jcls. Journal of Computational Literary Studies

Conference Reader

5th Annual Conference of
Computational Literary Studies

CCLS2026 Potsdam

Potsdam, May 28-29, 2026

Venue	University of Potsdam
Local Organizer	Digital Humanities Network, University of Potsdam
Web	https://jcls.io/site/ccls2026/
Contact	digitalhumanities@uni-potsdam.de
Hashtag	#CCLS2026
JCLS Editors	Evelyn Gius, Christof Schöch, Peer Trilcke
JCLS Editorial Assistants	Zita Baronnet, Svenja Guhr, Julian Häußler, Élodie Ripoll, Henny Sluyter-Gäthje
Contact JCLS	info@jcls.io | https://jcls.io/site/contact/

Conference Programme

Thursday | May 28, 2026

Venue: Wissenschaftsetage im Bildungsforum, Potsdam

12:00 p.m. to 12:45 p.m. | Arrival, Registration & Coffee

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

1:00 p.m. to 2:30 p.m. | Session 1

- Botond Szemes. Contrasting Verbal Prominence and Network Centrality. A Typology of Dramatic Characters
- Agnes Hilger, Anton Ehrmanntraut: Coreference Resolution for Full German Novels Using Large Language Models
- Esther Shizgal, Eitan Wagner, Omri Abend, Renana Keydar: Character Development in Oral Testimonies: Computational Modeling of Religiosity in Holocaust Narratives

2:30 p.m to 3:00 p.m. | Time for Talks & Coffee

3:00 p.m to 4:00 p.m. | Session 2

- Ze Yu, Federico Pianzola, Lanping Zhang: Echoes of Emotion: Linking Narrative and Reader Response of Web Novels in Chinese and English
- Marijn Koolen, Joris J. Van Zundert, Peter Boot, Silvia Lilli, Katja Tereshko: The Anatomy of the Online Book Review

4:00 p.m to 5:00 p.m. | Session 3

- Arthur Freitas Ramos: Grace as a Formal Turning Point: Computational Detection in Flannery O'Connor's Short Fiction
- Ella Montgomery, Alexandra Montgomery: Much Ado about Meaning: Shakespeare in Localization

5:00 p.m. to 6:00 p.m. | Poster Session: Opening & More Coffee

6:00 p.m. to 7:00 p.m. | Keynote: Ruth Ahnert, Close Reading in the Age of AI

7:00 p.m. to 10:00 p.m. | Poster Session: Finale. Snacks & Drinks at the Rooftop Terrace

Friday | May 29, 2026

Venue: University of Potsdam, Campus Am Neuen Palais, Building 8

9:00 a.m. to 9:30 a.m. | Good Morning Coffee

9:30 a.m. to 11:00 a.m. | Session 4

- Andrew Piper: 200 Years of Children in the Novel. On their Visibility, Value, and Agency
- Antonina Martynenko, Artjoms Šeļa, Petr Plechac: Where Empires End: Geography of the Poetic Formula 'from A to B'
- Maria Levchenko: Stylometry or Embeddings? Authorship Attribution for Russian and Italian Poetry

11:00 a.m. to 11:30 a.m. | Time for Talks & Coffee

11:30 a.m. to 12:30 p.m. | Session 5

- Emilio Maria Sanfilippo, Claudio Masolo, Alessandro Mosca, Gaia Tomazzoli: Modeling and Reasoning over Observations: An Ontology for Literary Criticism
- Federico Gabriel Cortés: From Literary Criticism to Literary Studies: Topic Modeling Argentine Academic Journals (1982–2024)

12:30 p.m. to 1:00 p.m. | Closing

Keynote: Ruth Ahnert, Close Reading in the Age of AI

Abstract

This talk examines the evolving place of close reading within computational literary studies. Early digital humanities debates framed ‘distant reading’ in opposition to close reading and sometimes even as a form of ‘not-reading’ (Moretti, 2000). In practice, however, scholars continued to rely on close reading as a methodological checkpoint: trends and anomalies in quantitative results prompted returns to specific texts. Yet the movement from aggregate pattern to textual interpretation typically required a manual shift into ‘reading mode’.

I suggest that recent developments in large language models enable a more systematic integration of close reading within computational workflows and encourage a reconceptualisation of the hermeneutic process. Drawing on research from Living with Machines and Text Machine: Computing Literary Innovation, I will present a method for identifying moments of textual ‘surprise’ at the sentence-level using mask-and-predict models trained on historically specific corpora. This method proceeds simultaneously at the level of distant and close reading: the scale of the output (the sentence) is designed for close reading; but we can triage those outputs (perhaps millions of sentences) through classification. As such, I will suggest how computational methods can scale the exploratory practices through which literary scholars identify the arguments, examples, and the interpretive stakes that structure critical research.

About the speaker

Ruth Ahnert is Professor of Digital Humanities and Literary History at Queen Mary University of London, specializing in early modern culture, and computational humanities. She has led several large collaborative projects, including Living with Machines (PI), Networking Archives (Co-I), and Text Machine: Computing Literary (PI). She is the author of "The Rise of Prison Literature in the Sixteenth Century" (2013), and four further co-authored books: "The Network Turn" (2020); "Tudor Networks of Power" (2023, winner of the Richard Deswarte Digital History Prize 2024); "Collaborative Historical Research in the Age of Big Data" (2023); and "Living with Machines: Computational Histories of the Age of Industry" (2026).

Poster Session

- Ingo Börner: Agentic Access to Linked Open Data: An MCP Server for Chat Interaction with the “DraCor Open Knowledge Graph”
- Akintoye Samson Japhet, Ismail Olaitan Afolabi: “Animal Farm” as a Textual Motif in Orwell’s “Animal Farm”
- Merten Kröncke, Agnes Hilger, Jana Eckardt: A Canonicity Score for Your (German-language) Corpus
- Elena Hamidy: Can LLMs Effectively Detect Poetry? Identifying Quality Differences Among Local, Free, and Commercial Models
- Pensalfini Martina, Sabatino Lorenzo: Copyright, Computation, and Literary Space: A Linked Open Data Pipeline for Contemporary Authors
- Gilad Gutman: A Computational Method for Analyzing Figurative Language: Classifying the Body Politic Metaphor in Early Modern English Tragedies
- Luise Prager: Corpus Construction and Preprocessing Evaluation of Contemporary German Prize-Winning Novels (2010-2025)
- Maxim Demin, Mark Schwindt, Mariia Menshikova: Digital Analysis of Studies in Soviet Thought / Studies in East European Thought (1961–2020): The Transformation of the Expert Perspective
- Yann Audin: Design questions all the way down: A General Methodological Model of Distant Reading
- Jeffrey Clapp, Lau Chaak Ming: Does Autofiction in English Exist? Generic and Narrative Measures
- Tina Ternes: Emotional Links between Text and Discussion in Shared Reading Sessions
- Simran Bhimjyani, Shanmugapriya T: Green CLS: Introducing Environmental Sustainability in Computational Literary Studies
- Pascale Feldkamp, Kristoffer L. Nielbo, Yuri Bizzoni: Language in Expansion: Diachronic Variation across Genres in Danish Newspapers
- Lisa Gollner: Machine Learning-Based Extraction and Clustering of Recurring Characters in Historical Periodicals
- Svenja Guhr, Irem Kurtdemir, Hayden L. Nurnberg, Adikkya Rahman, Shannon Thornton, David Bamman: Measuring Suspense in English-Language Short Fiction
- Matilde Innocenti: Modeling Institutional Literary Networks: A Relational Infrastructure for Nineteenth-Century Theatre Archives
- Khoa Van Tuan Le: Operationalizing Narrative Entrapment: Predictive Affective Trajectories via Contextual Sentence Embeddings in Kafka’s The Metamorphosis
- Laura Duparc, Camille Bertrand, Ami Nagai: Shadows: A Computational Knowledge Graph of Mythological and Literary Characters for Cross-Cultural Narrative Analysis
- Marc Barcelos: Style After Success: A Multidimensional Analysis of Context-Driven Expressive Drift in US-Published Creative Writing in the Long 20th Century
- Keli Du, Julia Röttgermann, Christof Schöch: Test for Uniformity of P-values: a reproduction of results

