encoding 671
for Chinese LLaMA and Alpaca”. In: *arXiv preprint arXiv:2304.08177*. [https://a 672](https://arxiv.org/abs/2304.08177)
[rxiv.org/abs/2304.08177](https://arxiv.org/abs/2304.08177). 673
- Fleiss, Joseph L. and Jacob Cohen (1973). “The equivalence of weighted kappa and 674
the intraclass correlation coefficient as measures of reliability”. In: *Educational 675*
and psychological measurement 33.3, 613–619. [https://doi.org/10.1177/001316 676](https://doi.org/10.1177/001316676447303300309)
[447303300309](https://doi.org/10.1177/001316676447303300309). 677
- Guo, Jie (1995). “Zhongguo gudian xiaoshuo zhong shiwen ronghe chuantong de 678
yuanyuan yu fazhan 中國古典小說中詩文融合傳統的淵源與發展”. In: *Zhongguo 679*
wenxue yanjiu 中國文學研究 2, 11–17. [CNKI:SUN:ZWX.Y.0.1995-02-001](https://doi.org/10.1007/978-7-309-00100-1). 680

- Hegel, Robert E. (1985). "Distinguishing Levels of Audiences for Ming-Ch' ing Vernacular Literature: A Case Study". In: *Popular Culture in Late Imperial China*. Ed. by David Johnson, Andrew J. Nathan, and Evelyn S. Rawski. Berkeley and Los Angeles: University of California Press, 112–142.
- Huang, Chihan and Xiaobo Shen (2025). "PoemBERT: A Dynamic Masking Content and Ratio Based Semantic Language Model For Chinese Poem Generation". In: *Proceedings of the 31st International Conference on Computational Linguistics*, 50–60. <https://aclanthology.org/2025.coling-main.5/>.
- Konle, Leonard, 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* 2.1, 1–22. [10.48694/jcls.3604](https://doi.org/10.48694/jcls.3604).
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (2019). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *CoRR abs/1910.13461*. <http://arxiv.org/abs/1910.13461>.
- Liu, Chao-Lin, Thomas J. Mazanec, and Jeffrey R. Tharsen (2018). "Exploring Chinese Poetry with Digital Assistance: Examples from Linguistic, Literary, and Historical Viewpoints". In: *Journal of Chinese Literature and Culture* 5.2, 276–321. <https://muse.jhu.edu/article/724617>.
- Luo, Ye (1965). *Zuiweng tanlu* 醉翁談錄. Taipei: Shijie shuju.
- Rao, Longsun (2023). "Mingdai xiaoshuo zhong de shiciqu jisheng shuo bianyi—yu Zhao Yishan xiansheng shangque 明代小说中的詩詞曲寄生說辨義—與趙義山先生商榷". In: *Zhongshan daxue xuebao (shehui kexue ban)* (中山大學學報(社會科學版)) 63.01, 29–37. [10.13471/j.cnki.jsysusse.2023.01.003](https://doi.org/10.13471/j.cnki.jsysusse.2023.01.003).
- Reiter, Nils, Marcus Willand, and Evelyn Gius (2019). "A Shared Task for the Digital Humanities Chapter 1: Introduction to Annotation, Narrative Levels and Shared Tasks". In: *Journal of Cultural Analytics* 4.3. <https://doi.org/10.22148/16.048>.
- Rouzer, Paul (2001). *Articulated Ladies: Gender and the Male Community in Early Chinese Texts*. Cambridge, MA: Harvard University Asia Center.
- Vierthaler, Paul (2016). "Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature". In: *Journal of Cultural Analytics* 1.1. <https://doi.org/10.7910/DVN/4ZVSKA>.
- Walsh, Melanie, Anna Preus, and Maria Antoniak (2024). "Sonnet or Not, Bot? Poetry Evaluation for Large Models and Datasets". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. <https://aclanthology.org/2024.findings-emnlp.914.pdf>.
- Williams, Adina, Nikita Nangia, and Samuel Bowman (2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

- New Orleans, Louisiana: Association for Computational Linguistics, 1112–1122. <http://aclweb.org/anthology/N18-1101>. 723
724
- Yeh, Chia-ying (2004). “Mantan Honglouloumeng zhong de shici 漫談紅樓夢中的詩詞”. In: *Shaanxi shifan daxue xuebao (zhexue shehui kexue ban)* 陝西師範大學學報: 哲學社會科學版 33.3, 58–64. doi:CNKI:SUN: SXSS.0.2004-03-010. 725
726
727
- Yu, Haoyang, Chang Gao, Xingsen Li, and Lingling Zhang (2024). “Ancient Chinese Poetry Collation Based on BERT”. In: *Procedia Computer Science* 242, 1171–1178. <https://doi.org/10.1016/j.procs.2024.08.179>. 728
729
730
- Zhang, Jiaxing, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen (2022). “Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence”. In: *CoRR abs/2209.02970*. <https://doi.org/10.48550/arXiv.2209.02970>. 731
732
733
734
735
736
737
- Zhang, Zhejun (2015). “Proven with a poem: Confucian Classic Origins and Functions of the Phrase in Novels and Traditional Operas (Youshi weizheng: Xiaoshuo, Xiqu taoyu de jingxue yuanyuan yu gongneng 有詩為證: 小說、戲曲套語的經學淵源與功能)”. In: *Journal of Macao Polytechnic University* 澳門理工學報 3, 154–161. 738
739
740
741
742
- Zhao, Yanwei (1966). *Yunlu manchao* 雲麓漫鈔. Beijing: Zhonghua shuju. 743
- Zhao, Yishan (2014). *Mingdai jisheng ciyu yanjiu* 明代寄生詞曲研究. Beijing: Shangwu yinshuguan. 744
745

Citation

Natalie M. Houston (2025). "Rhymefindr. An Historical Poetics Method for Identifying Rhymes in Nineteenth-Century English Poetry". In: CCLS2025 Conference Preprints 4 (1). 10.26083/tuprints-00030148

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-02-11

Keywords

computational poetics, historical poetics, rhyme, Rstats

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. This paper describes a new approach to rhyme identification that is grounded in the critical tradition of historical poetics. Rhymefindr comprises a set of R scripts designed to identify rhymes in nineteenth-century English poetry by operationalizing the rules presented in an 1824 edition of John Walker's *A Rhyming Dictionary*, one of the leading references on rhyme throughout the nineteenth century. By using an historical dictionary as a data source, Rhymefindr is sensitive to changes in pronunciation as well as changing theories about rhyme. As a corpus-independent method it can be used to identify rhymes in corpora of any size.

1. Introduction

Although poetic language is made up of words and sentences, and many text analysis methods can therefore be fruitfully applied to poetry, poetry also displays a number of distinctive formal features, including lineation, stanza patterns, meter, and rhyme, which can enrich text analysis and be the object of study themselves. Rhyme is of particular interest because it not only connects individual words through their shared sounds, but also connects poetic lines within stanzas. The patterns created by rhyme are thus integral to both the structure and the sound of poetry.

The predominant placement of rhyme words in modern English poetry is at the end of poetic lines. At the simplest level, rhyme can be defined as the relationship between "two syllables at line end . . . that have identical stressed vowels and subsequent phonemes but differ in initial consonant(s) if any are present – syllables that, in short, begin differently and end alike" (Greene et al. 2012, 1184). The availability of rhymes is determined in part by linguistic structures: highly inflected languages, for example, produce many more possible rhymes than are available in English.

But the very same reference work also notes that "the definition of what counts as rhyme is conventional and cultural: it expands and contracts from one national poetry, age, verse tradition, and genre to another" (Greene et al. 2012, 1185. Both rhyme practice and rhyme theory change throughout history: although perfect rhymes (cat, hat) have always been used, at different points in English poetic history, poets and critics have variously accepted or rejected other forms of rhyme. Some of these alternate forms include near rhymes, where the vowel sounds are close in sound, but not identical (soul, all); eye rhymes, where words are orthographically similar but pronounced differently (good, food); and the repetition of a given word. Additionally, historical changes in

pronunciation mean that some rhyme words used in the sixteenth century, for example, are no longer pronounced similarly (love, prove).

The development of computational methods for the analysis of literary texts have flourished in recent decades, spurred by the increasing availability of digitized text corpora. The ability to analyze features of poetic language across large corpora supports research in distant reading. As Franco Moretti suggests, shifting the focus of analysis to “units that are much smaller or much larger than the text” brings forth new kinds of knowledge about the literary “system in its entirety” (Moretti 2000, 57). Rhyme is a fundamental component of English poetry and understanding the connections it draws among words and ideas can contribute to research in many areas of poetics.

Previous work on the identification of rhyme words within English poetry include phonetic dictionary-based approaches, sometimes paired with text-to-speech generation, to identify words with matching final syllables (McCurdy et al. 2015, Heuser et al. 2018); an unsupervised expectation maximization algorithm to generate rhyme schemes (Reddy and Knight 2011; and a collocation-based method for identifying rhyme pairs in large corpora based on the frequency of their co-occurrence within individual poems (Plecháč 2018). However, these approaches do not directly address the variations in how rhyme has been defined and used throughout literary history, and particularly in the nineteenth century.

This paper describes a new approach to rhyme identification that is grounded in the critical tradition of historical poetics, which contextualizes the study of literary form in the theories and assumptions that poets and readers of past historical periods would have encountered and absorbed (Jarvis 2014, 98). By combining an historical poetics concept with computational criticism, this project makes it possible to model historical works of poetic theory and test them against collections of texts beyond the specific examples cited in those theories. This expands the work of historical poetics beyond the conception of its founders, a collaborative group of literary scholars focused on theoretical, not applied scholarship. Yopie Prins, for instance, says that “practical application is not the point of historical poetics” (Prins 2008, 233).

In contrast, this paper suggests that computational analysis provides a method for understanding nineteenth-century theories of rhyme through examining their relationship to actual historical poetic practice. By historicizing those rules as part of the analytic process, this project seeks to reconcile the multiple subjectivities of humanist knowledge with methods of quantitative analysis, responding to Johanna Drucker’s call for a “radical critique to return the humanistic tenets of constructed-ness and interpretation to the fore” of digital humanities scholarship (Drucker 2011, 1). This paper describes translating a specific historical theory about rhyme — one critic’s set of rules for understanding and evaluating rhyme — into code that can be processed by a machine. Although this iteration of the code only utilizes one historical dictionary, additional rhyme dictionaries will be added in the future for further comparison and analysis.

The structure of this paper is as follows: section 2 discusses the historical context of English rhyme and rhyme dictionaries in the nineteenth century. Section 3 discusses previous approaches to rhyme identification. Section 4 presents Rhymefindr, a set of R scripts designed to identify rhymes in nineteenth-century English poetry by operational-

izing the rules presented in an 1824 edition of John Walker's *A Rhyming Dictionary*, one of the leading references on rhyme throughout the nineteenth century (Walker [1775] 1824). By using Walker's dictionary as the basis for rhyme matching, this method is grounded in the theories of rhyme that were contemporary with the nineteenth-century poetry being analyzed. This method provides the opportunity to compare historical rhyme theory with historical rhyme practice by assessing how Walker's rules for rhyme compare to actual poetic usage. Section 5 presents an evaluation of this approach using gold standard data from the Chicago Rhyming Poetry Corpus (Reddy and Sonderegger 2011). Section 6 discusses the findings and Section 7 notes future enhancements planned for this project.

2. Nineteenth-Century Rhyme and Rhyme Dictionaries

Readers today often come to the study of rhyme with assumptions drawn from the aesthetic values of the twenty-first century. In an era that elevates free verse, structured verse forms are often seen as old-fashioned and contemporary critics and poets often assume rhyme constrains poetic expression (Cohen-Vrignaud 2015, 995). But in the nineteenth century, as Peter McDonald suggests, "the legitimacy of rhyme as a mode of writing was not in serious question . . . rhyme was a shared idiom, without which the lyric was all but unthinkable. To that extent, a rhymed poem did not really represent, in any useful sense, a decision to use rhyme." (McDonald 2012, 6–7). Almost all nineteenth-century English lyric poems are rhymed, and some dramatic and narrative poems use rhyme as well.¹ Rhyme was so prevalent in nineteenth-century verse that it would likely "feel to poets and readers as though it were something like a feature of the language itself" (Jarvis 2011, 36). This larger context of rhyme pairs that would have been familiar to many readers shaped a poet's choice of specific rhyming words, whether they were typical or unusual.

The central nineteenth-century critical debate about rhyme focused on whether imperfect rhymes were acceptable in poetry. Imperfect rhymes are today frequently termed near rhymes: words that are not pronounced exactly the same, as in a perfect rhyme, but are closely similar in sound. A second, sometimes overlapping, category of imperfect rhymes are eye rhymes: words whose endings are spelled the same, but pronounced differently. The history of English poetry from every century includes examples of words that do not sound the same, but are nonetheless interpreted as rhyme pairs because of the structural context in which they are placed. For example, because Alexander Pope's 1711 poem *An Essay in Criticism* is written in heroic couplets, the reader understands that "take" and "track" are presented as rhyme words in this passage about the necessity of poetic license:

If, where the rules not far enough extend,
(Since rules were made but to promote their end)
Some lucky LICENCE answers to the full
Th' intent propos'd, that licence is a rule.
Thus Pegasus, a nearer way to take,

1. For example, 95% of the 108,182 nineteenth-century poems in the Chadwyck-Healey English Poetry database are rhymed.

May boldly deviate from the common track. 110
 (Pope 1831, 8) 111
 112

As Pope suggests here, poets do not always follow the rules. An historical poetics 113
 approach to rhyme seeks to understand this variability both in how rhyme was used 114
 and how it was theorized. 115

Rhyme dictionaries, which became quite prevalent from the 18th century onward, offer 116
 a valuable resource for understanding the changing idiom of nineteenth-century rhyme 117
 and the history of rhyme theories. These dictionaries reflected poetic practice, often 118
 quoting examples of specific rhymes in the works of major poets, and they also prescribed 119
 particular rules and values around rhyme. The two most popular rhyme dictionaries 120
 for the eighteenth and nineteenth centuries, Edward Bysshe's 1714 *The British Parnassus* 121
 and John Walker's 1775 *A Rhyming Dictionary*, both draw on examples from canonical 122
 English poets to justify their inclusion of imperfect rhymes (Bysshe 1714, Walker [1775] 123
 1824). Walker, for example, claims that: "The delicate ears of a Pope or an Addison, 124
 would scarcely have acquiesced in the usage of imperfect rhymes, and sanctified them 125
 so often by their practice, if such rhymes had been really a blemish" (Walker [1775] 126
 1824, 635). But later in the century, when many competing rhyme dictionaries were 127
 published, Tom Hood would instruct the reader of his 1869 *The Rules of Rhyme* that 128
 "he must use such rhymes only as are perfect to the ear, when correctly pronounced" 129
 (Hood 1869, xii). Hood's emphasis on correct pronunciation reflects the association of 130
 pronunciation with social class in England. Like any reference work, rhyme dictionaries 131
 are not neutral: they often reveal how class and education shaped the aesthetic values 132
 associated with rhyme. Rhyme was frequently a touchpoint for larger cultural concerns 133
 during a time period in which increasing quantities of poetry were being published, not 134
 only in book form, but also in periodicals and newspapers. 135

3. Related Work 136

Brown et al. (2024) conduct a mapping review of 89 studies on rhyme identification 137
 algorithms, demonstrating increasing interest in this area of research since the 1960s. 138
 While the identification and analysis of rhymes has remained a continued thread of 139
 research, many recent studies have focused rhyme generation and have shifted from 140
 poetry to rap lyrics as the sample texts (Malmi et al. 2016, Popescu-Belis et al. 2023). 141
 This section highlights key topics in rhyme identification and analysis relevant to the 142
 historical poetics approach outlined in Rhymefindr. 143

3.1 Characteristics of Rhyme 144

As discussed previously, poetic rhyme is understood as the relationship between two or 145
 more words that terminate in syllables with similar sounds. Rhyme relationships exist 146
 within many kinds of natural language use, but within poetry and song lyrics, we find 147
 "foregrounded phonetic repetition" due to the placement of rhymes at the end of lines 148
 and within patterned stanza structures (Rickert 1978, 35). Similarly, Condit-Schultz 149
 suggests that rhyme should be understood as "a perceptual phenomenon which is 150
 evoked by phonemic parallelism" (Condit-Schultz 2016, 132). Poetic rhyme occurs 151

within particular structures and patterns that encourage listeners or readers to perceive certain words as rhyme words. Conversely, two words that share the same rhyme sound but are widely separated (ie, by 50 lines within a long poem) may not be perceived by the reader as a rhyme because of the temporal distance in perception. Thus studies of rhyme as a poetic phenomenon within specific texts may operationally define a window within which two lines may be considered to rhyme (Plecháč 2018, 86); studies of rhyme as a larger linguistic phenomenon may be interested in all words with shared endings, regardless of placement within the text.

In texts where a given rhyme sound is shared by more than two words, it is customary to understand those relationships as forming a rhyme chain (Joyce 1979, 129; Condit-Schultz 2016, 132). Although poetic lines are sequentially presented in a poem, and the proximal paired word would presumably have the most impact, the rhyme relationships accumulate, such that in a poem containing lines ending in "day," "stay," and "away," three rhyme pairs would be counted for the syllable "ay". Thus rhyme relationships can be considered as a graph structure. Joyce (1979) models the rhyme relationships within one long Middle English poem as a directed graph to maintain the sequential component of these chains. Sonderegger (2011) constructs an undirected rhyme graph for a large corpus of modern poetry and finds that its connected components reflect pronunciation, suggesting that rhymes could be used as supporting information for studies of historical pronunciation changes. Baley (2023) applies graph theory to the problem of evaluating inter-annotation agreement on rhymes in Chinese poetry.

3.2 Rhyme as a Stylistic Feature of Poetry

Many text analysis approaches treat rhyme as a stylistic feature of poetry. Kaplan and Blei (2007) include four different types of rhyme among the 89 features of poetic style they modeled to compare the work of American poets. Mayer et al. (2008) use rhyme along with text statistics to classify music lyrics by genre. Hirjee and Brown (2010) train a probabilistic model to identify rhymes as part of a stylistic study of rap lyrics. Kao and Jurafsky (2012) use a logistic regression model over 16 features of contemporary poetry, including rhyme, to distinguish between the work of amateur and professional poets. Pérez Pozo et al. (2022) compare a rule-based system, decision trees, and neural network approaches to classifying 46 defined stanza types in Spanish poetry based on verse length, rhyme structure, and rhyme pattern.

3.3 Pronunciation

Because rhyme relationships are constituted by similar word sounds, rhyme has been used as the basis of studies of historical pronunciation (Sonderegger 2011, List et al. 2017) and references on pronunciation are used as support for rhyme identification (Plamondon 2006).

Many researchers, like Kaplan and Blei (2007), Kao and Jurafsky (2012), and McCurdy et al. (2015) rely on the open-source machine-readable Carnegie Mellon University Pronouncing Dictionary, which provides phonetic transcriptions for 134,000 words in North American English (*The CMU Pronouncing Dictionary* n.d.). This dictionary is widely available but was not designed for literary analysis. Its vocabulary is also

skewed towards contemporary English. McCurdy notes the limitations of the CMU dictionary's vocabulary and extends it by use of letter-to-sound rules and syllable segmentation algorithms (McCurdy et al. 2015, 17). Popescu-Belis et al. (2023) uses the CMU dictionary to construct synthetic rhyme data to fine tune a GPT-2 model to generate rhymed verse. Other researchers have incorporated text-to-speech technologies into rhyme identification workflows (Heuser et al. 2018, Plecháč 2018).

3.4 Rhyme identification

Because rhyme describes a relationship, the task of rhyme identification has been defined either as the discovery of stanzaic rhyme schemes (ie, ABAB, ABBA) or as the discovery of rhyme pairs.

Noting the limitations of using phonetic transcription for historical texts, Reddy and Knight (2011) proposed identifying rhyme schemes through an unsupervised expectation maximization algorithm trained on a corpus of 93,014 lines of English poetry from 1450-1950 and 26,543 lines of French poetry from 1450-1650 with rhyme annotations. This approach starts with a predefined set of 462 possible stanza rhyme schemes drawn from the training corpus. The algorithm builds on the intuition that rhyming words within a given stanza are also likely to co-occur within a large corpus. Adding a measure to account for orthographic similarity improved the performance of their model, as did using a hidden Markov model to condition each stanza on the previous one in the poem. Other related approaches to rhyme scheme identification include Addanki and Wu (2013), who use a hidden Markov model with nine rhyme patterns for an unsupervised approach to detecting rhyme schemes in rap lyrics.

Building on the work of Reddy and Knight, but noting the limitations of their stanza-based approach, Plecháč (2018) focuses on the discovery of rhyme pairs in large poetic corpora. The model is first trained with the collocation of rhyme word pairs throughout the corpus. Then text-to-speech corpus transcription is used to obtain the phonetic elements of the rhyme words and learn the "rhyme probabilities between particular vowels (syllable peaks) and consonant clusters," with an added probability for orthographic similarity (Plecháč 2018, 84). Plecháč shows that this collocation approach generally outperforms Reddy and Knight's maximization approach on their corpus of English and French poetry and on a corpus of 2.5 million lines of Czech poetry (Plecháč and Kolár 2015). A recent supervised approach to the identification of rhyme pairs uses Siamese Recurrent Networks to identify rhyme pairs in German, English, and French poetry Haider and Kuhn 2018.

One challenge in identifying rhymes in historical texts are changes in how rhyme was defined and used. An historical poetics approach to rhyme does not assume that rhyme relationships are static. Using specific historical guides to rhyme as the basis for rhyme identification allows for the discovery of rhymes that may not be identified by phonetic matching with contemporary dictionaries, particularly given the variability of national pronunciation differences and historical changes in pronunciation. Rhymefindr has been designed to support stylistic analysis by identifying features related to rhyme words and rhyme syllables. As a rule-based approach, Rhymefindr does not require a large training corpus, as do the expectation maximization and collocation approaches.

4. Rhymefindr

237

The Rhymefindr approach to rhyme identification presented here is grounded in rules of rhyme that were relevant for poets and readers in the nineteenth century. Specifically, this approach utilizes John Walker's *A Rhyming Dictionary*, which was highly influential throughout the nineteenth century, particularly in its documentation of imperfect rhymes that were acceptable in English verse. Walker's dictionary also offers a window onto historical British pronunciation of English words that is valuable for analyzing rhyme.

Many nineteenth-century poets deliberately experimented with rhyme and other formal structures in their poetry. Rhymefindr does not utilize knowledge of a particular literary corpus or of specific stanza rhyme patterns, so it is not limited to finding rhyme only in works that conform to the literary tradition, or in works written by familiar canonical poets. Although Walker's dictionary includes quotations from poetry to support his views on near rhymes as compared with perfect ones, the rhyme data contained in the dictionary's entries are completely distinct from any poetic tradition. In arguing for distant reading as an alternative to close reading, Franco Moretti argued that traditional literary scholarship "necessarily depends on an extremely small canon. . . . you invest so much in individual texts only if you think that very few of them really matter" (Moretti 2000, 58). As a corpus-independent method, Rhymefindr supports research in non-canonical poetics and can be used to identify rhymes in corpora of any size, thereby contributing to a wide range of research situations.

Rhymefindr currently comprises a key-value table created from an historical rhyme dictionary; an endword extraction script; and a rhyme identification script. The *find_rhymes* script performs a series of attempts to match the rhyming words within a poem based on the different kinds of rhyme expressed in that historical dictionary. Although the current iteration of the project utilizes only one dictionary, future versions will incorporate additional rhyme dictionaries to enable comparative analysis of rhyme theories as well as rhyme practice in the nineteenth century.

4.1 Dictionary Data

264

John Walker's *A Rhyming Dictionary; Answering, at the Same Time, the Purposes of Spelling and Pronouncing the English Language, on a Plan not Hitherto Attempted* was selected as the data source for the dictionary component of this project because it was one of the most popular rhyme dictionaries throughout the nineteenth century. (Byron and Tennyson both owned copies, as did many other poets.) It was first published in 1775 and reprinted and expanded in both British and American editions throughout the nineteenth century. A Google-digitized file created from a Harvard University copy of the 1824 edition published in London by W. Baynes and Son was used to prepare the data for this project.

Walker's dictionary is structured in two parts, both of which focus on the endings of English words. Walker argued that his work was more than a "mere rhyming dictionary" or "resource for poetasters"; rather, his "dictionary of terminations subservient to the art of spelling and pronouncing" would provide a new perspective on the structures of the English language: "In this arrangement of the language, we easily discover its

idiomatic structure, and find its several parts fall into their proper classes, and almost every word as much distinguished by its termination as by its sense" (Walker [1775] 1824, v–vi). The first part of the volume, titled a "Syllabic Dictionary," lists English words with brief definitions, as one might find in other dictionaries. However, Walker lists these words according to reverse-spelling order ("s" in these entries indicates nouns, or substantives):

Elf A fairy; a devil, s. 285
 Delf A mine; quarry, earthen ware, s. 286
 Shelf A board to lay things on; a sand bank in the sea; hard coat of earth under the mould, s. 288
 (Walker [1775] 1824, 186) 289

Later editors changed the title of the dictionary to make this innovation clear: *The rhyming dictionary of the English language: in which the whole language is arranged according to its terminations* (Walker [1775] 1894). Walker argued that presenting its contents in reverse-spelling order would help teach the rules for English spelling, which he calls "an insuperable difficulty for foreigners" and an "eternal source of dispute and perplexity for ourselves" (Walker [1775] 1824, vi). This reverse-spelling presentation makes groups of rhyming words readily visible on the page.

But Walker also recognized that readers accustomed to other rhyme dictionaries would want an easier way of finding rhymes. So the second part of the volume consists of an "Index of Perfect and Allowable Rhymes" containing entries for the final syllables of English words, arranged alphabetically by their first letters (elf, elk, elm, elp) as the editors of previous rhyme dictionaries had done (Poole 1657, Bysshe 1714). What distinguished Walker's index from those earlier dictionaries was his decision to document and include imperfect rhymes, which he renamed "allowable" rhymes, documented with "authorities for their usage from our best poets" (Walker [1775] 1824, 635). By renaming what earlier critics had called "imperfect" rhymes as "allowable," Walker emphasizes the capacious quality of his approach to rhyme. Walker's generous definition of allowable rhyme became the standard theory of rhyme for many nineteenth-century readers and poets, even after the resurgence of stricter definitions of perfect rhyme in competing rhyme dictionaries published in the 1860s. Walker's "Index of Perfect and Allowable Rhymes" serves as the basis for the dictionary portion of this project.

Entries in the "Index of Perfect and Allowable Rhymes" begin with a rhyming syllable, followed by a list of words that include the key syllable, or that rhyme perfectly with it. Some of these lists are ostensibly comprehensive, but others end with an "etc" suggesting that the reader would be able to come up with additional rhyming words. After the perfect rhymes, Walker occasionally notes what he terms "nearly perfect" rhymes, and then lists the allowable rhymes:

EM 317
 Gem, hem, stem, them, diadem, stratagem, &c. Perfect rhymes, condemn, contemn, &c. Allowable rhymes, lame, tame, &c. team, seam, theme, phlegm, &c. 319
 (Walker [1775] 1824, 655) 321

Where the allowable rhymes are especially controversial, Walker provides references to

specific rules in his Preface and quotations from the works of English poets who use the rhyme. Within the entries there are also a number of cross-references: entries for some syllables consist entirely of a cross reference to a homophone, and cross references are also included within the lists of perfect or allowable rhyme syllables.

4.2 Creation of a key-value dictionary

A key-value dictionary was created to represent Walker's index of rhymes, with each rhyme syllable that heads an entry in the dictionary defined as a key and matched with the values listed in Walker for perfect rhyme syllables, perfect rhyme words, allowable rhyme syllables, and allowable rhyme words. The small number of words Walker labels "nearly perfect" were included with the perfect rhymes.

Although the intention behind this project is to create an historically sensitive rule base for rhyme from Walker's rhyme dictionary, that historical document contained some inconsistencies in its presentation of data, so in some instances strict fidelity to Walker's text had to be modified in order to make the key-value dictionary fully operational. For example, many cross-referenced rhyme syllables are listed under both headings in Walker, but in some cases only one is cross-referenced: the entry for EIGHT says "see ATE" but the entry for ATE does not point to EIGHT. To standardize the data for this project, all cross-referenced rhyme syllables were duplicated for both key entries. The other modification to the historical data obtained from Walker's dictionary was to add modern spellings for one-syllable past participles (adding missed where Walker lists miss'd) to make the key-value dictionary applicable to a wider range of nineteenth-century texts.

4.3 Endword extraction script

The *get_endwords* R script is included in the project repository to facilitate the extraction of endwords from a directory containing plain text files of poems. Because this script is designed for the analysis of rhyme, hyphens are removed and hyphenated words are put together. Thus the common nineteenth-century spelling "to-day" becomes "today" rather than "to day." Although this decision produces some odd-looking word forms, like "garretroom," overall it produces more accurate results in the rhyme analysis stage.

In addition to the vectors of endwords for each poem that are required for the rhyme discovery script, the *get_endwords* script also outputs several poetic features useful for exploratory text analysis, including the number of stanzas and lines in the poem.

4.4 Rhyme identification script

The *find_rhymes* R script is designed to work with an input csv containing a text id and a character vector of endwords for each poem. The final syllable of each endword is extracted with regular expressions based on the orthographic principles of English and is used as the basis for a series of lookups in the key-value table created from Walker's dictionary. For each endword, the script looks first to match it with a perfect rhyme syllable or rhyme word in Walker; if one isn't found, it checks the allowable rhyme syllables and words listed in Walker. As rhyme matches are found, a vector indicating the rhyme sequence is constructed. Capital letters are conventionally used for this purpose

in the study of poetics, and are applied to all of the endwords in the poem, including any non-rhyming lines. A final lookup checks for orthographic matches among the rhyme syllables in the poem that have not been matched to rhymes in Walker’s dictionary; however, these matches are currently limited to identical matches, or perfect rhymes.

It should be noted that all of the entries in Walker’s rhyme dictionary are for single rhyme syllables. The majority of rhymes used in nineteenth-century English poetry (and indeed, English language poetry from any period) are monosyllabic rhymes, in large part because of the predominance of iambic meter in both natural English speech and especially in English poetry. An iambic metrical foot consists of an unstressed syllable followed by a stressed syllable; thus most lines of iambic poetry end with a stressed syllable, which is the focus of the rhyme. Although the *find_rhymes* script thus only identifies single syllable rhymes, many bisyllabic rhyme pairs can also be identified through this approach.

After all the rhymes have been identified, the ratio of unique rhymes to the total number of rhymes in the poem is calculated to assess the likelihood of whether the poem is rhymed or not, using the first 75 lines of longer poems and the entire text for poems with fewer than 75 lines. For nineteenth-century English poetry, an operationally successful range of ratios was defined as: rhymed poems have a ratio smaller than .70; ratios of possibly rhymed poems fall between .70 and .86; and unrhymed poems have a ratio greater than .86. These ranges account for the likelihood that even ostensibly unrhymed poems, like long poems written in blank verse, will contain some rhymes across many hundreds of lines.

For each poem, the script outputs the rhyme scheme, a categorical indicator of the likelihood of the poem being rhymed, and a vector indicating which of the rhymes are perfect rhymes according to Walker’s dictionary.

5. Evaluation

Rhymefindr was tested using the gold standard annotated data for English poetry in the Chicago Rhyming Poetry Corpus (Reddy and Sonderegger 2011) which was the same English corpus used by Reddy and Knight (2011) and Plecháč (2018). Because rhyme constitutes a relationship between two or more words, different approaches to evaluating rhyme discovery have been applied in previous work and are used here for comparison.

5.1 Gold standard data

The English language component of the Chicago Rhyming Poetry Corpus contains annotated rhyme data for 11,613 stanzas containing 93,014 lines of poetry by 32 poets (Reddy and Sonderegger 2011). The gold standard data files are separated into five 100-year spans from 1450-1950. These files contain an entry for each stanza in the corpus poems that consists of its end words and a numeric sequence indicating its rhymes.

Because Rhymefindr is based on a rhyme dictionary popular in the nineteenth century, it is relevant to consider the representation of nineteenth-century poets in the gold standard data subgroups for 1750-1850 and 1850-1950. Although no information is pro-

vided in the corpus repository about how poets or poems were selected for the rhyme corpus, all of the poets included in the English selections in the 1850-1950 chronological period overlap with the list provided in Sonderegger (2011), which describes compiling a rhyme corpus of “poetry written by English authors around 1900” Sonderegger 2011, 657. As seen in Table 1, which arranges the list of poets by date of birth, the Chicago Rhyming Poetry Corpus includes poets who were mostly active during the Romantic and Edwardian eras, skipping over poets from the Victorian period (1837-1900).

Chicago Rhyming Poetry Corpus

Sub-corpus	Poet	Lifespan
1750-1850	Oliver Goldsmith	1728-1774
1750-1850	Charlotte Turner Smith	1749-1806
1750-1850	William Wordsworth	1770-1850
1750-1850	Samuel Taylor Coleridge	1772-1834
1750-1850	Lord Byron (George Gordon)	1788-1824
1750-1850	Percy Bysshe Shelley	1792-1822
1850-1950	A. E. Housman	1859-1936
1850-1950	Thomas Crosland	1865-1924
1850-1950	Rudyard Kipling	1865-1936
1850-1950	G. K. Chesterton	1874-1936
1850-1950	Edward Thomas	1878-1917
1850-1950	Rupert Brooke	1887-1915

Table 1: Poets included in the 1750-1850 and 1850-1950 sub-corpora in Reddy and Sonderegger (2011)

In the process of working with the Chicago Rhyming Poetry Corpus, 102 entries in the published gold standard files were found to have incomplete data and were discarded from the evaluation; an obvious typographical error was corrected in one additional entry.² This resulted in a total of 11,511 stanzas, distributed over the five chronological sub-groups as shown in Table 2.

Gold standard data files

Sub-corpus	Stanzas	Lines
1415_pgold	197	1250
1516_pgold	3786	35485
1617_pgold	2141	19683
1718_pgold	2546	20546
1819_pgold	2843	15408
totals	11513	92372

Table 2: Number of stanzas and lines in the gold standard data files used in the evaluation

5.2 Rhyme scheme evaluation metrics

As described in section 3, Reddy and Knight (2011)’s expectation maximization (EM) approach identifies rhyme schemes in separate stanzas of poetic texts. They define accuracy at the scheme level, indicating that a discovered rhyme scheme either does or does not match the gold standard rhyme scheme exactly. Table 3 shows Rhymefindr’s accuracy in discovering rhyme schemes according to Reddy and Knight’s definition

2. Details are available at: https://github.com/nmhouston/rf_eval.

and compares it to the performance of two of their models: their EM approach for separate stanzas with an initialization for orthographic similarity, and their hidden Markov model (HMM) approach which conditions for stanza dependencies (Reddy and Knight 2011, 81).

Rhyme scheme accuracy %			
	RK EM with orthographic	RK HMM	Rhymefindr
1450-1550	69.04	74.31	61.93
1550-1650	71.98	79.17	53.2
1650-1750	89.54	91.23	51.24
1750-1850	33.62	49.11	57.66
1850-1950	54.05	58.95	70.56

Table 3: Rhyme scheme accuracy percentage for Rhymefindr compared with Reddy and Knight’s EM and HMM approaches Reddy and Knight 2011, 81

Rhymefindr performs better according to this measure of rhyme scheme accuracy than the EM or HMM approaches for the chronological periods 1750-1850 and 1850-1950, which are the time periods for which Walker’s dictionary (first published in 1775) would be expected to have the strongest relevance. Notably, Reddy and Knight’s EM and HMM approaches perform significantly worse on poetry after 1750 than on poetry from the earlier subgroups. This may be due to the greater variety of stanza structures in later poetry or to the makeup of the training set data.

Reddy and Knight (2011) also calculate precision and recall at the stanza level: precision as the number of rhyming words within each stanza that are correctly discovered by the algorithm divided by the number of rhyming words output for the stanza, and recall as the number of correctly discovered rhyming words within the stanza divided by the number of rhyming words in the gold standard for the stanza. Words without rhyme pairs in a stanza are ignored. They total the precision and recall scores for all stanzas before calculating the F score for each chronological sub-group. Table 4 compares Rhymefindr’s precision and recall for rhyme schemes calculated in this way with Reddy and Knight’s EM approach with orthographic similarity and their HMM approach (Reddy and Knight 2011, 81).

Rhyme scheme F score			
	RK EM with orthographic	RK HMM	Rhymefindr
1450-1550	0.82	0.86	0.88
1550-1650	0.88	0.9	0.87
1650-1750	0.96	0.97	0.84
1750-1850	0.7	0.82	0.88
1850-1950	0.84	0.9	0.87

Table 4: Rhyme scheme F scores for Rhymefindr compared with Reddy and Knight’s EM and HMM approaches (Reddy and Knight 2011, 81)

Rhymefindr’s performance on poetry after 1750 improves on Reddy and Knight’s EM approach and is close to the performance of their HMM approach.

5.3 Rhyme pair evaluation metrics

As discussed earlier, Plecháč (2018) defines the task as the discovery of rhyme pairs, rather than stanza rhyme schemes, and uses a collocation approach to train a model with the phonetic probabilities of rhyme. Plecháč does not provide an accuracy metric in the evaluation, focusing instead on precision and recall, calculated with the total numbers of rhyme pairs in the output and gold standard. Table 5 evaluates Rhymefindr’s performance using this approach to precision and recall and compares it to the results of Plecháč’s collocation approach (Plecháč 2018, 89).

Rhyme pair F scores		
	Plecháč collocation	Rhymefindr
1450-1550	0.87	0.9
1550-1650	0.91	0.84
1650-1750	0.92	0.81
1750-1850	0.92	0.9
1850-1950	0.93	0.87

Table 5: Rhyme pair F scores for Rhymefindr compared with Plecháč’s collocation approach (Plecháč 2018, 89)

Rhymefindr’s performance according to this metric is notably better for poetry after 1750 than for 1550-1750, and while it does not match the performance of Plecháč’s collocation approach, its F scores are still good.

6. Discussion

As noted earlier, the definitions of acceptable poetic rhyme change over time and can be shaped by many factors, including changes in pronunciation and conventions of usage. Historical poetics emphasizes the importance of understanding that complexity. The question of whether a given pair of words rhyme may not always be possible to answer with a strict logical yes/no; sometimes the answer depends upon the historical period, expected national or regional pronunciation, and the literary context surrounding the words. Inspection of the rhyme vectors from the evaluation corpus with poor accuracy scores reveals three main causes for rhyme misclassification according to the gold standard data: plural nouns, historical pronunciation differences, and near rhymes.

Walker’s dictionary is inconsistent in its presentation of plural nouns, because he expected readers to be able to generalize from the singular noun to its plural. For example, the word “eyes” does not appear anywhere in Walker’s entries, but of course is very frequently used in nineteenth-century poetry. The Reddy and Sonderegger (2011) corpus includes rhymes between eyes/wise and eyes/dies, neither of which are marked as rhymes by Rhymefindr.

Pronunciation differences between nineteenth-century British English and contemporary English are another source for mismatches between the Reddy and Sonderegger (2011) annotations and the rhymes identified by Rhymefindr. For example, their gold standard data defines “anew/you” as a rhyme pair, which according to Walker (and most British pronunciation) have completely different vowel sounds. Walker’s dictionary was selected because it provides a guide to historical British pronunciation, which

was considered important for an historical poetics project focused on the nineteenth century. 479 480

Walker's inclusion of allowable, or near rhymes, is another source of mismatches with the gold standard data. For example, Walker says that ale/ell syllables are allowable rhymes, so Rhymefindr tags vale/hell as a rhyme, where the gold standard data does not. Future iterations of the project will give the user an option of selecting perfect and allowable rhymes, or only perfect rhymes, when making identifications, just as a reader of Walker's dictionary could have chosen for their own purposes. 481 482 483 484 485 486

Unfortunately, Reddy and Sonderegger do not provide documentation of their approach to creating the rhyme annotations in the Chicago Rhyming Poetry Corpus, and how they handled different kinds of ambiguous or non-perfect rhymes. Understanding historical rhyme usage requires taking into account the various ways in which our contemporary sense of rhyme may not align with historical poetic practice. By keying rhyme identification to the constraints of particular historical dictionaries, Rhymefindr reminds users that identifying and describing rhyme is always an act of critical interpretation. 487 488 489 490 491 492 493

7. Future Work 494

With the framework of this historical dictionary-based method in place, other dictionaries will be added to expand the capacities of Rhymefindr as a rhyme identification tool and to enhance the utility of this project for comparative historical poetics. Several different rhyme dictionaries were published in the nineteenth century, including J. E. Carpenter's *A Handbook of Poetry* (1868); Tom Hood's *The Rules of Rhyme* (1869); Samuel W. Barnum's *A Vocabulary of English Rhymes, Arranged on a New Plan* (1876); and Andrew Loring's *The Rhymers's Lexicon* (1905). Operationalizing multiple dictionaries would contribute not only to the computational analysis of rhyme, but would also enable new experiments that could test the application of different theories of rhyme over a large poetry corpus. 495 496 497 498 499 500 501 502 503 504

8. Data Availability 505

The Walker dictionary data needed to run Rhymefindr can be found here: <https://github.com/nmhouston/Rhymefindr>. Gold standard rhyme data from the Chicago Rhyming Poetry Corpus (Reddy and Sonderegger 2011) and outputs from the evaluation scripts can be found here: https://github.com/nmhouston/rf_eval. 506 507 508 509

9. Software Availability 510

The Rhymefindr scripts can be found here: <https://github.com/nmhouston/Rhymefindr>. The evaluation scripts can be found here: https://github.com/nmhouston/rf_eval. 511 512 513


References


- Addanki, Karteek and Dekai Wu (2013). “Unsupervised rhyme scheme identification in hip hop lyrics using hidden Markov models”. In: *International conference on statistical language and speech processing*, 39–50.
- Baley, Julien (2023). “Evaluating rhyme annotations for large corpora: Metrics and data”. In: *Cahiers de Linguistique Asie Orientale* 52.2, 137–162.
- Brown, Daniel G., Rebecca Hutchinson, and Carolyn E. Lamb (2024). *A systematic mapping review of algorithms for the detection of rhymes, from early digital humanities projects to the rise of large language models*. <https://uwspace.uwaterloo.ca/bitstream/handle/10012/20723/rhymesysrev.pdf?sequence=1>.
- Bysshe, Edward (1714). *The British Parnassus: or, a compleat common-place-book of English poetry: ... To which is prefix'd, A dictionary of rhymes*. J. Nutt.
- The CMU Pronouncing Dictionary (n.d.). <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- Cohen-Vrignaud, Gerard (2015). “Rhyme’s Crimes”. In: *ELH* 82.3, 987–1012.
- Condit-Schultz, Nathaniel (2016). “MCFlow: A digital corpus of rap transcriptions”. In: *Empirical Musicology Review* 11.2, 124–147.
- Drucker, Johanna (2011). “Humanities approaches to graphical display”. In: *Digital Humanities Quarterly* 5.1, 1–21.
- Greene, Roland, Stephen Cushman, Clare Cavanagh, Jahan Ramazani, and Paul Rouzer (2012). *The Princeton encyclopedia of poetry and poetics*. Princeton University Press.
- Haider, Thomas and Jonas Kuhn (2018). “Supervised rhyme detection with Siamese recurrent networks”. In: *Proceedings of the second joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, 81–86.
- Heuser, Ryan, J.D. Porter, Jonathan Sensenbaugh, Justin Tackett, Mark Algee-Hewitt, and Maria Kraxenberger (2018). *Poesy*. <https://github.com/quadrismegistus/poesy>.
- Hirjee, Hussein and Daniel G. Brown (2010). “Using Automated Rhyme Detection to Characterize Rhyming Style in Rap Music”. In: *Empirical Musicology Review* 5.4, 121–145.
- Hood, Tom (1869). *The Rules of Rhyme: A Guide to English Versification. With a Compendious Dictionary of Rhymes, an Examination of Classical Measures, and Comments upon Burlesque, Comic Verse, and Song-Writing*. James Hogg & Son.
- Jarvis, Simon (2011). “Why rhyme pleases”. In: *Thinking Verse* 1.2.
- (2014). “What is historical poetics?” In: *Theory Aside*. Ed. by Jason Potts and Daniel Stout. Duke University Press, 97–116.
- Joyce, James (1979). “Re-weaving the word-web: graph theory and rhymes”. In: *Annual Meeting of the Berkeley Linguistics Society*, 129–141.
- Kao, Justine and Dan Jurafsky (2012). “A computational analysis of style, affect, and imagery in contemporary poetry”. In: *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*. NAACL-HLT, 8–17.
- Kaplan, David M and David M Blei (2007). “A computational approach to style in American poetry”. In: *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE, 553–558.

- List, Johann-Mattis, Jananan Sylvestre Pathmanathan, Nathan W. Hill, Eric Baptiste, 558
and Philippe Lopez (2017). “Vowel purity and rhyme evidence in Old Chinese 559
reconstruction”. In: *Lingua Sinica* 3, 1–17. 560
- Malmi, Eric, Pyry Takala, Hannu Toivonen, Tapani Raiko, and Aristides Gionis (2016). 561
“Dopelearning: A computational approach to rap lyrics generation”. In: *Proceedings* 562
of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data 563
Mining, 195–204. 564
- Mayer, Rudolf, Robert Neumayer, and Andreas Rauber (2008). “Rhyme and Style Fea- 565
tures for Musical Genre Classification by Song Lyrics”. In: *ISMIR 2008, Proceedings of* 566
the 9th International Conference on Music Information Retrieval. 567
- McCurdy, Nina, Vivek Srikumar, and Miriah Meyer (2015). “Rhymedesign: A tool 568
for analyzing sonic devices in poetry”. In: *Proceedings of the Fourth Workshop on* 569
Computational Linguistics for Literature. Association for Computational Linguistics, 12– 570
22. 571
- McDonald, Peter (2012). *Sound Intentions: The Workings of Rhyme in Nineteenth-Century* 572
Poetry. Oxford University Press. 573
- Moretti, Franco (2000). “Conjectures on world literature”. In: *New left review* 2.1, 54–68. 574
- Pérez Pozo, Álvaro, Javier de la Rosa, Salvador Ros, Elena González-Blanco, Laura 575
Hernández, and Mirella De Sisto (2022). “A bridge too far for artificial intelligence?: 576
Automatic classification of stanzas in Spanish poetry”. In: *Journal of the Association* 577
for Information Science and Technology 73.2, 258–267. 578
- Plamondon, Marc R. (2006). “Virtual verse analysis: Analysing patterns in poetry”. In: 579
Literary and Linguistic Computing 21.suppl, 127–141. 580
- Plecháč, Petr (2018). “A Collocation-Driven Method of Discovering Rhymes (in Czech, 581
English, and French Poetry)”. In: *Taming the Corpus: From Inflection and Lexis to* 582
Interpretation. Ed. by Masako Fidler and Václav Cvrček. Springer International Pub- 583
lishing, 79–95. [10.1007/978-3-319-98017-1_5](https://doi.org/10.1007/978-3-319-98017-1_5). 584
- Plecháč, Petr and Robert Kolár (2015). “The corpus of Czech verse”. In: *Studia metrica et* 585
poetica 2.1, 107–118. 586
- Poole, Josua (1657). *The English Parnassus: Or, A Helpe to English Poesie*. Tho. Johnson. 587
- Pope, Alexander (1831). *The Poetical Works of Alexander Pope*. Vol. 2. London: William 588
Pickering. 589
- Popescu-Belis, Andrei, Alex R. Atrio, Bastien Bernath, Étienne Boisson, Teo Ferrari, 590
Xavier Theimer-Lienhardt, and Giorgos Vernikos (2023). “GPoeT: a language model 591
trained for rhyme generation on synthetic data”. In: *Proceedings of the 7th Joint* 592
SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, 593
Humanities and Literature. Association for Computational Linguistics. 594
- Prins, Yopie (2008). “Historical poetics, dysprosody, and the science of English verse”. 595
In: *PMLA* 123.1, 229–234. 596
- Reddy, Sravana and Kevin Knight (2011). “Unsupervised discovery of rhyme schemes”. 597
In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:* 598
Human Language Technologies, 77–82. 599
- Reddy, Sravana and Morgan Sonderegger (2011). *The Chicago Rhyming Poetry Corpus*. 600
<https://github.com/sravanareddy/rhymedata/blob/master/README.txt>. 601
- Rickert, William E. (1978). “Rhyme Terms”. In: *Style* 12.1, 35–46. 602
- Sonderegger, Morgan (2011). “Applications of graph theory to an English rhyming 603
corpus”. In: *Computer Speech Language* 25.3, 655–678. 604

- Walker, John [1775] (1824). *A Rhyming Dictionary; Answering, at the Same Time, the Pur-* 605
poses of Spelling and Pronouncing the English Language, on a Plan not Hitherto Attempted. 606
 William Baynes and Son. 607
- [1775] (1894). *The rhyming dictionary of the English language: in which the whole language* 608
is arranged according to its terminations. Ed. by John Longmuir. Rev. and enlarged. 609
 George Routledge and Sons. 610

Opening Worlds: Narrative Beginnings and the Role of Setting

Katrin Rohrbacher¹ 

1. Department of Digital Humanities and Social Studies (DHSS), FAU Erlangen-Nürnberg , Erlangen, Germany.

Citation

Katrin Rohrbacher (2025).
“Opening Worlds: Narrative
Beginnings and the Role of Set-
ting”. In: *CCLS2025 Conference
Preprints 4* (1). [10.26083/tuprints-00030149](https://doi.org/10.26083/tuprints-00030149)

Date published 2025-06-17

Date accepted 2025-04-18

Date received 2025-02-08

Keywords

narrative beginnings, German
fiction, space, machine learning,
computational narratology

License

CC BY 4.0 

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 4th Annual
Conference of Computational
Literary Studies at Krakow,
Poland, in July 2025. Please
check <https://jcls.io> for the
final journal version.