jcls. Journal of Computational Literary Studies

conference version

Contrasting Verbal Prominence and Network Centrality

A Typology of Dramatic Characters

Botond Szemes[1]

1. DigiTS Research Group, University of Tartu, Tartu, Estonia.

OPEN ACCESS

Citation
Botond Szemes (2026). "Contrasting Verbal Prominence and Network Centrality. A Typology of Dramatic Characters". In: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7996

Date published 2026-05-05 (preprint)
Date accepted tbc
Date received 2026-01-07

Keywords
computational drama analysis, character types, count-based and network-based metrics, keyword analysis, Shakespeare

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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. This paper proposes a typology of dramatic characters based on different dimensions of prominence, conceptualizing it as a multidimensional and potentially divergent phenomenon. The study offers new perspectives on comparative drama analysis and character description through a contrastive analysis of count- and network-based metrics. While the number of words and the number of speech acts are commonly used measures of verbal prominence, betweenness centrality is chosen for its selectivity—assigning high values to only a small subset of characters—and because it differs most from the other two metrics, thereby supporting the contrastive approach. Using these metrics, key positions within the play's communicative system such as main producers and main transmitters can be identified. The analysis distinguishes three main character groups: Dominant characters, who rank highly across all measures; Speakers, who are verbally prominent but structurally less central; and Connectors, who are structurally central but less verbally dominant. A case study of Shakespeare demonstrates the interpretative value of this typology. In a subsequent step, keyword analysis and the distribution of selected word categories are used to compare the linguistic profiles of Speakers and Connectors. The results show that their discursive patterns align with their structural roles, suggesting a systematic relationship between functional position and language in Shakespeare's plays.

1. Introduction

This paper proposes a typology of dramatic characters based on distinct dimensions of prominence. Prominence is understood as substantial contribution to the plot, the discourse, or the social world of a play, which distinguishes a small subset of characters from the rest of the cast. The paper aims to produce analytically meaningful groups that transcend traditional categories to motivate new ways of comparative drama analysis and character description.

Computational studies typically rely only on features of discourse, presence and interaction – i.e., how and how much characters speak, present on stage and with whom they interact – as observable proxies for characterization. These features are used to infer higher-level properties such as actantial function in the plot – e.g., protagonists (Fischer et al. 2018; Reiter et al. 2018; Masías et al. 2017), schemer (Beine 2024, Krautter and Pagel 2024), or "virtuous daughter" and "tender father" (Krautter et al. 2020) –,

conference version

or to external traits such as gender (Keith et al. 2025, Hicke and Mimno 2024), age, or social class (Krautter et al. 2023, Murphy et al. 2020). Character prominence is, in this context, described by centrality in a character network (e.g. Moretti 2011, Masías et al. 2017) or as verbal (e.g. Hicke and Mimno 2024) and, to lesser extent, scenic prominence (best elaborated in Krautter et al. 2018). The latter two aspects are operationalized using count-based,[1] while the former is captured by network-based metrics (Fischer et al. 2018). In this paper, I focus on count-based metrics of verbal prominence and network centrality measures in co-occurrence networks, as the latter also relates to scenic prominence,[2] which is typically analyzed only as an additional factor in character description.

Much research has demonstrated the strength of combining these aspects in the identification of character types based on action/plot function or external traits (e.g. Reiter et al. 2018). Building on this work but rather than developing a similarly multidimensional classification model, this paper takes a step back and contrasts the two main approaches to character prominence: verbal and network-based. Consequently, rather than treating protagonism as the result of combining multiple features within a single model, the paper examines how different dimensions of prominence – discursive and relational – interact and, crucially, where they diverge across characters. In this sense, a character can be prominent without being a protagonist (for discussion of protagonism see Krautter et al. 2018).

To ensure that the resulting groups remain interpretable and relatable to drama theory, as well as to sociology, I limit the analysis to three metrics: the number of words, the number of speech acts (utterances), and betweenness centrality (BC). All of these are commonly used in computational character studies and capture different positions in the communicative system of plays. The first two metrics reflect the extent to which characters contribute to and shape the observable discourse of the play. Among network centrality measures, BC captures the most distinct aspect of prominence (see Feature selection), thereby making a contrastive approach more fruitful. BC is often used to reveal information transmitters or brokers (Hudson and Stiller 2005), which links the metric to established roles in dramatic texts that exhibit high structural prominence but fall outside traditional categories of protagonists –such as messengers, helpers, or conspirators (Algee-Hewitt 2017), as well as schemers (Beine 2024) – and analogous roles in real-world social organizations (e.g., Goffman 1959's "go-between" or mediator, Simmel 1950's "mediating third", Granovetter 1973's weak ties, or brokerage as described by Burt 1995.) The selected metrics can thus highlight two substantial, but potentially divergent roles in the inner communication system of a play: the main producers and the main transmitters of information.

On this basis, I identify characters who rank highly in terms of these metrics, and distinguish three groups by combining binary distinctions in a structuralist manner—whether

1. Krautter et al. 2018 and Reiter et al. 2018 also take into account the topical content of character speech; however, they struggle to relate the uncovered topics to semantically coherent units and therefore give them less importance in their interpretation. Their analysis, on the other hand, also underscores the importance of number of words as a count-based metric of prominence.
2. Since these networks are constructed based on co-occurrences within the same scene, most of the measures correlate strongly with the key indicator of scenic prominence—the number of scenes in which a character appears. In particular, weighted degree closely aligns with the metric (Pearson correlation: 0.94 in the Shakespeare corpus). Betweenness centrality, which is used in this study, also shows a substantial correlation (0.79).

a character ranks highly or not with respect to each measure. Dominant characters speak the most and are also central to the social world of the drama; Connectors are central bridges in the social network but do not speak the most; Speakers have many lines but are less central in the network. In other words, the method isolates two theoretically motivated dimensions and examines their combinations.[3]

The case study presented here analyses plays of Shakespeare, although the method is applicable to other corpora as well. The evaluation of the resulting character groups is qualitative and interpretative, but not arbitrary: it is based on their relation to established interpretation of the texts.

This approach offers several advantages:

1. The method makes explicit the assumptions underlying computational approaches to character analysis by showing that different metrics operationalize distinct perspectives on prominence, which capture different aspects of character function (also highlighted but not systematically analyzed by Fischer et al. 2018 and Jannidis 2014). Consequently, the approach treats character prominence not as a unified property to be inferred from multiple features, but as a field of potentially diverging dimensions corresponding to distinct observable phenomena and focuses explicitly on how these dimensions converge on—or separate from—the same characters, in line with recent studies of characters in narrative fiction (Bourgois et al. 2026, Mian et al. 2026).

2. From an interpretative perspective, this leads to character groupings that cut across established categories such as protagonist or other roles. It highlights other aspects of plot function, thereby makes new connections visible and enables comparative analyses of plays and characters from a different angle. A Connector or Speaker, for instance, may simultaneously be male or female, young or old, and may fulfill any function within the plot (e.g., protagonist, lover, helper). Such a regrouping can support alternative approaches to literary analysis by highlighting similarities across texts as well as differences in the individual elaboration of shared patterns. The case study of Shakespeare's plays serves as an example of this comparative perspective—for example, in the detection of gender imbalance of some groups, or in the comparison of tragic and comic speakers (e.g., Hamlet and Lear versus Nick Bottom and Falstaff), which may point to the comic dimensions of tragic heroes.

3. Finally, the relative simplicity of the method also contributes to its interpretability. In contrast to machine-learning classification procedures, the characteristics of each group can be more readily traced back to individual aspects and metrics. The restricted number of features used in the study also contributes to this transparency.

While some studies also incorporate semantic aspects of character speech—such as topic or sentiment—to establish character groups (Krautter et al. 2018, Reiter et al. 2018),

3. This procedure is similar to the contrast of eigenvector and betweenness centrality in the foundational paper by Algee-Hewitt 2017, although I focus on characters and not on the level of play: "Together, the two metrics, EC and BC, characterize two types of plays: plays with a central character who both focalizes the interactions and bridges the individual factions, and plays where the bridging function is displaced from those characters who participate in the most interactions with characters." (759.)

conference version

these features are less directly related to character prominence theoretically and are therefore excluded from the clusterization. However, once characters have been grouped, the question arises whether they also exhibit distinct linguistic and semantic patterns. Accordingly, in a second step of the analysis, I examine whether these functionally distinct groups differ in what they talk about and how they express themselves. In this context, keyword analysis using the Zeta method serves to characterize the language of the groups, an approach closely related to other corpus-linguistic studies of Shakespeare (Archer et al. 2005, Culpeper 2009, Archer and Findlay 2020, Murphy et al. 2020).

This step is primarily exploratory; however, some of the hypotheses derived from keyword analysis are also tested in the paper. Specifically, the analysis compares the frequency of selected word categories—beyond simple keywords—in the speeches of Speakers and Connectors. In this way, methodologically distinct approach to Zeta based on independently defined semantic categories is also introduced for comparing character groups, including the application of statistical tests.

The results show that clearly definable and interpretable character groups emerge that also differ in their language use. Speakers not only speak more but also tend to refer more frequently to communication itself and to highlight their positions within the communicative situation. Connectors, by contrast, tend to describe elements of the external world, using a more referential mode of language.

2. Related Work

In describing characteristics of roles in a plot, gender, age or social status, most research uses three types of metrics (network-, count-based and semantic), to varying degrees. The position of characters within the social world of a play is most often captured using metrics derived from co-occurrence or interaction networks. Jannidis 2014, for example, relies on the weighted degree of characters in such networks to identify protagonists, although in the context of novels: "This metric is most intuitively interpretable with regard to the importance of characters in a fictional world." Others, especially in the field of computational drama analysis, emphasize also the importance of centrality measures such as betweenness, closeness, or eigenvector centrality (e.g. Masías et al. 2017), especially in the case of character prominence as suggested by pioneering study of Moretti 2011: "the 'protagonist,' far from being a fundamental reality of dramatic construction, is only a special instance of the more general category of 'centrality'" (9). Eigenvector centrality was used to describe "protagonism" (Algee-Hewitt 2017), betweenness centrality is often used to identify "keystone characters" who are regulators of the information flow (Hudson and Stiller 2005, Beine 2024).

Network analysis can be supplemented by count-based metrics of utterances: how many words a character says and how many times they speak in a play. Hicke and Mimno 2024 define influence, more precisely, as "the product of the number of words in a given section of text and how many characters that section of text is addressed to" (92) which, they argue, allows „for comparisons between different scenes in a play, a dimension which is invisible in typical networks." (93) However, count- and network based aspects are often used together, as in the case of Fischer et al. 2018 and Reiter et al. 2018, who argue that a protagonist cannot be characterized by only one of them.

conference version

conference version

Other studies place greater emphasis on the semantic aspects of language, an approach that is particularly common in analyses of gender representations. Keith et al. 2025 trained a BERT model to examine male and female speech patterns in Calderón's comedies. Their findings suggest that women and men interact more frequently within their own groups; however, women tend to speak more about domestic matters, while men more often address military topics. Šeļa et al. 2024 arrive at similar conclusions (in a research of character speech distinctiveness, similar to Vishnubhotla et al. 2019) by examining utterances across different historical corpora using stylometric techniques: women's language tends to be more domestic and interpersonal, whereas men's speech is more strongly oriented toward politics. Murphy et al. 2020 analyze the keywords associated with Shakespeare's characters and conclude that women, beyond being socially constructed through family relations, tend to express grief and sorrow more often, while men are more likely to act as architects of events.