Abstract. Beginnings are central to narrative structure, shaping the reader’s engagement with the storyworld. This study examines the role of setting in narrative openings, using a large-scale dataset of German-language fiction and non-fiction. Drawing on Herman’s concept of “worldmaking” and Hoffmann’s phenomenological model of space, we classify settings into four types: *Aktionssraum* (action space), *gestimmter Raum* (space reflecting mood and atmosphere), *Anschauungsraum* (field of vision), and “descriptive space”. Using a multiclass text classification model, we analyze their distribution across narrative time, historical time, and genre focusing specifically on their prominence in story openings. Our findings show that openings tend to prioritize establishing what the depicted world feels and looks like, emphasizing affect and visual description before shifting toward movement and the mobilization of setting through dynamic character interaction. Comparative and historical analyses reveal that these trends are unique to fiction and have increased over time. By leveraging computational models, we provide an empirical foundation for understanding how narrative beginnings structure fictional worlds.

1. Introduction

In his book on Narrative Theory, David Herman highlights the importance of “narrative beginnings,” arguing that “story openings prompt interpreters to take up residence [...] in the world being evoked by a given text” (Herman 2009, 14). Beginnings are “meant to be noticed,” as they introduce and establish the essential features of the narrative, serving as “points of entry into the narrative and [orienting] the reading” (Mikkonen 2020, 4). They mark the moment when readers are introduced to the narrative world and its key characteristics are first established.

This paper proposes a novel method to approximate beginnings in the context of narrative worldmaking by examining the concept of “setting” and its role in the plot. Although much has been written about the significance of textual openings (see Polaschegg 2020, Romagnolo 2015, Richardson 2008, Said 1968, Miller 1965), only a few studies to date have attempted large-scale quantitative analyses that include a focus on beginnings (see Piper et al. 2023, Boyd et al. 2020). However, little research has specifically investigated the importance of setting in relation to narrative beginnings.

Herman views the notion of “worldmaking” as central to storytelling itself, identifying it as one of the four fundamental elements of narrative, emphasizing its particular

importance in story openings (Herman 2009, 112f). To capture the concept of “setting” (or worldmaking in Herman’s terms), we draw on a framework proposed by the German literary theorist Gerhard Hoffmann (1978). This model accounts for the “lived space” in stories, that is, space as experienced through the perceptions of characters. Rather than treating setting as a static background for a story’s plot or simply distinguishing between static description and dynamic events, as past critics often have, Hoffmann’s model allows for an exploration of the “dynamic” interplay and “coexisting simultaneity” of setting, characters, and events (Hones 2011). He divides setting into three distinct categories, each corresponding to different ways characters perceive their environment, both built and natural, within a story. These categories are *Aktionsraum* (“action space”), *gestimmter Raum* (“space reflecting mood and atmosphere”), and *Anschaungsraum* (“field of vision”).

Research in cognitive psychology has emphasized the importance of “setting” in the telling of oral narratives. Studies suggest that narratives commonly begin with an “orientation,” introducing the participants (or characters, in this case) along with “the time, the place, and the initial behavior” — answering the questions of “who,” “when,” and “where” (Labov and Waletzky 1997, 4). In this way, setting functions as a key component of narrative openings by grounding the story in a specific context and aiding readers’ orientation.

Gustav Freytag (1895) argues that exposition, or the establishment of a story’s setting, serves as the basis for its overall structure. Testing this theory in their study on “narrative arcs” in English-language narratives, Boyd et al. (2020) found that the use of “staging-related words” (in their case, a higher percentage of articles and prepositions) is indeed most frequent at the beginning of a narrative and then decreases as the story progresses (4). While their article relies solely on surface-level textual features, identifying linguistic cues to measure how story openings differ linguistically from the rest of the narrative, the model used in this paper introduces a more nuanced and interpretive dimension to the analysis of setting in narratives. Rather than focusing on syntactic markers as proxies for setting (or “staging” in their study), this study examines setting based on its narrative role and interaction with characters and events. This approach provides deeper insights into how and why these openings are distinct.

To study this, we apply a multiclass text classification model fine-tuned on Hoffmann’s categories to German fiction and non-fiction works, using the model and method developed in Rohrbacher (forthcoming). This allows us to estimate, on a large scale, how different types of settings change over narrative time. Additionally, by leveraging metadata on “genre” and “sub-genre” from our dataset, we test Herman’s claim that the “distinctive protocols for worldmaking” followed by story openings are genre-specific (see Herman 2009, 112). Incorporating non-fiction works such as history books and travelogues, we also examine to what extent the composition and structure of story openings are unique to fiction. Finally, considering how “beginnings” may have changed over historical time and literary periods, we analyze whether — when taken in aggregate — historical differences can be detected in the spatial composition of literary openings.

In Hoffmann’s model, *Aktionsraum* refers to spaces where characters move in a goal-directed manner — or face obstacles to their movement — while interacting with the environment around them. *Gestimmter Raum*, in contrast, involves a pre-conscious at-

mospheric space that can be sensed and “felt” through sensuous impressions, such as sounds, tastes, or smells, or through atmospheric markers such as weather phenomena, light and darkness, or the “expressive” qualities (“Ausdrucksstärke”) that things may have. Characters may be emotionally or physically affected by this space without engaging with it functionally, as they do in *Aktionsraum*. Finally, *Anschaunungsraum* refers to a distanced space (*Fernraum*) that a character observes from a static viewpoint. Unlike *Aktionsraum*, where space is appropriated through movement and touch, *Anschaunungsraum* is characterized by the appropriation of space through vision and sight.

This model is grounded in a phenomenological understanding of space, which refers to how space is experienced and perceived by the subject in a bodily way. To account for the description of setting, Rohrbacher (forthcoming) introduced a fourth category, namely “descriptive space” into the model. This category differs from the others in that it does not involve a space that affects the experiencing subject in any way but rather focuses on ornamental details or the narrative’s arrangement of subjects and objects within a given space. Refer to Table 1 for an overview of the model, including the definition of each category as well as an example.

Type	Definition	Example
Aktionsraum	Space as moved through; directional; appropriated by touch; things and objects that characters interact with serve a functional role	As he tried to leave the cabin, because the sea was pressing in and he was up to his knees in the water, he found the door closed (Dauthendey (2012[1912]))
gestimmter Raum	Space as sensorially experienced (e.g., sounds, smells, taste); setting is associated with affect and emotional resonance; setting contributes to mood and atmosphere; anthropomorphic notion of space	It was very quiet in the large house, but even in the hallway, one could sense the scent of fresh bouquets of flowers (Storm (2018[1874]))
Anschaunungsraum	Distanced space (<i>Fernraum</i>); viewed from a static position	Among the things I saw from the stones, there was often a man of a peculiar kind (Stifter (2022[1853]))
Descriptive space	Not related to a subject’s agency; shows how things are positioned in space	The sled, a simple sleigh with a wicker carriage covered by a so-called “plan,” stood calmly the entire time on the roadway, right by the opening of a snow wall that had been piled up here (Fontane (2014[1878]))
No space	Negative examples; no concrete spatial relationship is present	With trembling hands, she gathered the folds of the torn shirt over her chest (Ganghofer (2023[1900]))

Table 1: Overview of the model as outlined in Rohrbacher (forthcoming), with examples and definitions for each category. (All translations are mine unless otherwise indicated.)

In the framework described, we can roughly distinguish between a functional/tactile (immediate), i.e., *Aktionsraum*; affective/emotional (absorbed), i.e., *gestimmter Raum*;

and visual/pictorial (detached), i.e., *Anschauungsraum* relationship that structures the perceptual environment of a character. This framework is character-bound. “Descriptive space”, on the other hand, is not part of this relationship; it functions mainly as ornamental. The model thus differentiates between a scene presented as “just” a description, without the acting subject experiencing it, and a character being at the “perceptual center” of things.

While newer terminologies concerning narrative space (e.g., Dennerlein 2009) or space and its sociological role more generally (e.g., Löw 2001) have been developed in recent work, we limit our analysis specifically to the notion of setting, focusing only on the concrete and “lived spaces” that characters interact with. Hoffmann’s model is particularly useful in this regard, as it allows us to understand better the relationship between setting and character behavior.¹

To get a better grasp of how this model might play out in fiction, consider the opening passage from Kafka’s *The castle* (1922):

It was late evening when K. arrived. The village lay deep in snow. There was nothing to be seen of Castle Mount, for mist and darkness surrounded it, and not the faintest glimmer of light showed where the great castle lay. K. stood on the wooden bridge leading from the road to the village for a long time, looking up at what seemed to be a void. Then he went in search of somewhere to stay the night. People were still awake at the inn. The landlord had no room available, but although greatly surprised and confused by the arrival of a guest so late at night, he was willing to let K. sleep on a straw mattress in the saloon bar. K. agreed to that. (Kafka (1992))

From the outset, the narrative immediately directs our attention to the character, briefly sketching the scene in which he finds himself with just a few details before quickly shifting to an action-oriented element — his search for a place to sleep. Despite a few spatial markers and concrete details, the depiction of setting remains rather sparse. While we learn that K. arrives in a town where a castle holds importance for him, little additional detail is provided of what the scenery actually looks like.

What is rendered instead is a distinct atmosphere or *Stimmung*, shaped by his perception — or rather, the lack thereof — of the castle. Interwoven here, is not merely a description of the setting but an evocation of the place’s *feel*, where character, setting, and events feed into each other.

Herman described “feltness” or “what it’s like,” alongside the notion of worldmaking, as another defining feature of narrative. While he views “feltness” in broad terms, describing it as “the experience of living through [a] storyworld-in-flux,” (Herman 2009, 1) our model aligns *gestimmter Raum* with the idea that setting can have a distinct “feel” that a character perceives affectively or emotionally, rather than merely in a direct, functional way — such as through touch and movement, as in *Aktionsraum*.

Let’s contrast this with the opening passage of Stifter’s novel *Der Nachsommer* published in 1857:

1. While Dennerlein’s terminology of narrative space also focuses on concrete rather than metaphorical or symbolic space, it does not account for direct, experiential relations between characters and setting.

My father was a merchant. He lived in part of the first floor of a moderately sized house in the city, which he rented. In the same building, he also had his shop, an office, along with storage rooms for goods and other items necessary for running his business. Besides us, only one other family lived on the first floor — a pair of elderly people, a man and his wife — who dined with us once or twice a year. We would visit them, and they would visit us on festive occasions or days traditionally reserved for paying visits or offering well wishes. My father had two children: me, his firstborn son, and a daughter who was two years younger than I.

Typical of Stifter and realist literature more generally, the opening passage of Stifter's novel is rich in description. While the novel is told from a first-person perspective, in contrast to Kafka's passage above, the beginning presented here feels rather detached. The sole purpose of this passage, it seems, is to be informative, to lay out the scene in which the story is set. The description of the character's father goes hand in hand with the description of the house and social world the family lives in.

While Kafka's passage includes some descriptive elements, it transitions much more quickly into action, while also conveying a sense of what the world the reader is about to enter feels like. This "embodied situatedness" is largely absent in Stifter's beginning, which is told from a detached, external point of view.

Stifter's passage aligns with the common assumption that realist works rely more heavily on descriptive elements, whereas Modernist works adopt a more character-centered approach, placing an emphasis on how the setting is experienced by the character. Indeed, while perceptual elements are present from the outset in Kafka's passage, they are entirely absent in Stifter's opening world, which focuses more on what the world looks like that the character inhabits.

To determine whether a formulaic structure specific to fiction can be detected across a large and diverse array of texts, one that prioritizes certain modes of setting throughout a narrative, we apply our model to different datasets. Specifically, we analyze different categories within our datasets, including fictionality (fiction vs. non-fiction), canonicity or prestige (canon corpus), and genre (e.g., fairy tales vs. historical crime). A detailed description of the datasets follows in the next section.

2. Data and Methods

2.1 Data

The corpora used in this study are derived from the German and American Gutenberg libraries. Beginning in 1780, they include books in the German language up until 1940, as outlined in Rohrbacher (forthcoming). In addition to metadata on "year," "author," and "title," they also include information on genre and subgenre.

The fiction corpus consists of 4,577 books spanning 12 genres (with "novels, novellas, and short stories" making up the majority), authored by 1,140 unique writers, and comprising a total of 17,130,609 sentences. Additionally, we construct a non-fiction corpus, drawn from the same German digital edition of the Gutenberg corpus from

which most fiction works were gathered. This non-fiction corpus spans six genres and is comparatively smaller, including 754 books by 413 unique authors, with a total of 3,064,259 sentences. Sentence segmentation was performed using the Stanza library for German-language texts (Qi et al. 2020).

We also include a sample of canonical fiction, drawn from Brottrager et al. (2022). Since their dataset ends with the year 1914, we manually supplemented the dataset with canonical works from subsequent years to ensure that the sample approximates the timeframe covered in our corpus. 202 texts were manually added, resulting in a corpus of 677 canonical works of German-language fiction. Refer to Table 2 for an overview of the different corpora.

	1780-1800	1800-1840	1840-1900	1900-1940
Fiction	108	269	1,749	2,451
Non-Fiction	43	78	264	396
Canon	46	132	256	243

Table 2: Overview of Corpora with the number of books for each range of years and each corpus.

The majority of the books in both fiction and non-fiction fall within the 19th and early 20th century, with the number of works published in the early 20th century making up the largest segment. Similar to the fiction data, we have also included the more detailed genre metadata provided by the digital edition for the non-fiction corpus. Figure 1 shows the count of books and sentences categorized by genre, where we can see that “travelogues” and “history” constitute the largest segments of the dataset, followed by “philosophy.”

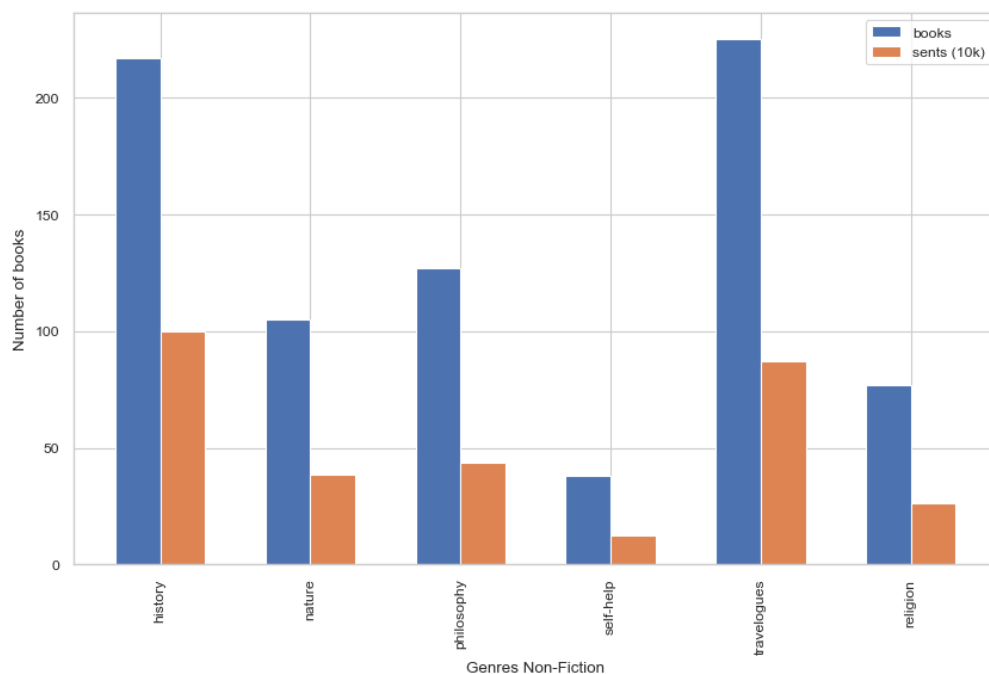


Figure 1: Overview of different genres by number of books in the non-fiction corpus

2.2 Model and Validation

180

We apply the BERT-based multiclass classification model developed in Rohrbacher (forthcoming), which was fine-tuned on a hand-annotated training set of approximately 2,800 sentences from German fiction, categorized according to the different types of settings outlined above, to the datasets presented here. Applying the classifier to our corpora, each sentence is mapped to a unique space type — the one assigned the highest probability by the classifier. While validity scores indicate strong model performance in categorizing sentences correctly ($F_1 > 80\%$ for all categories), we conduct a further manual inspection of the output, focusing on passages that were not part of the training or test sets. To do this, we revisit the examples provided above and examine the classifier’s output shown in Figure 2.²

181
182
183
184
185
186
187
188
189
190

Franz_Kafka_-_Das_Schloß.txt

```
Das erste Kapitel Es war spät abends, als K. ankam. <descriptiv> Das Dorf
lag in tiefem Schnee. <\descriptiv> <perceived> Vom Schloßberg war nichts zu
sehen, Nebel und Finsternis umgaben ihn, auch nicht der schwächste Lichtschein deutete das
große Schloß an. <\perceived> <visual> Lange stand K. auf der Holzbrücke, die
von der Landstraße zum Dorf führte, und blickte in die scheinbare Leere empor. <\visual>
<action> Dann ging er, ein Nachtlager suchen; <\action> <action> im
Wirtshaus war man noch wach, der Wirt hatte zwar kein Zimmer zu vermieten, aber er wollte,
von dem späten Gast äußerst überrascht und verwirrt, K. in der Wirtsstube auf einem
Strohsack schlafen lassen. <\action> K. war damit einverstanden. <action>
Einige Bauern waren noch beim Bier, aber er wollte sich mit niemandem unterhalten, holte
selbst den Strohsack vom Dachboden und legte sich in der Nähe des Ofens hin. <\action>
```

Adalbert_Stifter_-_Der_Nachsommer.txt

```
Der Nachsommer Eine Erzählung von Adalbert Stifter Inhalt: Die Häuslichkeit
Der Wanderer Die Einkehr Die Beherbergung Der Abschied Der Besuch Die Begegnung Die
Erweiterung Die Annäherung Der Einblick Das Fest Der Bund Die Entfaltung Das
Vertrauen Die Mitteilung Der Rückblick Der Abschluß Die Häuslichkeit Mein Vater war
ein Kaufmann. <descriptiv> Er bewohnte einen Teil des ersten Stockwerkes eines mäßig
großen Hauses in der Stadt, in welchem er zur Miete war. <\descriptiv>
<descriptiv> In demselben Hause hatte er auch das Verkaufsgewölbe, die Schreibstube
nebst den Warenbehältern und anderen Dingen, die er zu dem Betriebe seines Geschäftes
bedurfte. <\descriptiv> <descriptiv> In dem ersten Stockwerke wohnte außer uns
nur noch eine Familie, die aus zwei alten Leuten bestand, einem Manne und seiner Frau,
welche alle Jahre ein oder zwei Male bei uns speisten, und zu denen wir und die zu uns
kamen, wenn ein Fest oder ein Tag einfiel, an dem man sich Besuche zu machen oder Glück zu
wünschen pflegte. <\descriptiv> Mein Vater hatte zwei Kinder, mich, den
erstgeborenen Sohn, und eine Tochter, welche zwei Jahre jünger war als ich.
<descriptiv> Wir hatten in der Wohnung jedes ein Zimmerchen, in welchem wir uns
unseren Geschäften, die uns schon in der Kindheit regelmäßig aufgelegt wurden, widmen
mußten, und in welchem wir schliefen. <\descriptiv>
```

Figure 2: Classifier output showing the model’s color-coded tags for Franz Kafka’s *Das Schloß* (1922) and Adalbert Stifter’s *Der Nachsommer* (1857). For the color codes: green for “Anschauungsraum,” purple for “gestimmter Raum,” red for “Aktionsraum,” white for “no space,” and turquoise for “descriptive space.” “No space” is represented by white and is not labeled with a tag.

2. From both outputs, we can see that the sentence segmenter treats the first sentence, which includes the chapter title (Kafka), and the chapter overview in Stifter’s case, as a single unit. This raises a potential limitation in terms of possible noise resulting from cleaning the corpus. While it’s difficult to clean texts in a way that entirely excludes titles and other non-narrative elements such as chapters, it is important to raise the issue here, as this could falsely suggest a predominance of “no space” in the very first sentence(s) of a book. Since we examine sections or larger sentence windows in the subsequent analysis of this study, we are confident, however, that overall, the classifier is able to capture the predominance of one type of setting over the other.

This is the translated text from the opening of Kafka’s novel including the output of the classifier: 191

<descriptive> It was late evening when K. arrived. The village lay deep in snow. </descriptive> <perceived> There was nothing to be seen of Castle Mount, for mist and darkness surrounded it, and not the faintest glimmer of light showed where the great castle lay. </perceived> <visual> K. stood on the wooden bridge leading from the road to the village for a long time, looking up at what seemed to be a void. </visual> <action> Then he went in search of somewhere to stay the night. </action> People were still awake at the inn. <red> The landlord had no room available, but although greatly surprised and confused by the arrival of a guest so late at night, he was willing to let K. sleep on a straw mattress in the saloon bar. </action> K. agreed to that. <action> Several of the local rustics were still sitting over their beer, but he didn’t feel like talking to anyone. </action> <action> He fetched the straw mattress down from the attic himself, and lay down near the stove. </action> 192

Per contrast the output of Stifter’s opening: 207

My father was a merchant. <descriptive> He lived in part of the first floor of a moderately sized house in the city, which he rented. </descriptive> <descriptive> In the same building, he also had his shop, an office, along with storage rooms for goods and other items necessary for running his business. </descriptive> <descriptive> Besides us, only one other family lived on the first floor — a pair of elderly people, a man and his wife — who dined with us once or twice a year. </descriptive><descriptive> We would visit them, and they would visit us on festive occasions or days traditionally reserved for paying visits or offering well wishes. </descriptive> My father had two children: me, his firstborn son, and a daughter who was two years younger than I. 208

As shown above, the classifier generally performs well in distinguishing between the different space types outlined in the model used. For instance, in Kafka’s opening, we can observe a narrative shift between more descriptive notions (comprised of “descriptive space” and *Anschaunungsraum* in our model), which relate to the visuality and verisimilitude of a scene, and atmospheric and perceptual notions (such as the character being unable to see due to the fog and darkness). This then transitions into action elements, as the character moves to find a place to sleep and interacts with the space around him — for example, fetching the straw mattress to lie down to sleep. 219

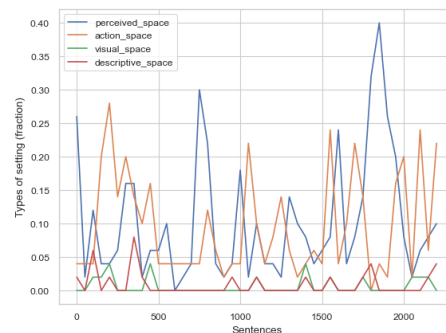
As the examples above show, the categories do not always fit neatly into an either/or scheme — even at the sentence level, different types may overlap. For instance, the sentence from Kafka’s passage, “There was nothing to be seen of Castle Mount, for mist and darkness surrounded it, and not the faintest glimmer of light showed where the great castle lay,” blends visual perception with atmospheric elements. Since our classification system is mutually exclusive — i.e., only one category can be assigned — the model selects the label with the highest probability. In ambiguous cases like this, during annotation, we assigned the label that was most prominent or of central focus. 227

Here, although vision is involved, the emphasis is on the mood and atmosphere rather than on what the character sees. We therefore agree with the model's classification of this sentence as *gestimmter Raum* rather than *Anschauungsraum*.

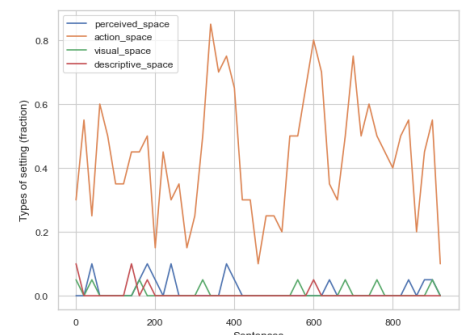
It's important to note that in the analysis presented here, we are primarily concerned with the "concrete space" in which characters exist, reflecting how characters might react to and interact with the built or natural environment around them. Thus, while the first sentence presented here, "It was late evening when K. arrived," does indeed imply that the character has arrived at a place, we don't yet know where. This is not explicitly stated in terms of a concrete presence, so, in our view, the classifier is correct in labeling it as "no space" rather than *Aktionsraum*, which would indicate movement.

Similarly, in Stifter's passage, as outlined in the close reading above, we see that the classifier correctly identifies the sentences in this passage as "descriptive space," aligning with our own interpretation. This further reinforces the model's reliability in capturing how different types of spaces function within a text.

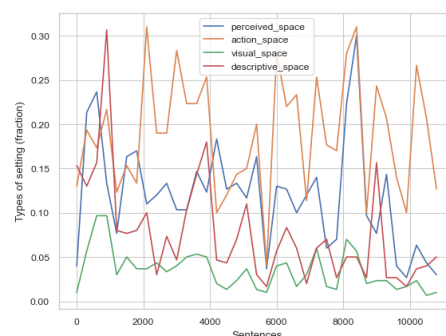
Expanding our analysis beyond individual passages, we can gain further insights by examining how different types of setting evolve across the story time of entire books. A closer look at works by various authors reveals distinct differences in how setting is portrayed over narrative time. Figure 3 shows the distribution of the different types of setting at the individual book level.



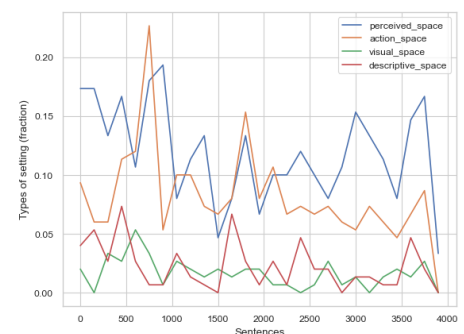
Johann Wolfgang von Goethe, *Die Leiden des jungen Werther*, 1774



Heinrich von Kleist, *Michael Kohlhaas*, 1810



Adalbert Stifter, *Der Nachsommer*, 1857



Rainer Maria Rilke, *Die Aufzeichnungen des Malte Laurids Brigge*, 1910

Figure 3: Distribution of the different types of setting across individual books.

We can see for example that Goethe's *Werther* which is characteristic of the proto-romantic movement of *Sturm und Drang*, features a high frequency in *gestimmter Raum*,

already dominates at the very beginning of the story. This speaks to the strong emotions rendered in the book, often expressed through the depiction of nature, as a reflection of the protagonist's emotional state. In contrast, Heinrich von Kleist's novella *Michael Kohlhaas* (1810) is almost exclusively comprised of *Aktionsraum*. This might come as no surprise to anyone familiar with the book. Taking place in the 16th century, the story centers around Michael Kohlhaas, a horse dealer who is almost always on the move fanatically fighting for justice.

When plotting Stifter's *Der Nachsommer* (1857), we can observe a high frequency of "descriptive space", which at one point accounts for 30% of the narrative. Once again, this might not come as a surprise, given Stifter's reputation for lengthy descriptive passages. What's interesting, however, is the equal predominance of *gestimmter Raum*, indicating that, alongside "descriptive space" and *Aktionsraum*, the depicted settings in the novel — and those engaged with by the characters — are also imbued with qualities that contribute to mood and atmosphere. As we observed in the close reading, descriptive elements are especially prominent at the beginning of the story and then decline over time.

Rilke's *Die Aufzeichnungen des Malte Laurids Brigge* (1925), dominates in *gestimmter Raum*, especially at the beginning of the novel compared to the other types. Rilke's work can be seen as being illustrative of modernist works more generally, in which space is commonly thought of being filtered through the perceptions of the internal focalizer, focusing on sensory elements rather than presenting the external world that the character inhabits Buchholz and Jahn 2005. While *Aktionsraum* also features high, it only at one point surpasses the frequency of *gestimmter Raum* in the narrative. Based on the visualization above, we can discern a general trend of a decrease in spatiality overall over the course of the narrative.

When comparing the close readings as well as the plots of different individual books to one another, we can certainly detect historical (and stylistic) differences in how these spaces are described. To assess whether these individual observations align with broader patterns — and whether they support what critics have previously theorized, we can test them at scale. By analyzing a much larger and more diverse set of openings, including both canonical texts (such as those presented here) and non-canonical ones, we can determine whether these trends hold up in aggregate.

3. Results

3.1 Towards a quantitative analysis of narrative beginnings

To analyze the distribution of the different types of settings outlined in the model across the fiction corpus used in this paper, we begin by categorizing them based on their narrative placement, examining their frequency and distribution at various points within the text. For this analysis, we divide each text into multiple "chunks" or sections to track how the different modes of setting change over narrative time. We then aggregate these patterns across all books in the corpus. By quantifying the prevalence of specific types of setting in story openings versus later parts, we aim to assess the extent to which initial settings function as "anchors" or "narrative establishments" of the fictional world

and how this orientation shifts as the narrative progresses. Refer to Figure 4 for the distribution of each space type across narrative time. 298 299

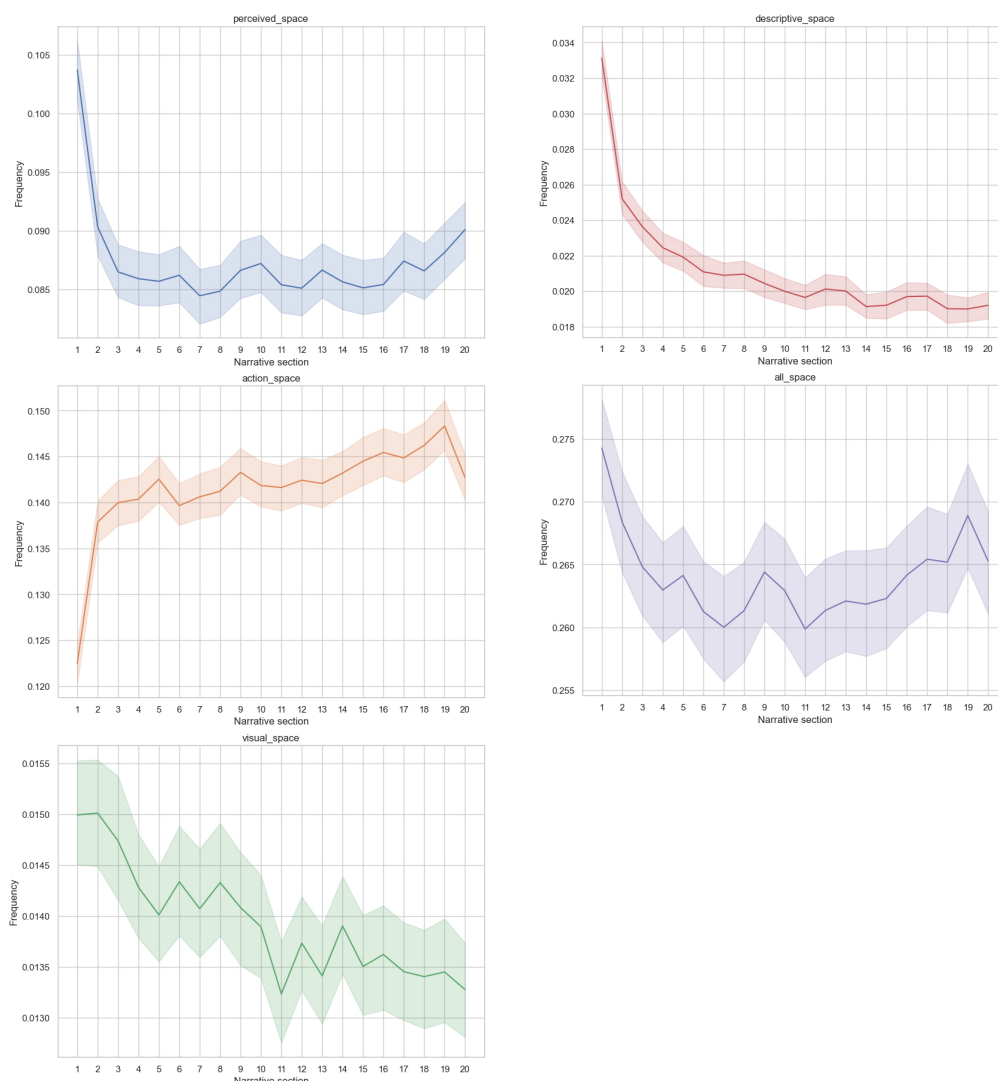


Figure 4: Distribution of the different types of setting across narrative time. Labels: “perceived space” = *gestimmter Raum*, “action_space” = *Aktionsraum*, “visual_space” = *Anschauungsraum*. All_space shows the aggregate of all types combined.

When looking at setting across narrative time, we observe significant differences between the ways the different types of settings are distributed. We can see that while “descriptive space,” *gestimmter Raum*, and *Anschauungsraum*, feature high in the very first sections of a book, they then gradually decrease (except for *gestimmter Raum*, which again increases slightly at the end of narratives). This is the exact opposite for *Aktionsraum*, which features low at the beginning and then rises while falling at the end of narratives. When looking at the “aggregate” of all space types, we can also see that overall narratives show a slighter higher distribution of settings in their opening (and closing) sections compared to the rest of narratives. ³

To further test this and to determine whether there is a statistically significant elevated distribution of *gestimmter Raum* and “descriptive space,” and *Anschauungsraum* at the

3. Interestingly, the ‘U-shape’ observed here in relation to setting across narrative arcs runs counter to the ‘inverted U’ commonly found in narrative arc structures (see Boyd et al. 2020).

beginning of stories, as the plots above suggests, we employ a time series regression analysis. This approach allows us to statistically model the relationship between individual space types and narrative time, providing insights into how different types of settings vary across sections of a book. We use sine and cosine components, along with a normalized time variable, to capture any cyclical trends in our dataset.

The statistical analysis reflects what we've already observed in the plot above. The regression analysis finds that "descriptive space" and *gestimmter Raum* are indeed significantly higher at the beginning of texts, compared to the other parts of a book. *Aktionsraum*, in turn, is lower at the beginning, and then increases across narrative time. The model suggests that *Aktionsraum* tends to increase over the course of narratives, with some slight cyclical fluctuations throughout. This effect is statistically significant ($\beta = 0.0539, p < 0.001$), suggesting that as the narrative progresses *Aktionsraum* becomes more prevalent. Descriptive space ($\beta = -0.0402, p < 0.001$) and *gestimmter Raum* ($\beta = -0.0636, p < 0.001$) in turn are more prevalent at the beginning of narratives and decrease over time. Both types exhibit slight cyclical patterns similar to *Aktionsraum*. The effect we've observed in aggregate, with "all space" being high at beginnings and then declining, is also statistically significant ($\beta = -0.0530, p < 0.05$). For *Anschauungsraum*, however, the statistical analysis found the perceived trend of a decline over the course of the narrative to be non-significant ($\beta = -0.0031, p = 0.324$).

When analyzing non-fiction books, we decided to separate the genre of "travelogues" from the other non-fiction genres due to the similarity between travelogues and fiction in their use of the phenomenological experience of space. In Figure 5, we observe that travelogues indeed exhibit a much higher frequency of different types of setting across narrative time compared to the other genres in our non-fiction corpus, which show minimal frequency overall. This trend is also evident when compared to the fiction corpus. Interestingly, despite *gestimmter Raum*, which remains relatively consistent throughout the narrative in the analyzed corpus, all other types of setting exhibit a lower frequency at the beginning of the stories. However, this perceived trend, based on the visualizations, is not supported statistically.

Although our model identifies significant sin and cosine variables—indicating a cyclical trend (i.e., different types of setting fluctuate across narrative time) for travelogues — we do not find any statistically significant effect suggesting a general increase or decrease in spatial depictions at the beginning of texts compared to later sections. This is also evident in the greater variance travelogues exhibit across the different types of spaces, whereas fiction works demonstrate more consistency, particularly at the very start of narratives. Compared to non-fiction, the depiction of setting in fictional texts over narrative time is highly generic, in that it follows a distinct pattern or shape that can be detected across a wide range of books. This kind of consistency is not evident in travelogues or other non-fiction works. Refer to the supplementary materials for the complete regression table results.

Viewed this way, one defining feature of fiction is the condensed and concentrated manner in which descriptions appear at the beginning of stories before nearly disappearing over the course of the narrative. This suggests that when descriptions do play a role, it is primarily at the very start of stories, serving an important function in establishing narrative worlds compared to their almost negligible role throughout the rest of the

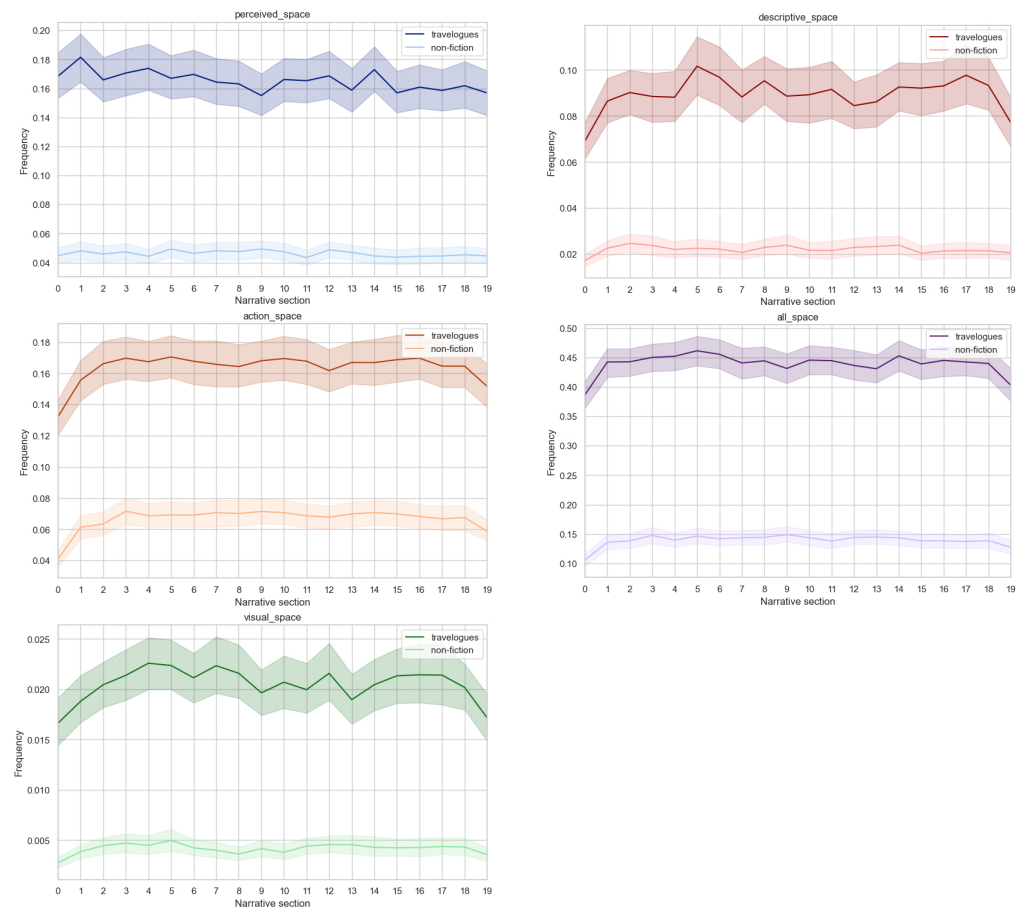


Figure 5: Distribution of the different types of setting across narrative time in travelogues and non-fiction.

narrative. While *gestimmter Raum* is overall more prevalent, it is also more frequent at the start of narratives. *Aktionsraum*, meanwhile, remains the most prevalent space type overall — even at the beginning of stories compared to the others — but rises significantly in importance as the story progresses.

It has to be noted, however, that the residuals in the data are not normally distributed, violating the so-called “normality assumption” required for this kind of statistical analysis. Given the distribution of the data, which is rather irregular and cannot be well approximated by any standard statistical model (at least to our knowledge), this could impact the reliability of the regression’s estimates. Exploring alternative models or transformations of the data that better accommodate its distinct distribution could thus yield more accurate results.⁴

3.2 “Protocols for worldmaking”: Beginnings and Genre

In his analysis of “worldmaking” Herman (2009) devotes a significant section in analyzing the importance of “worldmaking” for narrative beginnings. By comparing and contrasting story openings from different genres, he seeks to investigate how “part of the meaning of “genre” consists of distinctive protocols for worldmaking” (112). While Herman basis his analysis on two openings drawn from fiction (a short story and a science fiction novel) to analyze in what ways a set of “worldmaking procedures” is “inflected differently when different genres are involved,” (115) we can make use of a much larger, and more varied database of texts as the basis of our analysis.

To analyze how narrative beginnings differ across genres, we approximate the length of “openings” by using a 15-sentence window for each book in our corpus.⁵ Unlike the previous analysis across narrative time, where we investigated how the different types of setting evolve throughout the narrative and how the beginning of a story relates to the rest of the book, this section focuses exclusively on the beginnings.

By contrasting the beginnings of each genre in our fiction dataset, we observe that genres differ in how they use the different types of setting defined in our framework. Refer to the lineplot shown in Figure 6.

For example, “crime novels” exhibit a particularly high proportion of *Aktionsraum* at the very beginning of stories (as well as a high use of spatial descriptions overall), whereas the opposite is true for “young adult”, where setting is particularly low. “Horror,” on the other hand, stands out for its limited use of descriptive space in openings, while *gestimmter Raum* dominates. Interestingly, in the “science fiction” genre, “descriptive space” is noticeably more prominent compared to thematically similar genres such as “speculative fiction” or “horror”.

Applying a post-hoc pairwise Tukey test to this data, we confirm that most genres differ significantly in how they employ the various types of setting in their opening passages.

4. However, research has shown that, in datasets with large sample sizes (like ours), linear models tend to be robust to violations of the normality assumption. Transformations of the data might even have a detrimental effect, introducing new biases (See Schmidt and Finan 2018)

5. While we also tested smaller and larger windows (e.g., 8 and 30 sentences), we found that the results did not differ significantly. It’s important to note, however, that a fixed window length may not fully account for the variability in how different authors structure their beginnings or to capture what makes up an “opening scene”, which remains a limitation of this approach.

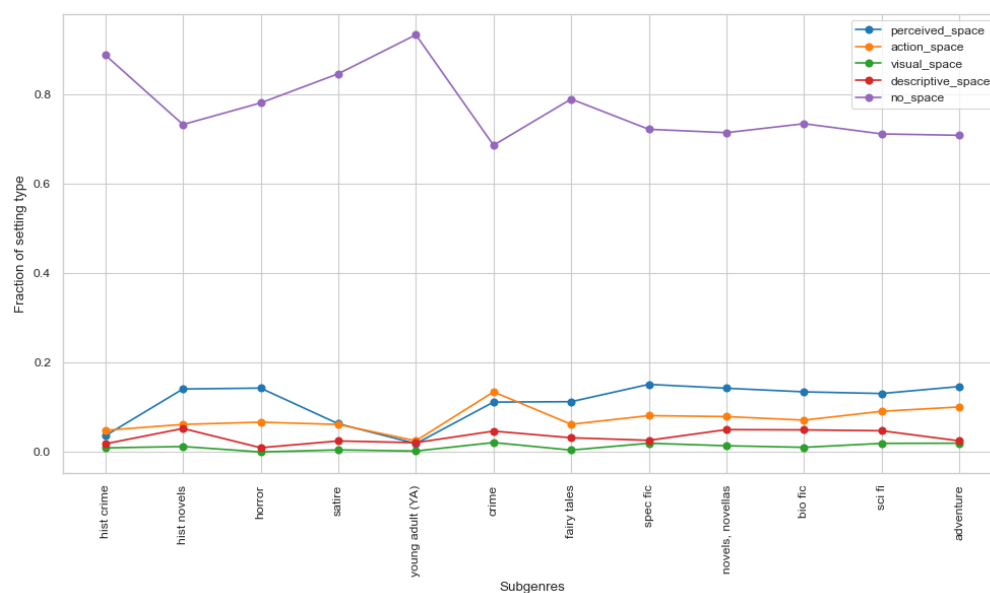


Figure 6: Spatial composition of setting in beginnings across genre.

For instance, compared to “young adult”, “historical crime” shows a 6.9-point higher frequency of *gestimmter Raum* ($p < 0.05$, $M = 6.9$). “Fairy tales”, by contrast, exhibit a 4.6-point and 3.8-point higher frequency of *Aktionsraum* compared to “historical novels” ($p < 0.05$, $M = 4.6$) and “crime” ($p < 0.05$, $M = 3.8$), respectively. Similarly, “adventure novels”, exhibit a 7.1-point higher frequency than novels and novellas ($p < 0.05$, $M = 7.1$) and a mean difference of 6.7 compared to “speculative fiction” ($p < 0.05$, $M = 6.7$). Comparing “fairy tales” and “adventure novels,” the mean difference in relation to fairy tales is almost negligible, amounting to just 0.5, suggesting a close similarity between these genres in how they make use of setting in the opening passages of books. This makes intuitive sense, as both genres emphasize action and mobility (as is typical of the quest narrative), which, as the analysis here suggests, already dominate in the opening scenes.

However, when it comes to “descriptive space,” no significant difference is found across most genre pairs, with mean differences not exceeding 0.8–1. This suggests that, while these genres share a similarly high reliance on description, the other spatial categories — including *Anschauungsraum* — are more indicative of how a specific genre renders its storyworld in its opening passages. Refer to the supplementary material for the output of the Tukey test.

While a more detailed, qualitative exploration of each genre’s openings lies beyond the scope of this study, our analysis reveals clear differences in how various types of setting are employed — particularly in genres such as “young adult,” “crime,” “adventure novels,” “horror,” and “fairy tales.” Others, such as “biographical fiction,” “historical novels,” and, interestingly, “science fiction,” align more closely with general trends observed in the larger dataset of “novels and novellas”.

3.3 The rising spatiality of beginnings across history

In his study on “narrative beginnings,” Mikkonen (2020) raises the question of a “historical narratology of literary openings” — one that examines how the beginning of stories may reflect the period in which they are set (14). Using quantitative methods, we can take up this question and analyze, by examining thousands of story openings, the ways in which they differ in their use of “worldmaking” across history. Applying the same window size previously used for different genres, we now explore how narrative openings evolve over historical time.

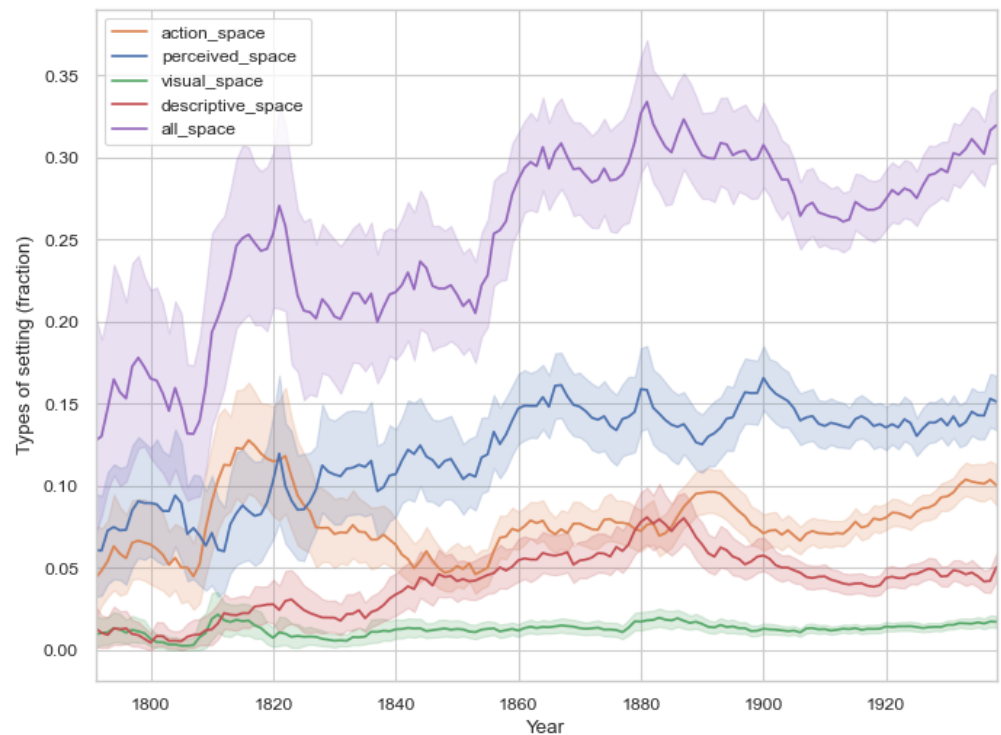


Figure 7: Distribution of the ratio for each type of setting in story openings across the analyzed timeframe.

As we can see in Figure 7, story beginnings tend to become more spatial within the analyzed timeframe. Looking at the individual modes, we observe, for instance, an increase in openings that make use of “descriptive space” between 1840 and 1900, which coincides with the period of Realism. This trend is especially prevalent between 1880 and 1890, where “descriptive space” nearly reaches the level and is on par with *Aktionsraum*.

The slight rise in *Aktionsraum* and the decrease in “descriptive space” during Modernism could indeed be interpreted as reflective of the *in media res* openings characteristic of literature from this period. Rather than spending extended time on scene-setting, the narrative tends to jump directly into portraying a “character-in-action.” Despite some larger fluctuations in earlier periods around 1800 — which may be partially due to the smaller amount of data available for those years and thus reflect an artifact of the dataset — *gestimmter Raum* is the most prevalent space type in story openings overall, followed by *Aktionsraum* and “descriptive space.”

To analyze the effect of canonicity in the way setting is employed in openings across history, we apply our model to the canonical dataset presented here. Research in CLS

has repeatedly shown that a manually curated dataset — i.e., one created based on certain distinct features — produces different results compared to a more diverse and heterogeneous one. Prior work has demonstrated that canonical works often differ in style and lexical diversity from non-canonical ones (see Algee-Hewitt et al. 2016; Brottrager et al. 2021; Koolen et al. 2020; Underwood and Sellers 2016).

For Pascale Casanova (2004), the canon “embodies the very notion of literary legitimacy,” representing the standard of what is “formally” acknowledged as Literature and serving as a benchmark (or “unit of measurement”) for evaluating other literary works (14). While our model focuses “just” on setting, can we observe any differences in how this concept is represented in the openings of canonical works compared to non-canonical ones?

Refer to Figure 8 for the historical distribution of each individual space type in the opening sections of canonical works compared to the larger dataset presented above, which includes both canonical and non-canonical works. While the overall trend remains the same, we can indeed see significant differences, particularly in the frequency with which different types of setting are employed in canonical versus non-canonical openings.

While the larger sample does not show an overall increase in *gestimmter Raum* during Romanticism, the canonical sample reveals a significant spike in this type of setting between 1800 and 1830. Other peaks are also more pronounced in the canonical sample — for instance, *Aktionsraum* around 1820 and “descriptive space” during the Realist period. The confidence intervals (CIs) indicate that the number of outliers in the canonical sample is much larger than in the non-canonical one. While overall spatiality appears to increase in the non-canonical sample, it tends to decrease in the canonical one, particularly in the years following 1840. This trend, however, seems to be primarily driven by the sharp rise in both *gestimmter Raum* and “descriptive space” during Realism.

To further investigate which works might be driving these spikes, we plot the outliers, texts that exhibit a particularly high frequency of a specific type of setting in their openings, at the book level within the canonical sample. For readability, we include only outliers from the years in which the canonical and non-canonical samples differ significantly. Figure 9 shows that certain works by individual authors display especially high frequencies for specific spatial types. During the Romantic period, the works of Ludwig Tieck, E. T. A. Hoffmann, and Joseph von Eichendorff cluster together — all three being prominent representatives of the Romantic canon.

Aktionsraum, in turn, is dominated by Heinrich von Kleist, and again Hoffmann and Tieck when looking at the years between 1800 and 1830, where the frequency of this space type in beginnings is particularly high. Regarding the spike in “descriptive space”, which is more pronounced in canonical works during Realism, Stifter’s oeuvre is particularly prominent.

A closer look at outliers from specific literary periods suggests that at least some of the differences compared to the larger corpus are driven by a few highly canonical authors who stand apart from their contemporaries, many of whom display a more consistent and uniform use of the various types of setting. While Tieck’s or Hoffmann’s frequent use of *gestimmter Raum* in Romanticism may resonate with readers – given its characteristic idyllic or atmospheric depiction of nature – this brief examination of

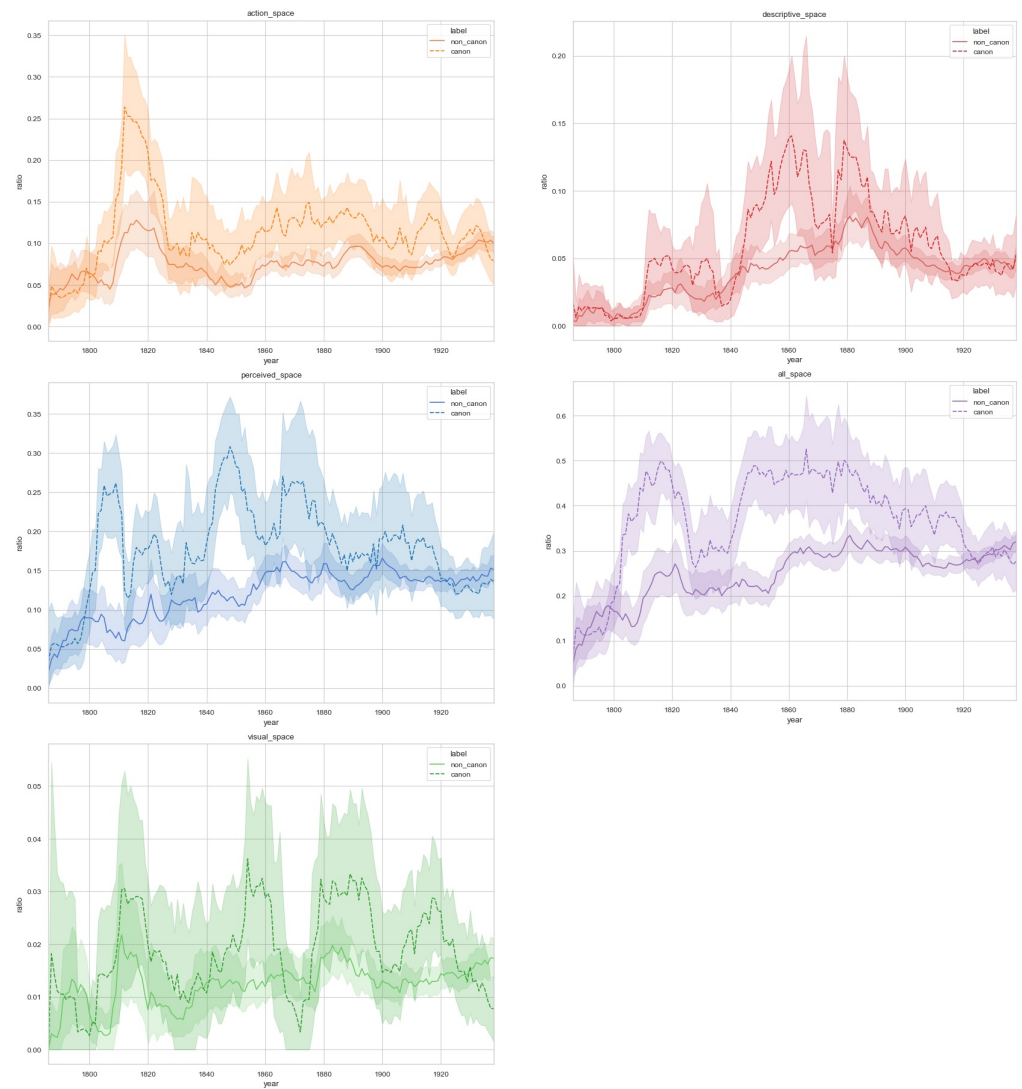
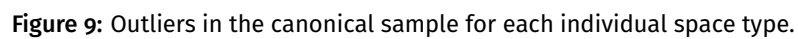


Figure 8: Distribution of the different types of setting across narrative time in canonical and non-canonical works.



canonical works suggests that the “selectiveness” of these works might contribute to certain assumptions about stylistic features specific to a given period.

When examining the broader historical trendline, however, these assumptions become more complex. The peak of *gestimmter Raum* during Romanticism, which is evident in the canonical sample, disappears entirely in the larger one. Similarly, while “descriptive space” does experience a noticeable increase during Realism, the peak is much less pronounced. Analyzing a larger, more heterogeneous dataset that includes both canonical and non-canonical works, we can identify significant structural patterns in literary history that more traditional, narrower approaches may fall short of detecting.

4. Discussion

Consistent with previous critical arguments, we found that story beginnings exhibit a higher distribution of setting overall, providing further evidence that setting plays a crucial role in establishing the fictional world at the start of narratives. This helps readers orient themselves within the spatial and sensory framework of the story. Importantly, however, when looking at the different types, we found significant differences as to how beginnings play out at the individual space type level. When aggregating the different space types for each individual book in our corpus and analyzing how these trends manifest across all books, we found that it is primarily at the beginning of stories that a more descriptive focus on space appears, emphasizing pictorial representation and atmosphere over action and movement. Based on the data used in this study, this effect is unique to fiction and varies across fictional sub-genres.

The “privileged position” of an opening, according to Mikkonen (2020), is due to a narrative’s potential for “referential grounding” — the introduction of a text’s initial set of referents (5). This aligns with classic models like Gustav Freytag’s pyramid, where exposition precedes rising action. Given the framework employed here, we can state that more generally, narrative openings engage in anticipation, accommodating characters in a pre-established environment. Narratives thus shift from establishing a narrative world (the way how it looks and feels) to mobilizing it (through the actions of the characters). Over the course of the narrative, the settings that characters inhabit and traverse become less associated with affective and atmospheric markers or static description, and instead take on a more functional role, emphasizing movement and interaction — especially through tactile engagement with the things and objects that make up space.⁶

Historically, in the analyzed dataset, we have observed a general trend of increasing spatiality in literary openings, with some pronounced fluctuations in the individual types of setting. Specifically, we found a noticeable increase in “descriptive space” during Realism and a rise in more action-centered elements during both Realism and Modernism. *Gestimmter Raum*, while slightly increasing overall and remains relatively stable throughout the analyzed timeframe. The overall predominance of *gestimmter Raum* highlights, above all, the importance of beginnings in fleshing out a narrative universe’s “feel” and atmosphere, allowing the reader to become attuned to the affective qualities of the setting depicted, rather than focusing primarily on concrete action or

6. For the importance of things as “narrative props” in fiction, as well as their “infrastructural” role, Piper and Bagga (2022).

visual detail.	525
When comparing this to a canonical sample, we found that some of these fluctuations	526
are more pronounced, and new spikes appear that were absent in the larger dataset.	527
While canonical works generally employ a higher frequency of spatial elements and	528
seem to decline (rather than rise) after the period of Realism, further inspection suggests	529
that at least some of the individual spikes detected might be driven by a few select,	530
well-known authors.	531
Future work investigating the role of narrative beginnings could consider employing	532
different analytical frameworks to account for their uniqueness in storytelling. While our	533
analysis has focused primarily on setting and the spatial composition of beginnings in	534
terms of narrative worldmaking, other aspects of narrative openings remain unexplored.	535
For instance, future studies might examine the interaction between setting and narrative	536
perspective or the role of temporality in world-building. Comparative analyses across	537
different languages or cultural contexts could also shed light on the extent to which	538
these findings are culturally specific.	539
5. Data Availability	540
Data can be found here: https://github.com/katrinrohrb/narrative-beginnings .	541
6. Software Availability	542
Code can be found here: https://github.com/katrinrohrb/narrative-beginnings .	543
7. Acknowledgements	544
I would like to thank the anonymous reviewers for their constructive suggestions and	545
comments. This article is based on research conducted for my PhD thesis. I would also	546
like to thank my advisor, Andrew Piper, for his valuable input, feedback, and support.	547
8. Author Contributions	548
Katrin Rohrbacher: Conceptualization, Formal analysis, Investigation, Methodology,	549
Writing – original draft, Writing – review & editing.	550
References	551
Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti,	552
and Hannah Walser (2016). <i>Canon/Archive: Large-scale Dynamics in the Literary Field</i> .	553
Stanford Literary Lab.	554
Boyd, Ryan L, Kate G Blackburn, and James W Pennebaker (2020). “The Narrative Arc:	555
Revealing Core Narrative Structures through Text Analysis”. In: <i>Science Advances</i>	556
6.32, eaba2196.	557

Brottrager, Judith, Annina Stahl, and Arda Arslan (2021). "Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and-Intrinsic Features." In: <i>Computational Humanities Research</i> 2021, 195–205.	558 559 560
Brottrager, Judith, Annina Stahl, Arda Arslan, Ulrik Brandes, and Thomas Weitin (2022). "Modeling and Predicting Literary Reception". In: <i>Journal of Computational Literary Studies</i> 1.1, 95. 10.48694/jcls.95 .	561 562 563
Buchholz, Sabine and Manfred Jahn (2005). "Space in Narrative". In: <i>Routledge Encyclopedia of Narrative Theory</i> , 551–555.	564 565
Casanova, Pascale (2004). <i>The World Republic of Letters</i> . Harvard UP.	566
Dauthendey, Max (2012[1912]). <i>Der Geist meines Vaters</i> . Project Gutenberg.	567
Dennerlein, Katrin (2009). <i>Narratologie des Raumes</i> . Walter de Gruyter.	568
Fontane, Theodor (2014[1878]). <i>Vor dem Sturm: Roman aus dem Winter 1812 auf 13</i> . Project Gutenberg.	569 570
Freytag, Gustav (1895). <i>Technique of the Drama: An Exposition of Dramatic Composition and Art</i> . S. Griggs.	571 572
Ganghofer, Ludwig (2023[1900]). <i>Die Mühle am Fundsee</i> . Project Gutenberg.	573
Herman, David (2009). <i>Basic Elements of Narrative</i> . John Wiley & Sons.	574
Hoffmann, Gerhard (1978). <i>Raum, Situation, erzählte Wirklichkeit. Poetologische und historische Studien zum englischen und amerikanischen Roman</i> . J. B. Metzler.	575 576
Hones, Sheila (2011). "Literary Geography: Setting and Narrative Space". In: <i>Social & Cultural Geography</i> 12.7, 685–699.	577 578
Kafka, Franz (1992). <i>The Castle: Introduction by Irving Howe</i> . Everyman's Library.	579
Koolen, Corina, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout (2020). "Literary Quality in the Eye of the Dutch Reader: The National Reader Survey". In: <i>Poetics</i> 79, 101439.	580 581 582
Labov, William and Joshua Waletzky (1997). "Narrative Analysis: Oral Versions of Personal Experience." In: <i>Journal of Narrative & Life History</i> .	583 584
Löw, Martina (2001). <i>Soziologie des Raumes</i> . Suhrkamp.	585
Mikkonen, Kai (2020). "'The marquise went out at 5 o'clock': Novel Beginnings and Realistic Expectations". In: <i>Frontiers of Narrative Studies</i> 6.1, 4–17.	586 587
Miller, Norbert (1965). <i>Romananfänge: Versuch zu einer Poetik des Romans</i> . Literarisches Colloquium Berlin.	588 589
Piper, Andrew and Sunyam Bagga (2022). "A Quantitative Study of Fictional Things." In: <i>Computational Humanities Research</i> 2022, 268–279.	590 591
Piper, Andrew, Hao Xu, and Eric D Kolaczyk (2023). "Modeling Narrative Revelation". In: <i>Computational Humanities Research</i> 2023, 500–511.	592 593
Polaschegg, Andrea (2020). <i>Der Anfang des Ganzen: Eine Medientheorie der Literatur als Verlaufskunst</i> . Wallstein Verlag.	594 595
Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning (2020). "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations</i> .	596 597 598 599
Richardson, Brian (2008). <i>Narrative Beginnings: Theories and Practices</i> . U of Nebraska Press.	600 601
Rohrbacher, Katrin (2025). "'Lived Space': A Computational Study of Setting in Fiction". In: <i>Comparing Landscapes. Approaches to Space and Affect in Literary Fiction</i> . Ed. by Robin M. Aust, Giulia Grisot, and Berenike Herrmann. Bielefeld University Press.	602 603 604

- Romagnolo, Catherine (2015). *Opening Acts: Narrative Beginnings in Twentieth-Century Feminist Fiction*. U of Nebraska Press. 605
606
- Said, Edward W (1968). "Beginnings". In: *Salmagundi* 2.4 (8), 36–55. 607
- Schmidt, Amand F and Chris Finan (2018). "Linear Regression and the Normality Assumption". In: *Journal of clinical epidemiology* 98, 146–151. 608
609
- Stifter, Adalbert (2022[1853]). *Bunte Steine*. Project Gutenberg. 610
- Storm, Theodor (2018[1874]). *Viola tricolor*. Project Gutenberg. 611
- Underwood, Ted and Jordan Sellers (2016). "The Longue Durée of Literary Prestige". In: *Modern Language Quarterly* 77.3, 321–344. 612
613

A. Appendix: Supplementary Material

Regression Results for *Aktionsraum*

Model:	OLS	Adj. R-squared:	0.003
No. Observations:	95420	Log-Likelihood:	96842.
Df Model:	5	F-statistic:	53.10
Df Residuals:	95414	Prob (F-statistic):	3.13e-55
R-squared:	0.003	Scale:	0.0076917

	Coef.	Std.Err.	t	P> t	[0.025	0.975]
const	0.0392	0.0238	1.6468	0.0996	-0.0074	0.0857
time_tf	0.0539	0.0151	3.5762	0.0003	0.0244	0.0834
sin	0.0284	0.0034	8.2800	0.0000	0.0217	0.0351
cos	0.0749	0.0232	3.2345	0.0012	0.0295	0.1203
sin2	-0.0120	0.0046	-2.5822	0.0098	-0.0211	-0.0029
cos2	0.0114	0.0016	7.2115	0.0000	0.0083	0.0145

Regression Results for *gestimmter Raum*

Model:	OLS	Adj. R-squared:	0.002
No. Observations:	95420	Log-Likelihood:	1.0171e+05
Df Model:	5	F-statistic:	40.97
Df Residuals:	95414	Prob (F-statistic):	2.87e-42
R-squared:	0.002	Scale:	0.0069460

	Coef.	Std.Err.	t	P> t	[0.025	0.975]
const	0.2062	0.0226	9.1266	0.0000	0.1619	0.2505
time_tf	-0.0636	0.0143	-4.4435	0.0000	-0.0917	-0.0356
sin	-0.0305	0.0033	-9.3676	0.0000	-0.0369	-0.0241
cos	-0.0937	0.0220	-4.2574	0.0000	-0.1369	-0.0506
sin2	0.0159	0.0044	3.6055	0.0003	0.0072	0.0245
cos2	-0.0109	0.0015	-7.2292	0.0000	-0.0138	-0.0079

Regression Results for "descriptive_space"

Model:	OLS	Adj. R-squared:	0.012
No. Observations:	95420	Log-Likelihood:	2.0504e+05
Df Model:	5	F-statistic:	228.2
Df Residuals:	95414	Prob (F-statistic):	4.95e-243
R-squared:	0.012	Scale:	0.00079631

	Coef.	Std.Err.	t	P> t	[0.025	0.975]
const	0.0928	0.0076	12.1343	0.0000	0.0778	0.1078
time_tf	-0.0402	0.0048	-8.2972	0.0000	-0.0497	-0.0307
sin	-0.0136	0.0011	-12.3232	0.0000	-0.0158	-0.0114
cos	-0.0566	0.0075	-7.5895	0.0000	-0.0712	-0.0420
sin2	0.0092	0.0015	6.2007	0.0000	0.0063	0.0122
cos2	-0.0044	0.0005	-8.5624	0.0000	-0.0053	-0.0034

Regression Results for all_space

Model:	OLS	Adj. R-squared:	0.000
No. Observations:	95420	Log-Likelihood:	50887.
Df Model:	5	F-statistic:	9.261
Df Residuals:	95414	Prob (F-statistic):	7.92e-09
R-squared:	0.000	Scale:	0.020153

	Coef.	Std.Err.	t	P> t	[0.025	0.975]
const	0.3571	0.0385	9.2808	0.0000	0.2817	0.4326
time_tf	-0.0530	0.0244	-2.1737	0.0297	-0.1008	-0.0052
sin	-0.0162	0.0055	-2.9126	0.0036	-0.0270	-0.0053
cos	-0.0792	0.0375	-2.1130	0.0346	-0.1527	-0.0057
sin2	0.0139	0.0075	1.8460	0.0649	-0.0009	0.0286
cos2	-0.0038	0.0026	-1.4986	0.1340	-0.0088	0.0012

Regression Results for Anschauungsraum

Model:	OLS	Adj. R-squared:	0.001
No. Observations:	95420	Log-Likelihood:	2.4793e+05
Df Model:	5	F-statistic:	14.48
Df Residuals:	95414	Prob (F-statistic):	3.27e-14
R-squared:	0.001	Scale:	0.00032412

	Coef.	Std.Err.	t	P> t	[0.025	0.975]
const	0.0190	0.0049	3.8898	0.0001	0.0094	0.0285
time_tf	-0.0031	0.0031	-0.9864	0.3240	-0.0091	0.0030
sin	-0.0004	0.0007	-0.6217	0.5341	-0.0018	0.0009
cos	-0.0039	0.0048	-0.8130	0.4162	-0.0132	0.0055
sin2	0.0007	0.0010	0.7251	0.4684	-0.0012	0.0026
cos2	-0.0000	0.0003	-0.0602	0.9520	-0.0007	0.0006

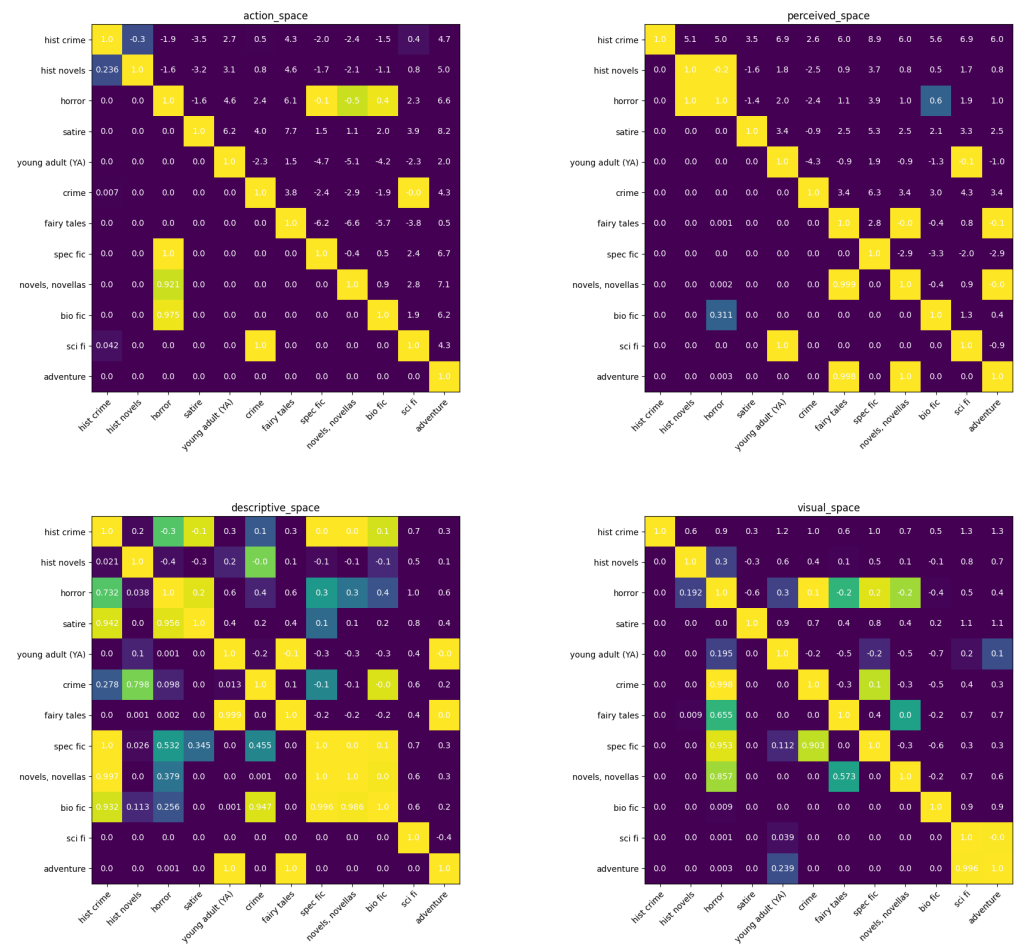





Figure 10: P-values (<0.001) and mean differences Tukey-Kramer (HSD) test. The values in the left diagonal correspond to the p-values, the values on the right the mean differences.

Event Detection between Literary Studies and NLP

A Survey, a Narratological Reflection, and a Case Study

Noa Visser Solissa¹ 
 Andreas van Cranenburgh¹ 
 Federico Pianzola¹ 

1. Center for Language and Cognition, University of Groningen , Groningen, The Netherlands.

Citation

Noa Visser Solissa, Andreas van Cranenburgh, and Federico Pianzola (2025). "Event Detection between Literary Studies and NLP. A Survey, a Narratological Reflection, and a Case Study". In: *CCLS2025 Conference Preprints 4* (1). 10.26083/tuprints-00030150

Date published 2025-06-17

Date accepted 2025-04-24

Date received 2025-02-07

Keywords

events, event detection, narratology, literature, news, historical texts

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. The presentation of narrative to the reader is a key aspect of fiction, as information gaps created by the ordering of events create narrative tension. Our objective is to develop a computational model that can detect *syuzhet*, the way the text presents events to the reader. We have created a theoretical model for the annotation of events in fiction in several languages. Automatic event detection has also been applied in several other domains, such as journalism and history. Due to the lack of consensus on the definition of event within and across domains, previous works demonstrate a wide range of approaches and applications of automated event detection. We give an overview of how previous works differ from each other, and how our model relates to it. We also compare our model to a storyline analysis framework developed for news. We show how our model is applicable on news as well.

1. Introduction

A reader does not only read a story to get to know what happens in a text, but also because of the manner in which this narrative is presented to them (Scheffel 2013). Information gaps created by the ordering and disordering of events according to logical and temporal links are what create narrative tension and make stories engaging (Baroni 2007; Sternberg 1992). Narrative organization is particularly important in fictional texts, as literariness in fiction can be related to semantic complexity and are more likely to portray higher levels of non-linearity in comparison to non-fiction (Piper and Toubia 2023; van Cranenburgh et al. 2019). Our objective is to develop a computational model able to detect the events of a fictional text in the way in which the reader has learned about them (Genette 1980; Scheffel 2013). An intermediate goal is to create a theoretical model for the annotation of events in fiction in several languages, keeping into account the many challenges posed by literary language and narrative strategies.

Automated event detection has been a field of interest in natural language processing (NLP), linguistics, journalism, history, and literature (Caselli and Bos 2023; Norambuena et al. 2023; Santana et al. 2023; Sprugnoli and Tonelli 2017). However, despite this broad interest in automated event detection, the definition of events differs greatly across scholarly works on events due to the different objectives for the task in the different fields and between different research projects.

In this article, we give an overview of related research and the manner in which these works are related to our theoretical model for the annotation of events in fiction. First, we give an overview of research on automated event detection in literature, news texts, and historical texts (See [section 2](#)). Then, we elaborate on the definition of *syuzhet* — the concrete order in which events are *presented* (Scheffel 2013) — and provide a theoretical background on the different definitions of events in narratology, concluding with our operationalization of literary events (See [section 3](#)). This approach differs from automated event detection in, for example, NLP and journalism, as these mainly focus on the *fabula*, the chronological or causal order of the events represented in the text (Scheffel 2013). After an overview of related works in literary event detection (see [section 4](#)), we will compare our framework to a narratology-based framework developed on news (Vossen et al. 2021) to demonstrate how our theoretical model for fiction differs from frameworks in news.

2. Related work

Event detection has been a research topic in a multitude of domains, such as journalism, history, and literature (Lai 2022), using NLP and information extraction (IE) techniques (Santana et al. 2023). Despite the wide range of research conducted on events, adapting previous work to a new domain is complex, for example, due to the scarcity of corpora annotated with temporal information in historical texts (Sprugnoli and Tonelli 2017).

Another challenge is the lack of a general definition of events in (Sprugnoli and Tonelli 2017) and across (Caselli and Bos 2023; Santana et al. 2023) domains. In NLP, event detection is defined as the task of finding all pairs of linguistic expressions $(w_i, w_j) \in D$, in which D is a given document, w_i is an instance of an event trigger, and w_j is an instance of an event participant (Caselli and Bos 2023). The event triggers are defined as linguistic expressions that depict the happening of something, or a state. The event participants are expressions concerning the actors, location, and time of occurrence. Thus, by this definition, events represent complex relationships between people, places, objects, actions, and states.

Because of the definition of events as complex relationships, events, and storylines can be expressed as knowledge graphs (Kishore and He 2024; Wadhwa et al. 2024; Yan and Tang 2023). Yan and Tang 2023 introduce EventTKG, a narrative graph generation framework which can be used to generate storylines based on news and other media streams. They distinguish events from complex events, where an event is defined as something that happens at a specific time and place, carried out by an individual or organization. Complex events are clusters of events concerning the same topic that are also considered to be the basic elements of a storyline as a storyline is a chronologically arranged sequence of events. Despite this broad definition of event, complex event, and storyline, Yan and Tang 2023 conclude that EventTKG can be applied only to a limited number of news datasets and that real-world events are also too complex for this framework. Therefore, the applicability of this framework to fiction appears limited.

Another approach is using large language models (LLMs) to generate event sequences based on an event knowledge graph with partial causal relations (Wadhwa et al. 2024) or track the context of sentences and events (Miori and Petrov 2024), which can then be

used in event knowledge graphs. Using LLMs in the development of event sequences and event knowledge graphs is promising, but the bias in an LLM can influence event extraction. For example, Kishore and He 2024 show that GPT-3.5 has a bias towards “AFTER” in a question-answer format concerning the chronological sequence of two events in a given text, whereas GPT-4 has a preference for “BEFORE.” When assessing truthfulness on the chronological order of events in a given text, GPT-3.5 has a bias towards “TRUE,” whereas GPT-4 tends towards “FALSE.”

In addition to these limitations, we need to consider that the definition of event in NLP as a linguistic expression of a relationship between a happening or state and an actor, location, or time of occurrence (Caselli and Bos 2023), do not provide a way to distinguish different sequences of the same events, e.g. the *fabula* vs. the *syuzhet*. In NLP, the goal of event extraction is mainly to derive and represent the events occurring in a text, so that the events and the text can be easily analyzed, visualized, and searched. The relationship between event triggers and event participants described by Caselli and Bos 2023 and applied in most NLP work, only links the *what* to actors (*who*), location (*where*), and time of occurrence (*when*). However, when focusing on how a reader learns about the events in a text, the construction of narrative events is rather modeled as the relation between the fictional world (i.e., the *what* of narration) and its representation in the text (i.e., the *how* of narration) (Gius and Vauth 2022), for which using event categories based on their eventfulness are more suitable (Gius and Vauth 2022; Hühn 2009, 2013).

This basic theoretical difference makes it difficult to compare and relate previous work in NLP event detection to the goal of event detection in fiction. However, in subsection 2.1 and 2.2, we discuss different NLP techniques used in event extraction on news and historical texts, with the goal of showing more in detail to what extent these works can complement a narratological approach. In section 3 an introduction to literary events is given and an overview of event extraction from literature is discussed.

2.1 News

In this section, we discuss the issues identified in the comparability between different works in two recent surveys (Caselli and Bos 2023; Norambuena et al. 2023) and by discussing the data structure of events proposed in Vossen et al. 2021, since it is one of the most elaborate ones.

Caselli and Bos 2023 find that variation in the definition of events and the annotation of linguistic realization (s), and the assignment of events to specific semantic classes, make most of the event-labeled corpora incompatible with each other. They give an overview of six event-annotated news corpora, which all use a different event definition. The majority of these corpora restrict the annotation of events by solely annotating events that occur in given event classes. These restrictions make these frameworks unsuitable for fiction. For example, ACE (Doddington et al. 2004) only annotates events in news articles when occurring in one of eight semantic classes (Life, Movement, Conflict, Business, Contact, Personnel, Justice, Transaction). In contrast, TimeML (Pustejovsky et al. 2005) rejects restrictions on semantic classes and linguistic realizations of events, as annotations are based on the lexical aspect and their contextual syntactic structure. As TimeML is aimed to portray events as temporal expressions relative to each other,

this approach is not applicable in the analysis of *syuzhet*. 107

Norambuena et al. 2023 identify two fundamental units in news narrative extraction, 108
events and *entities*, i.e., the actions and happenings in the text and the characters and other 109
entities that are related to the events. Focusing on the former, they define *computational* 110
narrative representation as a discrete story structure, such as a graph or a timeline of events. 111
They observe that the most common and simple way to computationally represent a 112
narrative is as a linear sequence of events, such as a timeline. 113

Since the survey only analyses research using news corpora, they assume that each 114
text (news article) focuses on one single main event. Previous or secondary events, 115
which for example can be used to link articles together, are not taken into account in this 116
survey. As previous and secondary events are crucial in fiction, this assumption is not 117
applicable to literary event detection. They identify three scopes: events as sentences, 118
events as entire documents, and events as a cluster of documents. This is a broader view 119
of events than in many other approaches, for example, TimeML defines events as more 120
specific to an action, such as a perception. 121

Among these seemingly incompatible approaches, there are also two that leverage 122
insight coming from narratological scholarship. The first one is Vossen et al. 2021, 123
who propose a framework informed by narratology and argue that a plot structure is 124
composed of three elements: (1) an *exposition*, in which the characters and the setting 125
are introduced, (2) a *predicament*, which consists of a set of struggles or problems that an 126
actor has to go through (3) and the *extrication*, which is the end of the predicament. The 127
predicament itself consists of three elements: (1) *rising action*, which consists of events 128
that increase the tension (2) *climax*, which consists of events where the tension reaches 129
its maximum, (3) *falling action*, which consists of events that resolve the climax and 130
lower tension. Besides these dynamic patterns, they define also three data structures: 131
the timeline based on the *fabula*, which they define as a chronological timeline, the 132
causeline, related to the plot, which they define as a set of loose and strict causal relations 133
and the *storyline*, which they define as a set of (pairwise) relations between events 134
according to the patterns mentioned above and is associated to the plot structure. The 135
storyline includes the explanatory causal relations between events that are related to a 136
climax event that have the strongest connection with the climax event. The events in the 137
storyline are chronologically ordered. In annotation, every event mention is associated 138
with a temporal expression or direct temporally related to other events in the timeline. 139
In the causeline only events that express a loose causal relation are included. Based on 140
the causelines, the storyline depicts explicit additional explanatory relations, that may 141
lead to a climax event. In section 5 we will compare this framework to our approach of 142
analyzing narrative events. 143

The second NLP work looking at narratology — as well as at Critical Discourse Analysis 144
(CDA) — is by Huang and Usbeck 2024, who propose a theoretical framework to 145
construct new narratives from an author-focused perspective. CDA considers news 146
narrators as a dominant group that shapes a narrated world encoded in language, in 147
which real-world events are portrayed, to the public. Therefore, the focus is on how 148
real-world events are organized to shape a narrated world, using an adapted definition 149
of *fabula* and *discourse* by Gervás and Calle 2024. They consider the information flow 150
from a real-world event to a news item as follows: first, based on a real-world event a 151

subset of an organized event sequence forms the fabula, then the discourse is created through narrative composition, simplified as causal relations between the events in the fabula, and lastly, the discourse is used to form textualized narratives in natural language. They define fabula as “the actual sequence of events, that is chronologically and causally ordered” and discourse as “the product of the telling, which reorganizes the chronological and causal order of this sequence. They view the narrated world as event-event causal relations and narration as a function that shapes the narrated world. They consider events as the smallest unit in a narrative, but do not consider all events in a text to be part of the narrated world. Indeed, they make a distinction between constituent and supplementary events, of which only the former are represented as event-event causal relations. The proposed theoretical framework represents this information flow as the narrated world logic, which can be used to extract the core story of events told by a news narrator. As this is a proposal for a theoretical framework that has not been evaluated yet, it is unclear how effective it is and whether this framework is applicable to fiction.

To conclude, in the task of event detection in news there is no general consensus on the definition of event. This lack of consensus shows that relying on existing frameworks and corpora does not lead to broadly applicable annotations, as the different corpora are hard to compare and relate to each other (Caselli and Bos 2023). Moreover, most corpora restrict events to some semantic event classes but this is too restrictive for a comprehensive analysis of the *syuzhet*.

2.2 Historical texts

The lack of a general consensus on the definition of events does not only occur with event extraction in news texts, but also with historical texts. Additionally, the aim of event extraction from historical texts is not focused on information extraction only, but also on the analysis and interpretation of events. To solve the difference of objectives between fields, and to make NLP techniques applicable to historical texts in such a way that it will lead to a more homogeneous usage of event extraction in historical research, Sprugnoli and Tonelli 2017 suggest using the expertise of historians for the linguistic annotation of events.

Sprugnoli and Tonelli 2019 conclude from their discussions with historians that the semantic type of an event is the most relevant information for annotation, that multi-token annotation of event phrases should be possible, and that events can have different syntactical forms and grammatical classes. Accordingly, they define 22 relevant semantic classes, based on the semantic categories of the Historical Thesaurus of the Oxford English Dictionary (HTOED), aiming to avoid too much granularity while at the same time ensuring broad informativeness. The latter is important due to the diverse topics and genres in historical texts.

They consider three different types of events spans: (1) single-token, (2) multi-token, and (3) discontinuous expressions. Events can be verbs, past participles, present participles, adjectives, nouns, and pronouns. Multi-token events are restricted to seven types of linguistic construction, such as phrasal verb constructions, final and non-finite verbs, and nouns.

The resulting annotated corpus, the Histo Corpus, is used to train two different classifiers: 195
 CRF classifiers and a BiLSTM. Two CRF classifiers were implemented: one to identify 196
 the event span and the other to predict the correct event class on unseen text. The 197
 BiLSTM is used for sequence tagging as well as event detection and event classification. 198
 Overall, the BiLSTM outperforms the CRF classifiers in event classification, except for 199
 the event class physical sensations. 200

In another project (Verkijk and Vossen 2023) historians have been involved in the de- 201
 velopment of an ontology that can be used for event extraction from the archives of 202
 the Dutch East India Company (VOC). The ontology should enable the extraction of 203
 implied events, as this is deemed to be important by experts. They used the CEO on- 204
 tology (Segers et al. 2017), which models semantic circumstantial relations between 205
 event classes, as the basis for the definition of event classes, since they want to be able 206
 to annotate static events. They identified three relevant types of observable events: ship 207
 movement, trade, and (geo)political/social relations. More detailed classes, for example, 208
 whether an action is legal or illegal, depend on the context and the interpretation of an 209
 expert, and are therefore not considered an observable event. Building on FrameNet 210
 (Ruppenhofer et al. 2016) and CEO, they define participants specific to each event class. 211
 Other event arguments are spatial or temporal. Roles can be recycled from one event to 212
 another, for example, the Agent in an *Attacking* event is a Patient in the state *BeingInCon-* 213
flict. Results show good agreement between human annotators for the labeling of event 214
 triggers, but poor performance of fine-tuned models for automated event detection 215
 (Verkijk et al. 2024). 216

From this type of research, we can observe that event annotation in historical texts differs 217
 greatly from approaches to annotate events in literature. Both Sprugnoli and Tonelli 218
 2019 and Verkijk and Vossen 2023 use predefined semantic classes and themes to identify 219
 and analyze events, while considering a multitude of syntactical forms and grammatical 220
 classes. However, for research on literature, all events in the text are relevant because 221
 they can fulfill different functions that cannot be defined in advance (Pianzola 2018). 222
 Some events contribute to creating the setting for the story, other events contribute to 223
 the progression of the plot, others contribute to show the personality of the fictional 224
 characters. All events potentially play a role in the cognitive and aesthetic processing of 225
 literary text that readers do (Caracciolo 2014). 226

3. Literary events 227

Our goal is a definition of narrative event that can be broadly operationalized (Pichler 228
 and Reiter 2022) for the automatic detection of events in literary texts. Thus, similar 229
 to Sprugnoli and Tonelli 2017, we need to create a domain-specific framework that 230
 contributes to bridging the gap between NLP research and its techniques to analyze 231
 events on the one hand, and our domain, computational literary studies, on the other. 232
 Additionally, it would be ideal to define narrative events in a manner that is operational- 233
 izable across different languages. Many scholars in literary studies and narratology 234
 have addressed the concept of event, trying to define its constitutive properties and 235
 the role of events in stories. The main difference from NLP research is probably the 236
 conceptualization of different event categories (see subsection 3.1) and event sequences 237

(see [subsection 3.2](#)).

3.1 Event categories

Events can be considered the smallest units that make up a narrative. An event can also be seen as a change of state, any type of expressed change that contributes to the narration (Hühn 2013). To distinguish what can be considered to be a change of state, and therefore an event, Hühn 2013 distinguishes two types of events, based on the context in which the concept of event is used: (1) “a type of narration that can be described linguistically and manifests itself in predicates that express changes (event I), and (2) an interpretation- and context-dependent type of narration that implies changes of a special kind (event II), on the other.” Both *event I* and *event II* portray a basic type of narration and are characterized by a change of state, the transition from one situation to another, usually in relation to a character. *Event I* and *event II* are distinguished by the degree of specificity of the change of state. *Event I* changes of state consist of any change of state that contributes to the narrative, defining narrativity as the “relation of changes of any kind” (Hühn 2013). *Event I* concerns every type of change of state expressed in a text, whereas *event II* refers to specific changes of state that meet additional conditions, such as changes that are decisive, unpredictable turns in the narration or a deviation from the norm of what is expected. The evaluation of the additional conditions of *event II* is a matter of interpretation, and therefore *event II* is a hermeneutic category. On the contrary, *event I* can be evaluated rather objectively.

The definition of narrativity used in *event II* differs from the definition of narrativity used in *event I*. In *event II*, narration is considered to be the “representation of changes with certain qualities” (Hühn 2013). Whether these qualities are present is dependent on context and interpretation of the events in relation to the whole text. For example, “Mary stepped onto the ship” contains a type I event, namely the change of state of the character Mary by moving from the bank to the ship, resulting in a change of surroundings. However, in the context of a particular literary or cultural context, such as emigration, this can also be a type II event. Emigration can be seen as a new beginning and is therefore a deviation from what is expected. Therefore this example can also be a *event II* change of state, depending on the literary and cultural context. *Event II* changes of state are considered to be more or less eventful, according to what extent they meet the following five criteria: relevance, unpredictability, effect, irreversibility, and non-iterativity (Hühn 2013). These additional criteria are also predominantly dependent on cultural, historical, or literary context. Therefore the eventfulness of a change can be interpreted differently by different readers. Besides different event types, different event sequences have been conceptualized too.

3.2 Fabula and Syuzhet

The Russian formalist Viktor Shklovsky introduced the terms *fabula* and *syuzhet* (Scheffel 2013), by analyzing the difference between chains of events in “actual life” and in art. Shklovsky argues that to understand the “aesthetic laws” of artistic narrative, the distinction between *fabula* and *syuzhet* is necessary. He defines *syuzhet* as “the material of the *fabula* in the artistic form.” In other words, the *fabula* represents what has happened or what was in the narrated world, whereas *syuzhet* is the artistic form in which the

fabula is presented to the reader. *Fabula* is defined as “the material for *syuzhet* formation,”
a chronological chain of events.

The *fabula/syuzhet* distinction is similar to the *story/plot* and *histoire/discourse* distinction
(Pier 2003; Scheffel 2013). *Story* is “a narrative of events arranged in their time sequence”
(Scheffel 2013). For example, dinner comes after breakfast and Tuesday after Monday.
Plot is a narrative of events focused on causality, for example “The king died, and then
the queen died of grief.” In the *plot* a causal relation between events is established,
whereas in the *story* the relationship is only chronological. More broadly, *plot* involves
the transformation of “happenings” to a sequence of structured events that form a
narrative (Xin 2022).

Similar to Shklovsky, Todorov identifies two aspects of literary works: *histoire* and
discourse. A literary work is

at the same time a story [*histoire*] and a discourse [*discours*]. It is story, in the
sense that it evokes a certain reality [...]. But the work is at the same time
discourse [...]. At this level, it is not the events reported which count but
the manner in which the narrator makes them known to us (Scheffel 2013).

The difference between *fabula/syuzhet* and *histoire/discours* is mainly found in the artistic
value prescribed to the different terms. Todorov considers both *histoire* and *discourse*
as important aspects of a literary work, as the *histoire* is necessary to create a certain
reality for the reader. *Discours* is important, as literariness is not solely about the events
reported, but also about the manner in which the narrator presents them to the reader.
Discours also considers features such as perspective, style, and mode, whereas the
syuzhet primarily focuses on the order of events represented in a text. Additionally
histoire contains the continuum of the narrated world, in contrast to *fabula* that only
contains the parts of the narrated world that are relevant to the plot. Due to their
broader definitions, *histoire* and *discours* are considered to be of equal literary value,
whereas the *fabula* is considered not to be of literary value, and the artistic value of a
text is represented solely in the *syuzhet*. Moreover, the interplay of the two sequences,
with flashback and anticipations, generates a narrative tension, the narrativity that
keeps readers engaged (Baroni 2007; Sternberg 1992). The automatic detection of both
sequences is a difficult task, but computational literary studies have a unique interest
in the way in which events are presented and can complement efforts done in NLP to
detect the “*histoire*” of news. However, it is also worth noting that there are NLP works
interested in some aspects of the “*syuzhet*,” mostly in relation to the framing of events
(Hamborg 2023; Minnema et al. 2022a).

3.3 An operationalization of literary events

Given the specific interest of computational literary studies in the way in which events
are presented, an operational model for the automatic detection of events in literary
texts should enable the extraction of information not only about the semantics of events
but also their rhetorical, narratological, and literary functions. To this end, Gius and
Vauth 2022 started from operationalizing the narrativity of event representation at the
level of discourse, using German prose as a case study.

Gius and Vauth 2022 define four different event categories that can be called *event I* in

the context of Hühn:	324
1. Changes of state are physical or mental states' changes of animate or inanimate entities	325 326
"Gregor Samsa one morning from uneasy dreams awoke"	327
2. Process events are actions or happenings that do not result in a change of state, such as moving, thinking, feeling	328 329
"found he himself in his bed into a monstrous insect-like creature transformed"	330
3. Stative events are physical and mental states of animate or inanimate objects	331
"His room lay quietly between the four well-known walls"	332
4. Non-events do not relate to facts in the narrated world, such as general statements, questions or hypothetical situations	333 334
"She would have closed the door to the apartment".	335
The four different types of events were chosen in order to differentiate them for narrativity analysis and define events as "any change of state explicitly or implicitly represented in a text." The events are ordered by degree of narrativity, in which <i>changes of states</i> have the highest degree of narrativity and <i>stative events</i> the lowest narrativity. <i>Non-events</i> do not contribute to narrativity, but are included for comprehensive annotation. Gius and Vauth 2022 consider the whole text when annotating events. However, they aim to avoid "relatively strong interpretations necessary when primary relating to the story world 'behind' its representation in the narrative" (Gius and Vauth 2022). The hierarchy of narrativity of the four types of event categories ensures that the representation of eventfulness in discourse is reflected in the annotation. This indirect annotation of eventfulness is more aligned with the different types of eventfulness related to <i>event II</i> (Hühn 2013). One of the approaches of eventfulness discussed by Hühn 2013 requires that a change actually takes place in the narrated world (thus is a fact in the narrated world) and that it reaches a conclusion (thus the change cannot only be described to have begun or be in progress). This definition of eventfulness is similar to Gius and Vauth 2022's definition of <i>change of state</i> . However, as they annotate every event occurring in the text, and additionally non-events, their overall definition of event categories is broader than that allowed by <i>event II</i> and in line with <i>event I</i> .	336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353
Aiming at a broader applicability of the model of Gius and Vauth 2022, we have modified their guidelines with extra examples and edge cases from English fiction. Our multilingual corpus consists of fiction, specifically fanfiction. Four of the added examples can be found below:	354 355 356 357
1. Change of state:	358
[The baking sheet sighed a bit,] ₁ [beginning to relax.] ₂ PROCESS EVENT ₁ + CHANGE OF STATE ₂	359 360
2. Process event:	361
[you are on a path in the woods]	362

3. Stative event: 363
[unsure what to make of a scene] 364
4. Non-event: 365
[“You **need** to make friends, Ryeowook ah,”]₁ [he **had** said over the 366
dinner table]₂ NON EVENT₁ + PROCESS EVENT₂ 367

The first example shows the importance of the duration of a motion, as the first part of 368
the sentence “the baking sheet sighed a bit” is a process event, whereas the second part 369
“beginning to relax.” is a change of state. As sighing is a short-lasting motion, it is a 370
process event. In the second part, the finite verb is *beginning*, which implies that this 371
phrase marks a longer-lasting change in the character state, namely relaxation. 372

Our corpus also displays a great variety in the type of narrators used. For example, in a 373
text written from the point of view of a second person narrator, the sentence “you are 374
on a path in the woods” (example 2) is a *process event*. The finite verb in this sentence 375
is *are*, which implies that the character in the sentence (*you*) is in motion, because the 376
next sentence in the text is “at the end of the path is a cave,” which suggests that the 377
characters have reached the cave. 378

A third notable case we observed in our corpus is the use of implied verbs. Despite 379
the missing verb in example 3, “unsure what to make of a scene,” this has still been 380
annotated, as the words “he was” are implied in the context of the full text. The inclusion 381
of implied verbs is particularly important for the applicability of a definition of events 382
to multiple languages as not all languages are as verb-focused as English and German. 383
For example in Bahasa Indonesia, it is possible to form a grammatically correct sentence 384
that does not contain any verbs, as auxiliary verbs do not exist in Bahasa Indonesia. 385

The fourth example shows the influence of dialogue in fiction, where the first part of the 386
sentence “You need to make friends, Ryeowook ah” is spoken. Since this is an opinion 387
stated by the speaker, this is a non-event, as the sentence does not relate to a fact in the 388
narrated world. The second part of the example, “he had said over the dinner table,” is 389
a process event, as the verb focuses on the action of saying the first sentence, which is 390
an action. 391

The fact that events reported in dialogue are labeled as non-events is quite limiting, 392
because it is not unusual that readers get to know about happenings in the story world 393
through the voices of different characters. The four different categories of events pro- 394
posed by Gius and Vauth 2022 are not enough for a complete account of all events in a 395
story. To fill this gap, we have introduced complementary labels for speech and thought. 396
Since the goal is to give a fine-grained representation of *how* events are presented, we 397
decided to work with *four* additional secondary labels that distinguish between direct 398
and indirect reports: 399

1. Direct speech. Example: “Man, am I tired!” 400
2. Indirect speech. Example: Man, was he tired! 401
3. Direct thought. Example: “I’m tired!” he thought. 402
4. Indirect thought. Example: He thought he was tired. 403

The aim of adding these four extra labels, is to be able to analyze in what way speech and thought are used to present events and narrate a story. For example the phrase: “–You looked through my phone!” would get two labels: *process event* and *direct speech*. The usage of speech and thought in a narrative also influences the certainty and uncertainty of the occurrence of an event. In the phrase: “But they don’t want to be friends with me, Appa”, the reader will perceive it as the speaker’s opinion that they don’t want to be friends with them. If this information would be stated by the narrator, this would be perceived as a fact. These complementary labels could therefore be interesting in the analysis of framing and the presentation of information in fiction, but also in news.

Additionally, these labels can be combined with other narrative features for a more nuanced analysis. For example, with labels for the type of narrator (first-, second-, third-person narrator) or focalization, the different points of view from which the action is looked at (Jahn 2021). The presentation mode of events can influence the reader’s epistemic stance towards their occurrence. For instance, when events are conveyed through speech, thoughts, or dreams, the reader’s confidence that they actually took place may be diminished. Having distinctive labels for thoughts is useful as thought presentation occurs in two contexts. First, it can show that the narrator had direct access to relevant thoughts (Semino and Short 2004), either as a third-person omniscient narrator expressing the thoughts and mental states of the characters in a text, or a first-person narrator presenting their own thoughts and mental states. In the second context, the narrator does not have such access, but infers the character’s thoughts based on external evidence, such as a person’s speech, facial expressions, and actions (Semino and Short 2004). Thought presentation, in particular indirect thought, is also associated with the creation of feelings of closeness and empathy by the reader for the characters. Thus, adding these four extra labels to the event categories, enables a more thorough analysis of the *syuzhet* of a text. The perception of events in the *syuzhet* is influenced not only by their narrativity but also by presentation modes and focalization. Operationalizing the annotation and classification of events in literary texts taking into account all these variables would be the best-case scenario for a computational narratology of events. However, this has not been done yet by research on literary event detection.