Burrows and Craig 2012's analysis is particularly instructive for the present study, as they contrast character speeches in early modern plays without pre-grouping or labeling them (i.e., they do not perform supervised classification). Although their primary aim is not to identify underlying groups, but rather to demonstrate the presence of distinct character voices *within* an author's style, their work nonetheless reveals similarities and differences between characters across different plays. The interpretative value of their results lies in the fact that they transcend traditional categories such as established character function, gender, or genre, thereby offering a fresh perspective on the plays:

> "It seems that some Shakespeare characters are more concerned with individuals and more questioning than Fletcher's ever are. Beatrice [from M*uch Ado About Nothing*] would seem to have these characteristics in the extreme, reflecting her disposition as interrogative, responsive, not given to thinking collectively, and not given to the hierarchical and intimate relationships where *thou*, *thee* and *thy* are used. (...) Desdemona [*Othello*] and Mistress Ford [*Merry Wives of Windsor*] are in this group, but there are males as well: the quirky Fluellen [*Henry V*], the manipulative Pandarus [*Troilus and Cressida*], and Beatrice's sparring partner Benedick. There are also two clowns (Touchstone [*As You Like It*] and Feste [*Twelfth Night*]), a servant (Dromio of Syracuse [*The Comedy of Errors*]), and a protagonist of farce (Master Page from *Merry Wives of Windsor*). These characters generally react to situations, rather than setting out a considered point of view. For them the dramatic moment is dominant, whether in witty, unrehearsed exchange, or, as in the case of Desdemona, in response to an unfolding tragedy." (298)

The notable work of Nils Reiter, Benjamin Krautter, Janis Pagel and Marcus Willand should be also mentioned, who demonstrated the combined application of all the three aspects in a series of papers (Reiter et al. 2018, Krautter et al. 2018, Krautter et al. 2020, Krautter et al. 2023, Krautter and Pagel 2024). In addition to metrics of co-occurrence networks and number of words / speech acts, their studies also took into account the sentiment of the characters' utterances, their topic (based on topic models and predefined word fields), their style (average utterance length and TTR), the verbs they used, and their stage presence (active and passive – this latter is when they are only mentioned in a scene). Somewhat surprisingly, in the case the detection of schemers,

they conclude that "the calculated network metrics (...) have only limited importance for the classification decision." (Krautter and Pagel 2024, 137.) In other contexts, network metrics have proven to be important factors, and given their contribution to interpretations of character's structural roles, they remain central in computational studies in this domain.

Finally, recent scholarship in character studies of narrative fiction has argued, similarly to this paper, that "while such aggregation [of different perspectives] provides a compact summary of character attributes, it also comes with important limitations. By collapsing heterogeneous information into a single unified representation, much of the nuance of individual attributes is lost, and the resulting vectors can prove difficult to interpret" (Bourgois et al. 2026). In response, the authors analyze the distribution of different semantic fields in texts independently and examines how they relate to one another and to character prominence. Mian et al. 2026 develop a similar approach, but additionally contrast metrics derived from different types of networks (e.g., co-occurrence and interaction networks, as well as directed networks based on references by other characters) with component-based frequency measures.

conference version

3. Method

In the present study, I identify three types of prominent characters by contrasting verbal prominence and network centrality: (1) *Dominant characters*, who are central in the social world of the play and contribute substantially to the discourse; (2) *Connectors*, who are central but speak relatively little; (3) *Speakers*, who speak frequently and at length but are not among the most central characters. This is, of course, not an exhaustive typology of character prominence, but rather a contrastive classification, analogous to structuralist descriptions of characters based on binary oppositions (cf. Ubersfeld 1999, 78; Pfister 1988, 163–164).

3.1 Feature selection

Verbal prominence is operationalized using two count-based metrics: the number of words and the number of speech acts. These highlight the main producers in the communication system of the drama. Taken together, these metrics balance between two extremes: a character who delivers long monologues but interacts only rarely with others (high number of words, low number of speech acts), and a character who interacts frequently but only briefly (low number of words, high number of speech acts). Both extremes can reflect a certain kind of prominence.[4] Combining these measures makes it possible to identify characters who occupy a substantial portion of the discourse through both extended and frequent utterances.

4. In this paper, I examine the combined effect of the two metrics in detail. However, additional groupings can also be defined based on the divergence between them: Orators, who rank highly in terms of the number of words but not in speech acts, and Interactors, who show the opposite pattern. In the Shakespeare corpus, Orators (as identified by the clustering described below) include Helen (*All's Well That Ends Well*), Falstaff (*Henry IV, Part 1*), York (*Henry VI, Part 2*), Warwick (*Henry VI, Part 3*), Faulconbridge (*King John*), Adriana (*The Comedy of Errors*), Troilus, and Viola (Twelfth Night). Interactors, by contrast, include Parolles (*All's Well That Ends Well*), Cleopatra, Prince Hal (*Henry IV, Part 1*), Plantagenet York (*Henry VI, Part 2*), Henry VI (*Henry VI, Part 3*), King John, Don Pedro (*Much Ado About Nothing*), Othello, Pandarus (*Troilus and Cressida*), Cressida, and Valentine (*Two Gentlemen of Verona*).

Network centrality is measured using betweenness centrality (BC) in co-occurrence networks of the plays. Other measures, such as degree, closeness, or eigenvector centrality, could also be considered in this context. BC was chosen for its distinctiveness: it captures a structurally distinct dimension of centrality compared to other measures, which tend to capture forms of central presence, reach, or influence that are often more closely aligned with interaction volume or visibility. Rather than participation or embeddedness, BC emphasizes mediation: the extent to which a character connects otherwise separate subgroups or occupies a brokerage position within potential pathways of interaction (Hudson and Stiller 2005). This links the metric to plot development and established character roles in drama theory, such as messengers, helpers, or conspirators (Algee-Hewitt 2017), as well as schemers (for this reason, BC receives particular attention in Beine 2024's identification of schemers). These characters are most often not categorized as protagonists, but they nonetheless possess structural prominence, highlighting how the present approach diverges from traditional discussions of protagonism. These distinct roles are also discussed in sociological approaches of mediation and brokerage, including weak ties by Granovetter 1973, mediating third by Simmel 1950, go-betweens by Goffman 1959, and structural holes by Burt 1995.[5]

The distinctiveness of BC is also reflected in its correlation with other measures. Chen et al. 2019 argues, for example, that for major characters, BC is more weakly correlated with other centrality measures, elaborating a different structural aspect of a character. A similar test was performed on the Shakespeare corpus in relation to verbal prominence. Figure 1 shows the correlation between the number of words and speech acts and various centrality measures (Pearson correlation), calculated for characters with the five or seven highest word counts in each play. Correlations were computed separately for the plays in order to avoid biases arising from differences in cast size, network density, and centrality distributions. The resulting per-play correlation coefficients were then aggregated across the corpus by taking their mean. While the absolute values differ only slightly, the relative ordering of the centrality measures remains stable across thresholds (with larger differences emerging when the number of "major" characters is reduced). The stronger a measure correlates, the more it reproduces the same signal. BC consistently shows the weakest correlation with verbal activity, making it particularly well suited for contrasting verbal prominence and network centrality.

Furthermore, the distributions of centrality measures among characters differ substantially (Figure 2). BC is heavily skewed, with most characters exhibiting very low values and only a few attaining high scores, thereby occupying strongly mediating positions. This makes BC the most selective metric in terms of prominence. Eigenvector centrality, by contrast, shows a concentrated distribution at higher values, suggesting a relatively homogeneous level of embeddedness in the network and that relatively higher scores are attained by a larger number of characters. Closeness centrality, similarly, shows generally high values across characters. These differences indicate that the measures vary in their ability to differentiate between character roles. In particular, BC is well

5. Simmel's "mediating third" highlights the structural position of a third party who can arbitrate or benefit from connecting two others; Goffman's "go-between" emphasizes the interactional and strategic role of an intermediary managing impressions and information across social situations; Granovetter's "weak ties" focuses on bridges linking different social clusters, showing how actors with ties across groups enable novel information flow; and Burt's "structural holes" argues that actors spanning gaps between disconnected groups gain informational and control advantages (while using local and ego-network based measures instead of the global metric of BC).

conference version

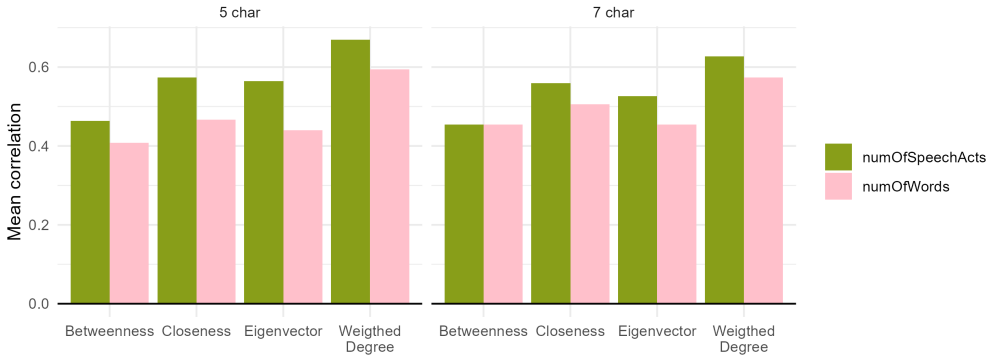


Figure 1: The correlation of centrality scores with number of words and speech acts.

suited for identifying structurally distinctive positions.

On the other hand, other centrality measures could be used in a similar analysis, potentially extending the typology of character prominence. Algee-Hewitt 2017, for example, contrasts eigenvector centrality with betweenness centrality to distinguish protagonists from mediators in interaction networks. In such networks, characters have high eigenvector centrality when they speak to those who are themselves speak with many others, indicating structural influence. At the same time, extending the typology on the basis of binary oppositions would pose a challenge in conceptualizing the resulting groups. Categories such as "influential characters" could, for example, be defined by high eigenvector centrality combined with relatively low BC and verbal prominence; however other combinations would be harder to interpret.

3.2 Clusters

One of the most challenging tasks is to determine the threshold for including a character among those with the highest values in each category. Taken every character from every play together, raw numerical values depend heavily on the size and structure of a network, as well as on the length of a play. By normalizing count-based values with respect to the length (total number of words),[6] and by applying additional z-score normalization for every metric within each play, more comparable results can be obtained (Figure 3). Characters can then be clustered based on these values using different approaches (e.g., k-means clustering, quartiles, or deciles; see Table 1). However, these results remain influenced by differences in word distribution and network structure across plays, especially when not taking into account genre differences (e.g. comedies tend to be more dense, producing smaller BC scores – see Szemes and Vida 2024). The present analysis therefore focuses on character prominence within individual plays, that is, on relations relative to other characters in the same drama, independently of cross-play comparisons.

For this reason, it is more appropriate to rely on ranks, which capture a character's relative position in a play. At the same time, relying solely on rankings can also be misleading. For example, in *Othello*, Othello has 274 speech acts and Iago 273, while the

6. Note that the version of betweenness centrality used here is already normalized by the total number of possible shortest paths, which implicitly accounts for the size of the network.

conference version

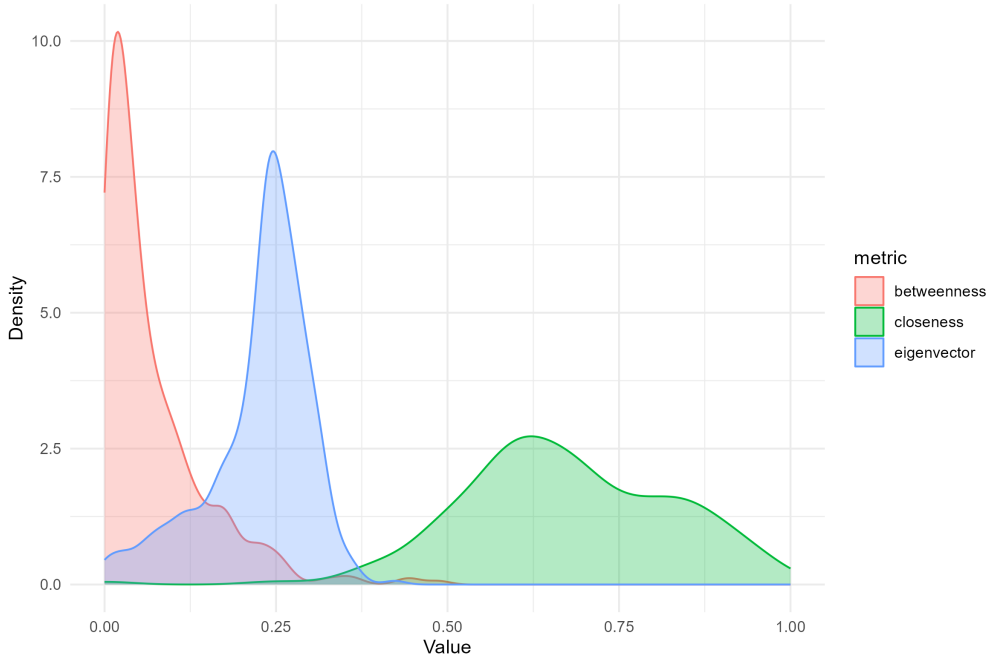


Figure 2: Density plot of centrality measures among Shakespearean characters speaking more than 500 words.

conference version

third-highest value—Desdemona's—drops to 165. Ranking Iago second and Desdemona third fails to capture the fact that the meaningful difference lies between her and the two male characters, who, on the other hand, have almost identical numbers of speech acts.

One solution is to calculate quartiles of the metrics, as in Algee-Hewitt 2017's analysis of English plays. But dividing the range of the data into four categories may not be sensitive enough to differences between characters. In *A Midsummer Night's Dream*, for example, this approach places five characters in the cluster with the highest number of speech acts— Bottom (59), Lysander (50), Theseus (49), Demetrius (48), and Hermia (48)—without accounting for the more than 15% difference between Bottom and the others. This method, thus, fails to produce clusters with skewed sizes that correspond to our general understanding of character prominence, i.e. differentiate just a small number of characters from the entire cast (cf. Woloch 2003, also highlighted by Masías et al. 2017). This limitation motivated Fischer et al. 2018 to use deciles instead of quartiles, i.e. dividing the range of values into ten equal parts. In this case, Bottom would appear alone in the highest cluster, followed by the other four characters in the second cluster.