4. Literary event detection

In this section, we provide an overview of event detection in literature, discussing whether and how the various approaches could be applied to several languages and complemented by other methods.

The model with four event types has been used by Vauth et al. 2021 to annotate four German prose texts (Vauth and Gius 2021) and automatically classify events by following a two-step process. First, they extract verb phrases, which are labeled with an event type in the second step. Since the annotation guidelines in Vauth and Gius 2021 focus on the finite verb of the sentence, the verb phrase extraction is done by selecting the finite verbs in each sentence using a pre-trained tagger. Then, for each verb, the dependency tree of a pre-trained parser is used to identify all tokens they cover, by traversing the tree. Relative clauses are not considered when moving down the dependency tree, and neither are conjunctions if their children consist of full verbs. On unseen data, the model

reached a 0.71 F1 score in identifying the correct span and a 0.78 F1 score in classifying the event type. However, Vauth et al. 2021 only used German prose and therefore relied on a German pre-trained tagger and parser. Suitable pre-trained taggers and parsers will need to be selected for other languages, to test this approach.

4.1 Literary events as *realia*

Sims et al. 2019 define a literary event as an event that is actually happening in the text (*realis*), with the goal of analyzing the narrative plot. In this model there are no stative events, they only consider activities, achievements, accomplishments, and changes of state, following Vendler 1957. A phrase is considered an event by Sims et al. 2019 if either (1) a change of state has occurred, (2) the cause of a state can be deduced, or (3) the phrase refers to an acute mental state, such as acute short lasting responses like *shocked* or *astonished*. This specification of the types of events is in line with *event II*, in which an event is defined as a “representation of changes with certain qualities” (Hühn 2013). As a consequence of these requirements, phrases are more often considered not to be events than in Vauth and Gius 2021 or our guidelines. For example, the sentence “at the end of the path is a cave” is a stative event for us, but it would not be an event according to Sims et al. 2019.

Similar to the guidelines by Vauth and Gius 2021 and ours, events must have occurred, thus negations are not considered to be events, nor are possible future events. Generic phrases are also considered not to be events in all three guidelines. However, Sims et al. 2019 do not treat hypothetical phrases considering wishes and desires as events, whereas Vauth et al. 2021 and us consider the act of wishing as stative events. Another difference is that Sims et al. 2019 consider single words as events, whereas for Vauth et al. 2021 and us, all words that can be assigned to a finite verb are included in the annotation span of an event. Lastly, Sims et al. 2019 define their event triggers more broadly, including not only verbs but also adjectives and nouns. This approach has the advantage of being extendable to languages whose syntax does not rely on verbs as much as English does, but it also has the limitation that Vendler’s verb classes are not applicable to many languages.

For event detection, they use an LSTM and five BiLSTMs both on the annotated verbs only (baseline) and on a featurized model containing six extra features of information of the token: lemma, part of speech tag, context, syntax, WordNet synset and hyponymy information, word embeddings, and bare plurals as subjects. The differences between the BiLSTMs are the context included in the BiLSTM, including a sentence CNN, document context, and BERT contextual representations. The BiLSTM with BERT representations on the featurized model has the highest performance, with an F-score of 73.9.

4.2 Hylistic analysis

Pannach 2023 analyses events in folktales using the hylistic theory (Zgoll 2020). A hyleme is an individual statement containing events and states in chronological order. For example the statement ‘Orpheus came to his end by being struck by a thunderbolt’ results in the following hyleme sequence, which consists of three parts: (1) “Orpheus is struck by a thunderbolt,” (2) “Orpheus dies,” and (3) “Orpheus is dead.” This model

does not include aspects related to how the events are presented, it rather focuses on achieving the best possible comparability between different variants of the same folktale, even across languages. That is why the events are translated into present-tense statements that describe precise actions or states. Additionally, Pannach uses four main categories in her hylestic analysis: single-point (punctual), durative-constant, durative-initial and durative-resultative, which are mainly associated to verbs in a phrase. Single-point hylemes consist of active actions, passive experiences, reactions, perceptions, or feelings. The beginning and end of the event are both included in the hyleme sequence. Durative hylemes are true for part of the sequence or the entire hyleme sequence. Durative-initial hylemes are true at the beginning of a sequence, durative-constant are true during the entire sequence, and durative-resultative at the end of the sequence.

Pannach 2023 compares this approach to the event model of Gius and Vauth 2022 and Vauth and Gius 2021. The change of state event category of Gius and Vauth 2022 corresponds to the single-point category used by Pannach 2023. Process events are also considered to be single-point. However, when the property of the event is iterative, such as “Charon works the sails,” the phrase would be considered to be durative-constant. Stative events correspond to the three durative hylestic classes, which class it belongs to depends on the context. Non-events are not annotated in the hylestic classes.

The vast majority of the annotated data consists of single-point statements. Due to the unequal distribution, as well as the similarity between the three different durative hylestic classes, a multinomial naive Bayes model was used, with a TF-IDF vectorizer. Three classifiers were implemented, one binary classifier distinguishing single-point and durative hylemes, and one classifying durative-initial, durative-constant and durative-resultative hylemes and one considering all four classes. The binary classifier has a 0.79 F1 score for the durative hylemes and a 0.92 F1 score for the single-point hylemes. The second classifier has a 0.32 F1 score of the durative initial statements, a 0.85 F1 score for the durative constant, and a 0.56 F1 score for the durative-resultative. It is important to note that 69% of this test set consists of durative-constant hylemes and 24% of durative-resultative hylemes. For the third classifier, the durative-initial hylemes have a 0.25 F1 score, the durative-constant hylemes a 0.69 F1 score, a 0.43 F1 score on the durative-resultative hylemes, and a 0.93 F1 score on the single-point statements. In this test set, distribution across the different classes is again unbalanced, as the test set only contains 30 durative initial hylemes and contains 1,151 single-point statements. As it is unclear if the class imbalance in the test set of the second and the third classifier is reflected in the respective training sets, it is hard to determine how this imbalance has influenced the results, and if this influences the strong preference for the single-point hylemes by the third classifier.

4.3 Analyzing narrative discourse with Large Language Models

Piper and Bagga 2024 uses LLMs to analyze narrative discourse within the framework of Genette 1980’s narrative triangle concerning story, discourse, and narrating instance. They use three categories to analyze narrative discourse: (1) “POV (Point of View),” focused on the experiencing agent; (2) Time, including use of tense, anachrony, flashbacks, eventfulness, and event sequences; and (3) Setting, including location and concreteness

(realized and tangible space). Thus, they explicitly use event sequences and eventfulness as features to capture dimensions of time.

They prompt LLMs to estimate the degree of presence of a given feature using a three-point scale. The dataset of Piper and Bagga 2022 is used to collect 13,543 passages from 18 genres, including contemporary novels, short stories, folktales, and non-fiction such as memoirs and stories from AskReddit. The experiments were run on a subset of passages with a manually annotated narrativity score higher than 3.0. The evaluation consists of four steps: (1) replication, (2) honeypot, (3) inter-annotator agreement, and (4) model performance. First, 15 iterations are run on half of the validation data. For the best model, 95.6% replication occurs in all documents. Secondly, “honeypot” a nonsensical feature is used of which the answer should never be positive. This feature is used to measure to what extent a model is randomly guessing. In the best model, all nonsensical prompts were answered negatively. Thirdly, three annotators answered identical prompts the models’ received. The inter-annotator agreement is fair, with a Fleiss’s kappa= 0.38 and a universal agreement rate of 43%. Lastly, the model’s accuracy is evaluated by comparing the model’s results to the majority vote of the human annotators and the minimum match, where the results are compared with any human answer regardless of majority vote.

There is a variance in the overall model’s F1 score from 0.28 to 0.79 of the majority vote, but a higher performance for the minimum match, with a highest F1 score of 0.95 and four out of six models with an F1 score of 0.87 or higher. The annotator agreement correlates strongly with model performance. Thus, LLMs are a promising tool in the analysis of narrative discourse, specifically since the results show that the features event sequences and eventfulness can have different weights in classifying narrative. As the high variance across models is also seen between human annotators, the results emphasize the subjectivity and ambiguity in the task.

5. Comparison of computational narratology in NLP and literary studies

To better illustrate the differences between approaches, we compare our narratological model to a narratology-inspired approach to NLP event extraction (Vossen et al. 2021) (see subsection 2.1), which proposes three data structures (or sequences): timelines, causelines, and storylines. For the comparison, we annotate a sample of news (originally used in Vossen et al. 2021) and a sample offiction from our corpus. The goal is to show how the domain-specific interests of computational literary studies and NLP for news analysis can lead to different operationalizations of narratological concepts.

5.1 Timeline, causeline and storyline

Figure 1 shows the news sample from Vossen et al. 2021. The temporal relation between all events is expressed in the timeline, whereas only the loose causal relations are included in the causeline, and only explicit explanatory relations that may lead to the climax event are included in the storyline. Figure 2 shows the fiction sample, annotated according to Vossen et al. 2021. Figure 2 shows that timelines, causelines, and storylines

Police **say**_{e1} that on Saturday around 11:30 p.m. Kimani Gray was **standing**_{e2} outside his home with five other young men before **splitting off**_{e3} when he **noticed**_{e4} two plainclothes officers in an unmarked car. After he “**adjusted**_{e5} his waistband and continued to **act**_{e6} in a suspicious manner,” officials **say**_{e7} the cops got **out**_{e8} of their car and **approached**_{e9} Gray — who allegedly **turned**_{e10} toward them with a loaded .38-caliber revolver in hand. The 30-year-old sergeant and 26-year-old **fired**_{e11} shots.

- *timeline*: [NOW] → includes → **say**_{e1}; **say**_{e1} → before → **say**_{e7}; **say**_{e1} → after → [Saturday around 11:30 p.m.]; [Saturday around 11:30 p.m.] → includes → **standing**_{e2}; **standing**_{e2} → before → **splitting off**_{e3}; **splitting off**_{e3} → simultaneous → **noticed**_{e4}; **noticed**_{e4} → before → **got out**_{e8}; **act**_{e6} → before → **got out**_{e8}; **got out**_{e8} → before → **approached**_{e9}; **approached**_{e9} → simultaneous → **turned**_{e10}; **turned**_{e10} → before → **fire**_{e11};
- *causelines*: **act**_{e6} → circumstantial → **approached**_{e9}; **splitting off**_{e3} → circumstantial → **noticed**_{e4}; **turned**_{e10} → circumstantial → **fire**_{e11}
- *storyline*: **noticed**_{e4} → rising_action → **splitting off**_{e3} → rising_action → **adjusted**_{e5} → rising_action → **act**_{e6} → rising_action → **approached**_{e9} → rising_action → **turned**_{e10} → rising_action → **fired**_{e12[climax]};

Figure 1: Example of the timeline, causeline and storyline framework applied on news from Vossen et al. 2021

do not fully reflect the story presented in fictional texts. Firstly, fiction contains more 574
description (of for example surroundings) than news. The timeline of the news sample 575
shows a clear temporal order of events in the text, whereas the temporal order for the 576
description of the grove and the way in which the wolf is stretched out are not explicitly 577
expressed. It can be assumed that the splitting of the grove was created before the stone 578
was placed there, however, it is also possible that the stone was first placed there and 579
the trees grew around it. In genres such as science fiction and fantasy, the environment 580
is not necessarily static, thus complicating expressing all events in a timeline. 581

Secondly, the causeline does not contain the description of the grove, the stone and the 582
way in which the wolf is stretched out. Therefore, this description is not included in the 583
storyline, as the storyline is based on the causelines. However, despite not being part 584
of the causal relations between events, the description of the grove, the stone, and the 585
wolf does contribute to the narrative, since it helps the reader to imagine the scene and 586
contributes to the build-up of suspense, the tension leading to the climax. 587

Lastly, the storyline that can be derived from the causeline stops at the event **froze**_{e7}, 588
which is the climax of the storyline. Half of the events occurring in the sample, namely 589
those related to the description of the grove and the wolf, are not included in the storyline. 590
However, due to the emphasis on the description of the grove, the stone and the wolf, 591
the wolf dying appears to be crucial to the narrative. The description of the scene also 592
contributes to the build-up of suspense, thus the event **froze**_{e7} is not actually a climax 593
(according to Vossen et al. 2021), as there is no falling of action or resolution afterward. 594
Additionally, readers could conclude from this excerpt that the death of the wolf is more 595
important to the narrative than Wilson walking towards and discovering the dead wolf, 596
whereas the storyline only portrays the movements of Wilson. 597

When Wilson first **heard**_{e1} the sounds of the dying wolf through the corpse of trees, he had **pulled**_{e2} the hunting rifle off his shoulder and **approached**_{e3} warily, **expecting**_{e4} the scene to include fighting foxes, or a stray dog that had **wandered**_{e5} into a snake nest. When he **saw**_{e6} the huge shape of the wolf, he **froze**_{e7}, **unsure**_{e8} of what to make of the scene. The grove of trees **split**_{e9} into a small clearing, and in the center of the circle of grass was a **stone**_{e10} about as tall as Wilson's waist. The wolf was **stretched out**_{e11} over the top of the stone, head pointed one direction, feet in the other. The stone was **covered**_{e12} in enough blood that it had **dripped**_{e13} down the side of the stone and **coated**_{e14} the dirt around the rock.

- *timeline*: [NOW] → after → **heard**_{e1}; **heard**_{e1} → before → **saw**_{e6}; **pulled**_{e2} → after → **heard**_{e1}; **pulled**_{e2} → simultaneous → **approached**_{e3}; **heard**_{e1} → simultaneous → **expecting**_{e4}; **expecting**_{e4} → simultaneous → **wondered**_{e5}; **saw**_{e6} → before → **froze**_{e7}; **saw**_{e6} → after → **stretched out**_{e11}; **saw**_{e6} → simultaneous → **unsure**_{e8}; **split**_{e9} → before → **heard**_{e1}; **stone**_{e10} → before → **heard**_{e1}; **stretched out**_{e11} → before → **covered**_{e12}; **covered**_{e12} → before → **dripped**_{e13}; **dripped**_{e13} → before → **coated**_{e14};
- *causelines*: **heard**_{e1} → circumstantial → **pulled**_{e2}; **heard**_{e1} → circumstantial → **expected**_{e4}; **saw**_{e6} → circumstantial → **froze**_{e7}; **saw**_{e6} → circumstantial → **unsure**_{e8}; **covered**_{e12} → circumstantial → **dripped**_{e13}; **dripped**_{e13} → circumstantial → **coated**_{e14};
- *storyline*: **heard**_{e1} → rising_action → **pulled**_{e2} → rising_action → **approached**_{e3} → rising_action → **saw**_{e6} → rising_action → **froze**_{e7} [*climax*]

Figure 2: Timeline, causeline and storyline framework by Vossen et al. 2021 applied on fiction

[Police **say**]_{process} [that on Saturday around 11:30 p.m Kimani Gray **was** standing outside his home with five other young men before splitting off]]_{stative & indirect speech} [when he **noticed** two plain officers in an unmarked car.]_{process & indirect speech} [After he “**adjusted** his waistband”]_{process & direct speech} [and **continued** to act in a suspicious manner,”]_{process & direct speech} [officials **say**]_{process} [the cops **got** out of their car]_{process & indirect speech} [and **approached** Gray]_{process & indirect speech} — [who allegedly **turned** toward them with a loaded .38 revolver in hand.]_{process} [The 30-year old sergeant and 26-year-old **fired** shots [...]]_{change of state}

Figure 3: News example of the annotation of narrative events. The bold verbs are the finite verbs per annotation span. Note that the word event is omitted from the annotation labels for abbreviation.

[When Wilson first **heard** the sounds of the dying wolf through the corpse of trees,]_{process} [he **had** pulled the hunting rifle off his shoulder]_{change of state} [and **approached** warily,]_{process} [**expecting** the scene to include fighting foxes,]_{non} [or a stray dog that **had** wandered into a snake nest.]_{non} [When he **saw** the huge shape of the wolf,]_{stative} [he **froze**,]_{change of state} [unsure what to make of the scene.]_{stative} [The grove of trees **split** into a small clearing,]_{stative} [and in the center of the circle of grass **was** a stone about as tall as Wilson's waist.]_{stative} [The wolf **was** stretched out over the top of the stone,]_{stative} [head pointed one direction,]_{stative} [feet in another.]_{stative} [The stone **was** covered in enough blood]_{stative} [that it **had** dripped down the side of the stone]_{change of state} [and **coated** the dirt around the rock.]_{change of state}

Figure 4: Fiction example of the annotation of narrative events. The bold verbs are the finite verbs per annotation span. Note that some events contain implied finite verbs and that the word event is omitted from the annotation labels for abbreviation.

5.2 Narrative events

598

In figures [Figure 3](#) and [Figure 4](#), the two samples are annotated following our definition 599 of narrative events. When comparing the storylines of the news and fiction sample to 600 the annotation of narrative events, it is evident that the build-up and rise in action to a 601 climax (as defined by Vossen et al. 2021) can be related to the narrative events model. 602 According to this model all events are processes, except for Gray standing outside and 603 the firing of the shots. Thus, process events in the text seem to build up to the same 604 climax event, which is annotated as a change of state. In [Figure 1](#) the storyline starts 605 with **noticed**_{e4}, whereas [Police **say**] is annotated as a process event. 606

The firing of the shots is described as a change of state, which puts the emphasis on the 607 police agents shooting at Gray. It is a change of state as the finite verb of the sentence is 608 *fired*. One of the distinguishing properties between changes of state and process events 609 is irreversibility. If the finite verbs consider an irreversible change, the corresponding 610 phrase is a change of state, as the irreversible change has led to a permanent property 611 change of an entity. Firing shots is such an irreversible change, as one cannot reverse 612 firing a shot. An alternative phrasing of the event reported in the last sentence could 613 have a different event type. For example, the same event could be presented from the 614 perspective of Gray (like in the second sentence “he noticed”): “Gray heard gunshots.” 615 This sentence would be annotated as a process event, as the finite verb is *heard* and 616 emphasizes describing a perception. 617

This can be related to research in which semantic frames are used to analyse perspective 618 and framing in news (Minnema et al. 2022b). For example, in the following headline 619 “Cyclist, 70s, seriously injured following collision in Dublin,” the word collision triggers 620 the frame “impact,” showing that the main event in the sentence describes the impact 621 on the cyclist. The same event has also been described with the following sentence: 622 “Driver hits pedestrian with his car, sending the 70-year old man to hospital with heavy 623 injuries.” In this headline, hits is the trigger of the frame “cause_impact,” which shows 624 that the main event in this headline is expresses the cause of the impact, namely the 625 driver causing the injures. 626

The first headline would be annotated as a stative event according to our framework, 627 as the finite verb *injured* describes the physical state the cyclist is in. The low level of 628 narrativity corresponding with this narrative event also corresponds with the frame 629 “impact,” as the impact is described without naming the agent that has caused the 630 accident. The second headline is a process event, as the finite verb is *hits*, which describes 631 a motion. This corresponds with a higher level of narrativity, which is suitable with the 632 frame “cause_impact,” as this emphasizes the action that caused the impact. 633

In the fanfiction sample, the different event categories fluctuate (see [Figure 4](#)). The text 634 starts with a process event, then the level of narrativity moves up to a change of state, 635 and then goes down again to a process event and two non-events. Next, a stative event 636 is followed by a change of state. Then several stative events and two changes of state 637 conclude the paragraph. This fluctuation in level of narrativity cannot be seen in the 638 storyline in [Figure 2](#) as only the first change of state is shown in the storyline. 639

6. Discussion

To sum up, our model of narrative events can be applied to fiction as well as non-fiction, such as news, and covers both semantic aspects (event types) as well as rhetorical and narratological aspects (presentation modes) that play a crucial role in how events are perceived by readers. Our goal was to propose a general model for the automatic detection of narrative events, as the overview of related work shows that the lack of consensus on a definition of events in NLP has led to a wide variety of frameworks and applications that are hard to compare and relate to each other, making it difficult to adapt an existing approach for events in news to literary texts. Whereas research on historical events has mainly focused on developing frameworks that enable the application of NLP research and techniques on historical texts, we have focused on developing a broad definition of narrative events that can be used by literary scholars, as well as other domains. The current limitation is that we still focus on verbs to select the textual span of an event. We are currently experimenting with using our guidelines for annotations on six more languages (Bahasa Indonesia, Dutch, Italian, Korean, Mandarin Chinese, Spanish) and we will modify the guidelines to be applicable more broadly.

Our comparison between the framework by Vossen et al. 2021 and our model of narrative events shows that the annotation of narrative events can be applied to news and is similar to the rise in action to a climax point as described in the storyline. On the contrary, Vossen et al. 2021's framework has strong limitations when applied to fiction, as the rise in action portrayed in the storyline does not align with the fluctuation in action and level of narrativity seen in fiction.

In the future, it would be interesting to analyze further to what extent our model of narrative events can be applied to various languages and domains. Specifically, we showed that the analysis of narrative events as part of the *syuzhet* can contribute to research on framing in news. This line of research has the potential to show how computational literary studies can make a meaningful contribution to NLP research that goes beyond the semantics of texts.

7. Acknowledgements

This work is part of the Graphs and Ontologies for Literary Evolution Models (GOLEM) project, a 5-year (2023-2027) research project funded by the European Commission.

8. Author Contributions

Noa Visser Solissa: Conceptualization, Methodology, Writing – original draft

Andreas van Cranenburgh: Supervision, Writing – review & editing

Federico Pianzola: Supervision, Methodology, Writing – review & editing

References

Baroni, R. (2007). *La tension narrative : suspense*. Paris: Seuil.


- Caracciolo, Marco (2014). *An Enactivist Approach*. Berlin, Boston: De Gruyter. ISBN: 9783110365658. <https://doi.org/10.1515/9783110365658>.
- Caselli, Tommaso and Johan Bos (2023). "Investigating interoperable event corpora: limitations of reusability of resources and portability of models". In: *Language Resources and Evaluation* 57.3, 1107–1137.
- Doddington, George R, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel (2004). "The automatic content extraction (ACE) program-tasks, data, and evaluation." In: *Lrec*. Vol. 2. 1. Lisbon, 837–840.
- Genette, Gérard (1980). "Narrative discourse: An essay in method". In: *Cornell UP*.
- Gervás, Pablo and José Luis López Calle (2024). "Representing Complex Relative Chronology Across Narrative Levels in Movie Plots." In: *Text2Story@ ECIR*, 65–76.
- Gius, Evelyn and Michael Vauth (2022). "Towards an Event Based Plot Model. A Computational Narratology Approach". In: *Journal of Computational Literary Studies* 1.1.
- Hamborg, Felix (Feb. 2023). *Revealing media bias in news articles*. 1st ed. Cham, Switzerland: Springer International Publishing.
- Huang, Junbo and Ricardo Usbeck (2024). "Narration as Functions: from Events to Narratives". In: *Proceedings of the The 6th Workshop on Narrative Understanding*, 1–7.
- Hühn, Peter (2009). *Event and eventfulness*. De Gruyter, 159–178.
- (Sept. 2013). *Event and eventfulness*. <https://www-archiv.fdm.uni-hamburg.de/lhn/node/39.html>.
- Jahn, Manfred (2021). *Narratology 2.3: A Guide to the Theory of Narrative*. www.uni-koeln.de/~ame02/pppn.pdf.
- Kishore, Sindhu and Hangfeng He (2024). "Unveiling Divergent Inductive Biases of LLMs on Temporal Data". In: *arXiv preprint arXiv:2404.01453*.
- Lai, Viet Dac (2022). "Event extraction: A survey". In: *arXiv preprint arXiv:2210.03419*.
- Minnema, Gosse, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim (Nov. 2022a). "Dead or Murdered? Predicting Responsibility Perception in Femicide News Reports". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang. Online only: Association for Computational Linguistics, 1078–1090. 10.18653/v1/2022.aacl-main.79. <https://aclanthology.org/2022.aacl-main.79/>.
- (2022b). "SOCIOFILLMORE: a tool for discovering perspectives". In: *arXiv preprint arXiv:2203.03438*.
- Miori, Deborah and Constantin Petrov (2024). "Narratives from GPT-derived networks of news and a link to financial markets dislocations". In: *International Journal of Data Science and Analytics*, 1–25.
- Norambuena, Brian Keith, Tanushree Mitra, and Chris North (2023). "A survey on event-based news narrative extraction". In: *ACM Computing Surveys* 55.14s, 1–39.
- Pannach, Franziska (2023). "'Orpheus Came to His End by Being Struck by a Thunderbolt': Annotating Events in Mythological Sequences". In: *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, 10–18.
- Pianzola, Federico (2018). "Looking at Narrative as a Complex System: The Proteus Principle". In: *Narrating Complexity*. Ed. by Richard Walsh and Susan Stepney. Cham:

- Springer International Publishing, 101–122. ISBN: 978-3-319-64714-2. [10.1007/978-3-319-64714-2_10](#). 723–724
- Pichler, Axel and Nils Reiter (Dec. 2022). “From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities”. In: *Journal of Cultural Analytics* 7.4. 725–727
- Pier, John (2003). “On the Semiotic Parameters of Narrative: A Critique of Story and Discourse”. In: *Questions and Answers Regarding the Status of a Theory*. Ed. by Tom Kindt and Hans-Harald Müller. Berlin, New York: De Gruyter, 73–98. ISBN: 9783110202069. [doi:10.1515/9783110202069.73](#). [https://doi.org/10.1515/9783110202069.73](#). 728–731
- Piper, Andrew and Sunyam Bagga (2022). “Toward a data-driven theory of narrativity”. In: *New Literary History* 54.1, 879–901. 732–733
- (2024). “Using Large Language Models for Understanding Narrative Discourse”. In: *Proceedings of the The 6th Workshop on Narrative Understanding*, 37–46. 734–735
- Piper, Andrew and Olivier Toubia (2023). “A quantitative study of non-linearity in storytelling”. In: *Poetics* 98, 101793. 736–737
- Pustejovsky, James, Robert Ingria, Roser Sauri, José M Castaño, Jessica Littman, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani (2005). *The Specification Language TimeML*. 738–740
- Ruppenhofer, Josef, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk (2016). *FrameNet II: Extended theory and practice*. Tech. rep. International Computer Science Institute. 741–743
- Santana, Brenda, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes (2023). “A survey on narrative extraction from textual data”. In: *Artificial Intelligence Review* 56.8, 8393–8435. 744–746
- Scheffel, Michael (Sept. 2013). *Narrative Constitution*. [https://www-archiv.fdm.uni-hamburg.de/lhn/node/57.html#](#). 747–748
- Segers, Roxane, Tommaso Caselli, and Piek Vossen (Aug. 2017). “The Circumstantial Event Ontology (CEO)”. In: *Proceedings of the Events and Stories in the News Workshop*. Ed. by Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, Martha Palmer, Eduard Hovy, Teruko Mitamura, and David Caswell. Vancouver, Canada: Association for Computational Linguistics, 37–41. [10.18653/v1/W17-2706](#). [https://aclanthology.org/W17-2706/](#). 749–754
- Semino, Elena and Mick Short (2004). “Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing”. In: Routledge. Chap. Thought presentation in the corpus. 755–757
- Sims, Matthew, Jong Ho Park, and David Bamman (2019). “Literary event detection”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3623–3634. 758–760
- Sprugnoli, Rachele and Sara Tonelli (2017). “One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective”. In: *Natural language engineering* 23.4, 485–506. 761–763
- (2019). “Novel event detection and classification for historical texts”. In: *Computational Linguistics* 45.2, 229–265. 764–765
- Sternberg, Meir (1992). “Telling in time (II): Chronology, teleology, narrativity”. In: *Poetics Today* 13.3, 463. 766–767

- van Cranenburgh, Andreas, Karina van Dalen-Oskam, and Joris van Zundert (2019). "Vector space explorations of literary language". In: *Language Resources and Evaluation* 53.4, 625–650.
- Vauth, Michael and Evelyn Gius (2021). "Richtlinien für die Annotation narratologischer Ereigniskonzepte". In: *Zenodo*.
- Vauth, Michael, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann (2021). "Automated Event Annotation in Literary Texts." In: *CHR*, 333–345.
- Vendler, Zeno (1957). "Verbs and times". In: *The philosophical review* 66.2, 143–160.
- Verkijk, Stella, Pia Sommerauer, and Piek Vossen (June 2024). "Studying Language Variation Considering the Re-Usability of Modern Theories, Tools and Resources for Annotating Explicit and Implicit Events in Centuries Old Text". In: *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*. Ed. by Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Marcos Zampieri, Preslav Nakov, and Jörg Tiedemann. Mexico City, Mexico: Association for Computational Linguistics, 174–187. [10.18653/v1/2024.vardial-1.15](https://aclanthology.org/2024.vardial-1.15/). <https://aclanthology.org/2024.vardial-1.15/>.
- Verkijk, Stella and Piek Vossen (2023). "Sunken Ships Shan't Sail: Ontology Design for Reconstructing Events in the Dutch East India Company Archives". In: *CEUR Workshop Proceedings*. CEUR Workshop Proceedings, 320.
- Vossen, Piek, Tommaso Caselli, and Roxane Segers (2021). "A narratology-based framework for storyline extraction". In: *Computational Analysis of Storylines: Making Sense of Events* 125, 125–140.
- Wadhwa, Somin, Oktie Hassanzadeh, Debarun Bhattacharjya, Ken Barker, and Jian Ni (2024). "Distilling Event Sequence Knowledge From Large Language Models". In: *International Semantic Web Conference*. Springer, 237–255.
- Xin, Wendy Veronica (2022). "Plot". In: *Fictionality and Literature: Core Concepts Revisited*. Ed. by Lasse R. Gammelgaard, Stefan Iversen, Louise Brix Jacobsen, James Phelan, Richard Walsh, Henrik Zetterberg-Nielsen, and Simona Zetterberg-Nielsen. The Ohio State University Press. Chap. 3.
- Yan, Zhihua and Xijin Tang (2023). "Narrative graph: Telling evolving stories based on event-centric temporal knowledge graph". In: *Journal of Systems Science and Systems Engineering* 32.2, 206–221.
- Zgoll, Christian (2020). "Myths as Polymorphous and Polystratic Erzählstoffe". In: *Mythische Sphärenwechsel: Methodisch Neue Zugänge zu antiken Mythen in Orient und Okzident*, 9–82.

Towards a perspectival moral history of the novel using LLMs

Andrew Piper¹ 

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

Citation

Andrew Piper (2025). "Towards a perspectival moral history of the novel using LLMs". In: *CCLS2025 Conference Preprints 4* (1). [10.26083/tuprints-00030151](https://doi.org/10.26083/tuprints-00030151)

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-01-30

Keywords

novels, large language models, fiction, narrative archetypes, ethical criticism, world literature, wikidata

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. This paper introduces a new framework for studying the moral history of the novel through the lens of large language models (LLMs). Drawing on over 9,000 Wikipedia plot summaries of 20th- and 21st-century novels, it demonstrates how LLMs can surface the implicit life lessons—or story morals—encoded in narrative summaries at scale. Building on recent work in moral inference and narrative abstraction, the study proposes a reflexive, perspectival approach that emphasizes interpretation over taxonomy. To account for the semantic variability of LLM-generated morals, the study employs a randomized prompt assignment strategy and analyzes the resulting moral keywords using co-occurrence networks and hierarchical clustering, enabling the identification of latent moral communities and comparison across modeling approaches and time. Taken together, the findings argue for the value of LLMs not only in extracting narrative values, but in enabling a new, culturally situated view of literary history through computational means.

1. Introduction

A long tradition of literary criticism has emphasized the fundamental importance of understanding the ethical concerns of stories. As Wayne Booth has argued, "All stories teach" (Booth 1998, 354). Indeed, the didactic function of storytelling – that stories have a moral or lesson to impart – is one of the oldest known functions of storytelling (Gregory 2010). Aesop's Fables are the best known version in the West, but similar types of tales exist in both Hindu (Panchatantra) and Buddhist (Jatakas) traditions that date back to around the fifth century BCE.

While we typically associate the concept of "story morals" with such traditional genres, critics like Booth (1998) and Nussbaum (1998) have argued that values-driven schemas are intrinsic to narratives more generally. As Russell and Van Den Broek (1992) argue, "Narrative schemas enable individuals to organize and represent experiences and/or events as meaningful wholes that function as the bases for comprehension and behavior." In this sense, stories need not explicitly communicate moral sentiments (e.g. "Kindness is good" or "Thou shalt not murder"). Rather, they can address general life lessons that may draw from, reinforce, challenge or extend existing moral frameworks.

This project seeks to construct a perspectival moral history of the novel by leveraging large language models to distill the central values encoded in narratives. By "moral history" I mean the implicit or explicit general life lessons conveyed by stories and storytellers over time. What does fiction teach us? And how is this historically and culturally

inflected? I use the term “perspectival” here to capture a sense of the interpretive nature of the project, that narrative values and lessons are not independent of observation but are *seen* and derived *from some point of view*.

Capturing story morals is thus tied to the longstanding narratological focus on understanding narrative archetypes or schemas (Brewer and Lichtenstein 1980; Campbell 2008; Frye 2020; Genette 1992; Propp 1968; Thompson 1955). As cognitive scientists have argued, schemas are crucial ways through which we process experience (Berns 2022). Where much of this earlier work focused on content-driven questions (“what happened?”), the attention to narrative morals focuses more on the *values* and *intentions* of the storyteller, i.e. “why was this told?” Like any schema, the story moral aims to distill an organizing principle that governs the generation and selection of narrative events and narrative perspective.

Large Language Models (LLMs) offer a potentially valuable new resource for this task given the abstractive and synthetic nature of story morals. While LLMs still suffer from hallucination with respect to fact-based extraction (L. Huang et al. 2023), they have exhibited significant progress when it comes to abstractive reasoning tasks such as narrative summarization (Subbiah et al. 2024; Zhang et al. 2024) or topic labeling (Pham et al. 2024). Indeed, deriving a story moral is in many ways analogous to the tasks of narrative summarization or topic labeling, where a model is tasked with abstracting higher-level narrative messages that are not explicitly present in the text.

Another affordance of LLMs is that given their generative nature they allow researchers to infer story morals in an unsupervised fashion, i.e. from the “bottom-up.” Rather than apply a pre-existing taxonomy that may not account for the diversity of cultural behavior, as Dundes (1962) long ago criticized, LLMs enable researchers to surface a much broader array of values and practices. This does not mean, however, that LLMs are neutral observers. They are of course “pre-trained.” They introduce yet another layer of perspective into the interpretive process that we need to account for.

In this paper, I outline a workflow for this project I am calling a perspectival moral history of the novel (Figures 1-3). It is crucial to remind ourselves of Underwood’s dictum that we do not yet have a clear understanding of the broad outlines of literary history, including the moral landscape of the modern novel (Underwood 2019). To undertake this project I engage in a series of steps of LLM-assisted narrative interpretation that move towards increasing levels of generality and structure (Figure 1). Beginning with stories themselves as interpretations of the world, it proceeds through summarization and moralization and ends with the identification of latent moral structures using co-occurrence networks and hierarchical clustering as two possible exploratory methods. As I will demonstrate, each step involves an act of perspective-taking that we need to build into the workflow.

This project utilizes Wikidata as its principal source of data, with plot summaries in particular as the primary data object. While traditional criticism may balk at using Wikipedia for literary study (or plot summaries for that matter), recent work in computational literary studies has illustrated Wikipedia to be an important resource for the study of literature, especially comparative literature. It provides one kind of “lay reader” view of literary history. As Wojcik et al. (2023) write in their preface to the special issue,

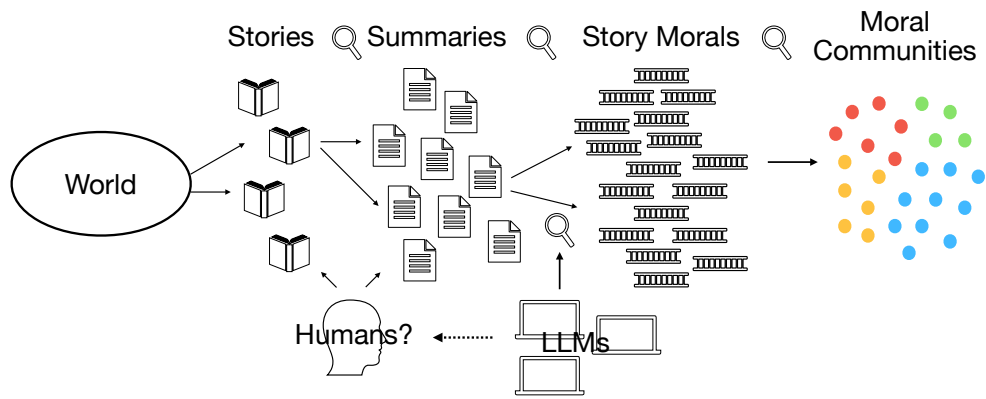


Figure 1: Overview of the story moral extraction task.

“Wikipedia, Wikidata, and World Literature”: “Despite the longstanding debate over the canon, what Wikipedia and Wikidata show us is that there is no monolithic canon, but many canons, depending on the data you choose to examine.”

The biases of Wikipedia contributors in terms of demographic distribution, for example, are well known (Wikipedia contributors 2024). As I show in Figure 5 (section 4), this affects the kinds of genres represented in the data, the time periods for which there is substantial data, and the choice of regions represented. But this is no less biased than a dataset generated by academic elites. Each provides a different perspective on literary history.

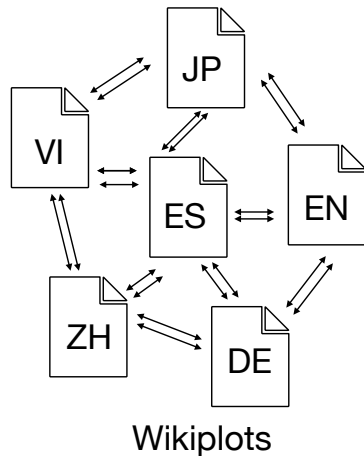


Figure 2: Schema of the many-to-many relationship of wikiplots between each language edition.

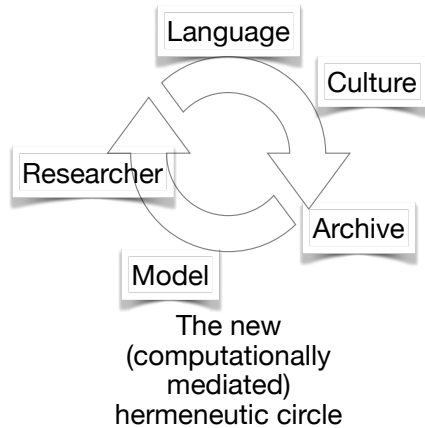


Figure 3: Schema of the new LLM-based hermeneutic circle.

But one of the fundamental affordances of Wikidata is its multilingual and multicultural nature. It is perspectival in its very nature by having multiple cultural *versions* that can potentially cover the same material. The plot summaries too play a crucial role here. They are not only instrumental for the LLM-generated story moral task. They also give us insights into cultural perspective: both in what works are chosen to be discussed and also how the stories are reflected through the act of summarization. Rather than provide a canonical summary of a canonical list of stories, Wikidata allows us to observe regional interpretations of story content through the practice of summarization and

selection (Figure 2). Each language Wiki provides a perspective not only of its own cultural artifacts (English-language summaries of stories originally written in English) but also other cultures (English-language summaries of stories originally written in Japanese and vice versa). Wikidata allows us to move past the idealized “view from nowhere” and instead contend with the idea of a “situated world literature” (Cheah 2015) (Figure 3).

For the purposes of this paper I will illustrate the workflow on a single Wikidata language set (English) and leave to future work the challenge of multilingual moral reasoning. The goal here is to demonstrate the ability of LLMs to generate common-sense based interpretations of story morals given narrative summaries as inputs. To do so, I build off of prior work validating LLMs’ capacity to generate story morals across numerous kinds of genres (Hobson et al. 2024; Zhou et al. 2024). In this paper, my focus will be on refining this workflow for this particular data and exploring the kinds of interpretive value this produces for literary historical analysis. As I hope to show, this method can generate novel insights about the moral landscape of novels at large scale.

2. Prior Work

The organization of stories into broad, overarching categories is deeply rooted in the field of narratology (Brewer and Lichtenstein 1980; Campbell 2008; Frye 2020; Genette 1992; Propp 1968; Thompson 1955). Despite addressing narratives at varying levels of abstraction, these models converge on a fundamental premise: stories inherently share common elements, and their selection is orchestrated by higher-level schemas that shape the narrative’s construction and interpretation.

One of the fundamental challenges for this work is deciding how to select and identify appropriate schemas as well as their level of generality. In the field of NLP, work related to labeling narrative schemas ranges widely across a diverse set of approaches. Early work by Chambers and Jurafsky (2009) focused on narrative schema detection focused on identifying related event chains (Sims et al. 2019; Vauth et al. 2021; Yan and Tang 2023). The chaining together of event schemas has been integral to operationalizing the concept of “plot” (Kukkonen 2014), including plot summaries and plotlines (Anantharama et al. 2022; Rashkin et al. 2020).

Other work has focused on detecting higher-level schemas such as “conflict” and “resolution” (Fermann et al. 2023), turning points (Ouyang and McKeown 2015; Piper 2015), folktale motifs (Karsdorp and Bosch 2013), story types such as “rags-to-riches” (Fudolig et al. 2023; Reagan et al. 2016), and the more traditional concept of “genre” (Dai and R. Huang 2021; Kundalia et al. 2020; Wilkens 2016).

The attention to story morals naturally draws connections to work on Moral Foundation Theory (Graham et al. 2013), one of the more popular frameworks in the social sciences for thinking about the moral perspectives of cultures. MFT posits that human moral reasoning is built upon a set of innate psychological foundations shaped by evolutionary processes. These foundations—such as care, fairness, loyalty, authority, sanctity, and liberty—underlie cultural variations in moral values and guide ethical decision-making. Work in NLP has attempted to surface moral foundations in texts such as tweets (Liscio

et al. 2022; Rezapour et al. 2019; Roy and Goldwasser 2021; Roy et al. 2023) and folk-
tales (Wu et al. 2023), as well as identifying the potential moral foundations of LLMs
(Abdulhai et al. 2023; Scherrer et al. 2024). Vida et al. (2023) provide a useful overview
of the use of “morals” as a concept within NLP research.

The key difference between the present work and prior work related to MFT or the
study of narrative archetypes is the absence of a pre-defined moral taxonomy. My aim
here is to uncover open-ended narrative-based moral frameworks using the generative
insights of Large Language Models. As Hobson et al. (2024) have shown, LLMs like GPT
produce interpretations that are both within the range of variance of human responses
and also most often preferred by independent human judges. As I will illustrate in the
next section, there are steps we can take to broaden the semantic variance generated by
LLMs to capture a wider cultural “perspective” from any given model. Future work
will have to consider the extent to which LLMs can approximate multi-lingual and
multi-cultural perspectives in their outputs. For now, however, I focus on examining
LLM reasoning about narrative morals in a single language.

3. Methods: Surfacing Story Morals Using LLMs

Hobson et al. (2024) have proposed and validated a workflow for story moral extraction
using LLMs. In that work, the authors define a “story moral” as *a general lesson that
the narrator wishes to impart to the audience about the world*. Central to this concept is the
focus on a higher order value: lessons are meant to encourage or discourage certain
behaviors, impart general wisdom to the reader, or influence their beliefs or worldview.
Story morals understood as lessons mean that they are not strictly synonymous with the
idea of moral “sentiments” (Vida et al. 2023). They focus instead on forms of behavior
and belief that may be integrated into or derived from pre-existing moral frameworks
but are not necessarily aligned with existing moral schemas.

To generate a story moral from a text, Hobson et al. (2024) use a two-level prompting
approach. They first ask the model to output the moral of a story in a single sentence
and then have the model output two keywords: one negative and one positive that
encapsulate the story moral. I modify this approach here in two ways that are relevant to
the data: first, I ask for three keywords instead of single positive and negative keyword
to allow for more overall semantic diversity; second, I include a catch for the model to
not output a story moral if the input is insufficient and also forbid the use of the word
empathy.¹ Table 1 (top) provides an overview of the base prompt structure.

One aspect not explored by Hobson et al. (2024) is the issue of variability in generative
outputs. Large language models are known to be sensitive to prompt formulation, with
even minor changes in phrasing often resulting in divergent outputs (Lu et al. 2022;
Reynolds and McDonnell 2021; Sclar et al. 2023; Webson and Pavlick 2022). This prompt
sensitivity poses challenges for both the interpretability and replicability of LLM-based
analyses, particularly in open-ended tasks such as narrative understanding or moral
reasoning.

1. While the exclusion of the word *empathy* may appear subjective, I have found that models have an over-
whelming and at times misleading affinity for this term. While this deserves further attention, as we will see
the models have no trouble substituting synonymous keywords for this value.

Prompt Structure Overview	
Unit	Prompt
Level 1	What is the moral of this story? State your answer as a single sentence. If not enough information, write NONE.
Level 2	Can you reduce this to three keywords? Don't use the word empathy.

Factorial Prompt Variants	
Factor	Levels / Description
Information Ordering	Story summary appears in Top or Bottom .
Role Framing	Present or Absent: Today, you are an expert story interpreter. I will give you a book summary and ask you a question about it.
Question Phrasing	Direct: What is the moral of this story? Interpretive: How might one interpret the moral of this story?

Table 1: Base prompting structure (Top) and experimental factors used in our 2×2×2 design (bottom) to evaluate model sensitivity to moral extraction prompts.

To assess the extent of prompt sensitivity for our moral extraction task, I conducted a controlled experiment using a random sample of 100 story summaries. Each summary was paired with eight prompting variants derived from a fully crossed 2×2×2 factorial design (N=800) (Table 1, bottom). This design systematically varied three factors that are independent of the base prompt meaning: (1) expert role framing, (2) information ordering, and (3) question phrasing. All prompts in the experiment were submitted to OpenAI’s gpt-4o-mini-2024-07-18 model via the API, using a temperature setting of 0.0 to minimize sampling variance. I show two examples of the factorial design prompt structure in Figure 4.

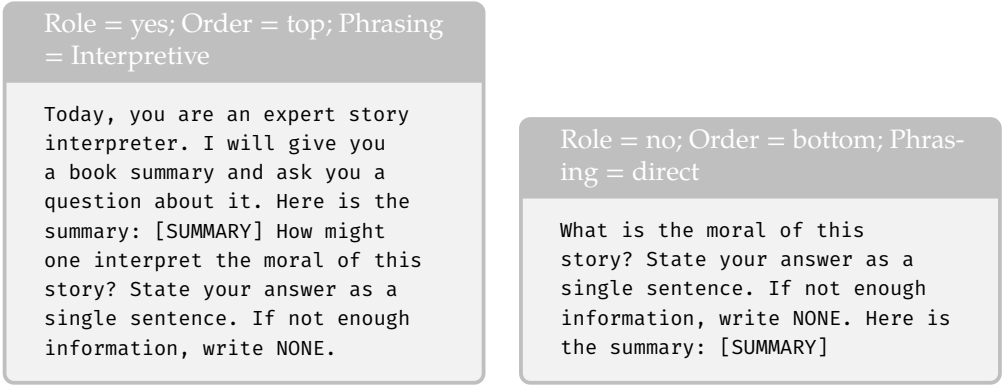


Figure 4: Examples of two prompt variants used in our 2×2×2 design. The left shows all three positive changes while the right is the original base prompt.

To quantify the effects of prompt variation, we computed pairwise Jaccard similarities between the keyword outputs generated by each prompt configuration for the same summary. This resulted in 28 pairwise comparisons across the 8 prompt variants for each of the 100 summaries. The mean Jaccard similarity across all prompt pairs was 0.38, with individual pairs ranging from 0.29 to 0.58, indicating that on average less than

40% of keywords overlapped between prompting runs on the exact same story set. The most divergent combination was `no_role-bottom-interpretive_phrasing` while the most convergent combination was `role-bottom-direct_phrasing`.

This high degree of variation across prompt types gives us a good indication of the interpretive problem LLMs introduce. Even with the same model and the same temperature, we can get divergent outcomes due to prompt structure. We can also expect this to be true at the level of the models themselves. Different models will likely provide different answers. To be sure, these answers are not independent of one another (i.e. random), nor are they in some sense inaccurate because of their variability. As Hobson et al. 2024 show, LLM story morals are generally within the variance of human responses and this consensus view is most often preferred by human judges. But the variability does tell us that the semantic space of responses is wider than a single answer would indicate. It gives us a first, valuable insight into the perspectival nature of LLMs as interpretive agents.

Given this variance, I opt not to rely on a single formulation for the moral extraction task (i.e. a “best prompt”). Instead, I aim to capture this interpretive diversity revealed by the pilot by aggregating across multiple prompt perspectives (referred to as “convergent validity” in the psychometric literature). To do this in a scalable way, I employed a randomized prompt assignment strategy: each story summary in the full dataset was paired with exactly one prompt variant, but all eight prompt types were evenly distributed across the corpus. As in the pilot experiment, prompts were submitted to OpenAI’s `gpt-4o-mini-2024-07-18` model via the API. This approach allowed me to preserve the diversity introduced by prompt wording while minimizing the computational and financial costs associated with full prompt replication across multiple models. Future work could explore additional prompt manipulation along with the inclusion of more models to capture the full range of possible interpretive diversity encoded in models. I provide examples of model outputs in Table 2.

4. Data: Wikipedia and Literary History

For the purposes of this project, I use a collection of 9,756 human-written story summaries drawn from English-language Wikipedia for books published since the beginning of the twentieth century. These summaries are based on Mark Riedl’s Wikiplots dataset (Riedl 2024), which consists of 112,936 story plots extracted from any English language article that contains a sub-header that contains the word “plot” (e.g., “Plot,” “Plot Summary,” etc.). David Bamman has provided a filtered version of this data that subsets by “creative works” and “books” to capture the genre of novels (Bamman 2024). I further filter Bamman’s version by removing any entries where the summary field is less than 40 words or overly long ($> 1,325$ words). Additionally, I filter for only publications published after 1900, which represents the bulk of summaries in the data. This yields summaries representing novels published in 116 countries, with the vast majority from English-speaking countries. The average summary is 372 words long with an inter-quartile range of 130 to 556 words. This dataset is accessible in our data repository and hereafter referred to as “Wikiplots_Novel_EN.” Figure 5 illustrates some descriptive statistics of the data.