At the same time, for casts with a median size of 36 characters (ranging from 17 to 74), a decile-based partition often produces too many groups to be analytically useful, particularly when clusters are compared across plays. A four-level scheme therefore appears more practical when combined with a deliberately skewed separation.[7] This combination offers a workable compromise: it preserves distinctions at the top of the distribution while keeping the overall typology readable. For this reason, I group

7. However, if we focus only on the top decile, similar—though not identical—results can be obtained using the proposed approach. In this case, the main difference is theoretical: the method explicitly incorporates the skewed distribution. At the same time, practical differences emerge when grouping characters with lower cluster ranks, this becomes evident in the example of characters with "second-order importance" (see below).

conference version

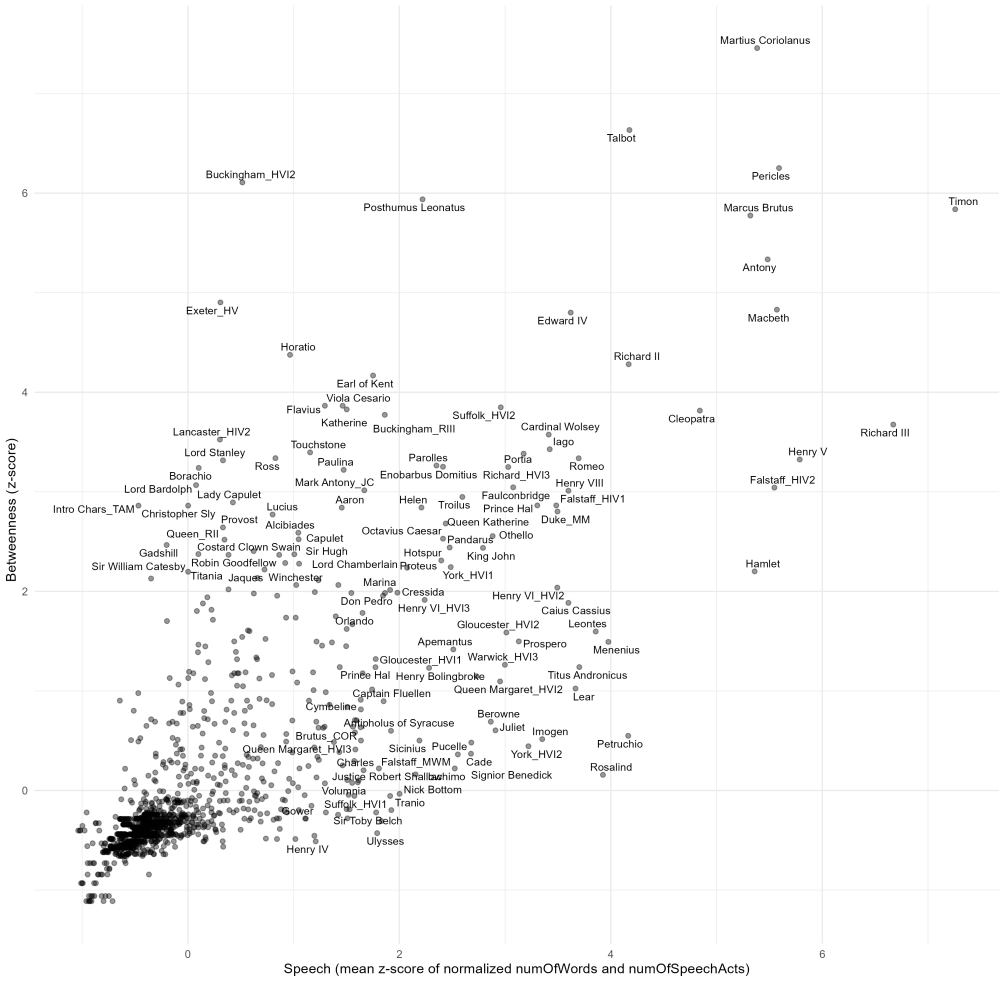


Figure 3: Position of Shakespearean characters based on z-scaled values of normalized verbal prominence and BC. Only labels are shown for characters in the top 5% of at least one metric.

characters into four categories such that the breaks between values follow a power-law–like distribution, resulting in skewed cluster sizes. Only one or two characters typically emerge as highly dominant or central (cluster 1), while the majority occupy structurally marginal positions, with further differentiation among clusters 2 to 4. Rather than treating differences between all characters as equally meaningful, this approach assumes that distinctions among highly prominent characters are more consequential than those among minor ones. Table 1 illustrates this effect and highlights the differences compared to other methods with a further example.

Character	Metric	Raw	Power	Q	D	zQ_genre	zD_genre	zK_genre	zQ_all	zD_all	zK_all
Romeo	Words	4729	1	1	1	2	5	1	2	5	2
	SpeechAct	163	1	1	1	2	4	1	2	4	1
	Betweenness	0.128	1	1	1	3	6	2	2	5	2
Juliet	Words	4350	2	1	1	3	6	2	3	6	2
	SpeechAct	118	3	2	3	3	6	2	3	6	2
	Betweenness	0.037	4	3	8	4	9	3	4	8	3
Capulet	Words	2195	3	3	6	4	8	2	3	8	3
	SpeechAct	50	4	3	7	4	9	3	4	8	3
	Betweenness	0.101	2	1	3	3	7	2	3	6	2
Nurse	Words	2264	3	3	6	3	8	2	3	8	3
	SpeechAct	90	3	2	5	3	7	2	3	7	2
	Betweenness	0.038	4	3	7	4	9	3	4	8	3
Mercutio	Words	2152	3	3	6	4	8	2	3	8	3
	SpeechAct	62	4	3	7	4	8	2	3	8	2
	Betweenness	0.002	4	4	10	4	10	4	4	10	4
Friar Lawrence	Words	2772	3	2	5	3	7	2	3	7	3
	SpeechAct	55	4	3	7	4	8	3	3	8	3
	Betweenness	0.066	3	2	5	3	8	3	3	7	3
Benvolio	Words	1179	4	4	8	4	9	3	4	9	3
	SpeechAct	63	4	3	7	4	8	2	3	8	2
	Betweenness	0.046	4	3	7	4	9	3	4	8	3
Lady Capulet	Words	895	4	4	9	4	9	3	4	9	4
	SpeechAct	45	4	3	8	4	9	3	4	8	3
	Betweenness	0.113	2	1	2	3	6	2	3	6	2

Table 1: Comparison of different methods of categorizing values, using the example of *Romeo and Juliet*. Clustering based on relative z-scores is performed either across all characters in the corpus (*_all*) or within genre-specific subsets (comedy vs. other, *_genre*), whereas the remaining methods are computed at the level of individual plays. Power = power-like distribution, Q = quartiles, D = deciles, K = k-means for 4 clusters, z = z-score normalization. Characters are filtered to include only those in clusters 1–3 in Q (play-level quartile) classifications for any metric.

In this study, after ranking characters separately for each metric, three character groups are derived by considering only those characters that attain the highest ranks in at least one of the dimensions. Characters who appear in cluster 1 across all three metrics form the Dominant group. Speakers are characters who appear in cluster 1 for the number of words and speech acts only, while Connectors appear in cluster 1 exclusively with respect to betweenness centrality. Other groupings are also possible, for example finding characters with "second-order prominence" who consistently grouped in cluster 2 or 3 in all metrics (like Demetrius, Lysander form *Midsummer* or Hotspur from *Henry IV Part 1*).

3.3 Keyword analysis

Once the groups are formed, their differences can also be investigated through patterns of language use. In this paper, I rely on keyword analysis that highlights words that are more distinctive to one group than to another. Specifically, I use the *stylo* package's *oppose* function, which is based on Hugh Craig's extension of Burrows's Zeta

conference version

conference version

method (Craig and Kinney 2009), a procedure that has also been developed in the context of Shakespeare scholarship. Recent research has shown that Zeta is among the best-performing approaches in cases where not primarily word frequency, but word dispersion is of central importance (Havrylash and Christof Schöch 2025). The method identifies keywords that are used by most characters within a group, rather than words that occur with very high frequency for only a few characters. In this way, subtle yet consistent differences in language use can be revealed.

In the analysis, I focus on the contrast between Speakers and Connectors, as these groups form opposing, complementary categories with clearly different characteristics. Dominant characters, by contrast, tend to share features with both groups, as they score highly in terms of speech distribution as well as betweenness centrality. Moreover, since keywords are often related to textual topics that are strongly influenced by genre (Christoph Schöch 2017), Speakers and Connectors are analyzed separately for comedies and for the remaining plays (tragedies and histories) which are treated as a joint category. As a preprocessing step, I pool the texts of each character group and genre and then divide the resulting corpus into 700-word segments. This segment size is substantially smaller than the default setting in *stylo* (3,000 words), but it is justified by the relatively small size of the corpora: Connectors in comedies comprise 19,829 words in total and 17,740 words in the other genres, while Speakers comprise 34,623 words in comedies and 45,186 words in the remaining genres (see Table 2 for individual word counts). Using 700-word segments allows multiple samples to be drawn from individual characters, which helps to mitigate the imbalances between the number of segments representing Speakers and Connectors (with 3,000-word segments, this would result in approximately 10–15 segments versus about 5). It also limits the disproportionate influence of individual characters on the distinctiveness scores, which would otherwise allow words unique to a single character to appear artificially "group-distinctive." With a larger number of segments, keywords must recur across many samples to be identified as typical of a group—although characters with exceptionally large amounts of text may still exert a stronger influence on the measurement. To assess the robustness of the results and reduce the influence of verbally extreme characters, the analysis was repeated also on versions of the texts in which the words were randomly reshuffled.

While keyword analysis is primarily exploratory, as the interpretation of distinctive words depends on the researcher, the hypotheses formulated in this section are tested using a methodologically distinct approach based on semantic categories. These categories capture the observed specificities of language use more comprehensively than individual keywords. For that, first, I determined seed terms, not exclusively derived from the keywords (e.g., "speak," "say," and "tell" for communication), and their semantically related terms. The latter are selected manually from the top 25 closest words to each seed term in the word-embedding space of the Shakespeare corpus (using *word2vec* algorithm). Semantic neighbors are independent of the predefined categories and reflect broader semantic relations within the corpus. The word lists for each category are provided in Appendix C. In the analysis, I measure the frequency of each category at the character level, calculate mean scores for character groups and genres, as well as perform Wilcoxon tests and calculate effect size to assess group distinctiveness.

4. Results

4.1 Character groups

Table 2 summarizes the three character groups across 37 plays from the Shakespeare Drama Corpus (cf. Fischer et al. 2019). The evaluation of the method is based on a qualitative assessment of this table: insofar as it presents a grouping that is meaningful, largely consistent with reading experiences, and suitable for further interpretation, the procedure can be used as an exploratory technique. Examining these groups and interpreting their members in relation to one another indeed proves illuminating: their shared —and, in some cases, well-known— structural features become visible in this way, while their distinctiveness can also be highlighted, particularly where shared patterns are realized in different ways. This also illustrates the value of computational methods for comparative studies, as a distant perspective makes it possible to grasp elements of a corpus in relation to one another.

Table 2: Members of the character groups.

Play	Character	Words
Dominant — Other		
antony-and-cleopatra	Antony	6105
coriolanus	Coriolanus	6578
henry-vi-part-1	Talbot	3191
henry-vi-part-3	King Edward IV	3558
henry-viii	Cardinal Wolsey	3252
julius-caesar	Marcus Brutus	5447
macbeth	Macbeth	5434
othello	Iago	8564
pericles	Pericles	4678
richard-ii	Richard II	6094
richard-iii	Richard	8979
romeo-and-juliet	Romeo	4729
timon-of-athens	Timon	6493
Dominant — Comedy		
measure-for-measure	Duke	6630
the-comedy-of-errors	Dromio of Syracuse	2109
the-merchant-of-venice	Portia	4666
the-tempest	Prospero	4790
two-gentlemen-of-verona	Proteus	3413
Speaker — Other		
cymbeline	Imogen	4324
hamlet	Hamlet	11832
henry-iv-part-2	Sir John Falstaff	5575
henry-v	Henry V, King of England	8454
henry-viii	King Henry the Eighth	3424
king-lear	Lear	5760

conference version

conference version

Play	Character	Words
titus-andronicus	Titus Andronicus	5817
Speaker — Comedy		
a-midsummer-nights-dream	Nick Bottom	2080
as-you-like-it	Rosalind	5793
loves-labors-lost	Berowne	4688
much-ado-about-nothing	Signior Benedick	3808
the-comedy-of-errors	Antipholus of Syracuse	2135
the-merry-wives-of-windsor	Falstaff	3745
the-taming-of-the-shrew	Petruchio	4714
the-winters-tale	Leontes	5020
twelfth-night	Sir Toby Belch	2640
Connector — Other		
cymbeline	Posthumus Leonatus	3379
hamlet	Horatio	2094
henry-iv-part-2	John of Lancaster	824
henry-v	Thomas, Duke of Exeter	980
henry-vi-part-2	Buckingham	571
king-lear	Earl of Kent	2636
richard-iii	Duke of Buckingham	2963
titus-andronicus	Lucius	1440
titus-andronicus	Aaron	2853
Connector — Comedy		
a-midsummer-nights-dream	Robin Goodfellow	1413
a-midsummer-nights-dream	Titania	1098
as-you-like-it	Touchstone	2406
loves-labors-lost	Costard Clown Swain	1560
measure-for-measure	Provost	1073
much-ado-about-nothing	Borachio	1014
the-comedy-of-errors	Dromio of Ephesus	1374
the-merry-wives-of-windsor	Sir Hugh	1874
the-taming-of-the-shrew	Katherine	1781
the-tempest	Ariel	1315
the-winters-tale	Paulina	2446
two-gentlemen-of-verona	Julia	2475

As for the Dominant characters, they are often the title characters: they actively shape the plot, and the story is primarily about them. They tend to speak extensively, interact with many other characters, and occupy a central position in the dramatic world. In *Romeo and Juliet*, this role is clearly fulfilled by Romeo (and not Juliet), in line with Masías et al. 2017's earlier findings that he consistently emerges as the main character of the drama from multiple perspectives. This observation also points to a broader pattern: with the notable exception of Portia in *The Merchant of Venice*, all Dominant characters are male, underscoring the gender imbalance of Shakespeare's dramatic world (while highlighting Portia's exceptional status as a character who both shapes and

reflects upon the events of the play.) This pattern also holds for the Speaker category, in which only Imogen and Rosalind appear as female characters; and partly for Connectors, where women occur in greater numbers only in comedies and are absent in other genres. In Shakespeare's world, mostly men are the prominent characters in a structural and verbal sense, while women tend to have a significant impact only as mediators between subgroups of the society.