Title	Label	Model Output
White Fang	Moral	<i>The moral of the story is that resilience, loyalty, and the capacity for change are powerful forces that can transform even the most dire circumstances.</i>
	Keywords	Resilience, Loyalty, Transformation
IQ84	Moral	<i>The moral of the story is that the quest for connection and understanding transcends the boundaries of reality and time, urging us to confront our past and the complexities of our world to find true love and purpose.</i>
	Keywords	Connection, Transcendence, Purpose
Play it as it lays	Moral	<i>The moral of the story is that the search for personal identity and meaning amidst the chaos of modern life can lead to self-destruction if not anchored by genuine connections and self-awareness.</i>
	Keywords	Identity, Chaos, Self-Destruction
The Golden Notebook	Moral	<i>The moral of "The Golden Notebook" is that the quest for personal and artistic wholeness requires confronting and integrating the fragmented aspects of one's identity and experiences.</i>
	Keywords	Integration, Identity, Wholeness

Table 2: Examples of story morals produced by GPT-4o-mini across well-known books.

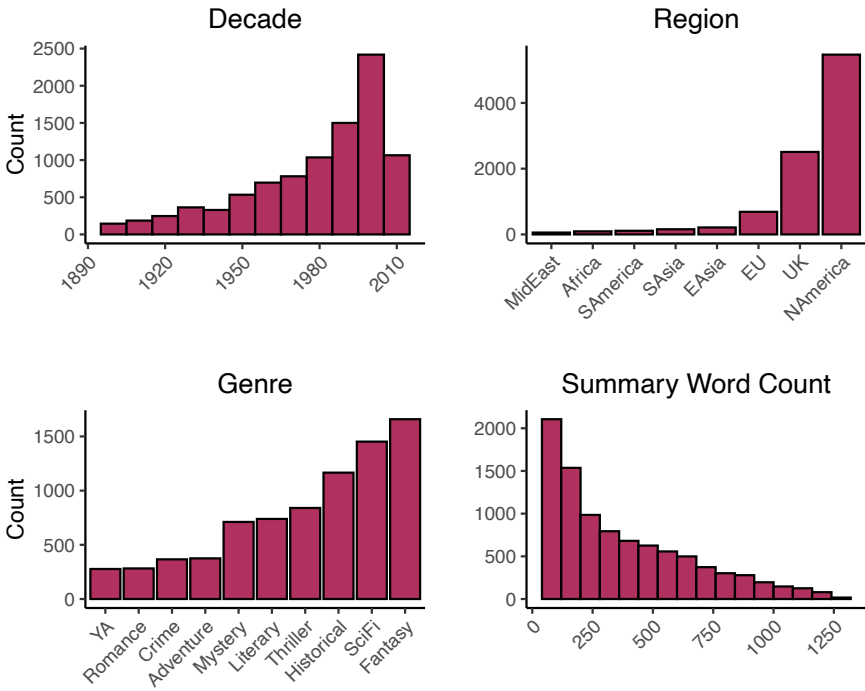


Figure 5: Overview of the Wikiplots_Novels_EN data used in this article.

One question we might ask moving forward is whether the summaries are themselves 221
reasonable representations of the books they claim to represent. As with all summa- 222
rization assessment, this is not an easy question to answer. There is no right or best 223
summary. Indeed, my research question is not principally interested in the morals of 224

the underlying books themselves, but rather the morals of the books *as they are captured* 225
by the human summaries in different Wikipedias. 226

That being said, in addition to the quality checks mentioned above (removing overly 227
short summaries and adding a prompt catch for low information) I also perform a 228
small validation study to estimate the quality of the summaries' relationship to their 229
source texts to get a rough estimate of the relationship between the summaries and their 230
sources. 231

For a subset of novels for which we have both the full text from Project Gutenberg and 232
corresponding summaries in our dataset ($N = 122$), I estimate the semantic similarity 233
between each novel and all candidate summaries. The underlying assumption is that an 234
accurate summary should be semantically closest to the book it describes, reflecting a 235
reliable condensation of its most salient content. 236

To measure semantic similarity, I divide each novel into 500-word chunks and embed 237
both the chunks and the summaries using the Sentence-BERT model `all-MiniLM-L6-v2` 238
from the `sentence-transformers` library. Each chunk is encoded into a 384-dimensional 239
embedding vector with L2 normalization enabled (`normalize_embeddings=True`) to 240
ensure comparability via cosine similarity. I then calculate the average cosine similarity 241
between all embedded novel chunks and each candidate summary, selecting the highest- 242
scoring match under both top-1 and top-3 conditions. The model achieves a top-1 243
matching accuracy of 72.80% and a top-3 accuracy of 87.20%. An error analysis of 244
mismatches suggests that length alone does not account for misattribution, indicating 245
that other factors may be influencing performance, including the coarseness of the 246
model itself. Nevertheless, this preliminary analysis suggests that an overwhelming 247
majority of summaries are indeed reflective of their source-texts and thus reasonable 248
proxies for the underlying books. 249

5. Results 250

I begin my analysis by looking at the distribution of moral keywords. The first thing we 251
can observe is the long-tailed nature of keywords with 1,383 unique moral keywords, 252
586 of those appearing just once, 408 appearing more than five times, and only 133 253
(10%) accounting for 80% of all occurrences. Table 3 provides a snapshot of the most 254
frequent keywords across the entire dataset. 255

Table 3: Top 10 most frequent moral keywords for `Wikiplots_Novels_EN`.

Keyword	Count
consequences	1138
resilience	909
identity	875
connection	837
love	779
understanding	710
courage	623
truth	592
loyalty	580
sacrifice	554

As we can see, our model and prompts provide novel insights into the high-level values associated with the modern novel as seen through the eyes of Wikipedians. One way to think about the contribution here is to contrast this taxonomy with the more traditional kinds of abstractive information such as topics that have traditionally been extracted from narratives. The story moral framework gives us a new lens to understand the narrative concerns of fiction over the past century.

One way we can deepen our understanding of this novelistic moral universe is by measuring and observing the co-occurrences of keywords for the same stories. By transforming moral co-occurrences into a network graph, we can better understand story morals at two levels of scale: 1) local semantic neighborhoods that can illustrate an individual term's meaning by identifying other terms it most often occurs with and 2) broader latent moral structures that may exist across the dataset.

To do so, I first construct a co-occurrence network from the model outputs, where nodes represent moral keywords and edges indicate how often two keywords co-occur within the same story. To improve interpretability, I trim the network by filtering low-frequency edges (< 10) and nodes (< 5) ($N=72$), and then apply multiple community detection algorithms to identify clusters of related moral concepts.

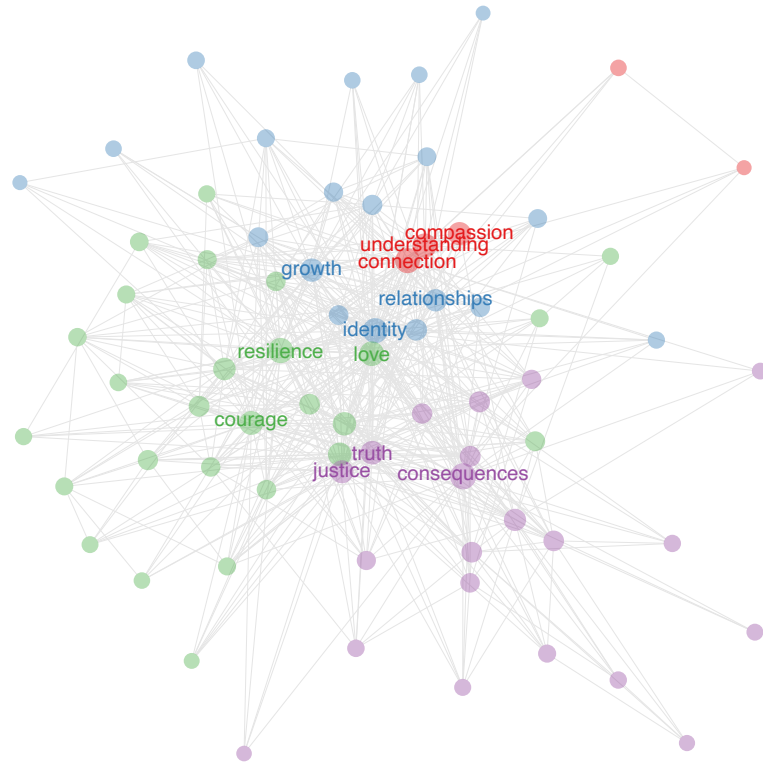


Figure 6: Co-occurrence network of moral keywords in the Wikiplots_Novels_EN corpus. Nodes represent moral concepts that appear together in story-level annotations, with edges weighted by the frequency of co-occurrence. The network is trimmed to include only edges with a frequency greater than 10 and nodes with at least five connections. Communities are identified using the Louvain method and labeled by color. Node size reflects the log frequency of each keyword, and labels illustrate the three most frequent keywords within each community.

To assess the robustness of the detected moral communities, I apply the following five

community detection algorithms to the co-occurrence network: the Louvain method yields the highest modularity (0.31) with four communities, followed closely by the Fast Greedy algorithm (0.30) which also identifies four clusters. Walktrap produces a slightly lower modularity (0.27) and divides the network into five communities. Both Infomap and Label Propagation produced only two communities and yielded the lowest modularity scores (0.17), suggesting a weaker fit to the network's structure. Overall, the convergence of Louvain and Fast Greedy on a four-community solution with relatively high modularity supports the presence of a stable latent structure within the moral co-occurrence network.

Figure 6 visualizes the co-occurrence network using a force-directed graph layout and Louvain community detection. I include the three most frequent labels for each community. The illustration helps us see greater clarity around the semantic associations of the different keywords along with larger frameworks to which they belong. If we take four communities as a reasonable estimate, we can infer high-level groupings around distinct areas of Truth/Justice, Resilience, Identity/Growth, and Compassion.

A network graph is of course only one way of surfacing latent structure within the co-occurrence matrix. Each method will shift our understanding of the moral communities by some degree. To explore the latent structure of moral keywords beyond discrete community detection, I also apply hierarchical clustering to the co-occurrence matrix (Figure 7). After filtering for keywords that appear in more than five stories ($N=408$), I compute pairwise cosine distances between normalized keyword vectors and perform agglomerative clustering using Ward's D.2 method. The resulting dendrogram reveals a multilevel hierarchy of moral groupings based on distributional similarity. To visualize how these groupings evolve across different levels of resolution, I generate a Sankey diagram showing how clusters at broader levels (e.g., $k = 2$) split into more refined subgroups at lower levels (up to $k = 6$). Cluster nodes in the Sankey diagram are labeled with their top three most frequent keywords, providing an interpretable summary of their semantic focus.

Here we see some further nuance to our network-based method. A *connection*, *love*, and *understanding* community emerges similar to the network, whereas *resilience* belongs to the *identity* and *growth* community rather than the *courage* and *perseverance* one. *Consequences*, the most frequent term overall, is located in a *power* and *ambition* cluster here with *truth* more squarely associated with its antonyms *deception* and *betrayal*.

Finally, I analyze changes in the prominence of moral clusters over time by comparing their relative frequency across decades (Figure 8). Using both network-based (Louvain) and hierarchical clustering methods, each moral keyword is assigned to a cluster and its frequency is tracked as a proportion of all moral keyword mentions in a given decade. The resulting time series visualization reveals a striking degree of stability: despite cultural and temporal shifts, the relative ordering of cluster prominence remains largely consistent within each method. Moreover, the comparison highlights important differences in semantic emphasis. In the hierarchical model, the cluster labeled by *connection* consistently dominates, suggesting a structurally central role for interpersonal and relational themes. By contrast, the network-based clustering foregrounds *resilience* as the most prominent and enduring cluster (with *resilience* second in the hierarchical model), pointing to a model of morality more centered on perseverance and individual

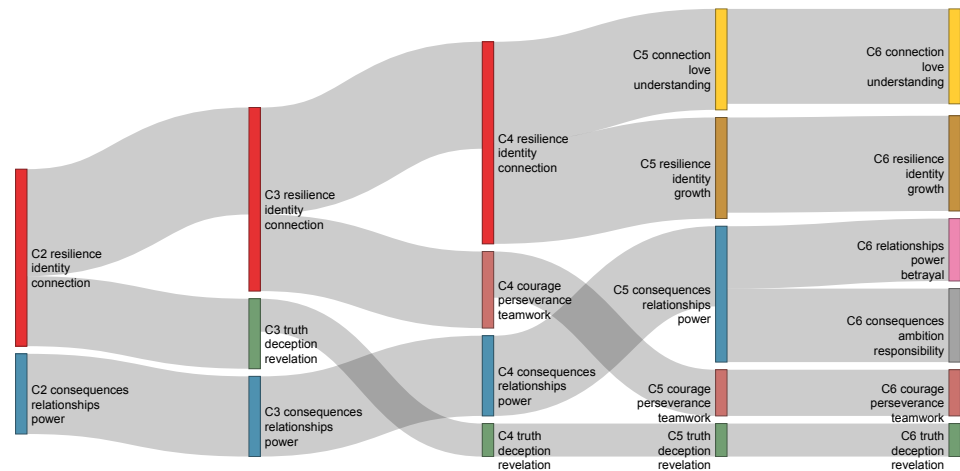


Figure 7: Hierarchical clustering of moral keywords visualized as a Sankey diagram. The diagram illustrates how clusters of moral concepts evolve across increasing levels of granularity, from $k = 2$ to $k = 6$. Each node represents a cluster of keywords identified through hierarchical clustering based on cosine distances between normalized co-occurrence vectors. Edges indicate how clusters at one level split into more fine-grained subgroups at the next. Nodes are labeled with the three most frequent keywords in each cluster. Cluster width reflects the average frequency of its top keywords.

strength. These contrasts illustrate how different modeling assumptions surface distinct moral contours within the same narrative data.

6. Conclusion

In this paper, I have endeavored to illustrate three salient points: the value of LLMs for extracting story morals at large scale, the value of Wikipedia for literary study, and the value of seeing literature through the lens of moral concerns. Each of these areas offers opportunities and challenges for future work.

As the work of Hobson et al. (2024) has shown and as we can see in section 5, LLMs offer us a reliable means of extracting high-level narrative representations that would have been unthinkable in the past. Nevertheless, even with the appearance of surface validity, it is worth pausing to ask in what ways LLMs interpretively orient us towards texts. Even though I have used a factorial variation approach to prompting and even though Hobson et al. (2024) show that LLM-generated morals are within the human range of labels, there are lingering questions about the overall semantic orientation of language models given their known cultural biases. Language models still *situate* us with respect to the text. Future work can focus on the effects of training data or fine tuning on the ways in which “story moral” inference depends on prior knowledge – and more specifically “whose knowledge.” To continue to foreground this issue of perspectivalism, we need to continue to better understand the intrinsic perspectives encoded in LLMs.

In a similar vein, there is still much more work to do to understand the large-scale insights offered by this methodology as it relates to the history of the novel. Even if we take at face value the moral outputs as reasonable approximations of “general” human judgments, what exactly do these commitments to “truth,” “resilience,” and

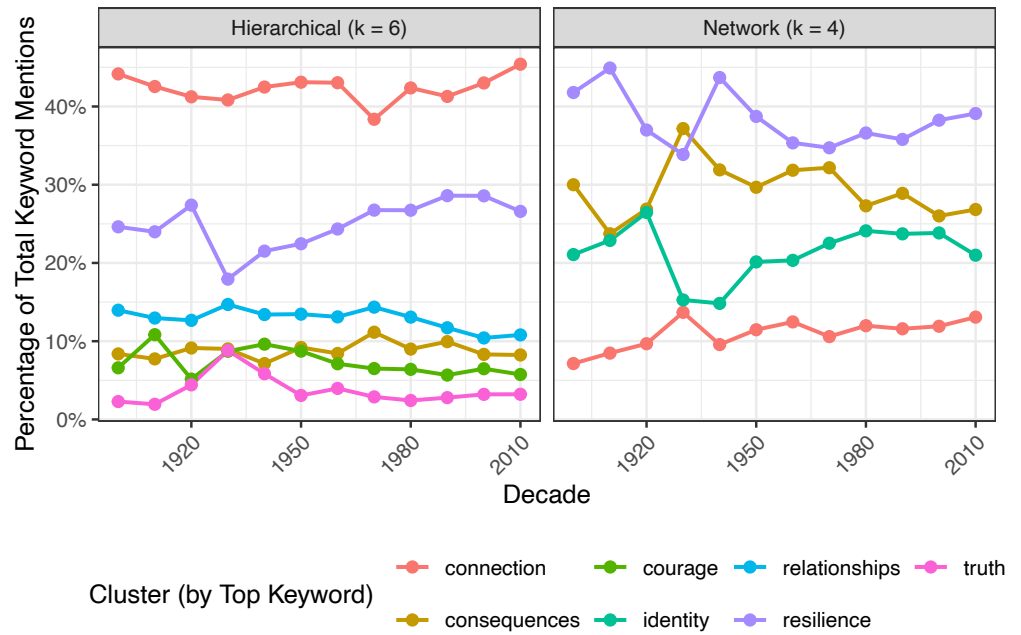


Figure 8: Relative frequency of moral clusters by decade, comparing hierarchical and network-based clustering methods. Each line represents a moral cluster labeled by its most frequent keyword, with vertical position indicating the proportion of total moral keyword mentions assigned to that cluster in each decade.

“connection” mean? Who are the principal agents of these stories? What are the common 343
settings, genres, or topics that are associated with such lessons? Are there nuances to 344
what it means to be “resilient” or who can exemplify it? And what if we go further down 345
the tree to understand novels of *redemption* or *sacrifice*? How many moral frameworks 346
are there according to the novel and how can we identify a more nuanced literary history 347
from this data? There is an opportunity here to explore methods for connecting the 348
large-scale structural insights we’ve been seeing to more granular understanding of the 349
moral concerns of novels. 350

Finally, to point in the other direction, how can we scale this workflow upwards to 351
encapsulate the multilingual level? What are the limitations and potential solutions 352
for working with less resourced languages than English when it comes to using LLMs? 353
How well can LLMs embody “cultural perspective”? Similarly, what limitations will we 354
encounter in the data when we collect multiple language versions of Wikiplots? 355

Despite these challenges, there is a tremendous amount of promise offered by LLMs for 356
the purpose of large-scale literary history and the moral history of the novel in particular. 357
Stories teach. Surfacing the kinds of lesson encoded in stories is an exciting prospect. 358
As we become less dependent on single, monolithic models, we can one day add-in a 359
further reflexive dimension where culturally specific models provide views of culturally 360
specific views of other cultures. Perspective all the way down. 361

7. Data Availability 362

Data can be found here: <https://figshare.com/s/b98d7be8802187344f81> 363

References



- Abdulhai, Marwa, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques (2023). "Moral foundations of large language models". In: *arXiv preprint arXiv:2310.15337*.
- Anantharama, Nandini, Simon Angus, and Lachlan O'Neill (2022). "Canarex: Contextually aware narrative extraction for semantically rich text-as-data applications". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3551–3564.
- Bamman, David (2024). *Wikiplots*. <http://yosemite.ischool.berkeley.edu/david/wikiplots.txt>. Accessed: May 29, 2024.
- Berns, Gregory (2022). *The self delusion: the new neuroscience of how we invent—and reinvent—our identities*. Basic Books.
- Booth, Wayne C (1998). "Why ethical criticism can never be simple". In: *Style*, 351–364.
- Brewer, William F and Edward H Lichtenstein (1980). "Event schemas, story schemas, and story grammars". In: *Center for the Study of Reading Technical Report; no. 197*.
- Campbell, Joseph (2008). *The hero with a thousand faces*. Vol. 17. New World Library.
- Chambers, Nathanael and Dan Jurafsky (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*, 602–610.
- Cheah, Pheng (2015). *What is a world?: On postcolonial literature as world literature*. Duke University Press.
- Dai, Zeyu and Ruihong Huang (2021). "A joint model for structure-based news genre classification with application to text summarization". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.
- Dundes, Alan (1962). "From etic to emic units in the structural study of folktales". In: *The Journal of American Folklore* 75.296, 95–105.
- Frermann, Lea, Jiatong Li, Shima Khanehazar, and Gosia Mikolajczak (2023). "Conflicts, villains, resolutions: Towards models of narrative media framing". In: *arXiv preprint arXiv:2306.02052*.
- Frye, Northrop (2020). *Anatomy of criticism: Four essays*. Vol. 69. Princeton University Press.
- Fudolig, Mikaela Irene, Thayer Alshaabi, Kathryn Cramer, Christopher M Danforth, and Peter Sheridan Dodds (2023). "A decomposition of book structure through ousiometric fluctuations in cumulative word-time". In: *Humanities and Social Sciences Communications* 10.1, 1–12.
- Genette, Gérard (1992). *The architext: An introduction*. Vol. 31. Univ of California Press.
- Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto (2013). "Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism". In: ed. by Patricia Devine and Ashby Plant. Vol. 47. *Advances in Experimental Social Psychology*. Academic Press, 55–130. <https://doi.org/10.1016/B978-0-12-407236-7.00002-4>. <https://www.sciencedirect.com/science/article/pii/B9780124072367000024>.
- Gregory, Marshall W (2010). "Redefining ethical criticism. The old vs. the new". In: *Proceed-*
- Hobson, David, Haiqi Zhou, Derek Ruths, and Andrew Piper (2024). "Story Morals: Surfacing value-driven narrative schemas using large language models". In: *Proceed-*

- ings of the 2024 Conference on Empirical Methods in Natural Language Processing, 12998–13032.
- Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. (2023). “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions”. In: *ACM Transactions on Information Systems*.
- Karsdorp, FB and APJ van den Bosch (2013). “Identifying motifs in folktales using topic models”. In.
- Kukkonen, Karin (2014). “Plot”. In: *The living handbook of narratology* 24.
- Kundalia, Kaushil, Yash Patel, and Manan Shah (2020). “Multi-label movie genre detection from a movie poster using knowledge transfer learning”. In: *Augmented Human Research* 5, 1–9.
- Liscio, Enrico, Alin E Dondera, Andrei Geadau, Catholijn M Jonker, and Pradeep K Murukannaiah (2022). “Cross-domain classification of moral values”. In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics (ACL), 2727–2745.
- Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp (2022). “Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8086–8098.
- Nussbaum, Martha Craven (1998). “Exactly and responsibly: A defense of ethical criticism”. In: *Philosophy and Literature* 22.2, 343–365.
- Ouyang, Jessica and Kathleen McKeown (2015). “Modeling reportable events as turning points in narrative”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2149–2158.
- Pham, Chau, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer (2024). “TopicGPT: A Prompt-based Topic Modeling Framework”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2956–2984.
- Piper, Andrew (2015). “Novel devotions: Conversional reading, computational modeling, and the modern novel”. In: *New Literary History* 46.1, 63–98.
- Propp, Vladimir (1968). “Morphology of the Folktale”. In: *U of Texas P*.
- Rashkin, Hannah, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao (2020). “PlotMachines: Outline-conditioned generation with dynamic plot state tracking”. In: *arXiv preprint arXiv:2004.14967*.
- Reagan, Andrew J, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds (2016). “The emotional arcs of stories are dominated by six basic shapes”. In: *EPJ data science* 5.1, 1–12.
- Reynolds, Laria and Kyle McDonell (2021). “Prompt programming for large language models: Beyond the few-shot paradigm”. In: *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 1–7.
- Rezapour, Rezvaneh, Priscilla Ferronato, and Jana Diesner (2019). “How do moral values differ in tweets on social movements?” In: *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, 347–351.
- Riedl, Mark (2024). *Wikiplots*. <https://github.com/markriedl/WikiPlots>. Accessed: May 29, 2024.

- Roy, Shamik and Dan Goldwasser (2021). "Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory". In: *Proceedings of the ninth international workshop on natural language processing for social media*, 1–13.
- Roy, Shamik, Nishanth Sridhar Nakshatri, and Dan Goldwasser (2023). "Towards few-shot identification of morality frames using in-context learning". In: *arXiv preprint arXiv:2302.02029*.
- Russell, Robert L and Paul Van Den Broek (1992). "Changing narrative schemas in psychotherapy." In: *Psychotherapy: Theory, Research, Practice, Training* 29.3, 344.
- Scherrer, Nino, Claudia Shi, Amir Feder, and David Blei (2024). "Evaluating the moral beliefs encoded in llms". In: *Advances in Neural Information Processing Systems* 36.
- Sciar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr (2023). "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting". In: *arXiv preprint arXiv:2310.11324*.
- Sims, Matthew, Jong Ho Park, and David Bamman (2019). "Literary event detection". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3623–3634.
- Subbiah, Melanie, Faisal Ladhak, Akankshya Mishra, Griffin Adams, Lydia Chilton, and Kathleen Mckeown (2024). "STORYSUMM: Evaluating Faithfulness in Story Summarization". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9988–10005.
- Thompson, Stith (1955). *Motif-Index of Folk-Literature, Volume 4: A Classification of Narrative Elements in Folk Tales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends*. Vol. 4. Indiana University Press.
- Underwood, Ted (2019). *Distant horizons: digital evidence and literary change*. University of Chicago Press.
- Vauth, Michael, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann (2021). "Automated Event Annotation in Literary Texts." In: *CHR*, 333–345.
- Vida, Karina, Judith Simon, and Anne Lauscher (2023). "Values, ethics, morals? on the use of moral concepts in NLP research". In: *arXiv preprint arXiv:2310.13915*.
- Webson, Albert and Ellie Pavlick (2022). "Do prompt-based models really understand the meaning of their prompts?" In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2300–2344.
- Wikipedia contributors (2024). *Wikipedians*. <https://en.wikipedia.org/wiki/Wikipedia:Wikipedians>. Accessed: 2024-05-29.
- Wilkens, Matthew (2016). "Genre, computation, and the varieties of twentieth-century US fiction". In: *Journal of Cultural Analytics* 2.2.
- Wojcik, Paula, Frank Fischer, Jacob Blakesley, and Robert Jäschke (2023). "Preface: World Literature in an Expanding Digital Space". In: *Journal of Cultural Analytics* 8.2.
- Wu, Winston, Lu Wang, and Rada Mihalcea (Dec. 2023). "Cross-Cultural Analysis of Human Values, Morals, and Biases in Folk Tales". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, 5113–5125. [10.18653/v1/2023.emnlp-main.311](https://aclanthology.org/2023.emnlp-main.311). <https://aclanthology.org/2023.emnlp-main.311>.

- Yan, Zhihua and Xijin Tang (2023). "Narrative graph: Telling evolving stories based on event-centric temporal knowledge graph". In: *Journal of Systems Science and Systems Engineering* 32.2, 206–221.
- Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen Mckeown, and Tatsunori B Hashimoto (2024). "Benchmarking Large Language Models for News Summarization". In: *Transactions of the Association for Computational Linguistics* 11, 39–57.
- Zhou, Haiqi, David Hobson, Derek Ruths, and Andrew Piper (2024). "Large Scale Narrative Messaging around Climate Change: A Cross-Cultural Comparison". In: *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, 143–155.

Exploring Measures of Distinctiveness An Evaluation Using Synthetic Texts

Julia Havrylash¹ 
Christof Schöch¹ 

1. Trier Center for Digital Humanities, Trier University , Trier, Germany.

Citation

Julia Havrylash and Christof Schöch (2025). "Exploring Measures of Distinctiveness. An Evaluation Using Synthetic Texts". In: *CCLS2025 Conference Preprints 4* (1). [10.26083/tuprints-00030152](https://doi.org/10.26083/tuprints-00030152)

Date published 2025-06-17

Date accepted 2025-04-15

Date received 2025-02-06

Keywords

evaluation, measures of distinctiveness, keyness, synthetic texts

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract. Measures of distinctiveness (aka keyness) are important tools for comparing groups of texts to identify each group's characteristic features. Evaluating these measures is essential to ensure their reliability and predictability. In our research, we developed and applied a new method for evaluating measures of distinctiveness. Our method uses a synthetically generated, homogenous text corpus to which we insert an artificial word whose frequency and dispersion are precisely manipulated. This approach allows us to determine each measure's sensitivity to variations in frequency and dispersion. Through our evaluation, we have uncovered previously unknown characteristics of these measures. Specifically, we discovered that the TF-IDF-based measure we used is more sensitive to dispersion variations than other dispersion-based measures. Moreover, we found that Eta cannot detect a word with a clear dispersion contrast when it has the same frequency in both the target and comparison groups. In our next steps, we aim to explore practical applications of this new knowledge about measures of distinctiveness.

1. Introduction

Comparing groups of texts to identify what is distinctive about each is a fundamental approach in many research contexts. In computational literary studies, such comparisons are particularly valuable for exploring literary style, genre conventions, authorial voice, or historical shifts in discourse. A key challenge in this task, alongside selecting the appropriate comparison corpora, is finding the most suitable measure and parameters for a specific research question and corpus composition. There is a wide range of measures available, and the list of most distinctive features they identify can vary considerably (as shown e.g. by Du et al. 2021a for the Zeta and Eta measures). While in principle, virtually any countable feature of texts may be submitted to a contrastive statistical analysis in order to identify distinctive features, we focus exclusively on lexical feature in this research, specifically on word unigrams. In this paper, we explore and evaluate various measures of distinctiveness, also known as keyness measures, which support such research from a quantitative perspective. Although we do not prescribe a particular measure for researchers to use, our paper offers valuable insights into the characteristics of these measures, helping researchers understand their behavior and the potential outcomes when applying different distinctiveness measures in their studies.

In the research we report on here, we focus on evaluating measures of distinctiveness

through an analysis based on synthetic texts¹. Our research proposes a new method for evaluating measures of distinctiveness, utilizing synthetically created text collections that reflect word frequencies as they would have occurred in a regular corpus built from the same original texts. Studies based on naturally occurring language must work around the fact that the frequency and dispersion of any word will vary and correlate to some extent. Our approach allows for precise, independent manipulation of word frequency and dispersion by inserting an artificial word. By conducting keyness analysis using synthetically created datasets and through inserting an artificial word with precisely manipulated frequency and dispersion into the synthetic dataset, we aim to systematically uncover the characteristics of different measures. Our goal is to determine the degree of sensitivity of each measure to variations in frequency and dispersion. Our method enables us to uncover new advantages and limitations of distinctiveness measures and compare their sensitivity to frequency and dispersion variations under consistent conditions.

The structure of our paper is as follows: We begin with an overview of previous work in the evaluation of measures of distinctiveness (section 2). Next, we describe our dataset (section 3) and provide a detailed explanation of our methodology (section 4). We then outline our hypotheses (section 5) and present the results of our evaluation (section 6). Finally, we conclude by summarizing our key findings and discussing potential directions for future research (section 7).

2. Previous work: Evaluation of keyness measures

Evaluating measures of distinctiveness is challenging due to the fact that generating a gold-standard annotation, based on which performance measures such as precision and recall can be calculated, is very difficult. Distinctiveness is not an inherent characteristic of a word, nor does it depend only on local context; rather, it can only be detected in the context of the entire target corpus while considering it in comparison to another corpus. Therefore, alternative methods of comparison and evaluation of the measures of distinctiveness are required. To tackle this challenge, several studies have attempted to evaluate distinctiveness measures using various methods.

Kilgarrieff 2001 examined corpus similarity by reviewing the mathematical characteristics of various distinctiveness measures and argued that the Chi-squared test is the most suitable in finding the most characteristic words of a corpus. Paquot and Bestgen 2009 compared three different measures in their ability to identify frequent and well distributed keywords of academic prose as opposed to fictional prose and discovered that the t-test leads to the best results for their task. Lijffijt et al. 2014 explored a broad array of measures, focusing on the statistical characteristics of these measures to identify their sensitivity to differences in word frequencies and distributions. The authors randomly sampled a text corpus into two parts in order to minimize differences in both parts and then performed a test for uniformity of p-values. Egbert and Biber 2019 introduced a distinctiveness measure based on dispersion, combining a straightforward dispersion metric with a log-likelihood ratio test. They compare the effectiveness of this approach

1. We use the term "synthetic texts" to describe texts that have been generated from documents written by humans through a specific word-level sampling procedure. These texts are therefore different both from 'naturally-occurring' text and from text generated using generative LLMs.

with corpus-frequency methods for identifying distinctive words in online travel blogs. Their study demonstrates that the dispersion-based measure outperforms the other types of measures.

Sönning 2023 evaluated 32 metrics, categorized into four dimensions of keyness. Like previously mentioned researchers, he distinguished between two primary perspectives on keyness: frequency-based and dispersion-based measures. His study assessed the effectiveness of these metrics in identifying predefined key verbs in academic writing. The results reveal significant differences among the metrics, with the Wilcoxon rank-sum test and dispersion-based measures emerging as the most effective.

The research we report on here also builds on fundamental work on measures of distinctiveness by our “Zeta and Company” project group. We conducted an in-depth analysis of the qualitative characteristics of these measures (Schröter et al. 2021). To enhance accessibility and usability, we implemented nine measures of distinctiveness in the Python package *pydistinto* (Du et al. 2021b). With Du et al. 2021a, we then introduced a new dispersion-based measure called Eta and compared it with the existing Zeta measure to highlight the advantages and disadvantages of each. Our group also performed a quantitative evaluation of nine measures on natural texts, including several dispersion-based measures, using a downstream classification task (Du et al. 2022). Our approach involved first identifying a given number of distinctive words provided by each measure for novels of a specific genre, in comparison to other literary genres. These distinctive words were then used to classify the novels by genre, with the classification accuracy obtained being a measure of each word list’s distinctiveness (in the qualitative sense of discriminatory power). We concluded that dispersion-based measures are more effective than frequency-based measures in identifying characteristic words of a target corpus.

Overall, while previous studies have provided valuable insights into distinctiveness measures, their reliance on abstract statistical analyses, intuitive evaluations, or a narrow selection of measures underscores the need for further research. Our study addresses these limitations by introducing a controlled, synthetic approach with precise manipulation of word frequency and dispersion, while also incorporating a wide range of different measures to enable a more systematic and nuanced assessment of their sensitivity. We have already conducted several analyses using naturally-occurring texts. Now, with our approach using synthetic texts, we aim to test theoretical insights about the measures under specially controlled conditions, allowing for a clearer understanding of how each distinctiveness score is calculated.

We think that using a wide variety of evaluation strategies is most likely to result in robust results, as past experience has shown that even theoretically sound and convincing arguments may not hold up to empirical scrutiny, whether quantitative or qualitative (as a case in point, consider investigations of distance-based stylometric authorship attribution; Argamon 2007, Evert et al. 2017).

3. Data

Our research is conducted on a synthetic text collection generated through random sampling, at the word level, from a corpus of French contemporary novels. The foundation for this corpus is a balanced subset from our larger collection of French contemporary popular novels and consists of 320 novels from the 1980s and 1990s. This custom-built corpus maintains equal representation (in terms of the number of novels included), per decade and across four subgroups: literary fiction, sentimental novels, crime fiction novels, and science fiction novels.

The original text corpus comprises approximately 19 million words. We load the entire corpus as a single dataset and randomly sample synthetic "novels", each with a consistent length of 40,000 words. The sampling was performed at the word level. Our newly-generated corpus contains 320 synthetic "novels", matching the number of novels in the original corpus. This approach addresses two main objectives. First, it ensures that the generated corpus reflects the word occurrences and frequencies as they can be observed in the original corpus. Second, it results in a homogeneous corpus, purposefully eliminating subgenre differences because each text is sampled from the entire corpus.

4. Methods

The objective of our analysis is to assess the hidden properties and limitations of the measure of distinctiveness in identifying distinctive words. This is achieved by applying each measure to a homogeneous synthetic corpus to which an artificial word with a controlled frequency and dispersion has been added. Systematically varying the frequency and dispersion of this word, and observing how its keyness rank in the results varies as a result, shows us to what degree a given keyness measure is sensitive to differences in frequency and/or dispersion.

In our analysis, we have analyzed all nine measures of distinctiveness implemented in our Python package *pydistinto*. The following measures have been implemented in this package: Burrows Zeta, logarithmic Zeta, Eta (Du et al. 2021a), TF-IDF (Spärck Jones 1972), Wilcoxon rank-sum test, Welch's t-test, the Ratio of relative frequencies (RRF), the Chi-squared test, and the Log-likelihood ratio test (LLR)². The implemented measures can be categorized into three distinct groups based on their approach to identifying unique keywords when comparing a target and a comparison corpus. Within this framework, the techniques employed can be classified as follows:

1. Frequency-based measures: These measures primarily focus on the frequency of the target word in the corpus, treating the corpus as a "bag of words" and disregarding how the target word is distributed within the corpus. Examples of measures falling under this classification include the RRF, the Chi-squared test, and the LLR.
2. Distribution-based measures: Rather than just considering corpus-wide mean

2. More information about our rationale for implementing this set of measures in *pydistinto*, as well as detailed descriptions of each measure, can be found in Du et al. 2022.

word frequencies, these measures are based on the distribution of a word (described e.g. via its central tendency and variability) in the corpus. Unlike simpler frequency-based measures, then, these metrics also consider variability indicators, such as standard deviation. They are also quite flexible, in that some of them don't require a normal distribution, allowing for a more nuanced comparison across different distributions. Welch's t-test falls into this category.

3. Dispersion-based measures: These measures evaluate the extent to which the target word is evenly distributed, or dispersed, across a corpus. Measures within this category encompass Burrows Zeta, logarithmic Zeta, Eta, TF-IDF (our implementation of a TF-IDF-based keyness measure), and Wilcoxon rank-sum test (with certain restrictions).³

Our approach was as follows: As *pydistinto* requires a certain format of input data (CSV format including following columns: token, lemma and POS), the original French corpus was annotated with spaCy⁴ before randomization. For the analysis with *pydistinto*, we used lemmas as the feature type. At the beginning of the process, the synthetic corpus was divided into segments of equal length, each containing 5000 words, resulting in 8 segments per novel and a total of 2560 segments. This segmentation is essential for the calculation of certain measures, such as Zeta and Eta.

Subsequently, the entire corpus was randomly divided into two sub-corpora of equal size for each run of *pydistinto*: target and comparison corpus. An artificial word⁵ was then added to both the target and comparison corpus parts with a specified frequency and dispersion. To maintain a constant total word count while adding an artificial word, each instance of the artificial word replaces one instance of an existing word in the corpus.

Our experiment was conducted in two primary settings to investigate the impact of two criteria – the frequency and dispersion of the artificial word within a corpus – on its distinctiveness score, calculated by different measures.

In the first setting, we added an artificial word to only one segment of the target and comparison corpus, albeit with varying frequency. This setting enables us to analyze the influence of only one parameter, namely the frequency. The frequency of the artificial word was set to 10 in the comparison corpus and remained constant there, while varying from 10 to 2000 words in the target corpus. We used 12 different parameters for the frequency setting in the target corpus (10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1000, and 2000). For each parameter setting, *pydistinto* was run 100 times to mitigate the impact on the results of high scores for frequent words, which may arise as a result of variation that follows from the random sampling procedure and may in turn influence the distinctiveness score of an artificial word. The corpus was randomly divided into target and comparison parts at the level of the "novels" for each run. Given the fact that texts were built by randomly sampling words from the entire corpus, and the two subcorpora were built by randomly sampling "novels" from among all "novels", any

3. Note that these latter measures are based on measures of dispersion that are not entirely uncorrelated with frequency (see e.g. Gries 2022). Detailed information about these measures can be found in Du et al. 2022.

4. See: <https://spacy.io/>.

5. An artificial word is a specially created combination of letters and numbers that cannot occur in any natural language. An example of an artificial word used in this study looks like the following: untuning55886.

difference between the target and comparison corpora, apart from the artificial word, 179
can only be due to random variation. 180

In the second setting, we experimented with the dispersion of the artificial word. In 181
this case, the frequency of the artificial word was kept constant at 1000 occurrences in 182
both the target and comparison corpus, but its dispersion varied in the target corpus 183
while remaining constant in the comparison corpus. The idea was again to isolate one 184
parameter, in this case dispersion, and analyze its influence on the performance of the 185
different measures. For the comparison corpus we used the following settings: we 186
added 1000 instances of the artificial word to just 1 segment.⁶ Dispersion variation was 187
achieved by adding the artificial word with a specified, constant total frequency to the 188
target corpus, but with varying degrees of dispersion. We conducted distinctiveness 189
analyses with variations in the target corpus according to the following schema, where 190
the first number refers to the number of segments that receive the artificial word, and 191
the second to the number of times the artificial word is included in each of the selected 192
segments: 1/1000, 2/500, 5/200, 10/100, 20/50, 50/20, 100/10, 200/5, 500/2, 1000/1. The 193
product of the two values, and therefore the total frequency, remains constant at 1000 194
(and is therefore identical to the frequency of the word in the comparison corpus), but 195
the number of segments these occurrences are spread out over is varied systematically. 196
This resulted in a total of 10 parameter settings for the dispersion experiments. Again, 197
pydistinto was run 100 times for each parameter setting. 198

Following this step, the results for each parameter setting were combined into a single 199
dataframe. Subsequently, all words in the corpus were sorted based on their distinctive- 200
ness scores, and for each measure, the rank of the word following from its distinctiveness 201
score was recorded. Each measure's performance was evaluated based on the rank of 202
the artificial word (where a rank of 1 indicates the highest distinctiveness score). 203

5. Hypotheses 204

For this evaluation experiment, we developed the following hypotheses: 205

1. For dispersion-based measures (Eta, Zeta, and logarithmic Zeta, Wilcoxon rank- 206
sum test), we hypothesize that they should not show any variation in scores when 207
frequency changes while dispersion remains constant. 208
2. However, dispersion-based measures should be sensitive to even minimal vari- 209
ations in dispersion even when frequency remains constant, as the number of 210
segments containing the target word is crucial for their calculation. 211
3. We hypothesize that frequency-based measures (RRF, LLR, and chi-square tests) 212
will show high variations in distinctiveness scores even when the frequency differ- 213
ence of an artificial word between the target and comparison corpus is relatively 214
small. This assumption stems from the statistical nature of these measures, which 215
treat a corpus as a bag of words and do not account for word dispersion. 216

6. In the dispersion analysis, we also tested another scenario, in which we randomly selected 1,000 segments and added one instance of the artificial word to each of them in the comparison corpus, ensuring even dispersion. However, this scenario turned out not to provide significant or additional insights. Therefore, we are not providing further explanations or results here.

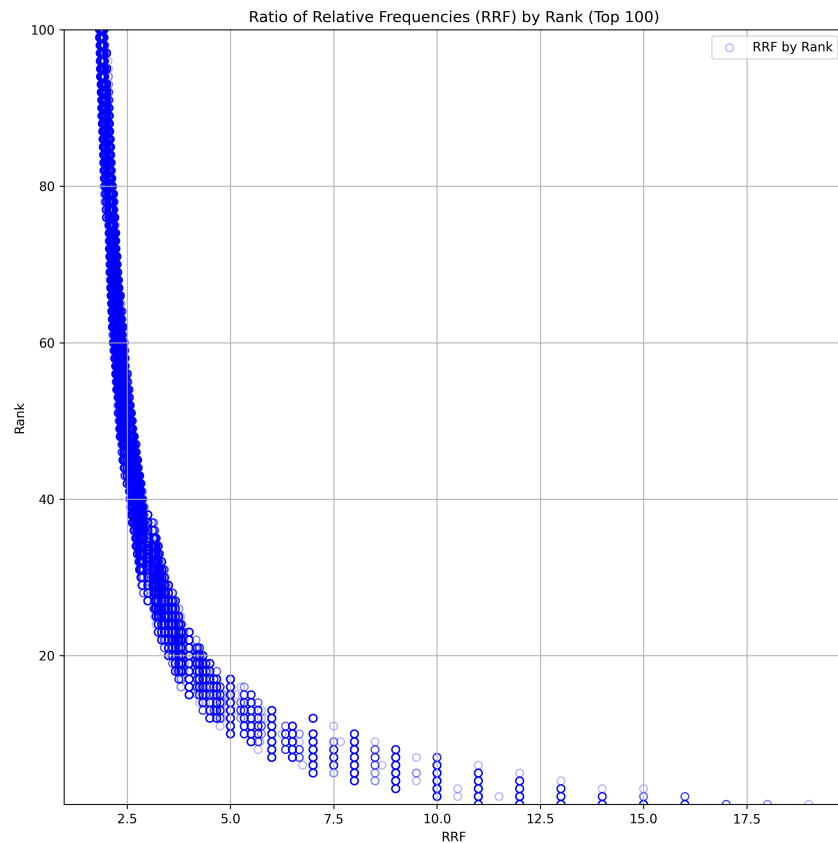


Figure 1: The correlation between the RRF score of the words and their ranks in the synthetic corpus.

4. When the frequency of an artificial word is the same in both the target and comparison while its dispersion changes, the scores of frequency-based measures should remain unchanged. 217 218 219
5. Regarding our TF-IDF-based measure, we expect it to exhibit moderate sensitivity in both frequency and dispersion manipulations. This is because TF-IDF is based on term frequency, but the number of segments containing the target word also significantly influences its calculation. 220 221 222 223
6. Regarding Welch’s test, we hypothesize that there will be minimal variations in the score in the case of frequency manipulation. This assumption is based on the fact that the calculation of Welch’s test relies on the mean and standard deviation of the frequency distributions, rather than on the raw frequency of the word. 224 225 226 227

6. Results 228

Because our corpus is based on naturally occurring word frequencies, we conducted an additional analysis to identify potential artifacts caused by random sampling effects in the synthetic texts without the artificial word. This analysis aimed to identify the

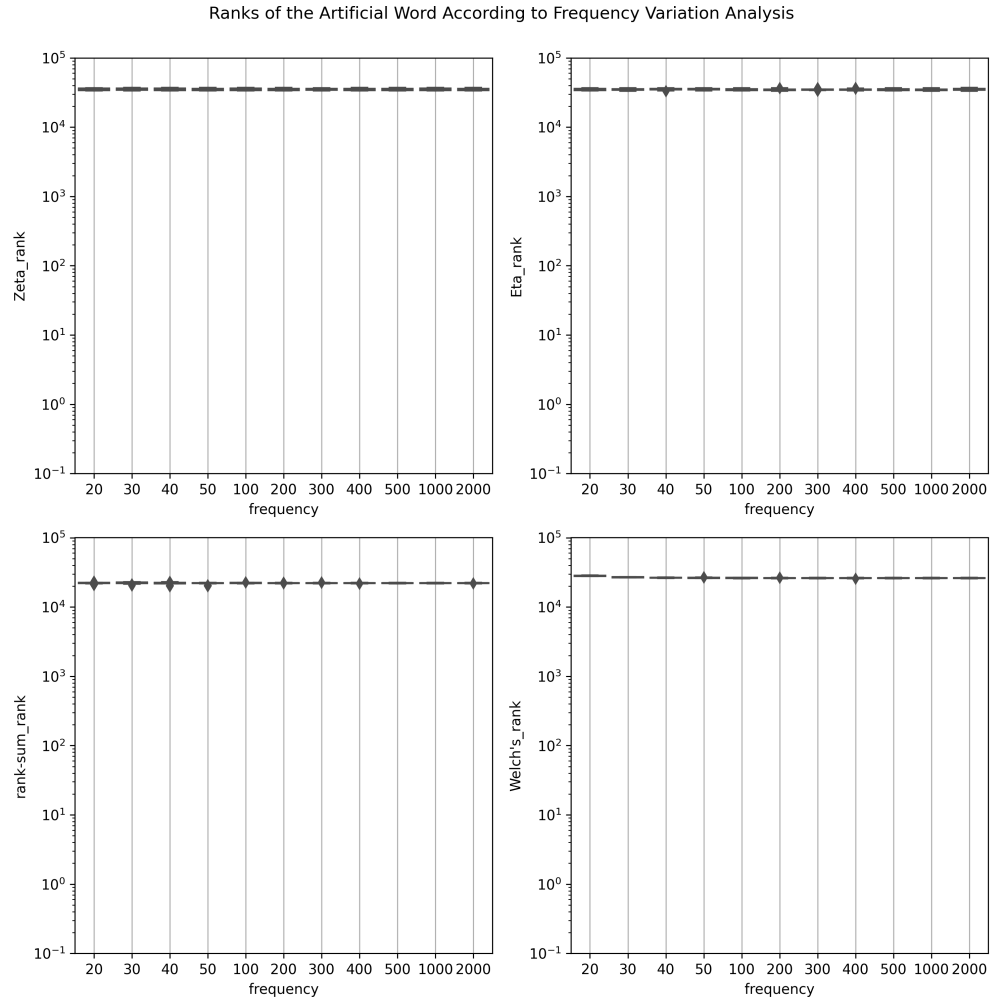


Figure 2: The relation between the frequency of the artificial word in the target corpus and its rank in the results, for Zeta, Eta, rank-sum test and Welch’s test.

frequency differences of words in the corpus across multiple runs. 232

Figure 1 illustrates the relationship between rank and the Ratio of Relative Frequencies (RRF) scores, based on 100 runs of randomly sampled synthetic corpora. As shown, the first rank is typically achieved with RRF scores ranging from 10 to 18. This suggests that, due to the natural variations in the frequencies of existing words, an RRF score below 10 for the artificial word is unlikely to secure the first rank. 233 234 235 236 237

As discussed in the section 4, we conducted our evaluation in two main settings: frequency variation of an artificial word and dispersion variation. First, we are going to discuss the results of the evaluation based on frequency variations. 238 239 240

6.1 Evaluation based on frequency variations 241

Concerning the impact of frequency variation on the performance of the measures, as described in section 5, we used 12 different parameters for the frequency settings. Figure 2 depicts the variation in the rank of the artificial word, as calculated by Zeta, Eta, the rank-sum test and Welch’s test, respectively, depending on its frequency in the 242 243 244 245

target corpus.⁷ The x-axis represents the frequency variation in the target corpus (from 20 to 2000 occurrences in one segment of the target corpus). On the y-axis, the rank of the artificial word is depicted. To enhance the readability of the figure, the values on the y-axis are presented on a logarithmic scale.

Dispersion-based measures including Zeta, logarithmic Zeta, Eta, Wilcoxon rank-sum test, as well as Welch's t-test, which we consider rather as a distribution-based measure, demonstrate very similar results. For these measures, the frequency variations of an artificial word in the target corpus don't play an important role. The rank of the artificial word consistently exceeds 10,000 for frequencies ranging from 20 to 2,000 in the target corpus, indicating a very low distinctiveness score according to these measures. The scores for Eta, Zeta, logarithmic Zeta, and the Wilcoxon rank-sum tests remain consistent, supporting Hypothesis 1 and validating our method. The scores from Welch's test show minimal variation, as expected in Hypothesis 6.

Frequency-based measures such as the chi-square test, LLR, and RRF exhibit high sensitivity to frequency variations, as expected, supporting Hypothesis 3. However, we can observe some interesting results here. When considering the RRF, the artificial word moves up in rank with increasing frequency from 20 to 100 (Figure 3). Starting from 200 artificial words in the target corpus, RRF-based rank is always 1, which means that the artificial word gets the highest score among all words in the corpus. As for LLR and chi-squared tests, both measures are even more sensitive to frequency variation compared to RRF. Starting at a frequency of just 40, we consistently observe the artificial word achieving the top rank.

TF-IDF is more sensitive to frequency variation than dispersion-based measures but significantly less so than frequency-based measures, aligning with our expectation in Hypothesis 5. With increasing frequency of the artificial word in the target corpus, its rank moves up. Figure 3 shows a continuous rise of the rank of the artificial word.

6.2 Evaluation based on dispersion variations

As previously described, the dispersion analysis was conducted with 1000 instances of the artificial word in one segment of the comparison corpus. Figure 4 illustrates the variation in the rank of an artificial word calculated by chi-square, LLR, RRF and Welch's test. The x-axis depicts the dispersion variation of the artificial word in the target corpus from 1/1000 to 1000/1, where the first number represents the number of segments and the second number represents the number of instances of the artificial word distributed over those segments. The dispersion of the artificial word in the comparison corpus remains constant, set at 1/1000, indicating 1000 words occurring in one segment. In these settings, the frequency-based measures produce results consistent with those predicted by Hypothesis 4. When the dispersion changes (while the frequency remains constant), the rank of an artificial word does not change significantly and consistently remains at the level between 10,000 and 100,000.

Regarding the results of Welch's test, when the frequency of the artificial word is identical in both the target and comparison corpora, the score consistently remains zero, resulting in a rank above 10,000. This indicates that, like the frequency-based measures, Welch's

7. Eta_log is not depicted in the figure, because its results are very similar to the Zeta results.

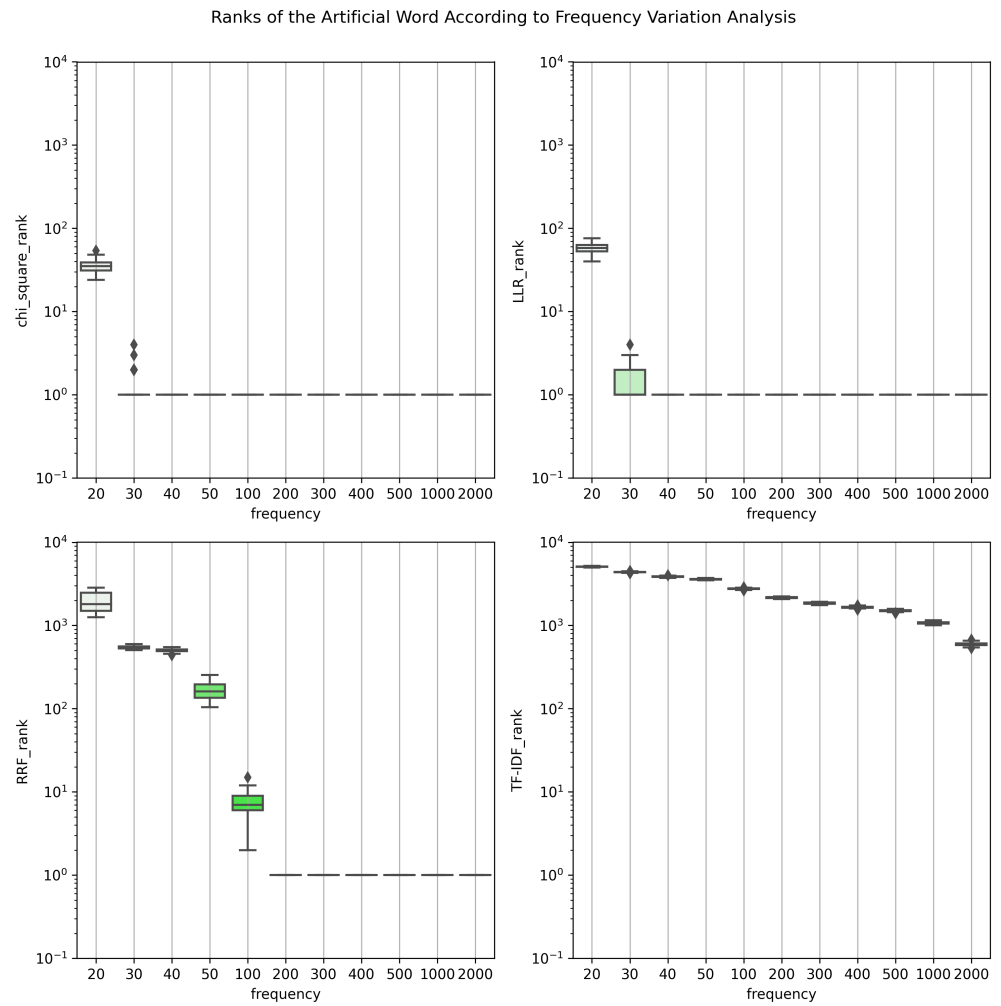


Figure 3: The relation between the frequency of the artificial word in the target corpus and its rank in the results, for RRF, chi-squared test, LLR and TF-IDF.

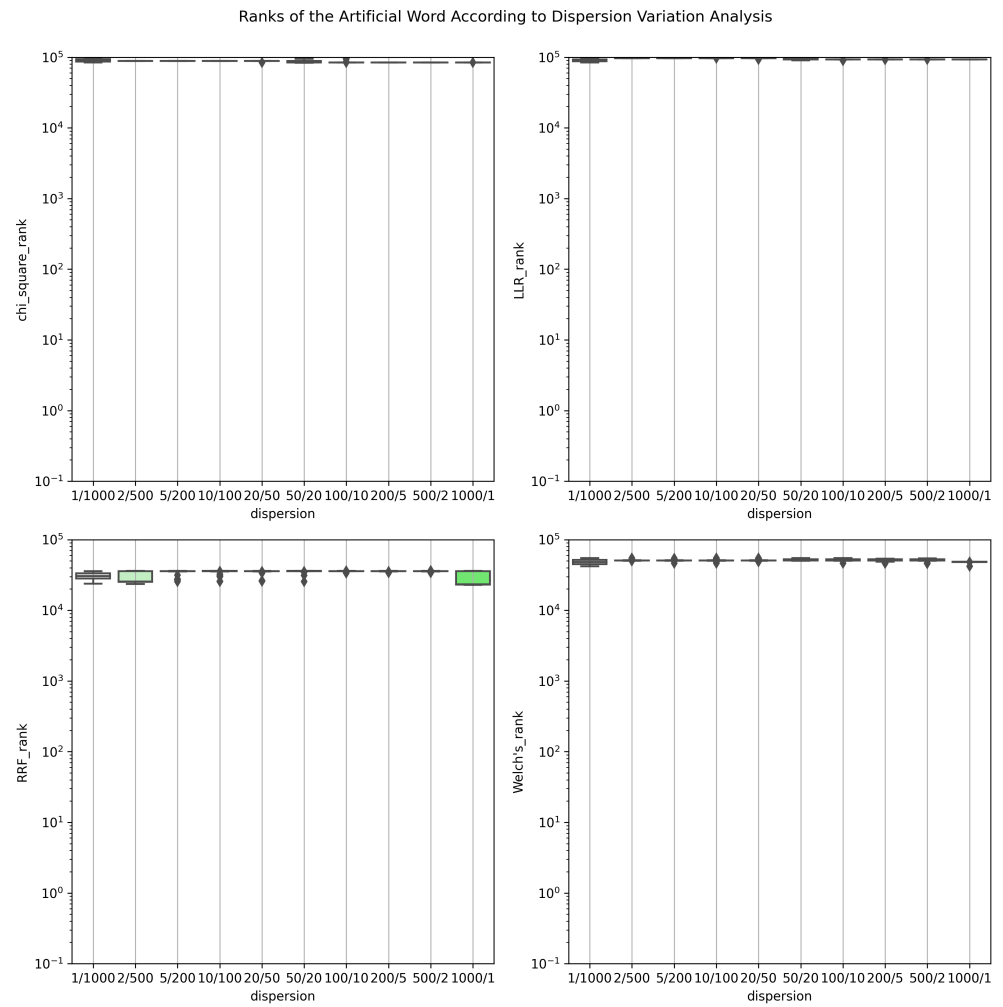


Figure 4: The relation between the dispersion of the artificial word in the target corpus and its rank in the results, for RRF, chi-squared test, LLR and Welch's. Dispersion in the comparison corpus 1/1000.

test is not sensitive to variations in dispersion within our settings. 288

An interesting result was obtained by Eta. As it is a dispersion-based measure, we 289
 expected Eta to effectively identify an artificial word as distinctive, especially when the 290
 word is evenly spread across a high number of segments. However, as the number of 291
 segments containing the artificial word in the target corpus increases, its scores remain 292
 consistently low compared to randomly assigned words. Only in the most extreme 293
 setting, with one occurrence in 1,000 segments in the target corpus, does the artificial 294
 word receive the top rank (Figure 5). 295

Regarding the remaining dispersion-based measures, such as both variants of Zeta and 296
 the rank-sum test, we observe expected results. With increasing numbers of segments 297
 containing the artificial word in the target corpus, the artificial word's rank moves up. 298
 Specifically, starting with 10 words in 100 segments, the artificial word consistently 299
 receives the top rank according to these three measures (Figure 5). This indicates that 300
 Hypothesis 2 is supported solely for these three measures. 301

Interesting results are also observed with TF-IDF. Here, we anticipated that as the 302
 dispersion becomes more even, the artificial word would receive a higher score, but 303
 only with a moderate rank improvement compared to other dispersion-based measures. 304
 In fact, we can observe that TF-IDF scores indeed increase as the number of segments 305
 containing the artificial word rises. However, the improvement in scores is not moderate; 306
 rather, TF-IDF appears to be highly sensitive to variations in dispersion, which partially 307
 rejects Hypothesis 5. We observed the artificial word achieving the top rank starting 308
 with a dispersion of just 100 words in 10 segments (Figure 5). This oversensitivity 309
 implies that the TF-IDF measure fails to distinguish between a dispersion of 100 words 310
 across 10 segments vs. one single word across 1,000 segments. 311

7. Conclusion 312

Conducting analyses of measures of distinctiveness based on synthetic texts, we cre- 313
 ated ideal conditions to uncover the hidden properties of a range of such measures. 314
 Through our experiment, we tested the sensitivity of these measures to variations in the 315
 frequency and dispersion of a specific word. In many cases, our hypotheses regarding 316
 the performance of the measures were confirmed. Frequency-based measures are not 317
 sensitive to variations in dispersion, while dispersion-based measures are not affected 318
 by frequency variations. These observations are not surprising, of course, but they do 319
 validate our method. 320

However, some hypotheses were partly rejected and we have also uncovered some previ- 321
 ously unknown (or at least undocumented) properties of measures of distinctiveness. In 322
 particular, we found that LLR and chi-squared tests are even more sensitive to frequency 323
 variation than RRF. For this reason, we generally do not recommend using the LLR and 324
 chi-squared tests, as they are highly sensitive to changes in frequency and are therefore 325
 not well-suited for keyness analysis aimed at identifying important content words. Both 326
 Zeta variations and the rank-sum test demonstrated similar scores and abilities to detect 327
 distinctive words, including when differences concern only the dispersion of words. 328
 Moreover, we discovered that TF-IDF is highly sensitive to dispersion differences of 329

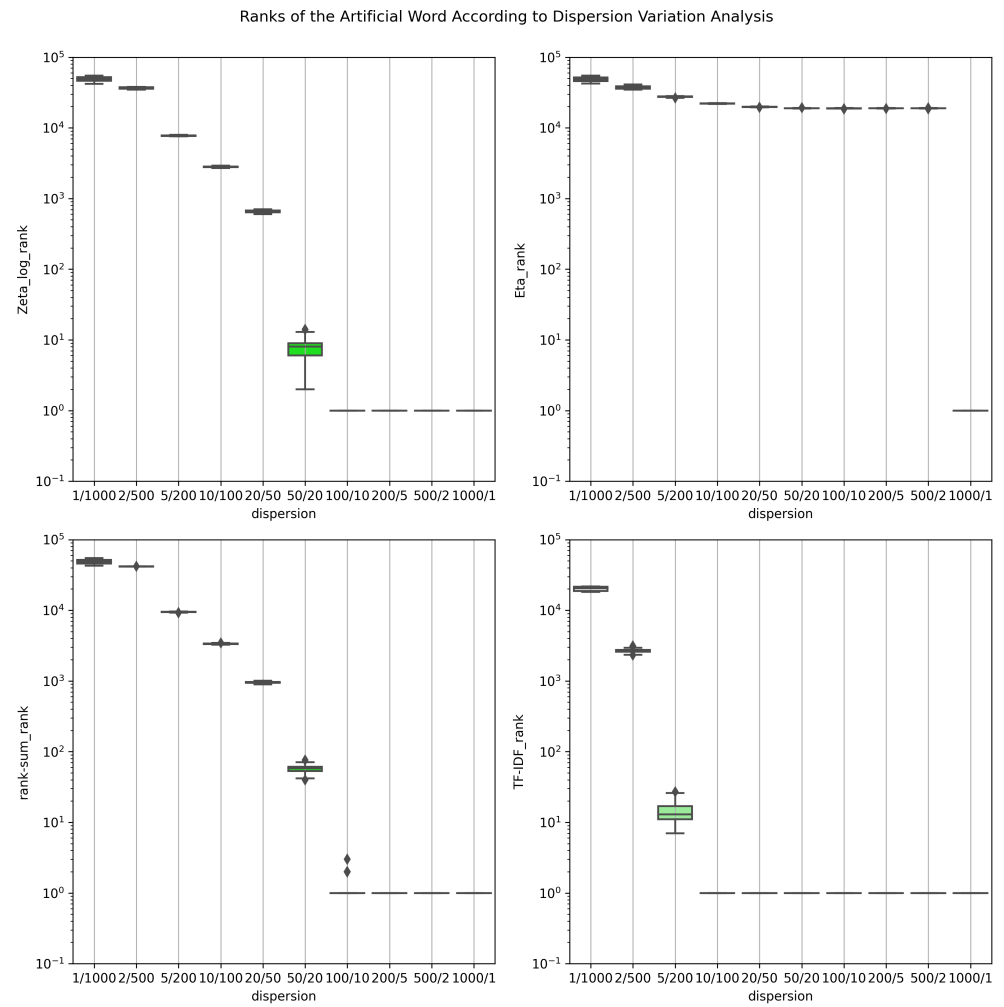


Figure 5: The relation between the dispersion of the artificial word in the target corpus and its rank in the results, for Zeta, Eta, rank-sum test and TF-IDF. Dispersion in the comparison corpus 1/1000

the target word, compared to other dispersion-based measures. Finally, we found that Eta cannot detect a word with a clear contrast in dispersion when its frequency is the same in both the target and comparison corpora. In our evaluation we observed words steadily moving up in rank with Zeta and rank-sum, while TF-IDF and Eta show more abrupt increases. We suggest that a gradual, continuous rank improvement is a desirable characteristic of a distinctiveness measure, as it indicates better sensitivity to slight variations in dispersion and is likely to produce more predictable results. For example, if a researcher is interested in identifying words that display contrasting dispersion within two subcorpora, without considering their frequency, then Zeta and the rank-sum test would be most appropriate for this task.

Despite the interesting observations derived from these analyses, there is significant potential for future work. One key step is to extend our framework by implementing additional measures of distinctiveness. Another area for future work involves expanding our analysis by implementing additional parameter settings that combine frequency and dispersion variations of the artificial word. Isolating dispersion or frequency often results in constant scores from the measures, but combining these parameters promises to provide new opportunities to uncover additional properties of these measures. A final, crucial step is to explore practical applications of this newfound knowledge about distinctiveness measures. Understanding the specific contexts and scenarios in which each of these measures can be most effectively utilized will open up new possibilities and enhance our ability to analyze and compare textual corpora more accurately.

8. Data Availability

Data can be found here: https://github.com/Zeta-and-Company/synthetic_texts_evaluation, <https://doi.org/10.5281/zenodo.15525428>.

9. Software Availability

Software can be found here: https://github.com/Zeta-and-Company/synthetic_texts_evaluation, <https://doi.org/10.5281/zenodo.15525428>.

10. Author Contributions

Julia Havrylash: Conceptualization, Data Curation, Methodology, Formal Analysis, Software, Visualisation, Writing – original draft, Writing – review & editing






Christof Schöch: Funding Acquisition, Supervision, Writing – review & editing

References

Argamon, Shlomo (2007). “Interpreting Burrows’s Delta: Geometric and Probabilistic Foundations”. In: *Literary and Linguistic Computing* 23.2, 131–147. [10.1093/llc/fqn003](https://doi.org/10.1093/llc/fqn003).

- Du, Keli, Julia Dudar, Cora Rok, and Christof Schöch (2021a). “Zeta & Eta: An Exploration and Evaluation of Two Dispersion-based Measures of Distinctiveness”. In: *Proceedings of Computational Humanities Research 2021* 1613, 0073. http://ceur-ws.org/Vol-2989/short_paper11.pdf (visited on 05/26/2025).
- Du, Keli, Julia Dudar, and Christof Schöch (2021b). *Pydistinto - a Python Implementation of Different Measures of Distinctiveness for Contrastive Text Analysis*. Version vo.1.1. Zenodo. [10.5281/zenodo.5245096](https://doi.org/10.5281/zenodo.5245096).
- (2022). “Evaluation of Measures of Distinctiveness: Classification of Literary Texts on the Basis of Distinctive Words”. In: *Journal of Computational Literary Studies* 1.1. [10.48694/JCLS.102](https://doi.org/10.48694/JCLS.102).
- Egbert, Jesse and Doug Biber (2019). “Incorporating Text Dispersion into Keyword Analyses”. In: *Corpora* 14.1, 77–104. [10.3366/cor.2019.0162](https://doi.org/10.3366/cor.2019.0162).
- Evert, St., Fotis Jannidis, Thomas Proisl, Steffen Pielström, Thorsten Vitt, Christof Schöch, and Isabella Reger (2017). “Understanding and Explaining Distance Measures for Authorship Attribution”. In: *Digital Scholarship in the Humanities* 23.suppl2. [10.1093/llc/fqx023](https://doi.org/10.1093/llc/fqx023).
- Gries, Stefan Th. (2022). “What Do (Most of) Our Dispersion Measures Measure (Most)? Dispersion?” In: *Journal of Second Language Studies* 5.2, 171–205. [10.1075/jsls.21029.gri](https://doi.org/10.1075/jsls.21029.gri).
- Kilgarriff, Adam (2001). “Comparing Corpora”. In: *International Journal of Corpus Linguistics* 6.1, 97–133. [10.1075/ijcl.6.1.05kil](https://doi.org/10.1075/ijcl.6.1.05kil).
- Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2014). “Significance Testing of Word Frequencies in Corpora”. In: *Digital Scholarship in the Humanities* 31.2, 374–397. [10.1093/llc/fqu064](https://doi.org/10.1093/llc/fqu064).
- Paquot, Magali and Yves Bestgen (2009). “Distinctive Words in Academic Writing: A Comparison of Three Statistical Tests for Keyword Extraction”. In: *Corpora: Pragmatics and Discourse*. Ed. by Andreas H. Jucker, Daniel Schreier, and Marianne Hundt. Brill | Rodopi. [10.1163/9789042029101_014](https://doi.org/10.1163/9789042029101_014).
- Schröter, Julian, Keli Du, Julia Dudar, Cora Rok, and Christof Schöch (2021). “From Keyness to Distinctiveness – Triangulation and Evaluation in Computational Literary Studies”. In: *Journal of Literary Theory* 15.1-2, 81–108. [10.1515/jlt-2021-2011](https://doi.org/10.1515/jlt-2021-2011).
- Sönning, Lukas (2023). “Evaluation of Keyness Metrics: Performance and Reliability”. In: *Corpus Linguistics and Linguistic Theory*. [10.1515/cllt-2022-0116](https://doi.org/10.1515/cllt-2022-0116).
- Spärck Jones, Karen (1972). “A Statistical Interpretation of Term Specificity and Its Application in Retrieval”. In: *Journal of Documentation* 28, 11–21.

A Computation Analysis of Character Archetypes in the Works of Calderón de la Barca

Allison Keith¹ 
 Antonio Rojas Castro² 
 Kerstin Jung¹ 
 Hanno Ehrlicher² 
 Sebastian Padó¹ 

1. Department of Natural Language Processing, University of Stuttgart , Stuttgart, Germany.
2. Faculty of Humanities, University of Tübingen , Tübingen, Germany.

Citation

Allison Keith, Antonio Rojas Castro, Kerstin Jung, Hanno Ehrlicher, and Sebastian Padó (2025). "A Computation Analysis of Character Archetypes in the Works of Calderón de la Barca". In: *CCLS2025 Conference Preprints 4* (1). 10.26083/tuprints-00030153

Date published 2025-06-17

Date accepted 2025-04-18

Date received 2025-01-30

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check <https://jcls.io> for the final journal version.

Abstract. The *Siglo de Oro* period of Spanish theater was marked by a rapid increase of production of theatrical pieces. These plays used clear patterns, notably for genre, plot and character even though there were no explicit drama conventions. Our study aims at characterizing such conventions in the work of Calderón de la Barca. We investigate the portrayal of character archetypes based on character speech, adopting a scalable reading approach that employs machine learning to aggregate empirical evidence from a large number of Calderón's works. Concretely, we develop a neural network model to predict the six main character archetypes. We analyze the predictions of this model, inspect the lexical material that determines these predictions, and visualize the representations that the model learns. We find that the model predicts character archetypes with some accuracy ($f1 = 0.47$), and that some character archetypes like *criadas*, are more standardized than others, like *reyes*.

1. Introduction

El Siglo de Oro, the Spanish Golden Age, is a period of time that begins with the Spanish imperial era in 1492. Its end is generally assumed to be the death of one of the period's last great playwrights, Pedro Calderón de la Barca, in 1681. This period in Spanish history is marked by broad cultural flourishing, and an immense productivity by playwrights such as Calderón, Lope de Vega or Tirso de Molina (Couderc 2012). These authors wrote hundreds of plays, eschewing some rules of classical theater such as the unity of time and place and adapting classical theater conventions to the more modern audience of the time (Ruggerio 1972).

Calderón was known for writing two types of plays specifically: Corpus Christi plays, which were one act plays featuring allegorical characters and characters from the bible, which to expressly convey religious values to the audience, and *comedias*. *Comedias* is the name for a diverse ensemble of three act plays that center around worldly events (as opposed to allegorical plays that were also popular at the time). The supercategory of *comedias* can be divided into the genres comedy or tragedy, although the distinction between the two genres is not as defined as it was in classical theater (Álvarez Sellers

2015), and the genres of some works are disputed. Main themes of the works include honor and romantic love, although different genres treat these themes differently and different types of characters interact with the themes in unique ways. Among his best known works are *La vida es sueño* (Life is a dream, 1636), *El médico de su honra* (The surgeon of his honor, 1636) or *El alcalde de Zalamea* (The Mayor of Zalamea, 1651).

One suggestion as to how authors were able to produce such vast quantities of work is that they relied so heavily on a set of theatrical conventions. A manifestation of these conventions is the characterization of different types of stock characters. For the first time, authors used characters from the real world, of different social classes to convey their stories, relying on character archetypes that share key traits (Elvira 2014). Traditionally, tragic pieces focused only on the noble class, while comedies told stories of lower class characters. This theatrical convention was another that was subverted during the *Siglo de Oro* (Elvira 2014). Notably, during the *Siglo de Oro*, the convention that only male actors could perform on stage was broken. Meaning that in this period, not only were realistic female characters being represented, but also actresses could perform the roles on stage, granting much more visibility to women. Theater troops were made up of a standard number of actors and actresses (3:2), meaning that the plays were written consistently with both male and female characters (Elvira 2014). However, the majority of characters were male. The authors of the *Siglo de Oro* gave visibility to female and male characters of different social statuses, which is why we can break down our exploration into three distinct classes (royalty, the nobility, and the servants).

The sheer productivity of these authors poses a significant challenge for traditional scholarship in analyzing recurring patterns in these works. In this article, we propose a scaled reading approach (Tracy 2016; Weitin 2017) to analyzing character archetypes in the work of Calderón de la Barca, from whose work over 120 three act plays and 80 Corpus Christi plays are available in digitized form via the DraCor project (Fischer et al. 2019). The scaled reading approach allows us to bring together theories on trends of literary works with a broad range of empirical evidence, contributing to an exchange of ideas and methods between skilled literary scholars and corpus linguistics.

Thus, we investigate the principal research question: What can we learn about character archetypes in the works of Pedro Calderón de la Barca using scaled reading across all of his digitized three act plays? We use three analysis methods to examine how defined these archetypes are:

1. To learn how different the speech of these character archetypes are to one another, we use automatic classification to determine how easy it is for a model to determine the character archetype.
2. We examine how disparate or cohesive the characters of a given archetype are to one another, telling us how uniform the character archetypes are, and how they relate to one another. In order to visualize the characters' relations to one another, we use dimensionality reduction.
3. We aim to identify which elements of speech contribute most to the characters' presumed class, i.e. what kind of speech specifically differentiates a king from a nobleman, or a nobleman from a servant. In order to do so, we use an attribution

model (Murdoch et al. 2019) to examine specific aspects of character speech and do qualitative analysis of the data.

2. Background

2.1 Character Archetypes

We examine the following character types: *rey*, *reina*, *galán*, *dama*, *criado* and *criada*. They are among the most commonly occurring characters *Siglo de Oro* works, and also represent three distinct social classes. Our research question rests on the assumption that these character archetypes are identifiable because of the differences in the way they speak and the topics they discuss. The reasoning here is two-fold: 1) stock character archetypes were somewhat static and shared certain traits, as will be discussed in the following few paragraphs, allowing authors to produce more works more quickly 2) Each character type has a distinct relationship with the key themes of the work, i.e., honor, which shapes the way that they speak and what they speak about.

Galán - the Nobleman *Galán* characters are particularly involved in conflicts surrounding *honor* (Couderc 2006). The concept of honor for the male nobility characters refers to their social standing and public perception of the character's virtue or power. These situations involve personal character or money, or property, leading the *galán* to a conflict in which he must preserve his honor or seek forgiveness (Lauer 2017).

Dama - the Noblewoman The *dama* is a broad category that captures women of high social class. While there are diverse different types of male characters in *Siglo de Oro* works, female characters are almost exclusively *damas* in the works of Calderón.

The main conflict of the *dama* character surrounds her love. For the *dama*, her relation to the honor code is her purity, and how her romantic or sexual behavior reflects on her father or partner. As stated by McKendrick in their 1974 study on women in the *Siglo de Oro*, the most ideal traits for a woman in 17th century Spain are 'virtue, humility, modesty, tenderness, silence, diligence, and prudence' (Lauer 2017; McKendrick 1974).

Generally, women were confined to domestic spaces, as the concept of the ideal woman was carried to theater as well, meaning women in theater often fit this role of remaining in the house (McKendrick 1974). While the prototypical female character would generally fall under this characterization, it is well known that Calderón represented several *dama* characters in ways that subvert gender norms (De Armas 2015). Notably, a principle character in his most famous work, *Rosaura* of *La vida es sueño*, is a *dama* who disguises herself as a man during many acts of the play in order to seek revenge on a man who dishonored her.

At this time women of noble class were educated and therefore their speech would reflect this fact, maybe containing literary or historical allusions (McKendrick 1974).

Rey - The King In the *Siglo de Oro*, the king (and queen) characters often act as the arbiter of honor (Lauer 2017). They settle disputes between characters, and grant forgiveness to noble characters who seek to earn or to restore their honor. Regarding

speech, the characters in the *Siglo de Oro* dramas use language in accordance with the appropriate social class (Mañero Lozano 2009). For kings and queens who are highly educated this might mean using flowery language, literary allusions, or references to the Bible, to convey their education and wisdom.

Reina- The Queen Like kings, there are multiple categories of queens that appear in *Siglo de Oro* theater: mythological, saintly, biblical, historical, and fictional (De Armas 2015). For example, Calderón's drama *La hija del aire* (the daughter of the air) centers around Semíramis, an Assyrian queen of the Bible. The characterization of queens is very diverse (Quintero 2017), often based on real Spanish and or other European queens, which might make it difficult to group them into one cohesive group.

Criado - The Male Servant The *criado* character is one that serves in both domestic labor role and serves the *galán*. He communicates with principle characters to reveal their thoughts and feelings to the audience (Ríos Carratalá 2022). The *criado* character speaks at a much more informal register compared to the *rey* and the *galán* and this, in part, serves to add comedic effect (Táuler et al. 2014). There is a special type of *criado* called the *gracioso*, present in many plays, who is a *criado* that plays a large role in the work, compared to other *criados* and serves a comic relief as well and involving the audience in the spectacle by commenting on the action of the play. The *criado*'s speech might be significantly different compared to other characters because of the presence of comedy.

Criada- The Female Servant As stated previously, the changing conventions of the *Siglo de Oro* (breaking the classical norms) allowed for a greater representation of both women, and low social class individuals (Elvira 2014). The *criada* is a character who, while relegated to the sidelines, and lacking visibility, simultaneously serves an important purpose in the works, by interacting with main female characters, allowing certain information to be revealed (L. G. Lorenzo 2008). We could assume that, because they are principally interacting with *damas*, that the topics the two discuss might overlap. There are not many *criada* characters in *Siglo de Oro* works (comparatively to the number of their male counterparts).

2.2 Classifying Character Types

There are some works that have described both the plays of Calderón and more specifically some of the characters in Calderón's works from a quantitative perspective, however none of these works have addressed unsupervised classification of character archetypes (Ehrlicher et al. 2020, Lehmann, Padó, et al. 2022, L. H. Lorenzo 2024).

The classification of character types with computational methods can be carried out on the basis of different types of information, of which two are particularly prominent. The first direction is based on the observation of typical contexts in which characters are mentioned in narrative passages. Along these lines, Bamman and colleagues extract informative contexts (adjectives and verbs) of character mentions and cluster them into archetypes such as 'hero' or 'love interest' (Bamman et al. 2013, 2014). The second important direction is the characterization of characters in terms of their social context, i.e., social networks, which are typically grounded in co-occurrence in the same scenes

(Beine 2024; Elson et al. 2010). This approach has been used to identify figures in German language drama (Krautter et al. 2020).

Both of these approaches present problems when applied to our current study. First, there is very little stage direction in *Siglo de Oro* drama (other than entrances and exits) or other information about the plot set down in the plays. This rules out the first family of methods that make use of information from narrative passages. The use of social networks to determine character types, on the other hand, typically involves the use of often intransparent network metrics, and puts a lot of theoretical weight on the notion of co-occurrence within one scene, which appears a successful, but fairly heuristic, assumption which we would like to avoid.

In this study, we propose instead to focus on the characters' *speech*. Arguably, in plays in this period and plays by Calderón, the majority of information is conveyed by the character's speech: Character archetypes interact with other characters in a standardized way, allowing playwrights to use formulaic language and topics to build plays quickly (cf. the characterization of the archetypes above). Therefore, in this experiment, we chose to focus exclusively on character speech as the classification criteria. For example, the fact that *reinas*, *reyes*, *damas*, and *galanes* were all educated characters, and therefore speak with a higher register, might then indicate that they would be easily differentiated from the servant characters. This approach has been used successfully to assign quotations to characters in literary narratives (Elson and McKeown 2010) and to classify character gender (Keith et al. 2024).

We represent characters speech via word embeddings. Word embeddings are numerical representations of words in a corpus that represent aspects of word usage, by using a word's context (i.e. the words surrounding it) to place each word into a high-dimensional semantic space. These embeddings can be used to carry out text classification. When combined with attribution models, models can also illuminate the most salient words for each category. Methods based on word embeddings can capture lexical but also grammatical and stylistic information (Tenney et al. 2019). This is particularly useful if we want to know what specific topics make a group unique. Because of the nature of the models in Spanish, which often breaks words into their subword units (Sennrich et al. 2016), this method is also useful if there are certain grammatical traits that are particular to certain groups, e.g., grammatical gender or politeness. In this way, the we can use an interpreter model to not only analyze the usage of words, but patterns in grammatical traits.

3. Methods

3.1 Data

We examine the 104 *comedias* of Calderon de la Barca, which are digitized in TEI format in CalDraCor, as part of the DraCor project (Fischer et al. 2019). These digitized dramas are orthographically modernized, which makes them amenable to analysis with NLP models trained on modern corpora. This data is enriched with gold standard labels on character type found directly in the original cast lists, sourced from the Calderon digital Project (Antonucci n.d.). We used the genre classifications by Simon Kroll (Kroll

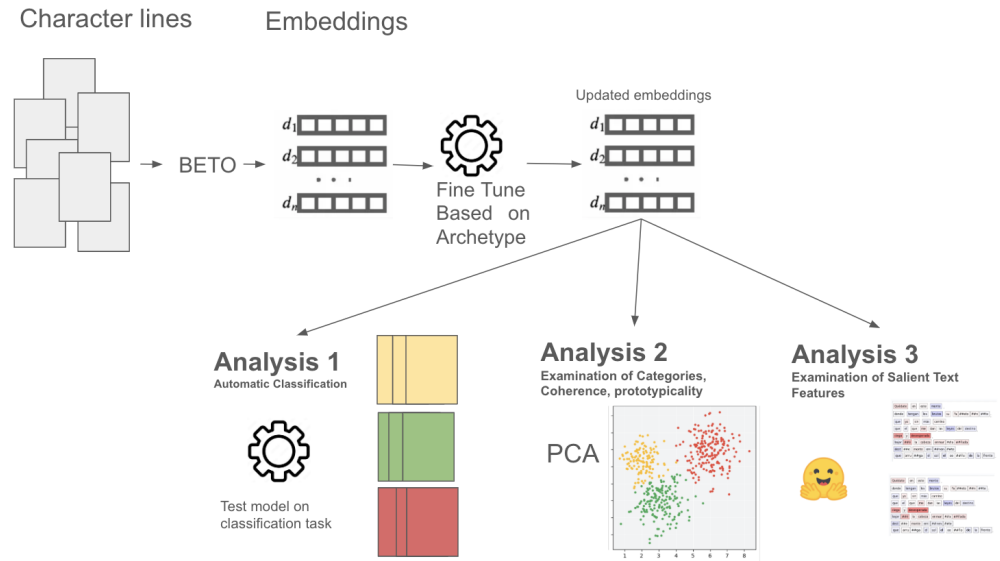


Figure 1: Analysis procedure: Fine-tune embeddings based on character archetypes, and conduct three analyses

2017), also included in CalDraCor, because they are more standardized than the original 184
DraCor genre classifications (8 different genres instead of 16). We split our corpus into 185
a training, a development and a test section (80%/10%/10%). 186

The complete corpus includes 147 *dama* characters, 103 *galán*, 96 *criado* characters, char- 187
acters, 78 *criada* characters, 48 *rey* characters, and 17 *reina* characters. The reason for the 188
predominance of *damas* is that nearly every female character is a *dama*, whereas there 189
are many different archetypes for male characters. 190

Previous work on automatic character type classification shows that models perform 191
best on this classification task when given access to the maximum amount of speech for 192
each character (Keith et al. 2024). Therefore, we use all of the character lines for the task, 193
up to the input limit of the model (512 tokens). One data point is equivalent therefore 194
to a unique character and includes the first 512 tokens that character speaks. 195

3.2 Analysis Procedure 196

Figure 1 shows our analysis procedure. In preparation of our study, we fine-tune a 197
Spanish embedding model (BETO) to classify character archetypes on the training set of 198
our Calderón corpus. This effectively updates the embeddings that the model produces 199
for the characters from their utterances. Our study then carries out three analyses on 200
the resulting model: We assess their effectiveness on the test set, we analyze the internal 201
representations through visualization, and we use attribution methods to understand 202
which textual features are most important according to the model. We now describe 203
these analyses in more detail. 204

Analysis 1: Character Archetype Classification Bidirectional encoder representations 205
from transformers, widely known as BERT Models, are presently the standard in lan- 206
guage modeling (Devlin et al. 2019). These models use large corpora to pre-train deep 207
numerical representations of the texts known as embeddings. These embeddings can 208

then be further trained, or fine tuned, to create task-specific embeddings. We use BETO (Cañete et al. 2023), a BERT-base model pre-trained on Spanish language web data, and fine-tune it to encode character speech into embeddings that can be classified into one of our six character archetypes. While generative models (like the GPT model family) (Radford et al. 2018) are especially useful in text generation tasks, BERT-based models excel at analysis and classification tasks such as the current one.

In our first analysis, our aim to determine to what extent the speech of character archetypes is distinguishable. To address this, we carry out classification with our fine-tuned model and assess correct predictions as well as errors, which can indicate which characters fall outside the norm for their archetype, as well as which character archetypes are more similar to one another.

Analysis 2: Visualizing the Embedding Space In our second analysis, we examine how the character archetypes relate to one another, and specific characters relate to their assigned archetype. We reduce the dimensionality of the embedding space to 2 dimensions in order to visualize the location of each character into the embedding space. When we implement dimensionality reduction, a method in which we reduce the embedding space of hundreds of dimensions to only a few dimensions, we can plot the archetypes using only the most salient dimensions of the embedding space into a human understandable way, allowing us to visualize the way that the data points relate to one another. Principle component analysis (PCA) is the process by which we can visualize the results of the archetype embeddings (Murphy 2012). Once we have trained the embeddings to differentiate each data point based on archetype, we then use PCA to identify the most principle components - the dimensions that vary most between classes. Plotting this way allows us to visualize both the coherence of different categories, and their distance from one another. It permits us to identify at a large scale which members of a cluster are prototypical and which are outliers. In the case of character type, we can use the dimensionality reduction method to examine the prototypicality or atypicality of specific characters without having to read the entire corpus. This method can give scholars a starting point from which to examine certain characters or themes.

Analysis 3: Attribution Model We would expect that thematic indicators, i.e., content words, and grammatical features of speech both play a role in setting the archetypes apart from one another. To examine which traits are specific to given archetypes, we utilize an attribution method. Attribution, a technique from the area of explainable AI, aims at capturing the extent to which the different part of an input to a machine learning model are crucial in determining the models' output, thereby turning 'black-box' models transparent (Murdoch et al. 2019). In our case, attribution methods tell us which input tokens are particularly important for the character archetype classification, which differs from traditional stylometric approaches (Culpeper 2014L. H. Lorenzo 2024). Specifically, we use the Transformers Interpret implementation (Pierse 2021) of the integrated gradients approach (Sundararajan et al. 2017), a method to create attributions that are guaranteed to fulfill a set of consistency axioms. In order to get the most salient tokens for each archetype, we measure the attribution score of each text for each label, telling us the contribution of each token in the text to that label, based on the embeddings of our fine-tuned model. A token with a high score means that this

	galán	dama	rey	reina	craido	criada	overall
Precision	0.50	0.57	0.30	0	0.80	0.83	0.50
Recall	0.18	0.81	0.50	0	0.80	0.66	0.44
F1	0.27	0.66	0.38	0	0.80	0.74	0.47

Table 1: Performance of neural network model for all archetypes and overall

Prediction → Gold Label ↓	galán	dama	rey	reina	craido	criada
galán	2	3	3	0	2	0
dama	1	13	2	0	0	0
rey	0	3	3	0	0	0
reina	0	0	2	0	0	0
craido	0	2	0	0	8	0
criada	1	2	0	0	0	5

Table 2: Confusion matrix of model predictions. Highlighted in green and bolded are correct predictions. Highlighted in yellow are the most frequent incorrect predictions.

token is more likely to be attributed to this label. Averaging all the tokens in the input text gives an attribution score of the whole text, where a high average means that that text is more likely attributed to the label, and a low average means the text is less likely to be attributed to the label. In order to find the most salient tokens for each category, we average the score of each token for each archetype, and examine the tokens with the highest average score. Words and tokens with a high score are those words that differentiate the archetypes from one another because the presence of those words in a text indicate that that text is more likely to be spoken by one archetype or another.

4. Results

4.1 Analysis 1: Character Archetype Classification

[Table 1](#) shows the performances of the model on the test set. The model shows a performance that is far from perfect ($F1=0.47$), but at the same time substantially above chance. The performance differs majorly between archetypes, with good performance for *criados*, *criadas* and *damas* and bad performance for *galanes*, *reyes* and *reinas*. This pattern can be explained to an extent by looking at the confusion matrix shown in [Table 2](#) (correct labels in rows, model predictions in columns). *Galanes* and *damas*, the two archetypes for which the model predictions are mostly the correct class, are also the two classes for which we have most data, while the category of *reina*, which is never predicted correctly, is the rarest class. This underlines the role of frequency in the model behavior.

However, we can also make observations that are interesting from a character analysis point of view. The most frequent incorrect guesses were frequently those of the same gender in an adjacent social class, or those of the wrong gender within the same social class. For example, *galanes* were most frequently incorrectly predicted as *damas* or *reyes*. *Criadas* were most frequently confused with *damas*. *Reyes* were more frequently guessed as *damas*. *Criados* were also frequently confused with either *damas*. Additionally, there were no incorrect guesses that transcended two social classes, i.e., there was no confusion between *reinas* and *reyes*, and *criados* and *criadas*, suggesting that there is

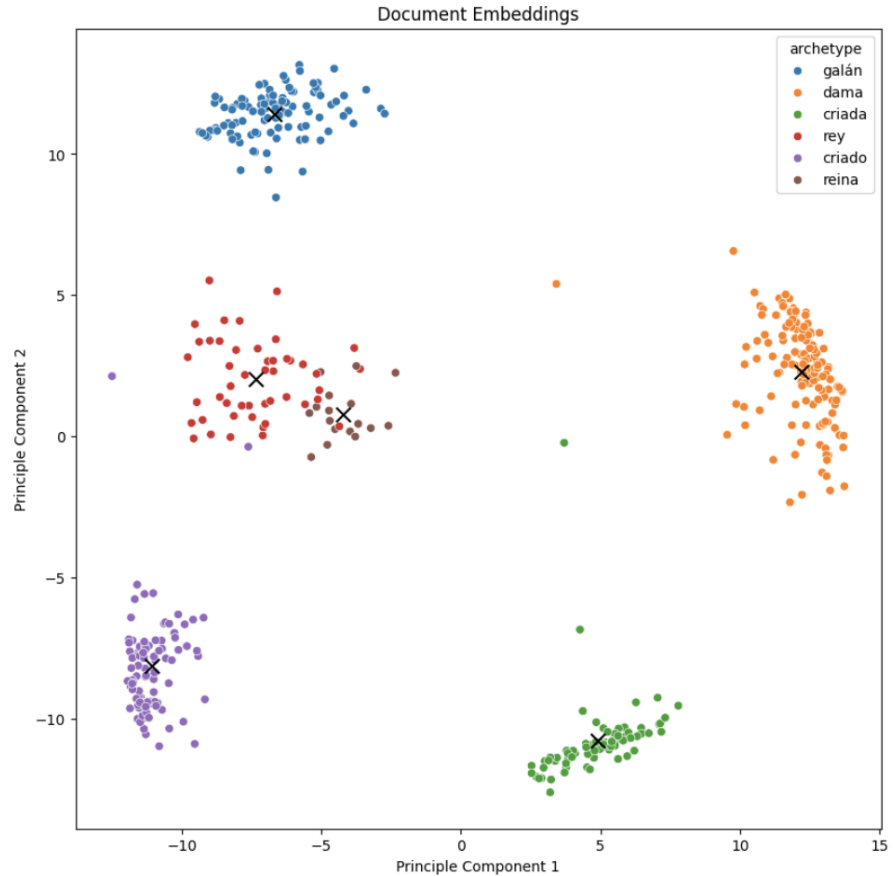


Figure 2: Principle component analysis of embeddings of character speech. Centroid Locations: *galán*:(-6.69, 11.40), *dama* (12.20, 2.30), *criada* (4.91, -10.79), *rey* (-7.35, 2.03), *criado* (-11.07, -8.14), *reina* (-4.22, 0.79)

a fundamental difference that makes the speech of the royal characters different to that of the servant characters. Although most of the incorrect guesses by the model were wrongly predicting a character as a *dama*, the *reinas*, whom we might expect to be predicted as *damas* as well due to the shared gender, were always predicted to be *reyes*. This indicates that the function of *reinas* in the works is so similar to that of *reyes*, that it over-shadows any influence of gendered speech.

One possible explanation for the fact that *criados* are predicted correctly at high frequency is that the *criado* characters were written in a formulaic way. The *criado* characters are less likely to be main characters compared to *galanes* or *damas*. Instead, they are more likely to serve a specific purpose as a plot device. There is less of a need for the *criados* to be unique individuals compared to main characters like *galanes* or *damas*, and therefore it may be likely that these characters follow a specific set of conventions compared to other character archetypes.

4.2 Analysis 2: Inspection of Categories via Dimensionality Reduction

Figure 2 visualizes the character embeddings, reduced to two dimensions with principal components analysis and colored by character archetype. We see that most of the character groups are distinct from one another with very little overlap, but that some

galán	dama	rey	reina	criado	criada
1.46	1.52	1.83	1.13	1.61	1.32

Table 3: Average distance of each character to its archetype, for all archetypes

archetypes have a couple of notable outliers, which we will mention below. We see a clear separation on the x-axis, Principle component 1, which appears to correlate to the characters' gender. There is an overlap between *reinas* and *reyes*. We also see that the archetypes from each 'social class' somewhat align with one another in the y axis in the second principle component.

We also calculate, for each archetype, the average Euclidean distance between each instance of the archetype and the archetype's centroid. In order to mitigate possibility that the number of characters in each archetype would affect the results, in the case that the model potentially learns better the archetypes with more instances, we used a sampling technique to calculate the centroid distances. We repeatedly sampled 17 unique characters of each archetype (equal to the number of queens in the sample, which was the least frequent class) so that all the classes were the same size. We then calculated the average distance from the centroid for each round of sampling, and averaged all 8 sampling rounds to obtain the average distance to the centroid for each category. Using this method, there was no correlation between the number of characters in a class and the coherence of that class. There was also no correlation between the average number of words spoken by each character archetype and the coherence of the category, indicating that the coherence of the category truly represents the coherence of that category and not the input length or the number of input instances.

We interpret these numbers, shown in Table 3, as a measure of archetype consistency: A low average distance indicates that an archetype has high coherence or specificity – i.e., its instances are all very similar to one another. In contrast, a high average distance can indicate that an archetype consists of several subtypes, or that its characters exhibit a large degree of individuality over and above their membership in an archetype.

The *galán* archetype was neither very cohesive nor dispersed (distance to centroid = 1.46). The most prototypical *galanes*, the *galanes* closest to the central point or 'prototype' are: Epafo of *El Faetonte*, Enrique of *El secreto a voces*, and Antonio of *Cual es mayor, perfección, hermosura, o discreción*. The least prototypical *galanes* are Petosiris of *los-hijos-de-la-fortuna-teagenes-y-cariclea*, Don Fernando of *Mañana será otro día*, and Álvaro from *Primero soy yo*.

The *dama* is also averagely cohesive compared to the other archetypes (distance to centroid = 1.51). The most prototypical *damas* were Tetis from *El Faeton*, Serafina of *Dicha y desdicha del nombre*, and Cintia of *Los dos amantes del cielo*. The *damas* farthest away from the prototype were: Leonor of *Con quien vengo, vengo*, Estela of *Amigo, amante, y leal*, and Violante of *También hay duelo en las damas*. Contrary to expectations, the cross-dressing characters, Rosaura of *La vida es sueño* and Claridiana of *El Castillo de Lindabridis*, were not among the top three 'atypical' *damas*.

King characters are more dispersed from a central point (distance to centroid = 1.81). The most prototypical kings in the corpus are the *rey* from *Amor, honor, y poder*, Basilio of *La vida es sueño*, and Sabinio from *Las armas de la hermosura*. The least prototypical

kings are: Ulises and *El rey* of *El monstruo de los jardines*, and Arsidas from *Amor se libra de amor*. 338 339

The *reina* character is the least dispersed (distance to centroid = 1.32), which seems to contradict the one source that described *siglo de oro* queens as being from many different types. Perhaps, while the queens are historical, biblical, mythological or fictional, the roles that they play in the works are much more defined. The most prototypical queens, the closest three queens from the central point, are Clodomira from *La exaltación de la cruz*, Admeta from *Los hijos de la fortuna: Teagenes y Cariclea*, Cristerna from *Afectos de odio y amor*. The queens that are the least prototypical are: Persina from *Los hijos de la fortuna: Teagenes y Cariclea*, Hianisbe from *Argenis y Poliarco* and Semíramis from *La hija del aire I*. 340 341 342 343 344 345 346 347 348

Criados were the second most dispersed character (distance to centroid = 1.61). The most prototypical *criados* were: Floro from *La señora y la criada*, Oton from *La selva confusa*, and Espinel *Bien vengas mal, si vienes solo*. The least prototypical *criados* were: Dinero from *Mejor está que estaba*, Poliarco from *Argenis y Poliarco*, and Turín from *Afectos de odio y amor*. In the DraCor cast list of *Argenis y Poliarco*, a lesser known work by Calderón, the titular Poliarco is listed as a *criado*. However in a Calderón Digital he is described as a French knight and the love interest of ArgenisAntonucci n.d. 349 350 351 352 353 354 355

Criadas were also medium dispersed (distance to centroid 1.32). The most prototypical *criadas* are: Sirena from *A secreto agravio, secreta venganza*, Inés from *No hay cosa como callar*, and Flora from *El postrer duelo de España*. The least prototypical *criadas* are: Ines from *Bien vengas mal, si vienes solo*, Flora from *El encanto sin encanto*, and Lesbia from *Afectos de odio y amor*. Perhaps another inconsistency in the corpus, Lesbia from *Afectos de odio y amor* is labeled as a *criada* in the DraCor cast list, however, in Calderón Digital she is described as a *dama*, and the ex lover of the king Sigismundo. 356 357 358 359 360 361 362

However, it should be noted that PCA only represents the two most salient dimensions in the embedding space, likely oversimplifying the results. The proximity of two categories in the PCA plot therefore is possibly an artifact of the information loss in the dimensionality reduction. 363 364 365 366

The PCA analysis corresponds to some findings from previous scholarship on character portrayal. It would be of interest therefore, to see if this finding replicates in other works of the time. The *criada* archetype seems to be the least diverse, possibly indicating that Calderón followed a more strict formula for writing the *criadas*, and could also signify that these characters had a stricter social role. Of course one possible interpretation of this apparent lack of diversity could be due to the strict social roles for women during this time period. Conversely, the *dama* archetype was less cohesive than many others. While female characters of the time did have a strict social role as discussed in subsection 2.1, we might attribute this finding to the fact that *dama* was the most widely used label in the corpus and therefore encompasses many different women characters who might be diverse, as opposed to male characters for whom there are many different labels used in the corpus. Previous work did find that certain *damas* and *reinas* who cross-dress in these works were more likely to be similar to male characters (Keith et al. 2024). This trend was not found in the present study, instead, these characters seemed to be no more or less typical than any other *damas*. 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381

Archetype	galán	dama	rey	reina	criado	criada
1	palace	-erti-	king	king	pink/rose	vest-
2	dei-	lover	dei-	queen	mountain	street
3	death	love	empire	empire	vin-	-erti-
4	die	fame	ray	crown	vest-	pia-
5	guard	gener-	crown	freedom	palace	-rade-
6	street	father	freedom	weapons	guard	door
7	dead	brother	weapons	ray	loc-	pink/rose
8	house	life	queen	-erti-	cover	speak
9	land / earth	freedom	blood	peace		sir, lord
10	deu-	honor				enam-

Table 4: Ten highest-scored words and subword tokens associated with each character archetypes (full words in English, subword tokens in Spanish)

We also considered the genre as a confounding factor. However, there appeared to be no clear pattern about the location of characters from different genres in the embedding space, and no correlation in the classification model.

4.3 Analysis 3: Attribution

We examine the words with the highest attribution scores for each category, meaning the words most likely to be spoken by a given character archetype. The purpose of this analysis is to assess the extent to which the model picks up on the cues that a domain expert would also consider as informative for the classification, as opposed to artifacts of the training data.

The current analysis is based on Table 4, which shows the top words, in order of highest average attribution score, associated with each character archetype. Due to space reasons, we only discuss English translations. Furthermore, while we interpret sub-word tokens that correspond to recognizable roots, we ignore sub-word tokens that do not carry semantic meaning or correspond consistently to unique semantic concepts. The full results in Spanish can be found in our GitHub repository (see section 6).

The top indicators for *galanes* are: **palace, dead, die, guard, street, death, house, land**. In the list were also two subword tokens: "*dei-*" which was associated with **deity** (occurring in the words *deidad* and *deidades*) and "*deu-*" which always occurred in the words meaning **debt** or debtor. These are all tokens that correspond with characterization of the *galán* archetype in previous literature. Specifically of interest is that *galanes* are more likely than other character types to discuss financial matters (**land, debt, house**), which corresponds to previous work stating that honor conflicts for *galanes* often involve property of some sort. The *galán* archetype does also discuss conflicts surrounding love, for example, but the theme of property is a key distinguishing factor between *galanes* and the other archetypes examined in this paper.

For *damas*, the highest attribution score was held by the subword token "*-erti-*" that, in this corpus, was always associated with some variation of the word *divertirse* meaning to **enjoy**. Next were: **lover, love, and fame**. Then the subword token "*gener-*" which always occurred in variations of the word *generous*. Then **father, brother, life, freedom, and honor**. Here we also see that the model builds its representations based on concepts that correspond to our understanding of the character archetypes.

Damas are more likely to mention the men that surround them (**lover, father, brother**) 413
and are more likely than other character types to be involved in conflicts where love is it 414
primary motivator (**lover, love**). Contrary to what we might expect, the word *honor* is 415
attributed more to *damas* than to *galanes*. Honor is a major theme for these two character 416
archetypes and appears commonly in the speech of both. 417

Words that were more likely to cause the model to predict the speaker as a *rey* were: king, 418
"dei-", **empire, ray** (meaning **ray of the sun, or lightning**), **crown, freedom, weapons,** 419
queen, and blood. The words most likely to be spoken by *reinas* were: **king, queen,** 420
empire, crown, freedom, weapons, ray, and "-erti-" (again, occurring in words related 421
to the verb to **enjoy**), and **peace**. We can see here that there is a great deal of overlap 422
between the words spoken by *reinas* and those spoken by *reyes* (**king, queen, empire,** 423
crown, freedom, weapons). However, interestingly, the word *blood* is more likely to be 424
associated with *reyes* while the word **peace** is more likely to be associated with *reinas*. 425
Also importantly, the prefix *dei-*, associated with deities, differentiates *reyes* from *reinas*, 426
with *reyes* being more likely to mention religious figures. 427

The words most likely to be spoken by *criados* were: **pink/rose, mountain, 'vin-'** (a root 428
of the verb *venir*, **to come**), **'vest-' palace, guard, "loc-"** (a root for the word **crazy, or** 429
craziness), and *cover*. As previously mentioned *criados* are frequently used to comment 430
on the action of the plays. In light of this, the roots *vin-* and *loc-* seem to indicate 431
commenting on dramatic action. 432

The top words for *criada* characters were: **'vest-'** a sub-word token that most often 433
occurred in words relating to dress or **getting dressed**, then **street**, the same subword 434
token "-erti-", another subword token **"pia-"** which occurred in words for **pious people** 435
(*piado* and *piados*), **"rade-"** which always occurred in variations of the word *agradecido* 436
meaning **thankful**. Then, **door, pink/rose, speak, sir/ lord**, and then **"enam-"** which 437
occurs as the root in words related to **falling in love**. Perhaps a theme that emerges 438
here is the standard for the ideal woman to be pious (**pious, thankful, Lord**). It seems 439
that here the *criada* is embodying the ideal 'womanly' traits of humility and piety (*pia-*, 440
-rade-, **sir**). We also know that *criadas* most frequently interact with *damas* and *criados* 441
in the works, which explains why there is some overlap between the speech of these 442
characters. 443

Many of the themes that appear in the most attributive words fall in line with themes 444
attributed to different character types in literary scholarship. *Damas* are discussing love 445
(McKendrick 1974), kings and queens are discussing their empires (Lauer 2017) etc. 446
There seems to be a great deal of overlap between the words most likely to be attributed 447
to *reinas* and the words most likely to be attributed to *reyes*, indicating that the *reinas* 448
and *reyes* are serving a similar purpose in the works. One expectation that was not 449
met, was that of different speech characteristics. The tokens that were most indicative 450
of each gender tended to be nouns, adjective, or verb roots, but were not specific to 451
any grammatical gender or verb tenses. This indicates that the key defining are lexical 452
items relating to primary themes of the works rather than grammatical features. Further 453
exploration should place specific emphasis on stylistic speech differences. It is likely 454
these distinctions do exist between the character archetypes, even if they are not among 455
the top differentiating tokens. 456

Limitations. One major limitation to this study is the sparsity of data for training a classifier model. We combated the issue by enacting measures to prevent the model over-fitting to the specific training data in order to improve generalization. However, the possibility that the data is not distinct enough to make reliable classifications remains. This work chose to focus solely on the works of Calderón de la Barca. However, a more complete investigation of the portrayal of character archetypes would benefit from including plays from other authors, both because more training instances make the classification model more robust, and because it would offer the option to make conclusions about a broad characterization of characters that's not specific to one author.

5. Conclusion

In this study, we proposed to use a scalable reading approach to analyze the representation of Calderonian character archetypes in a computational classification model from three complementary perspectives. Our study shows both the benefits and the limitations of this approach.

We were able to draw together information from more than one hundred dramas, using text classification essentially as an aggregation method. The success of the model, albeit limited, shows that it learned regularities about character archetypes, and our inspection of important inputs through the attribution method confirmed that these regularities are not merely artifacts of the training data. We were able to draw some interesting observations from the data. For example, we expected that gender would have some effect on the character prediction, however, it appeared to have no effect (wrong predictions by the model were no more likely to be the same gender).

By examining the dispersion of the character archetypes, we found some character types like *criadas* were more likely to adhere to a strict pattern of portrayal. Generally, archetypes seem to be strongly grounded in topics, which aligns well with observations from literary studies. We also found that, in all three analyses, there was a great deal of similarity between the *rey* and *reina* archetypes. These findings indicate that these characters fulfilled similar roles throughout the works and that any gender markers in the speech of these characters were outweighed by the content of the speech, which made these queens more similar to kings. The fact that the results align so closely to what we might expect, given our knowledge of character tropes of the Spanish baroque, points to the ways in which authors abide by dramatic norms.

We observe that (in-)frequency remains a challenge. Even taking all of Calderón's digitally available dramas into account, the dataset contained only seventeen *reina* characters, only two of which made it into the test set. Clearly, this set is too small to draw strong conclusions from. In fact, it is surprising that the results for the attribution analysis in [subsection 4.3](#) are as sensible results as they are – indicating that the grounding of character archetypes in their utterances provides access to rich information encoded in linguistic regularities even if the archetype has few instances.

In sum, we conclude that a scaled reading approach confirms descriptions by literary scholars, offering more evidence towards the depiction of character archetypes at a large scale. However, the strengths of the method would arguably profit from further scaling

up, beyond Calderón, towards a general analysis of character archetypes in *Siglo de Oro* 499
 dramas, including the work of other authors such as Lope de Vega. This would require 500
 overcoming practical hurdles, though, since other authors' works aren't generally as 501
 easily accessible and consistently represented as Calderón's in CalDraCor. 502

6. Data Availability 503

The Corpus used for the investigation, CalDraCor, is part of the DraCor Project. The 504
 project reflects the state of the Corpus available in a forked repo here: [https://github](https://github.com/allisonakeith/caldracor2025) 505
[.com/allisonakeith/caldracor2025](https://github.com/allisonakeith/caldracor2025). 506

7. Software Availability 507

The code used in this investigation can be found in the following repository: [https:](https://github.com/allisonakeith/calderon-archetypes) 508
[//github.com/allisonakeith/calderon-archetypes](https://github.com/allisonakeith/calderon-archetypes). 509

8. Acknowledgements 510

This work was conducted as part of the 'Identifying Regularities in the works of Pedro 511
 Calderón de la Barca ' project (PA 1956/10-1) funded by the Priority Programme / 512
 Schwerpunktprogramm 'Computational Literary Studies' (DFG SPP 2207). 513

9. Author Contributions 514

Allison Keith: Writing – original draft, Writing – review & editing, Investigation, Formal 515
 analysis 516

Antonio Rojas Castro: Conceptualization, Writing – original draft, Writing – review & 517
 editing 518

Kerstin Jung: Project Administration 519

Hanno Ehrlicher: Conceptualization, Project administration 520

Sebastian Padó: Conceptualization, Supervision, Project administration 521

References 522

- Álvarez Sellers, María Rosa (2015). *Tragedia, comedia y tragicomedia desde la preceptiva* 523
dramática: para una poética de los géneros en los siglos de oro. Biblioteca Virtual Miguel 524
 de Cervantes. 525
- Bamman, David, Brendan O'Connor, and Noah A Smith (2013). "Learning latent per- 526
 sonas of film characters". In: *Proceedings of the 51st Annual Meeting of the Association* 527
for Computational Linguistics (Volume 1: Long Papers), 352–361. 528
- Bamman, David, Ted Underwood, and Noah A. Smith (2014). "A Bayesian Mixed Effects 529
 Model of Literary Character". In: *Proceedings of ACL*, 370–379. [https://aclantholog](https://aclanthology.org/P14-1035/) 530
[y.org/P14-1035/](https://aclanthology.org/P14-1035/). 531


- Beine, Julia Jennifer (2024). "The Schemer Unmasked. Sketching a Digital Profile of the Scheming Slave in Roman Comedy". In: *Journal of Computational Literary Studies* 3.1. 10.48694/jcls.3670.
- Antonucci, Fausta (n.d.). <http://calderondigital.unibo.it>. Base de datos, argumentos y motivos del teatro de Calderón.
- Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez (2023). "Spanish pre-trained bert model and evaluation data". In: *arXiv preprint arXiv:2308.02976*.
- Couderc, Christophe (2006). *Galanes y damas en la comedia nueva: una lectura funcionalista del teatro español del Siglo de Oro*. Iberoamericana / Vervuert.
- (2012). *Le théâtre tragique au siècle d'or. Cristóbal de Virués, Lope de Vega, Calderón de la Barca*. Presses Universitaires de France.
- Culpeper, Jonathan (2014). "Keywords and characterization: An analysis of six characters in Romeo and Juliet". In: *Digital Literary Studies*. Ed. by David L. Hoover, Jonathan Culpeper, and Kieran O'Halloran. Routledge, 9–34.
- De Armas, Frederick A. (2015). "Sultanas, reinas, damas y villanas: figuras femeninas en la comedia ecfrástica del Siglo de Oro". In: *Hispanófila* 175, 49–61.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL*, 4171–4186. <https://www.aclweb.org/anthology/N19-1423.pdf>.
- Ehrlicher, Hanno, Jörg Lehmann, Nils Reiter, Marcus Willand, et al. (2020). "La poética dramática desde una perspectiva cuantitativa: la obra de Calderón de la Barca". In: *Revista de Humanidades Digitales* 5, 1–25. 10.5944/rhd.vol.5.2020.27716.
- Elson, David, Nicholas Dames, and Kathleen McKeown (2010). "Extracting Social Networks from Literary Fiction". In: *Proceedings of ACL*, 138–147. <https://aclanthology.org/P10-1015/>.
- Elson, David and Kathleen McKeown (2010). *Automatic Attribution of Quoted Speech in Literary Narrative*. <https://ojs.aaai.org/index.php/AAAI/article/view/7720>.
- Elvira, Ana Contreras (2014). "La criada maga en la comedia de magia del siglo XVIII, o de escenógrafas y pedagogas en el ocaso del Antiguo Régimen". In: *Cuadernos de Ilustración y Romanticismo: Revista del Grupo de Estudios del siglo XVIII* 20, 43–73.
- 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 Digital Humanities Conference 2019*.
- Keith, Allison, Antonio Rojas Castro, and Sebastian Padó (2024). "Computational Analysis of Gender Depiction in the Comedias of Calderón de la Barca". In: *arXiv preprint arXiv:2411.03895*.
- Krautter, Benjamin, Janis Pagel, Nils Reiter, and Marcus Willand (2020). "'Ein Vater, dächte ich, ist doch immer ein Vater': Figurentypen und ihre Operationalisierung". In: *Zeitschrift für digitale Geisteswissenschaften* 5. 10.17175/2020_007.
- Kroll, Simon (2017). *Las comedias autógrafas de Calderón de la Barca y su proceso de Escritura*. Peter Lang.
- Lauer, A. Robert (2017). "Revaloración del concepto del honor en el teatro español del Siglo de Oro". In: *Hipogrifo: Revista de Literatura y Cultura del Siglo de Oro* 5.1, 293–304.

- Lehmann, Jörg, Sebastian Padó, et al. (2022). "Clasificación de tragedias y comedias en las comedias nuevas de Calderón de la Barca". In: *Revista de Humanidades Digitales* 7, 80–103. [10.5944/rhd.vol.7.2022.34588](#).
- Lorenzo, Laura Hernández (2024). "Estilometría y género: aproximación a los personajes teatrales de Calderón y Sor Juana Inés de la Cruz". In: *Ínsula: revista de letras y ciencias humanas* 930, 32–36.
- Lorenzo, Luciano García (2008). *La criada en el teatro español del Siglo de Oro*. Vol. 171. Editorial Fundamentos.
- Mañero Lozano, David (2009). "Del concepto de decoro a la «teoría de los estilos». consideraciones sobre la formación de un tópico clásico y su pervivencia en la literatura española del Siglo de Oro". In: *Bulletin hispanique. Université Michel de Montaigne Bordeaux* 111-2, 357–385.
- McKendrick, Melveena (1974). *Woman and Society in the Spanish Drama of the Golden Age: A Study of the mujer varonil*. Cambridge University Press.
- Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu (2019). "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44, 22071–22080. [10.1073/pnas.1900654116](#).
- Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
- Pierse, Charles (2021). *Transformers Interpret*. Version 0.5.2. <https://github.com/cdpierse/transformers-interpret>.
- Quintero, Maria Cristina (2017). "Women and power in the Spanish Theatre of the Golden Age: The figure of the Queen". In: *Renaissance Quarterly* 70.1, 384–386.
- Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). *Improving language understanding by generative pre-training*. Tech. rep. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- Ríos Carratalá, Juan Antonio (2022). "Teatro del Siglo de Oro (curso 2021-2022)". In: *Teatro Español del Siglo de Oro*.
- Ruggerio, Michael J (1972). "Dramatic Conventions and Their Relationship to Structure in the Spanish Golden Age "Comedia"". In: *Revista Hispánica Moderna* 37.3, 137–154.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of ACL*, 1715–1725. <https://aclanthology.org/P16-1162/>.
- Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. JMLR.org, 3319–3328.
- Táuler, Álvaro Bustos, Elena Di Pinto, and José María Díez Borque (2014). *¿Hacia el gracioso?: Comicidad en el teatro español del siglo XVI*. Visor Libros.
- Tenney, Ian, Dipanjan Das, and Ellie Pavlick (2019). "BERT Rediscovered the Classical NLP Pipeline". In: *Proceedings of ACL*, 4593–4601. <https://aclanthology.org/P19-1452/>.
- Tracy, Daniel G (2016). "Assessing digital humanities tools: Use of scalar at a research university". In: *portal: Libraries and the Academy* 16.1, 163–189.
- Weitin, Thomas (2017). "Scalable Reading". In: *Zeitschrift für Literaturwissenschaft und Linguistik* 47, 1–6.

A. Appendix	623
A.1 Experimental Details	624
Classification We used an 80-10-10 split on the data for training, testing, and validation respectively. We used BETO create the embeddings as it is a multi-language embedding model specifically for Spanish. We implement early stopping and a dropout layer during training to combat over fitting. We use cross entropy loss and the Adam optimizer. The analysis in the results section is performed on the predictions of the model for the 10% of data points in the validation subset.	625 626 627 628 629 630
Dimensionality reduction and Attribution We utilized all of the same parameters to train the embeddings for dimensionality reduction, as with classification, but using all the character data so that all characters could be plotted. In order to visualize the data, we used principle component analysis (PCA) to reduce the dimensionality to the 2 most salient dimensions. We also use the interpreter model on all the data (not just the test data).	631 632 633 634 635 636

Encoding Imagism? Measuring Literary Imageability, Visuality and Concreteness via Multimodal Word Embeddings

Yuri Bizzoni¹ 
Pascale Feldkamp¹ 
Kristoffer L. Nielbo¹ 

1. Center for Humanities Computing, Aarhus University , Aarhus, Denmark.

Citation

Yuri Bizzoni, Pascale Feldkamp, and Kristoffer L. Nielbo (2025). "Encoding Imagism? Measuring Literary Imageability, Visuality and Concreteness via Multimodal Word Embeddings". In: *CCLS2025 Conference Preprints 4* (1). [10.26083/tuprints-00030154](https://doi.org/10.26083/tuprints-00030154)

Date published 2025-06-17

Date accepted 2025-04-17

Date received 2025-02-07

Keywords

imageability, concreteness, visuality, embeddings, imagism, multimodal embeddings

License

CC BY 4.0 

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 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025. Please check jcls.io for the final journal version.

Abstract.

This paper addresses the challenge of measuring "imageability" in literary texts – a concept from psycholinguistics that describes how words evoke sensory experiences. Imageability is connected to literature's immersive quality, but current methods face limitations due to vague definitions, poor lexical coverage, and difficulty scaling to sentence-level (/literary) analysis. To tackle this, we propose a data-driven approach using a multimodal model, alongside traditional word embeddings, to quantify imageability more effectively. Through three experiments: 1) a word-level analysis, 2) sentence-level comparison, and 3) a case study contrasting imagist and love poems – we test whether embeddings created with a multimodal model capture imageability, and the related features concreteness and visuality. We assess the extent to which multimodal embeddings capture imageability in literary texts, while considering compositionality and the multimodal nature of literary imagery.

1. Introduction

"The boy took the old army blanket off the bed and spread it over the back of the chair and over the old man's shoulders." – *The Old Man and the Sea*, Ernest Hemingway

Ernest Hemingway's prose is famously sparse, but conjures vivid mental images: Simple actions and objects – no florid descriptions, no overt emotional cues – yet the scene is immediately present, affective and immersive. One might say that the strength of literary texts lies in their *imageability*.

The concept of imageability originates in psycholinguistics, where it describes the ease with which words evoke sensory experiences (Paivio et al. 1968). However, when speaking of the imageability of literary texts, we are going beyond individual words, and rather touching upon implicit and evocative strategies, including imagery, narrated perception, and the overall immersiveness or experientiality of the text – strategies that have long been held to increase the appeal of texts and enhance the reading experience (Ellen J. Esrock 1994; Sharma Paudyal 2023).

These strategies are related. We can define imagery as language use that appeals to our senses – creating mental images (Lacey and Lawson 2013; Sharma Paudyal 2023). For example, Burroway (1987) notes that a certain use of nouns that evoke

sensory images and of verbs that represent visual actions makes writing “come alive”.¹⁸ This effect aligns with the broader concept of literary experientiality (Fludernik 1996).¹⁹ Experientiality describes how narratives prompt embodied engagement through a “quasi-mimetic evocation of real-life experience”, drawing on knowledge readers have acquired from their physical presence in the world. Psychological and reader-response studies further support this link, associating imagery with the intensity of emotional responses in reading and feelings of embodiment (Blackwell 2020; Goetz et al. 1993; Martínez 2024). By measuring imageability, we can gain insights into how texts evoke embodied responses, and better understand the immersive qualities that shape reader engagement.²⁷

When measuring the imageability of literary texts, studies typically rely on dictionaries developed in psycholinguistic research that assign imageability or concreteness scores to words (Feldkamp et al. 2024; J. T. Kao and Jurafsky 2015). However, the use of such word-based dictionaries presents issues. We identify three main limitations in current computational approaches to imageability: conceptual vagueness, poor lexical coverage, and lack of compositionality at the sentence level. First, imageability itself is not a straightforward property but tends to be inherently vague, with human judgments often diverging.¹ It spans literal and figurative language, object descriptions, and metaphorical expressions that rely on shared cognitive schemas – not merely knowledge of the visual world. The precise nature of what imageability encodes remains debated, and it is often conflated with related constructs such as visualness and concreteness. While some studies suggest that imageability primarily reflects visual features of language (Ellis 1991), it strongly correlates with concreteness and visuality (Brysbaert et al. 2014), making it difficult to isolate as a distinct linguistic property. To test the relationship to these related concepts, we would need to compare dictionaries for imageability, visuality and concreteness.⁴³

A second main issue is that dictionary-based scores struggle with sentence-level imageability. The accuracy of aggregated imageability, visuality, or concreteness scores at the sentence level is often poor (Verma et al. 2023), which may stem from the limited coverage of dictionaries and the questionable assumption that averaging word-level scores yields a meaningful sentence-level representation. A third issue is that dictionary-based approaches fail to account for compositionality – phrases such as “She painted a dark picture” and “She painted a picture in the dark” have different imageability despite containing similar words.²⁵¹

Regarding the latter issue, recent advances in natural language processing (NLP) have attempted to quantify imageability using multimodal models that integrate textual and visual information Verma et al. 2023; S. Wu and Smith 2023. However, approaches such as gauging imageability through text-to-image generation output show uneven relation to human judgement, especially for literary texts (S. Wu and Smith 2023).⁵⁶

In sum, existing methods for computing imageability scores face three main challenges:⁵⁷

1. In Verma et al. (2023), inter-annotator agreement for sentence-level imageability ratings was 0.45 (Krippendorff’s α). Note that this was for non-literary texts, where we might expect literary texts to effect an even lower agreement, which seems to be the case in annotation tasks for other concepts (Feldkamp et al. 2024).

2. Also, note that most dictionaries assign imageability scores at the lemma level, abstracting from the word-form level. Working from lemmas means that the variations in word forms – such as ‘painted’ vs ‘paint’ – have the same imageability score, even though differences in tense and part-of-speech category may evoke a different intensity and, in theory, a different set of sensory associations for human readers.

(1) the vague conceptualization of imageability, (2) the limited lexical coverage of dictionary-based approaches, and (3) the difficulty of generalizing from word- or phrase-level scores to sentence-level imageability. These issues are relevant for literary texts: imageability is a core stylistic and aesthetic device; and since literature constructs immersive sensory experiences through language alone, it is an ideal domain for testing computational models of imageability. Unlike instructional or descriptive texts, literary language frequently employs figurative expressions, ellipsis, and symbolic imagery—requiring more nuanced tools to capture sensory evocativeness at the sentence or paragraph level.

To address these limitations, we propose a data-driven, scalable approach that moves beyond static dictionary-based methods. Given the impact of imageability and concreteness on immersivity – and, by extension, reader appreciation – we explore automatic assessment techniques based on text representations. Prior work has demonstrated the visual knowledge of text-only models (Sharma et al. 2024), while recent advances in multimodal models (Radford et al. 2021) offer new opportunities for capturing the visual dimension of language.

Examining the shape of both text-based and multimodal embeddings, we test their ability to approximate imageability, concreteness, and/or visuality scores.

Specifically, we evaluate their efficacy in characterizing literary texts through three experiments:

- Word-level analysis: We assess the relationship between human imageability scores and metrics of multimodal word embeddings for dictionary entries of the imageability dictionary.
- Sentence-level analysis: We compare dictionary-derived scores with metrics of multimodal sentence embeddings for literary texts.
- Literary case study: We examine the discriminatory power of these embedding-based metrics in distinguishing between text types where imageability is expected to differ: imagist poems versus love poems.³

By systematically evaluating these approaches, we aim to develop a more robust framework for measuring imageability in literary texts—one that accounts for compositionality, sentence structure, and the multimodal nature of literary imagery.

While our study does not include human annotations, it represents an initial computational exploration aimed at (1) testing the relationship between imageability and related constructs such as concreteness and visuality, and (2) evaluating the potential of embedding-based metrics to model imageability beyond static, word-level ratings.

3. In this third experiment, we use imagist poems as a testbed to probe whether multimodal embeddings capture stylistic and sensory variation. This should not be taken to imply a direct equivalence between imageability and imagism, nor a reduction of poetic imagery to literal visual representation. Nonetheless, the historical emphasis of imagist poetry on economy, concreteness, and sensory immediacy makes it a useful comparative corpus for our purposes.

2. Related Works

2.1 Imageability in literary texts

The evocation of mental imagery in literary texts has been a debated topic in literary and psychological scholarship (Kuzmičová 2014). Despite its prominence in early 20th century literary movements like Imagism, where the emphasis was placed on clear, visual language and the rejection of abstraction (Pound 1913), the role of imagery and the imageability of literature was often overlooked in structuralist and New Criticist frameworks, prioritizing linguistic networks and meaning-making (Ellen J. Esrock 1994). However, in recent years, the concept of imageability has regained attention, with both literary scholars and psychologists increasingly examining its role in reader response (Kuzmičová 2014; Magyari et al. 2020; Martínez 2024; Sharma Paudyal 2023).

Imageability, defined as the ability of a text to evoke sensory experiences and mental imagery, is closely linked to the heightened emotional responses that images can provoke (Goetz et al. 1993) and the embodied nature of reading experiences (Martínez 2024). Literary passages that employ concrete, sensory language—those that engage the senses without explicit emotional cues—have been shown to elicit emotional responses from readers. For instance, Hemingway’s minimalist style (see our example above), which uses stark imagery without overt emotional direction, is perceived as emotionally charged by human readers, despite being classified as neutral by automatic sentiment analysis systems (Feldkamp et al. 2024). Furthermore, the evocation of interoceptive or physiological states can activate a reader’s embodied experience (Martínez 2024), while concrete language enhances emotional engagement and heightens suspense (Auracher and Bosch 2016).

While imagery, concreteness and imageability have long been concepts employed in literary analysis (Ellen J. Esrock 1994; Sharma Paudyal 2023), computational literary history studies have further shown how quantifying imageability can be employed to characterize certain literary texts. For instance, studies of poetry have shown that Imagism, with its focus on direct, visual language, is associated with higher levels of imageability (J. T. Kao and Jurafsky 2015). Additionally, a historical shift toward more concrete and imageable language in poetry has been observed, suggesting a gradual evolution of literary style over time (Ibid.).

However, quantifying imageability – along with related concepts like concreteness and visuality – is a challenge for computational literary analysis. These concepts are often defined and operationalized differently across domains. For instance, in more communicative texts, such as journalism, imageability is often tied directly to sensory or visual representations, where literary language frequently employs imagery in more abstract or symbolic ways, and may use it more strategically and with greater nuance. The concept of “implicit” expression – “show, don’t tell” – is particularly significant in literature, where imagery is often used to evoke affect without explicitly naming it (Feldkamp et al. 2024), and literary scholarship frequently use terms like ‘evocative’ or ‘understatement’ to describe authorial styles (Strychacz 2002, Daoshan and Shuo 2014), further emphasizing the subtlety of literary imageability as a strategic tool. Furthermore, literary studies have made efforts to distinguish between conceptually different types of

imagery (Kuzmičová 2014). 136

This implicit nature of literary expression poses additional challenges for computational 137
methods that rely on standardized lexical resources. Its subtleties may not align with 138
the operational definitions of imageability found in existing dictionaries or lexicons. 139

2.2 Dictionaries for imageability 140

The terms imageability, concreteness, and visuality are often used interchangeably, 141
though they capture distinct but overlapping dimensions of language. While concrete- 142
ness typically refers to the degree to which a word denotes a tangible entity, imageability 143
extends beyond the purely visual to include mental representations, including intero- 144
ceptive states (Dellantonio et al. 2014). 145

In literary studies, these constructs have been applied in different ways. For example, J. 146
Kao and Jurafsky (2012) use concreteness – or its reverse, abstractness – to assess literary 147
imagery, while J. T. Kao and Jurafsky (2015) measure “concrete imagery” through a 148
combination of object-word frequency, abstract-word frequency, and dictionary-based 149
concreteness and imageability scores. 150

Even when focusing specifically on imageability, various resources have been developed, 151
beginning in the 1960s with early psycholinguistic studies (Paivio et al. 1968). One of 152
the first large-scale lexicons, the MRC database, was compiled in the 1980s Coltheart 153
1981 and remains widely used despite its limited scope. More recent and expansive 154
dictionaries have been developed, such as the 40,000-lemma concreteness lexicon by 155
Brysbaert et al. (2014), which significantly surpasses the 4,800 lemmas found in the 156
MRC lexicon. However, coverage across different dimensions remains uneven, with 157
imageability, visuality, and concreteness lexicons varying in size and consistency. 158

2.3 Development of models for imageability 159

Recently, the visual aspect of imageability has gained significant attention in Natural 160
Language Processing (NLP), particularly in the context of text-to-image models like 161
DALL-E. These models rely on dual processing of text and images yet struggle when 162
dealing with long-form text containing spans of non-visual content. As a result, vi- 163
sualness has been proposed as a useful metric for characterizing the prompt prior to 164
generation, with the goal of improving the accuracy of text-to-image generation models 165
Chen et al. 2025; Verma et al. 2023. 166

For instance, Verma et al. (2023) introduces a binary classification task distinguishing 167
imageable from non-imageable sentences to enhance prompt characterization before 168
image generation. Similarly, Chen et al. (2025) explores the role of visualness in guiding 169
the generation process, aiming to refine model outputs. Apart from augmented image 170
generation, identifying imageable text might also have further downstream application 171
such as on the fly visuals (Liu et al. 2023) and image-assisted video navigation (Zhao 172
et al. 2019). However, while such binary classification is practical for generation and 173
other tasks, is not as usefull for describing nuanced data, where we ideally want to 174
maintain a level of granularity: gauging more or less imageable text. 175

To address the shortcomings of existing imageability dictionaries, S. Wu and Smith 176

Construct	Scope	Relation to imageability
<i>Concreteness</i>	Tangibility of a word’s referent	<i>Partial overlap.</i> Concrete words tend to be imageable, but abstract items such as “ <i>whirlwind</i> ” can evoke vivid scenes; conversely “ <i>road</i> ” is concrete yet often yields weak imagery in isolation.
<i>Visualness</i>	Strength of <i>visual</i> associations	<i>Proper subset.</i> Imageability spans <i>all</i> modalities (auditory, tactile, olfactory, ...), not vision alone.
<i>Imagery</i> (literary studies)	Textual clusters of sensory details	<i>Complementary.</i> Imagery is a textual feature; imageability is the reader-side potential.

Table 1: How imageability relates to creating mental constructs.

(2023) propose methods that incorporate sentence compositionality, aiming to better capture the nuances of how imageability evolves across different sentence structures. While their work shows promise in addressing fixedness in representations, a critical challenge remains: many texts can evoke strong mental imagery in readers without these images being strongly encoded in a culturally shared or visual sense. For instance, creative and poetic language can provoke vivid imagery that is not directly tied to shared or commonsense visual representations. When the consistency of generated images is used as a proxy for imageability (S. Wu and Smith 2023), this may actually measure the stability of a text’s representation rather than its inherent imageability. Conversely, a non-visual passage may still elicit a text-to-image model to generate superficially coherent images. With high uncertainty, text-to-image models often generate visually similar images (e.g., images of actual text) that may appear coherent but do not necessarily align with any human reader’s mental image of the text.⁴

Finally, methods for gauging imageability may show variability across genres. S. Wu and Smith (2023) finds an insignificant correlation between their imageability measure and human assessments of poem lines, yet a significant correlation for news sentences. This suggests that the effectiveness of imageability metrics may depend on the genre and its inherent stylistic and thematic characteristics.

3. What we mean by *imageability*

We follow the psycholinguistic tradition in defining imageability as the ease with which a linguistic expression evokes sensory representations in the mind of a typical reader, but we extend the *expression* from individual words to any contiguous span of text.

Given a reader r and a text span t , the imageability $I(t, r)$ is the *subjective vividness* of the multi-sensory mental imagery spontaneously elicited by t . In group studies we use the expected value $I(t) = \mathbb{E}_r[I(t, r)]$.

Two implications follow. First, imageability is a *psychological potential*, approximated by behavioural data or cognitively motivated proxies. Second, it tends to be **compositional**: the vividness of “*he smoked a crooked, emerald-green cigarette*” might not be a linear sum of the scores for its component words.

4. See examples of such visually similar but disjunct images in Verma et al. (2023), Fig. 5.

Level	Typical operationalisation	Limitations
<i>Word</i>	Psycholinguistic norms from 25–40 k-entry lexica (MRC, Lancaster, BLP)	Coverage gaps for literary vocabulary; ignores syntax and context.
<i>Sentence/line</i>	Human ratings (rare) or context-aware embeddings (this work)	Ratings costly; embeddings need interpretability.
<i>Whole text</i>	Aggregations (mean, max, entropy) over sentences	Sensitive to length and genre; reliant on robust lower-level scores.

Table 2: Granularity choices when measuring imageability.

Ideally, we would move from the *word* to the *sentence* level without sacrificing scalability. 206
 Because imageability is, by definition, a reader experience, ultimate confirmation demands sentence- or passage-level ratings. The present work should therefore be read 207
 as a *bridging effort*: dictionary-validated, embedding-based metrics ready for human 208
 calibration.⁵ 209
 210

4. Resources 211

4.1 Dictionaries 212

For Experiment I, we utilize lexicon-based resources to analyze the imageability of words, 213
 primarily relying on the **MRC Psycholinguistic Imageability Lexicon** (Coltheart 1981). 214
 This lexicon comprises 4,828 lemmas that have been rated for their capacity to evoke 215
 mental imagery. Additionally, we resort to two other well-established resources: the 216
Concreteness Lexicon Brysbaert et al. 2014, which assigns ratings to words based on 217
 their *perceived tangibility and sensory grounding*, and the **Lancaster Sensorimotor Norms** 218
 (Lynott et al. 2020), which provide detailed *modality-specific perceptual ratings* (e.g., visual, 219
 auditory, tactile associations). As the lexica of the three resources largely overlap, we can 220
 systematically compare how they conceptualize and quantify imageability, concreteness, 221
 and sensory experience. Given that previous research has noted a strong correlation 222
 between imageability and concreteness, but also some key distinctions between them 223
 (Paivio et al. 1968), our analysis seeks to clarify the extent to which dictionary-based 224
 imageability measures capture cognitive and perceptual properties distinct from general 225
word concreteness and *modality-specific sensory attributes*. 226

- The **MRC Psycholinguistic Imageability Lexicon** (Coltheart 1981). One of the 227
 earliest large-scale resources for word imageability, this lexicon contains 4,828 228
 lemmas, each rated based on the extent to which they evoke *mental imagery*. The rat- 229
 ings were collected from human participants, making it an empirically grounded 230
 resource for word-level imageability. We compare this resource with later expan- 231
 sions and refinements: (i) **Cortese and Fugett’s Imageability Ratings** (here, Imag. 232
 C) (Cortese and Fugett 2004): an updated version that increases the coverage 233
 of imageability scores and refines earlier ratings. (ii) **Reilly and Kean’s Formal** 234
Distinctiveness Model (here, Imag. R) (Reilly and Kean 2007): a lexicon that in- 235
 tegrates and updates multiple prior resources, including the MRC, while filtering 236

5. The final experiment of this paper already uses implicit human judgments by distinguishing two different literary genres.

out words with mid-range imageability ratings to focus on words that are strongly imageable or non-imageable. 237 238

- **The Visuality Lexicon of the Lancaster Sensorimotor Norms** (Lynott et al. 2020). 239
The Lancaster Sensorimotor Norms provide modality-specific sensory ratings 240
(e.g., visual, auditory, tactile, and motor associations) for 39,707 English words. 241
In our experiments, we use the *visuality* scores specifically. Unlike general image- 242
ability, *visuality* captures the extent to which a word evokes a visual percept. This 243
distinction is important because some words may be highly imageable but not 244
strongly visual (e.g., "fragrance" or "melody"). Comparing *visuality* to image- 245
ability allows us to examine how modality-specific sensory experience aligns with 246
broader notions of literary imagery . 247
- **The Concreteness Lexicon** (Brysbaert et al. 2014). This dataset provides con- 248
creteness ratings for 40,000 words, where concreteness is defined as the degree 249
to which a word refers to a tangible, physical entity. While concreteness and 250
imageability are often correlated, they are not identical concepts: some abstract 251
words (e.g., *freedom*) might be highly imageable due to their symbolic richness, 252
while some concrete words (e.g. *rock*) may elicit limited mental imagery despite 253
being physically tangible. By including concreteness as a comparative measure, 254
we assess how word-level concreteness and imageability interact, particularly in 255
literary contexts. 256

4.2 Literary texts 257

For Experiments II and III, we use full sentences from literary texts. Moving from single 258
words to entire sentences enables an assessment of how imageability, concreteness, and 259
visuality manifest in context. 260

For Experiment II, the dataset includes two modernist novels alongside a large-scale 261
corpus of fiction: 262

- *The Old Man and the Sea* by Ernest Hemingway (1952). This novel is characterized 263
by concise, concrete descriptions and a direct, unembellished prose, making it an 264
ideal candidate to evaluate imageability in an economical (yet vivid) narrative 265
style. 266
- *Mrs. Dalloway* by Virginia Woolf (1925). In contrast, Woolf's novel employs 267
stream-of-consciousness narration, featuring long, fluid sentences that foreground 268
subjective perception with immersive sensory detail. Its contrast with Heming- 269
way's style allows us to test whether opposite stylistic techniques correlate with 270
distinct levels of imageability. 271
- Sentences from the Chicago Corpus (1880–2000). A diverse dataset of 9,000 sen- 272
tences randomly sampled from 9,000 different novels, to ensure a broad coverage 273
of stylistic and historical variation in fiction. The corpus from which the sentences 274
are sampled includes works ranging from canonical literature to lesser-known 275
fiction, providing a representative snapshot of Anglophone prose writing across 276

the 19th and 20th centuries.⁶ For further details, see Bizzoni et al. (2024) and Y. Wu et al. (2024).

All sentences were tokenized using the SpaCy’s NLP library⁷. The inclusion of both single-author novels and a large, multi-author corpus allows us to assess how imageability varies both within and between different literary styles. In this context, Hemingway and Woolf serve as *controlled case studies* for contrasting narrative techniques: Hemingway’s prose is marked by concise, concrete descriptions of low abstraction, whereas Woolf’s stream-of-consciousness style favors immersive, introspective, but often highly imageable narration. These stylistic differences provide a useful basis for testing whether imageability metrics capture differences in literary technique. Their well-known status also makes them accessible and interpretable examples. Meanwhile, the Chicago Corpus, composed of diverse works spanning more than a century, offers a *broadly representative dataset* that enables generalization beyond the idiosyncrasies of individual authors.

	sentences	year	type
Hemingway	1,928	1952	prose
Woolf	3,578	1925	prose
Chicago corpus	9,000	1880-2000	prose
Imagist	1,195	1915	poetry
Modern Love	1,126	1896-1939	poetry

Table 3: Data

For Experiment III, we conduct a literary case study aimed at distinguishing *imagist poetry* from more topic-based, modern love-themed poetry using embedding-based metrics and dictionary-derived scores. The choice of Imagist poetry for testing the distinguishing power of dictionary- and embedding-based features was also made because of previous works supporting that Imagist poetry stands out in these dimensions (Gleason 2007, 2009; J. T. Kao and Jurafsky 2015). To this end, we utilize two distinct poetic datasets:

- *Some Imagist Poets* (1915), an anthology compiled by Ezra Pound, which includes 37 poems authored by six poets. Imagist poetry is characterized by its emphasis on precise and concrete imagery, minimalism, and a rejection of abstraction, making it an ideal test case for computational measures of imageability.⁸
- A selection of 74 *modern love-themed poems* by 22 authors, collected from *The Poetry Foundation’s* curated category ‘Love’.⁹

By juxtaposing these two corpora, we aim to assess whether computationally derived imageability and concreteness scores can effectively differentiate poetic traditions that

6. It should be noted that the dataset is predominantly English-language fiction, potentially limiting its generalizability to other linguistic traditions.
7. Specifically, we employed the SpaCy en_core_web_sm model, which provides robust sentence segmentation for literary texts, ensuring consistent parsing across different styles.
8. We use the 1915 Pound anthology primarily as a test case to explore computational representations of imageability. We do not claim that this selection fully captures the complexity or historical breadth of Imagism as a movement—nor that it represents the full range of poetic strategies associated with it. While alternative anthologies, such as those edited by Amy Lowell, might offer broader coverage, our aim is not to provide a literary-theoretical account of Imagism, but to use this corpus as a controlled experimental setting.
9. To maintain clear genre distinctions, we excluded poets in *Some Imagist Poets* from the love-themed poetry dataset. For a complete list of included poems and authors, see subsection A.2. The dataset is available at: <https://huggingface.co/datasets/merve/poetry>.

prioritize sensory evocation (Imagism) from those that may contain a broader range of abstract or figurative language (general love poetry). This comparison provides insight into the applicability of our methods for distinguishing stylistic and thematic variations in literary texts.

		imageability	concreteness	visuality
Hemingway		4359.2 ± 3241.6	38.1 ± 28.4	33.7 ± 26.0
	<i>normalized</i>	352.6 ± 38.0	2.8 ± 0.3	2.4 ± 0.3
Woolf		5156.8 ± 6167.0	45.5 ± 55.0	42.3 ± 51.5
	<i>normalized</i>	350.6 ± 47.7	2.7 ± 0.4	2.5 ± 0.4
Chicago		5173.8 ± 5292.2	47.3 ± 46.3	43.1 ± 43.1
	<i>normalized</i>	345.7 ± 37.1	2.7 ± 0.3	2.5 ± 0.3
Imagist		1870.6 ± 1027.3	17.1 ± 8.7	14.9 ± 7.9
	<i>normalized</i>	376.2 ± 64.5	3.0 ± 0.5	2.6 ± 0.6
Love		2112.53 ± 846.7	18.3 ± 6.6	16.2 ± 6.1
	<i>normalized</i>	363.7 ± 53.1	2.8 ± 0.5	2.5 ± 0.5

Table 4: Sentence-level average (and SD) imageability, concreteness, and visuality of datasets.

Note that in gauging the relationship between the dictionary-based features and the embeddings, we sum the imageability, concreteness, and visuality scores assigned via the dictionaries across sentences without normalizing for sentence length. This approach allows us to capture the total intensity of these features in the sentence, rather than averaging the intensity of individual words. We do this because we are interested in the overall presence or weight of these features in a given sentence. Literary texts may rely on the cumulative effect of imagery across sentences, and this approach allows us to reflect that broader, contextual presence, which we also expect the embeddings to capture as well. In Experiment III, where we compare Imagist and Love poetry lines, we do normalize for line length to replicate the methodology used in J. T. Kao and Jurafsky (2015), which assigns scores by dividing the summed imageability score of words with the number of words (extant in the dictionary).

It is important to note that whether features are assigned based on sums or length, normalized sum has significance when differentiating between groups or authors – which is what we do in Experiment III. For example, in the data summary (Table 4), we see that Love poetry, on average, has higher imageability than Imagist poetry. However, when normalizing the scores for line length, this trend is reversed, with Imagist poetry averaging 376.2 and Love poetry 363.7.

The decision to normalize feature scores when using dictionaries in literary experiments is an question of operationalization that we want to underline. For example: is a long sentence with many low- and high-imageable words *more or less imageable* than a shorter sentence with high-imageable words, if their summed scores are the same? In other words, do factors like density or brevity affect the perceived imageability of sentences? This question can only be answered by comparing human imageability judgments against both methods of score assignment. We leave this to future work and focus here only on the relationship between different systems (embeddings vs. dictionaries), not their relation to human judgment.

4.3 Embeddings

When it comes to embeddings, we employ the CLIP model (Radford et al. 2021), a multimodal vision language model trained on large-scale image-text pairs. CLIP is designed to align textual and visual representations, making it particularly suitable for capturing visual and concrete dimensions of language, which are directly relevant to our study of imageability.

Given its training objective, the semantic space of CLIP’s embeddings is expected to encode visual salience and concreteness more effectively than purely text-based models. This suggests that CLIP-based embeddings may provide a more explicit representation of imageability compared to traditional word embeddings derived from text-only corpora.

However, the extent to which multimodal representations differ from text-based embeddings in encoding sensory and imagistic properties remains an open question. To address this, we compare CLIP-based embeddings against those generated by a text-only model, specifically BERT. BERT embeddings provide a useful contrast, as they are derived solely from linguistic contexts without access to visual grounding.¹⁰ This comparison allows us to evaluate whether imageability-related features emerge naturally in textual embeddings or whether multimodal supervision enhances their representation.

It’s important to underline that psycholinguistic lexica used in Experiments I–III enter our pipeline only for validation. They provide a widely accepted yard-stick against which to gauge whether a candidate metric is even plausible. Crucially, the mapping from embedding shape to imageability is not fitted on dictionary scores—indeed no fitting is required, because norm and entropy are closed-form functions of the raw vectors. In downstream applications (e.g. analysing an unedited novel draft, or surfacing vivid quotations for digital exhibits) the dictionaries can be dropped entirely.

5. Methods

We analyze the *shape* of both word and sentence embedding representations to determine how imageability manifests in textual and multimodal embeddings. Specifically, we examine whether embeddings of highly imageable and concrete words exhibit distinct distributional properties compared to those of abstract or less imageable words.

Our initial hypothesis (H_1) is that words with higher imageability and concreteness might have more *localized values* in the embedding space, meaning that they might cluster more tightly within specific regions of the vector space. In contrast, embeddings of more abstract words, that might lend themselves to a larger array of contexts, may be more dispersed, leading to higher entropy in their distribution and lower norm (i.e., strength). To illustrate this, consider the word *dog*: especially in a multimodal model like CLIP, its embedding is likely to concentrate most of its information on specific dimensions, while a more abstract term like *beautiful* is less directly tied to a specific visual referent and may exhibit a broader, more diffuse representation across the semantic space.

10. We selected BERT due to its widespread use in word embedding research and literary analysis. In particular, the `bert-base-cased` model is frequently applied for semantic and stylistic investigations in computational literary studies Grisot et al. 2022; Paragini and Kestemont 2022; Silva et al. 2023. <https://huggingface.co/google-bert/bert-base-cased>

The difference in embedding structure between these two cases can be quantified by analyzing the distribution of activation values across all dimensions.¹¹

The opposite hypothesis is also a possibility (H_2). Under this view, concrete words such as dog or tree may activate a broader range of dimensions due to their rich sensory associations across multiple modalities. In contrast, abstract words like justice or hope may activate fewer, more specific dimensions, resulting in a sharp activation profile with, for example, higher norm but lower entropy — akin to a *spike* in certain representational axes. This could occur if embeddings for abstract concepts rely on a small number of high-level semantic features (e.g., valence, affect, discourse function), while embeddings for concrete words require a more distributed, multimodal representation that increases their variance and entropy.

To formally test this, we compute various vector shape metrics on the sentence embeddings. These include:

- **Norm** ($\|x\|_2$). The Euclidean norm measures the overall magnitude of the embedding vector. It provides a sense of how “large” the values in the vector are. This does not necessarily tell us about distribution across dimensions.

It is defined as:

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$

A sparse embedding (where most values are near zero and only a few are large) might have a lower norm, while more uniform activation leads to higher norm.

- **Entropy** ($H(e)$). Entropy is defined as:

$$H(\mathbf{e}) = - \sum_{i=1}^n p_i \log(p_i + \epsilon)$$

where \mathbf{e} is the embedding vector, and p_i are probabilities obtained by applying a normalization to the elements of \mathbf{e} . We primarily construct these probabilities by taking the absolute value of each dimension and normalizing to sum to 1, thus ensuring non-negative values interpretable as a pseudo-probability distribution over dimensions. This approach was chosen over the softmax transformation to avoid amplifying the largest embedding values exponentially, which can distort the distribution and reduce sensitivity to variations across smaller dimensions.¹² n is the number of components in the vector and ϵ is a small constant to avoid $\log(0)$. The entropy reflects how evenly the embedding’s values are distributed across dimensions.

11. This is also the hypothesis of Hessel et al. (2018, p. 2194) in quantifying visual concreteness, who write: “Intuitively, a visually concrete concept is one associated with more locally similar sets of images; for example, images associated with “dog” will likely contain dogs, whereas images associated with “beautiful” may contain flowers, sunsets, weddings, or an abundance of other possibilities.”

12. To verify the robustness of this approach, we also computed probabilities using a softmax transformation, which produced near-identical entropy values across our experiments. This confirms that our simpler absolute-value normalization provides a consistent and interpretable proxy for measuring the spread or concentration of embedding activations. While this method introduces nonlinearity, embeddings with primarily positive values (e.g., CLIP) may yield systematically different entropy scores. Nonetheless, this approach balances interpretability and computational simplicity.

- **Variance** (σ^2): It measures the spread of values in the embedding vector, indicating how much they deviate from the mean:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

where x_i is the value of the embedding at dimension i , n is the total number of dimensions, and \bar{x} is the mean of these values. Higher variance suggests a more dispersed representation, potentially reflecting greater contextual flexibility, while lower variance may indicate a more compact, feature-specific encoding. This is crucial in evaluating whether highly imageable embeddings are tightly clustered or broadly distributed in a semantic space.

- **Sparsity** ratio. Sparsity ratio can be defined as:

$$\text{sparsity_ratio} = \frac{\|\mathbf{e}\|_1}{\sqrt{n} \cdot \|\mathbf{e}\|_2}$$

where \mathbf{e} is the embedding vector, n is the number of components in the vector, $\|\mathbf{e}\|_1 = \sum_{i=1}^n |e_i|$ is the Manhattan norm (the sum of the absolute values of the elements of \mathbf{e}), and $\|\mathbf{e}\|_2 = \sqrt{\sum_{i=1}^n e_i^2}$ is the Euclidean norm (the square root of the sum of the squared elements of \mathbf{e}). The sparsity ratio gives us an idea of how densely populated the embedding is, with lower values indicating higher sparsity.

Note: For norm and entropy, which are less intuitive measures, we show the distribution of values over embedding dimensions for the embeddings with the highest/lowest entropy and norm of the Chicago corpus sentence samples (see [Appendix A](#), Figures 4-5).

To evaluate the validity of these embedding-based metrics as indicators of imageability, we correlate them with dictionary-based imageability, concreteness, and visuality scores. Specifically, we compare values derived from the lexical resources of the MRC Psycholinguistic Imageability Lexicon (Coltheart 1981), the Concreteness Lexicon (Brysbaert et al. 2014), and the Lancaster Sensorimotor Norms (Lynott et al. 2020) against our computed embedding shape properties (see Section 3). By computing Spearman correlations, we assess the degree to which embedding metrics reflect known psycholinguistic properties of the lexicon.

Since dictionary-based scores are primarily word-level ratings, we then extend our analysis to the sentence level by aggregating word-level values across sentences. Specifically, for each sentence, we compute the sum imageability, concreteness, and visuality of the sentence based on the words that are present in the dictionaries (see the end of Section 4.2 again for more details). This allows us to compare *sentence* embeddings with dictionary-based metrics computed over entire sentences.

It is important to note that the transition from word-level representations to sentence-level embeddings is non-trivial. The imageability of a sentence is not simply the sum of its individual words' scores; rather, it depends on syntactic structure, compositionality, and context-dependent meaning shifts which cannot be captured by sums of dictionary scores. For example, a sentence like “The sky darkened before the storm” contains words

with varying individual imageability scores, but their combined effect creates a vivid, 441
scene-setting description. Conversely, a sentence with highly imageable words may still 442
lack clear imagery if its structure is abstract or ambiguous. 443

At the same time, sentence embeddings are not simple averages of word representa- 444
tions: they incorporate contextual interactions, modifying the contribution of each 445
word depending on its grammatical role and semantic dependencies. This means that 446
embedding-based metrics may diverge significantly from dictionary-based scores when 447
applied at the sentence level. 448

6. Results 449

6.1 Experiment I: Correlations at the Word Level 450

Figure 1a presents the correlation matrix for dictionary-based and embedding-derived 451
metrics, computed over the MRC dictionary lemmas (i.e., the subset of words appearing 452
across all dictionaries used in this study). The figure illustrates the relationships both 453
within dictionary-based scores and within embedding-based metrics, as well as their 454
inter correlations. 455

Internal Correlations in Dictionary-Based Scores. We observe a strong mutual corre- 456
lation among dictionary-based metrics. Notably, **Cortese Imageability** (*imag R*) exhibits 457
a correlation with **MRC Imageability** (*imag*) comparable in magnitude to its correlation 458
with **Concreteness**. This suggests that, in practice, imageability and concreteness are 459
not sharply distinguished in these resources — at least at the word level. While both 460
measures are conceptually distinct, their empirical overlap aligns with prior research 461
indicating a strong connection between how vividly a word evokes imagery and how 462
concrete its referent is. 463

Interestingly, **Visuality** shows a weaker correlation with both **Imageability** and **Con-** 464
creteness, suggesting that the dictionary-based concept of imageability is more strongly 465
associated with tactile or sensorimotor properties than with purely visual modalities. 466
This reinforces the idea that imageability, as defined in psycholinguistic lexica, captures 467
a broader range of sensory experiences beyond just the visual salience. 468

Internal Correlations in Embedding-Based Metrics. Turning to the embedding-derived 469
metrics, we find an even stronger internal correlation structure. For instance, **norm** 470
and **entropy** exhibit an inverse relationship, indicating that embeddings with higher 471
activation magnitudes (higher **norm**) tend to have more localized values, while those 472
with lower **norm** tend to have more evenly spread, high-entropy distributions. Because 473
our entropy calculation involves absolute-value normalization, direct comparison of 474
entropy values across embedding types (e.g., CLIP vs. BERT) should be interpreted 475

cautiously, though relative trends within each model remain informative.¹³¹⁴

This is consistent with our hypothesis that abstract words may have sharp activation spikes in fewer dimensions, whereas concrete words may activate a broader set of features across the embedding space.

Correlations Between Dictionary and Embedding Metrics. Across all embedding-based metrics, we find a significant correlation with dictionary-derived scores, particularly with Imageability and Concreteness ($\rho = .55 - .61$). This suggests that embedding norms, entropy, and related properties encode information that aligns with human-annotated word imageability and concreteness ratings. As we transition to the sentence level in subsequent experiments, we investigate whether these correlations persist when compositional effects come into play.

norm	1	-0.97	1	-0.99	-0.54	-0.49	-0.52	-0.39	-0.59
entropy	-0.97	1	-0.97	0.99	0.55	0.5	0.53	0.4	0.61
variance	1	-0.97	1	-0.99	-0.54	-0.49	-0.52	-0.39	-0.59
sparsity	-0.99	0.99	-0.99	1	0.55	0.51	0.53	0.4	0.61
imag	-0.54	0.55	-0.54	0.55	1	1	0.88	0.6	0.83
imag R	-0.49	0.5	-0.49	0.51	1	1	0.85	0.6	0.82
imag C	-0.52	0.53	-0.52	0.53	0.88	0.85	1	0.65	0.85
visual	-0.39	0.4	-0.39	0.4	0.6	0.6	0.65	1	0.62
concrete	-0.59	0.61	-0.59	0.61	0.83	0.82	0.85	0.62	1
	norm	entropy	variance	sparsity	imag	imag R	imag C	visual	concrete

(a) Using the multimodal model.

norm	1	-0.29	1	-0.74	-0.08	-0.12	-0.09	-0.09	-0.05
entropy	-0.29	1	-0.29	0.72	-0.02	0.03	0.01	-0.04	0.02
variance	1	-0.29	1	-0.74	-0.08	-0.12	-0.09	-0.09	-0.05
sparsity	-0.74	0.72	-0.74	1	-0	0.05	0	-0.01	0.02
imag	-0.08	-0.02	-0.08	-0	1	1	0.88	0.6	0.83
imag R	-0.12	0.03	-0.12	0.05	1	1	0.85	0.6	0.82
imag C	-0.09	0.01	-0.09	0	0.88	0.85	1	0.65	0.85
visual	-0.09	-0.04	-0.09	-0.01	0.6	0.6	0.65	1	0.62
concrete	-0.05	0.02	-0.05	0.02	0.83	0.82	0.85	0.62	1
	norm	entropy	variance	sparsity	imag	imag R	imag C	visual	concrete

(b) Using BERT.

Figure 1: Comparison of embedding metrics and dictionary scores using the multimodal model and BERT. Numbers refer to the Spearman coefficient. Note that Imag C and Imag R refers to the two expansion dictionaries of imageability, see subsection 4.1, while Imag refers to the general MRC imageability dictionary.

Moreover, when comparing our results with a text-only model, we find that the previously observed correlations do not fully hold. While there are slight correlations between Imageability (MRC_Reilly) and embedding-derived metrics such as norm, entropy, and variance, these are notably weaker than those found using the multimodal CLIP model (Fig. 1b).

Additionally, the internal correlations between embedding-based metrics are less pronounced in BERT. Specifically, the relation between variance and entropy drops, as

13. See section 5 for details on entropy computation and normalization procedures. Note that: The relation between embeddings' norms and entropy is not necessary, but a by-product of CLIP's training objectives. The strong correlation between norm and entropy when using the multimodal CLIP model likely stems from the way the model processes text. Since CLIP relies on softmax at various stages to encode textual inputs, its embeddings inherently carry a probabilistic structure. When computing entropy, which also transforms embeddings into a probability distribution, this process can introduce an automatic dependency on the norm. Specifically, embeddings with higher norms tend to distribute probability mass differently, leading to a systematic correlation between norm and entropy. To avoid enforcing this relation, we ensured that we did not use softmax in the process of computing entropy, although the softmax approach was also tested. See section 5.

14. As with the relation between norm and entropy, we see the relation between variance and entropy strongly in the CLIP model, but not in the BERT model. Again, this may be related to the normalization process in generation CLIP embeddings, where a normalization will also fix the variance dependent to the entropy – both then relying on the scaling of the data.

well as the relation between norm and entropy (suggesting that the strong correlations between these in the CLIP model do relate to the model’s reliance on probability conversion, see footnotes 9 & 10).

Given the strong internal correlations of the embedding metrics in the CLIP model embeddings, we retain only norm and entropy for the next two experiments, while maintaining imageability, visuality, and concreteness as dictionary-based features.¹⁵

6.2 Experiment II: Sentence-level Analysis in Literary Texts

	data	imageability	visuality	concreteness
Norm	<i>Hemingway</i>	-0.61	-0.63	-0.62
	<i>Woolf</i>	-0.44	-0.46	-0.46
	<i>Chicago</i>	-0.31	-0.37	-0.36
Entropy	<i>Hemingway</i>	0.60	0.63	0.62
	<i>Woolf</i>	0.42	0.44	0.44
	<i>Chicago</i>	0.31	0.37	0.37

Table 5: Spearman correlations between dictionary scores of sentences and the norms and entropies of sentence embeddings across our literary data.

For correlations of metrics across sentences, we find very similar correlations as in Experiment I, presented in table 5. That is, across all of our 3 literary datasets, we find that Imageability has a negative relationship to embedding norm, and a positive relation to embedding entropy. The direction and strength of these correlations is the similar for Concreteness and Visuality. We find the strongest correlations within the Hemingway sentences (min. $\rho \pm 0.61$) and the weakest for the Chicago sentences (min. $\rho \pm 0.31$) (table 5). This suggests that embedding norm and entropy maintain their correlation direction with the dictionary-based features also when aggregating scores at the sentence level, but that the strength of the correlation might differ according to the type of literature. Among Chicago corpus sentences, where correlations between embedding- and dictionary-based metrics is the lowest, we might expect the diversity – across both genres, styles, and decade – to have an effect.

Still, within each group of literary data (Hemingway, Woolf, Chicago), we find more or less imageable sentences (according to the dictionary) have a relation to the level of norm and entropy of their embedding. To examine how norm and embedding distribute within the groups, and to illustrate this effect, we selected two example sentences as reference points for high and low imageability:

- Highly Imageable: The thin white surgical gloves he wore as he pumped the gas looked like pale skin.
- Non-Imageable: Wishful thinking as the saying goes.

These two sentences were among the top and bottom 10 in terms of dictionary-based imageability scores among all sentences in our data.

The contrast between these examples highlights the degree to which descriptive, sensory-

15. Alternative imageability dictionaries were excluded due to the smaller size of their lexica and due to seeing their high correlation with the MRC imageability dictionary, making them redundant for our analysis.

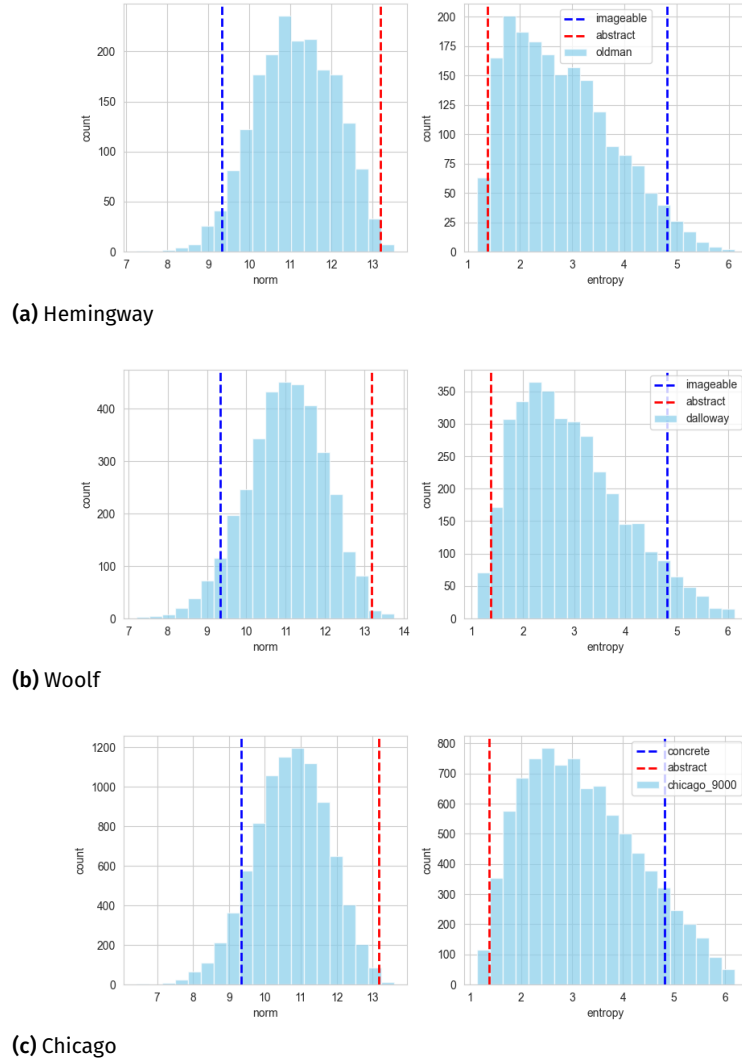


Figure 2: Histogram using the norms and entropies of example sentences among the most and least imageable sentences from the 9,000 sampled sentence of the Chicago Corpus as references.

rich language correlates with embedding structure, where highly imageable sentences appear to cluster in regions with lower entropy and higher norm.

As shown in Figure 2, our analysis indicates that all our 3 groups of literary data tend to be skewed toward lower entropy, meaning that we predominantly observe a tail distribution at the entropy levels of highly imageable sentences.

6.3 Experiment III: Comparing Poems

In our final experiment, we compare Imagist poems to an assorted set of Modern Love poems not constrained by any specific literary movement. Previous research by J. T. Kao and Jurafsky (2015) found that, compared to 19th-century poetry, Imagist poetry exhibits higher levels of object mentions, abstraction, imageability, and concreteness – particularly when measured using the MRC Imageability Dictionary Coltheart 1981 and the Brysbaert Concreteness Lexicon Brysbaert et al. 2014 – both of which are also used in this study. To ensure consistency with previous methodologies, we compute

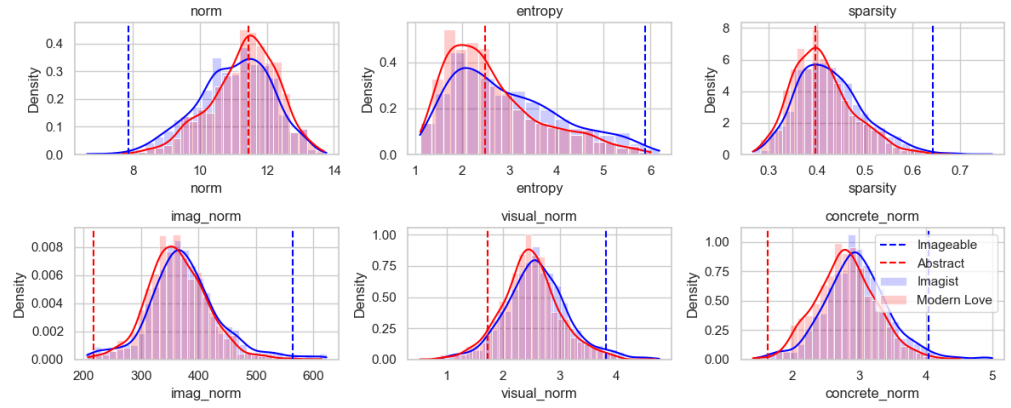


Figure 3: Imagist vs. love poems in terms of dictionary features and embedding-based metrics. We use the feature values of the example poem lines among the most and least imageable lines as reference points in each case.

imageability, visuality, and concreteness using the same approach as J. T. Kao and Jurafsky (2015).¹⁶

Figure 3 presents the juxtaposition of the two poetic traditions in terms of their embedding- and dictionary-based features. The lines indicate two poem lines which were found among the top 10 most imageable and bottom 10 least imageable lines of poetry in the full set (Imagist & Love poems). These were:

- Highly imageable: Homespun, dyed butternuts dark gold color.
- Non-imageable: Of insidious intent

Here we are showing where these two sentences are positioned in terms of each measure.

This comparison allows us to assess that embedding-based metrics do distinguish Imagist poetry, and that Imagist poetry does exhibit higher imageability and concreteness at the sentence level, aligning with prior findings on its heightened emphasis on sensory detail and concrete imagery. These findings are supported by conducting a t-test between the groups (Table 6).

	T-test	Mann-Whitney U
embedding norm	-6.69	567900.00
embedding entropy	<u>7.19</u>	<u>783363.00</u>
embedding sparsity	6.44	773000.00
imageability	5.07	734321.50
visuality	4.98	749496.00
concreteness	7.93	796327.00

Table 6: T-test and Mann-Whitney U results (for comparison) for dictionary-based features and embedding-based metrics between Imagist & Love poems groups. The largest statistic for each variable is in bold, and the second-largest is underlined. All tests were significant with $p < 0.01$.

16. Unlike previous studies, however, we conduct our analysis at the poem-line level rather than at the poem level. That is, we calculate the average imageability score across the words in each line that appear in the dictionary, rather than aggregating at the level of entire poems. This allows us to maintain higher granularity and a larger number of data points, providing a more detailed view of how Imagist poetic discourse — not just entire poems — manifests imageability.

7. Discussion

551

Our findings suggest that embedding structure meaningfully reflects psycholinguistic properties of words, particularly in relation to imageability, concreteness, and visuality. Across our experiments, we observe that embedding norm and entropy serve as reliable indicators of a text's sensory specificity, but in a manner contrary to our initial hypothesis (H_1). Instead, our results provide stronger evidence for H_2 , indicating that concrete and imageable words and sentences exhibit higher entropy and lower norm, while abstract words show lower entropy and higher norm. This suggests that highly concrete words are encoded in a more diffuse and broadly distributed manner, engaging multiple representational dimensions, whereas abstract words activate fewer dimensions more intensely, leading to sharper activation peaks (high norm) but a more compressed distribution (low entropy). We see this pattern when visualizing the embeddings with the highest and lowest norms, as well as the highest and lowest entropy, where low-entropy/high-norm embeddings appear to exhibit longer tails and more dimensions with zero values; while high-entropy/low-norm embeddings show values more evenly centered around zero (see [subsection A.1](#), Figures 4-5).

This pattern challenges the assumption that concrete words, due to their contextual constraints, would be more compactly represented in embedding space. Instead, it appears that concreteness leads to a more dispersed activation profile. This might reflect semantic affordances — that is, concrete words can be associated with a rich variety of semantic features, leading to a broader spread across dimensions. In contrast, abstract words tend to be semantically constrained to fewer, high-level conceptual dimensions, which results in embeddings with spiky, high-norm activations concentrated in a limited set of representational axes. This raises new questions about how different embedding models distribute meaning across dimensions — particularly whether multimodal training systematically encourages broader activation patterns for sensory-rich words compared to text-only embeddings (e.g., BERT).

At the sentence level, we observe a similar effect: literary texts exhibit a strong skew toward lower norm and higher entropy, with particularly imageable sentences spreading their representational load across more dimensions, while abstract sentences cluster in sharper, more concentrated regions of the embedding space.

Our analysis of Imagist vs. Modern Love poetry provides further confirmation that the shape of semantic embeddings encapsulates imageability-related psycholinguistic features. Consistent with prior research, we find that Imagist poetry exhibits higher overall imageability and concreteness, with embedding structures reflecting a more diffuse, multimodal distribution (low-norm/high-entropy). This finding reinforces the idea that imageability is not merely a product of genre conventions but is actively shaped by individual sentence composition. In contrast, Modern Love poetry — while still employing rich figurative language — tends to contain more conceptual abstraction and affective expression, which aligns with a more sharply clustered, locally spiky, embedding representation — reflected in the generally high-norm/low-entropy shape of their embeddings.

Taken together, these findings suggest that norm/entropy may act as a dictionary-free proxy for readers' experience of vividness. The Imagist–Love case study demonstrates

genre-level separability even under lexical control, indicating that the signal is not reducible to word-level concreteness counts. At the same time, benchmarking against legacy dictionaries offers an interpretable bridge to prior literature.

8. Conclusion

Our study suggests that the computational representation of sensory experience in embeddings follows distinct structural patterns for concrete vs. abstract language. Specifically, we find that highly concrete and imageable words exhibit greater entropy and lower norm, reflecting a more distributed, multimodal representation, whereas abstract words show lower entropy and higher norm, indicative of sharper, more localized activation patterns. These findings challenge our original assumptions about the compactness of concrete word representations and the way linguistic meaning is distributed across high-dimensional embedding spaces. Moreover, our results reinforce the role of multimodal models like CLIP in capturing sensorimotor properties, while text-only models like BERT appear to encode imageability less systematically.

Finally, if this approach is valid, it can constitute a method for dictionary-free inference on text imageability. Once the mapping from embedding shape (norm, entropy) to an imageability score is learned, no external lexicon is required at inference time. Any sentence—whether it contains out-of-vocabulary words, creative neologisms, or code-switched phrases—can be scored in a single forward pass through a pre-trained model.¹⁷

A key next step is to directly evaluate embedding-based metrics – alongside dictionary-derived features – against human judgments of sentence imageability. While our study establishes that embedding norms and entropies exhibit trends similar to dictionary-based imageability and concreteness, it remains unclear how well these computational features actually predict human-perceived sensory vividness.

This issue is particularly pressing because dictionaries, though widely used in psycholinguistics and NLP, are inherently limited – especially when extended to sentence-level interpretation, where contextual and compositional effects play an important role to their human interpretation. We have demonstrated the relationship between these metrics, but it remains an open question whether embeddings might actually outperform lexicon-based methods in capturing human imageability judgments – or whether they introduce biases or artifacts not present in traditional resources.

Further work should also explore a broader range of multimodal architectures, including models with more fine-grained visual-text alignment (e.g., DALL·E’s prior networks, BLIP, or fine-tuned vision-language transformers).

If multimodal embeddings systematically encode sensory experience, they could offer a scalable alternative to the hand-annotated psycholinguistic resources that are costly and relatively limited in scope. This is particularly relevant for literary studies, where large-scale human annotation of imageability, concreteness, or perceptual vividness

17. Scalability therefore stems from the billions of image–text pairs used to fit CLIP, not from the 5 k–40 k entries of psycholinguistic dictionaries. In this sense our approach is data-driven in deployment, and we use the legacy lexica only as a hold-out benchmark during evaluation. The distinction mirrors practice in automatic speech-recognition research, where acoustic models are trained on broadcast audio but validated against a much smaller, human-transcribed test set.

remains impractical outside of standard use of modern English and few other languages. 633

Further validation is needed before applying these embedding-based metrics to broader 634
literary studies – but this may perhaps also be said of applying imageability dictionaries 635
at the sentence level to broader literary studies. If proven reliable, these methods 636
could enable large-scale investigations into the evolution of prose styles, genre-specific 637
imageability trends, and historical shifts in literary sensory encoding. Additionally, it 638
would be valuable to compare literary texts to non-literary domains, such as journalistic 639
writing, political rhetoric, or scientific discourse, to better understand how imageability 640
and perceptual concreteness vary across communicative registers. 641

A. Data Availability 642

Data can be found here: [https://github.com/centre-for-humanities-computing/i 643](https://github.com/centre-for-humanities-computing/imageability_jcls)
[mageability_jcls 644](https://github.com/centre-for-humanities-computing/imageability_jcls)

B. Software Availability 645

Software can be found here: [https://github.com/centre-for-humanities-computi 646](https://github.com/centre-for-humanities-computing/imageability_jcls)
[ng/imageability_jcls 647](https://github.com/centre-for-humanities-computing/imageability_jcls)

C. Author Contributions 648

Yuri Bizzoni: Conceptualization, Methodology, Formal analysis, Validation, Resources, 649
Writing 650

Pascale Feldkamp: Conceptualization, Methodology, Formal analysis, Resources, Visu- 651
alization, Writing 652

Kristoffer L. Nielbo: Methodology, Formal analysis, Validation, Funding aquisition, 653
Writing 654

References 655

- Auracher, Jan and Hildegard Bosch (Dec. 2016). “Showing with words: The influence 656
of language concreteness on suspense”. In: *Scientific Study of Literature* 6.2, 208–242. 657
ISSN: 2210-4372, 2210-4380. [10.1075/ssol.6.2.03aur](https://doi.org/10.1075/ssol.6.2.03aur). [http://www.jbe-platform.c 658](http://www.jbe-platform.com/content/journals/10.1075/ssol.6.2.03aur)
[om/content/journals/10.1075/ssol.6.2.03aur](http://www.jbe-platform.com/content/journals/10.1075/ssol.6.2.03aur) (visited on 04/10/2024). 659
- Bizzoni, Yuri, Pascale Feldkamp Moreira, Ida Marie S. Lassen, Mads Rosendahl Thom- 660
sen, and Kristoffer Nielbo (May 2024). “A Matter of Perspective: Building a Multi- 661
Perspective Annotated Dataset for the Study of Literary Quality”. In: *Proceedings 662*
of the 2024 Joint International Conference on Computational Linguistics, Language Re- 663
sources and Evaluation (LREC-COLING 2024). Ed. by Nicoletta Calzolari, Min-Yen 664
Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, 665
Italia: ELRA and ICCL, 789–800. [https://aclanthology.org/2024.lrec-main.71 666](https://aclanthology.org/2024.lrec-main.71)
(visited on 09/28/2024). 667

- Blackwell, Simon E. (2020). "Emotional Mental Imagery". In: *The Cambridge Handbook of the Imagination*. Ed. by Anna Abraham. Cambridge Handbooks in Psychology. Cambridge: Cambridge University Press, 241–257. ISBN: 978-1-108-42924-5. [10.1017/9781108580298.016](https://www.cambridge.org/core/books/cambridge-handbook-of-the-imagination/emotional-mental-imagery/E619467FC94941DF59F5ED37434335DD). <https://www.cambridge.org/core/books/cambridge-handbook-of-the-imagination/emotional-mental-imagery/E619467FC94941DF59F5ED37434335DD> (visited on 02/04/2025).
- Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman (Sept. 2014). "Concreteness ratings for 40 thousand generally known English word lemmas". In: *Behavior Research Methods* 46.3, 904–911. ISSN: 1554-3528. [10.3758/s13428-013-0403-5](https://doi.org/10.3758/s13428-013-0403-5). <https://doi.org/10.3758/s13428-013-0403-5> (visited on 01/21/2025).
- Burroway, Janet (1987). *Writing Fiction: A Guide to Narrative Craft*. Little, Brown. ISBN: 978-0-316-11770-8.
- Chen, Yufeng, Guanghui Yue, Weide Liu, Chenlei Lv, Ruomei Wang, Fan Zhou, and Baoquan Zhao (2025). "Predicting Plain Text Imageability for Faithful Prompt-Conditional Image Generation". In: *PRICAI 2024: Trends in Artificial Intelligence*. Ed. by Rafik Hadfi, Patricia Anthony, Alok Sharma, Takayuki Ito, and Quan Bai. Singapore: Springer Nature, 89–95. ISBN: 978-981-9601-22-6. [10.1007/978-981-96-0122-6_9](https://doi.org/10.1007/978-981-96-0122-6_9).
- Coltheart, Max (1981). "The MRC psycholinguistic database". In: *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* 33A.4. Place: United Kingdom Publisher: Taylor & Francis, 497–505. ISSN: 1464-0740. [10.1080/14640748108400805](https://doi.org/10.1080/14640748108400805).
- Cortese, Michael J. and April Fugett (Aug. 2004). "Imageability ratings for 3,000 monosyllabic words". In: *Behavior Research Methods, Instruments, & Computers* 36.3, 384–387. ISSN: 1532-5970. [10.3758/BF03195585](https://doi.org/10.3758/BF03195585). <https://doi.org/10.3758/BF03195585> (visited on 01/21/2025).
- Daoshan, MA and Zhang Shuo (2014). "A Discourse Study of the Iceberg Principle in *A Farewell to Arms*". In: *Studies in Literature and Language* 8.1, 80–84.
- Dellantonio, Sara, Claudio Mulatti, Luigi Pastore, and Remo Job (July 2014). "Measuring inconsistencies can lead you forward: Imageability and the x-ception theory". In: *Frontiers in Psychology* 5, 708. ISSN: 1664-1078. [10.3389/fpsyg.2014.00708](https://doi.org/10.3389/fpsyg.2014.00708). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4097956/> (visited on 02/05/2025).
- Ellen J. Esrock (1994). *The reader's eye*. Johns Hopkins University Press. ISBN: 978-0-8018-4669-4. <http://archive.org/details/readerseyevisual00esro> (visited on 02/06/2025).
- Ellis, Nick (Jan. 1991). "Chapter 21 Word meaning and the links between the verbal system and modalities of perception and imagery or In verbal memory the eyes see vividly, but ears only faintly hear, fingers barely feel and the nose doesn't know". In: *Advances in Psychology*. Ed. by Robert H. Logie and Michel Denis. Vol. 80. Mental Images in Human Cognition. North-Holland, 313–329. [10.1016/S0166-4115\(08\)60521-X](https://doi.org/10.1016/S0166-4115(08)60521-X). <https://www.sciencedirect.com/science/article/pii/S016641150860521X> (visited on 01/21/2025).
- Feldkamp, Pascale, Ea Lindhardt Overgaard, Kristoffer Laigaard Nielbo, and Yuri Biz-zoni (2024). "Sentiment Below the Surface: Omissive and Evocative Strategies in Literature and Beyond". In: *Computaitonal Humanities Research 2024*. Forthcoming. CEUR Workshop Proceedings.
- Fludernik, Monika (1996). "Towards a 'Natural' Narratology". In: *JLSE* 25.2, 97–141. ISSN: 0341-7638, 1613-3838. [10.1515/jlse.1996.25.2.97](https://doi.org/10.1515/jlse.1996.25.2.97). (Visited on 07/07/2024).

- Gleason, Daniel W. (2007). "Seeing Imagism: A Poetics of Literary Visualization". Ph.D. Dissertation. Evanston, Ill: Northwestern University. 715
716
- (Sept. 2009). "The Visual Experience of Image Metaphor: Cognitive Insights into Imagist Figures". In: *Poetics Today* 30.3, 423–470. ISSN: 0333-5372, 1527-5507. 10.121 718
5/03335372-2009-002. <https://read.dukeupress.edu/poetics-today/article/30/3/423/20994/The-Visual-Experience-of-Image-Metaphor-Cognitive> (visited 719
on 02/05/2025). 720
721
- Goetz, Ernest T., Mark Sadoski, Michael L. Stowe, Thomas G. Fetsco, and Susan G. Kemp (Sept. 1993). "Imagery and emotional response in reading literary text: Quantitative and qualitative analyses". In: *Poetics* 22.1-2, 35–49. ISSN: 0304422X. 10.1016/0304-42 723
2X(93)90019-D. <https://linkinghub.elsevier.com/retrieve/pii/0304422X9390019D> (visited on 02/04/2025). 724
725
726
- Grisot, Giulia, Federico Pennino, and J. Berenike Herrmann (2022). "Predicting sentiments and space in Swiss literature using BERT and Prodigy". In: <https://pub.uni-bielefeld.de/record/2969114> (visited on 02/05/2025). 727
728
729
- Hessel, Jack, David Mimno, and Lillian Lee (June 2018). "Quantifying the Visual Concreteness of Words and Topics in Multimodal Datasets". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, 2194–2205. 10.18653/v1/N18-1199. <https://aclanthology.org/N18-1199/> (visited on 01/19/2025). 730
731
732
733
734
735
736
- Kao, Justine and Dan Jurafsky (June 2012). "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry". In: *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. Ed. by David Elson, Anna Kazantseva, Rada Mihalcea, and Stan Szpakowicz. Montréal, Canada: Association for Computational Linguistics, 8–17. <https://aclanthology.org/W12-2502/> (visited on 01/20/2025). 737
738
739
740
741
742
- Kao, Justine T. and Dan Jurafsky (Oct. 2015). "A computational analysis of poetic style: Imagism and its influence on modern professional and amateur poetry". In: *Linguistic Issues in Language Technology* 12.3. <https://aclanthology.org/2015.lilt-12.3/>. 743
744
745
- Kuzmičová, Anežka (2014). "Literary Narrative and Mental Imagery: A View from Embodied Cognition". In: *Style* 48.3. Publisher: Penn State University Press, 275–293. ISSN: 0039-4238. <https://www.jstor.org/stable/10.5325/style.48.3.275> (visited on 02/06/2025). 746
747
748
749
- Lacey, Simon and Rebecca Lawson (2013). "Introduction". In: *Multisensory Imagery*. Ed. by Simon Lacey and Rebecca Lawson. New York, NY: Springer, 1–8. ISBN: 978-1-4614-5879-1. 10.1007/978-1-4614-5879-1_1. https://doi.org/10.1007/978-1-4614-5879-1_1 (visited on 02/04/2025). 750
751
752
753
- Liu, Xingyu "Bruce", Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du (Apr. 2023). "Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. New York, NY, USA: Association for Computing Machinery, 1–20. ISBN: 978-1-4503-9421-5. 10.1145/3544548.3581566. <https://dl.acm.org/doi/10.1145/3544548.3581566> (visited on 01/21/2025). 754
755
756
757
758
759
- Lynott, Dermot, Louise Connell, Marc Brysbaert, James Brand, and James Carney (June 2020). "The Lancaster Sensorimotor Norms: multidimensional measures of percep- 760
761

- tual and action strength for 40,000 English words". In: *Behavior Research Methods* 52.3, 1271–1291. ISSN: 1554-3528. [10.3758/s13428-019-01316-z](https://doi.org/10.3758/s13428-019-01316-z). 762 763
- Magyari, Lilla, Anne Mangen, Anežka Kuzmičová, Arthur M. Jacobs, and Jana Lüdtké (2020). "Eye movements and mental imagery during reading of literary texts with different narrative styles". In: *Journal of Eye Movement Research* 13.3, 10.16910/jemr.13.3.3. ISSN: 1995-8692. [10.16910/jemr.13.3.3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7886417/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7886417/) (visited on 02/06/2025). 764 765 766 767 768
- Martínez, María-Angeles (Jan. 2024). "Imagining emotions in storyworlds: physiological narrated perception and emotional mental imagery". In: *Frontiers in Human Neuroscience* 18, 1336286. ISSN: 1662-5161. [10.3389/fnhum.2024.1336286. https://www.frontiersin.org/articles/10.3389/fnhum.2024.1336286/full](https://www.frontiersin.org/articles/10.3389/fnhum.2024.1336286/full) (visited on 02/04/2025). 769 770 771 772 773
- Paivio, Allan, John C. Yuille, and Stephen A. Madigan (1968). "Concreteness, imagery, and meaningfulness values for 925 nouns". In: *Journal of Experimental Psychology* 76.1. Place: US Publisher: American Psychological Association, 1–25. ISSN: 0022-1015. [10.1037/h0025327](https://doi.org/10.1037/h0025327). 774 775 776 777
- Paragini, Margherita and Mike Kestemont (2022). "The roots of doubt : fine-tuning a BERT model to explore a stylistic phenomenon". In: *Proceedings of the Computational Humanities Research Conference 2022 (CHR 2022), 12-14 December, 2022, Antwerp, Belgium*. CEUR workshop proceedings ; 3290. CEUR-WS.org, p. 72–91. <https://antwerp.ceur-ws.org/record/opacirua/c:irua:192413>. 778 779 780 781 782
- Pound, Ezra (1913). "A Few Don'ts by an Imagiste". In: *Poetry* 1.6, 200–206. ISSN: 0032-2032. <https://www.jstor.org/stable/20569730> (visited on 07/07/2024). 783 784
- Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (July 2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html> (visited on 01/21/2025). 785 786 787 788 789 790
- Reilly, Jamie and Jacob Kean (2007). "Formal Distinctiveness of High- and Low- Imageability Nouns: Analyses and Theoretical Implications". In: *Cognitive Science* 31.1. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1080/03640210709336988>, 157–168. ISSN: 1551-6709. [10.1080/03640210709336988. https://onlinelibrary.wiley.com/doi/abs/10.1080/03640210709336988](https://doi.org/10.1080/03640210709336988) (visited on 01/20/2025). 791 792 793 794 795
- Sharma, Pratyusha, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba (Jan. 2024). *A Vision Check-up for Language Models*. arXiv:2401.01862 [cs]. [10.48550/arXiv.2401.01862. https://arxiv.org/abs/2401.01862](https://arxiv.org/abs/2401.01862) (visited on 02/05/2025). 796 797 798 799
- Sharma Paudyal, Homa Nath (July 2023). "The Use of Imagery and Its Significance in Literary Studies". In: *The Outlook: Journal of English Studies* 14, 114–127. ISSN: 2773-8124, 2565-4748. [10.3126/ojes.v14i1.56664. https://www.nepjol.info/index.php/ojes/article/view/56664](https://www.nepjol.info/index.php/ojes/article/view/56664) (visited on 02/04/2025). 800 801 802 803
- Silva, Kanishka, Burcu Can, Frédéric Blain, Raheem Sarwar, Laura Ugolini, and Ruslan Mitkov (July 2023). "Authorship Attribution of Late 19th Century Novels using GAN-BERT". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. Ed. by Vishakh Padmakumar, Gisela Vallejo, and Yao Fu. Toronto, Canada: Association for Computational Linguistics, 310– 804 805 806 807 808

320. [10.18653/v1/2023.acl-srw.44](https://aclanthology.org/2023.acl-srw.44). <https://aclanthology.org/2023.acl-srw.44> (visited on 02/05/2025). 809
- Strychacz, Thomas (2002). ““The sort of thing you should not admit”: Ernest Hemingway’s Aesthetic of Emotional Restraint”. In: *Boys Don’t Cry? Rethinking Narratives of Masculinity and Emotion in the U.S.* Ed. by Milette Shamir and Jennifer Travis. Columbia University Press, 141–166. [10.7312/sham12034-009](https://doi.org/10.7312/sham12034-009). 810
- Verma, Gaurav, Ryan Rossi, Christopher Tensmeyer, Jiuxiang Gu, and Ani Nenkova (Dec. 2023). “Learning the Visualness of Text Using Large Vision-Language Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, 2394–2408. [10.18653/v1/2023.emnlp-main.147](https://aclanthology.org/2023.emnlp-main.147). <https://aclanthology.org/2023.emnlp-main.147> (visited on 01/19/2025). 811
- Wu, Si and David Smith (July 2023). “Composition and Deformance: Measuring Imageability with a Text-to-Image Model”. In: *Proceedings of the 5th Workshop on Narrative Understanding*. Ed. by Nader Akoury, Elizabeth Clark, Mohit Iyyer, Snigdha Chaturvedi, Faeze Brahman, and Khyathi Chandu. Toronto, Canada: Association for Computational Linguistics, 106–117. [10.18653/v1/2023.wnu-1.16](https://aclanthology.org/2023.wnu-1.16). <https://aclanthology.org/2023.wnu-1.16> (visited on 05/31/2024). 812
- Wu, Yaru, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo (Mar. 2024). “Perplexing Canon: A study on GPT-based perplexity of canonical and non-canonical literary works”. In: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*. Ed. by Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva, and Stan Szpakowicz. St. Julians, Malta: Association for Computational Linguistics, 172–184. <https://aclanthology.org/2024.latechclfl-1.16> (visited on 09/28/2024). 813
- Zhao, Baoquan, Songhua Xu, Shujin Lin, Ruomei Wang, and Xiaonan Luo (July 2019). “A New Visual Interface for Searching and Navigating Slide-Based Lecture Videos”. In: *IEEE Computer Society*, 928–933. ISBN: 978-1-5386-9552-4. [10.1109/ICME.2019.00164](https://www.computer.org/csdl/proceedings-article/icme/2019/955200a928/1cd0J1kUety). <https://www.computer.org/csdl/proceedings-article/icme/2019/955200a928/1cd0J1kUety> (visited on 01/21/2025). 814

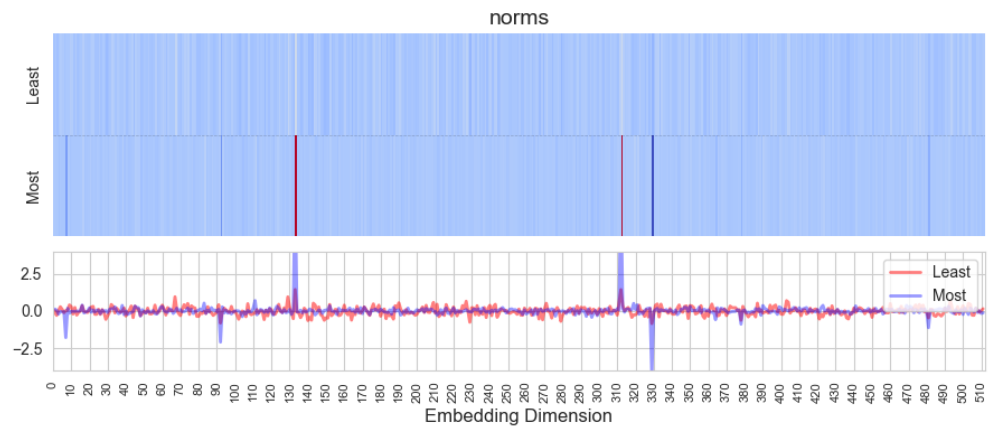
A. Appendix

839

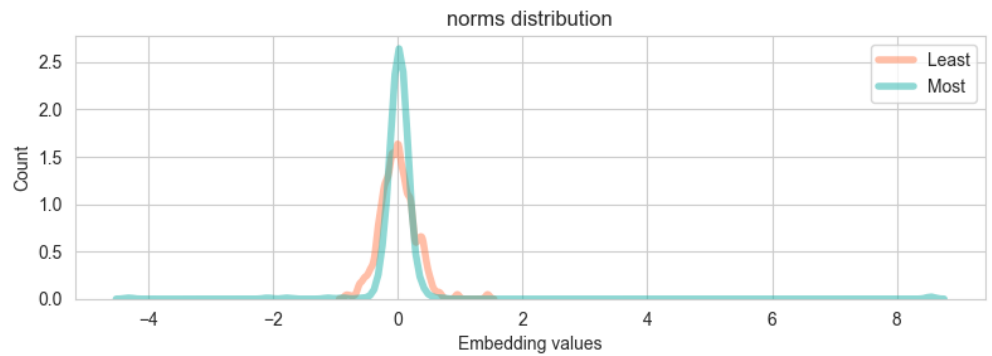
A.1 Norm & entropy of embeddings

840

We visualize the full embedding for extreme cases of norm and entropy values to give an idea of what these measures imply.

841
842

(a) Heatmap of the embedding values over dimensions, giving a sense of the strength and density of dimensions.



(b) Distribution (kde plot) of embedding values, giving a sense of the range.

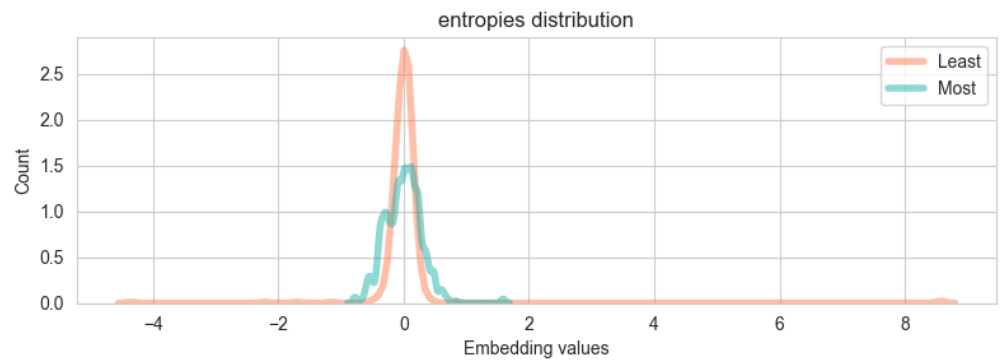
Figure 4: The two sentence embeddings with the **highest and lowest norm** of the 9,000 Chicago corpus sentences. Note how the embedding with *most norm* appears to contain stronger values (i.e., is bluer)(a) and more extreme values (in either direction)(b).

A.2 Experiment III Data

843



(a) Heatmap of the embedding values over dimensions, giving a sense of the strength and density of dimensions.



(b) Distribution (kde plot) of embedding values, giving a sense of the range.

Figure 5: The two sentence embeddings with the **highest and lowest entropy** of the 9,000 Chicago corpus sentences. Note how the embedding with *least entropy* appears to contain stronger values (i.e., is bluer)(a) and more extreme values (in either direction)(b), mirroring the shape of the embedding with most norm above.

Author	Poem
Michael Anania	Motet
Louise Bogan	To a Dead Lover
	Leave-Taking
	Juans Song
	Epitaph for a Romantic Woman
	Knowledge
	Song for the Last Act
	A Tale
Asil Bunting	from Odes: 30. The Orotava Road
Hart Crane	Voyages
	from The Bridge: Southern Cross
E. E. Cummings	as freedom is a breakfastfood
	i carry your heart with me(i carry it in)
	love is more thicker than forget
Paul Laurence Dunbar	The Old Front Gate
	A Negro Love Song
	Invitation to Love
	Night of Love
	Thou Art My Lute
	Song (Wintah, summah, snow er shine)
T. S. Eliot	Portrait of a Lady
	The Love Song of J. Alfred Prufrock
Kenneth Fearing	Aphrodite Metropolis (2)
	X Minus X
Ivor Gurney	Photographs
Stephen Spender	Song
James Joyce	Tutto Sciolto
D. H. Lawrence	Last Words to Miriam
	Gloire de Dijon
	Cruelty and Love
	Tortoise Gallantry
	The Bride
	Song (Love has crept...)
	Tortoise Shout
Edgar Lee Masters	Lydia Puckett
	Lucinda Matlock
	Mrs. Meyers
	Sarah Brown
Marjorie Pickthall	Adam and Eve
Carl Sandburg	Bilbea
	At a Window
	How Much?
Kenneth Slessor	New Magic
Gertrude Stein	The house was just twinkling in the moon light
	Idem the Same: A Valentine to Sherwood Anderson
Wallace Stevens	Hymn from a Watermelon Pavilion
	Peter Quince at the Clavier
Sara Teasdale	Union Square
	Spring in War-Time
	The Old Maid
	Since There Is No Escape
	The Look
	Over the Roofs
	Faults
	Eight O'Clock
	Old Love and New
	Debt
Louis Untermeyer	Infidelity
	Feuerzauber
Elinor Wylie	Wild Peaches
	Valentine
William Butler Yeats	When You Are Old
	Politics
	The Circus Animals Desertion
	He wishes his Beloved were Dead
	Never give all the Heart
	To an Isle in the Water
	Reconciliation
	The Cap and Bells
	Down By the Salley Gardens
	The Song of Wandering Aengus
	Adam's Curse
	No Second Troy
	A Drinking Song

Table 7: Poems and Authors of the 'Modern Love' poems

Author	Poem
Richard Aldington	Childhood
	The Poplar
	Round-Pond
	Daisy
	Epigrams
	The Faun sees Snow for the First Time
H. D.	Lemures
	The Pool
	The Garden
	Sea Lily
	Sea Iris
	Sea Rose
John Gould Fletcher	Oread
	Orion Dead
	The Blue Symphony
F. S. Flint	London Excursion
	Trees
D. H. Lawrence	Lunch
	Malady
	Accident
	Fragment
	Houses
	Eau-Forte
	Ballad of Another Ophelia
	Illicit
	Fireflies in the Corn
	A Woman and Her Dead Husband
Amy Lowell	The Mowers
	Scent of Irises
	Green
	Venus Transiens
	The Travelling Bear
	The Letter
	Grotesque
	Bullion
	Solitaire
	The Bombardment

Table 8: Poems and Authors of the ‘Imagist’ poems