At the same time, the Dominant group does not always coincide with the nominal or title character. In some cases, it includes characters who exert control over the title characters or conspire directly against them, thereby occupying their dominant position. Examples include Iago rather than Othello, Brutus rather than Caesar, Cardinal Wolsey rather than Henry VIII, or, in comedy, Dromio of Syracuse rather than his master (who unconsciously manipulates events). In such cases, these characters share the same structural significance as other Dominant characters by guiding, manipulating, or controlling the course of events – with the notable difference that their actions are shaped less by their own independent trajectories than by their relation to another character.

Of course, when the title character's dominance does not extend across the entire plot and community, it is not always replaced by a dominant scheming character such as Brutus or Iago. In some cases, a play instead exhibits a division between Speakers and Connectors: Speakers speak extensively and reflect on the events around them, but their actual influence on the dramatic world remains limited; while Connectors do not conspire against them—as dominant schemers would—but instead help them by maintaining ties between different subgroups of the community. Prototypical examples from tragedy include Hamlet and Horatio (whose structural importance was already described in Moretti 2011's study), or Lear—who deliberately renounces his central role in the community—and Kent.

Hamlet and Lear are indeed typical Speakers – at least their verbal prominence is often mentioned in Shakespeare scholarship and beyond. They constantly interpret events with great rhetorical power, yet for much of the time remain at a distance from direct action and the exercise of power. This emphasis on reflection, rather than action, seems necessary, since Speakers often confront intense emotional or psychological extremity from the outset of the plot. A similar pattern can be observed in the case of Leontes from *The Winter's Tale* who, after committing his fatal mistake, is removed from the center of the story—his main theme from this point on is mourning the past. Among histories *Henry VIII* can be mentioned, but there the king's counterpart is a dominant schemer, namely Cardinal Wolsey. Henry V is even more an exception in the Speaker's group. He consciously shapes his exercise of power in such a way that his commands are carried out even in distant locations without his physical presence; in the play, the loyal Exeter functions as the Connector who helps to maintain this form of authority. In this respect, Henry V may be the only king in the histories capable of sustaining such a community

conference version

structure (which also reflected in the play's famous beehive metaphor in I/2.) [8]

A different variation on the Hamlet–Horatio and Lear–Kent archetypes can be found in the pairing of Katharine and Petruchio in *The Taming of the Shrew*. Here, Katharine functions as a Connector insofar as suitors are drawn to her, rather than through any active effort on her part to bind the community together. Petruchio, by contrast, is highly active, but instead of engaging with the community as a whole, he concentrates his actions on its most distinguished member – namely, Katharine.

The comic dimension of Speakers is underscored by the comic counterparts of the tragic heroes in this group, such as Falstaff, who appears twice in the list (*The Merry Wives of Windsor* and *Henry IV, Part 2*). Falstaff is famously talkative but largely inactive; in *Henry IV, Part 2*, his Connector counterpart is the energetic yet far less verbose John of Lancaster. Read in this light, tragic Speakers such as Hamlet and Lear also take on aspects of the loudmouthed boaster. This impression is reinforced by other verbose characters ranging from Nick Bottom to Sir Toby Belch and Signior Benedick. It is important to note, however, that these clown-like characters are not clowns in the strict dramatic sense. True clowns such as Touchstone function more as Connectors, maintaining relationships between different parts of the community. In this sense, Touchstone functions exactly as he speaks, since his verbal wisdom promotes precisely the flexible movement between (conceptual) domains. Comic Speakers, by contrast, lack this reflexive wisdom: they are not witty observers of their own condition but appear instead as ridiculous characters in their persistent attempts to prove their importance. Again, this perspective can also be used to reconsider the seriousness of Speakers in tragedies.

Finally, as we have already seen, most Connectors are limal characters whose function is to establish connections between different levels of the dramatic world —and, at times, to generate chaos through this mediation by mixing different domains; think of, for example, Costard the Clown in *Love's Labour's Lost*. This role is particularly evident in characters such as Robin Goodfellow in *Midsummer* and Ariel in *The Tempest*, who mediate not only between lovers, opposing parties, or social classes, but also between the earthly and the magical realms. Servants often play a similar connective role, though in different ways. While Dromio of Ephesus primarily assists his master, Borachio in *Much Ado About Nothing* functions instead as an agent of intrigue. Other examples include Sir Hugh Evans in *The Merry Wives of Windsor*, who acts as a(n unsuccessful) broker of peace, as well as the Provost in *Measure for Measure* and Buckingham in *Henry VI, Part*

8.
Exeter
It follows, then, the cat must stay at home.
Yet that is but a crushed necessity,
Since we have locks to safeguard necessaries
And pretty traps to catch the petty thieves.
While that the armèd hand doth fight abroad,
Th' advisèd head defends itself at home.
For government, though high and low and lower,
Put into parts, doth keep in one consent,
Congreeing in a full and natural close,
Like music.

Bishop of Canterbury
Therefore doth heaven divide
The state of man in divers functions,
Setting endeavor in continual motion,
To which is fixèd as an aim or butt
Obedience; for so work the honeybees,
Creatures that by a rule in nature teach
The act of order to a peopled kingdom.
They have a king and officers of sorts,
Where some like magistrates correct at home,
Others like merchants venture trade abroad,
Others like soldiers armèd in their stings
Make boot upon the summer's velvet buds,
Which pillage they with merry march bring home
To the tent royal of their emperor,
Who, busied in his majesty, surveys
The singing masons building roofs of gold...

conference version

2, who serve as practical executors of power. This is the most diverse group in terms of gender, social status (although kings are missing here!), and ontology (fairies and humans); nevertheless, it appears functionally coherent, highlighting the usefulness of the approach to supplement existing typologies.

4.2 Keywords

This section investigates whether the character groups also differ linguistically beyond their functional-structural features. As a first step, I compare Speakers and Connectors across genres on the basis of their keywords (for setup see Appendix A, for cross-genre comparison see Appendix B)

The top 40 keywords for each group are presented in Figure 4. Although these keywords do not constitute a single, coherent semantic field, several consistent tendencies nevertheless emerge—primarily in connection with Speakers, whose keywords occur more frequently than those of Connectors (raw frequencies are indicated by the numbers in the plot). One particularly notable pattern is that Speakers not only speak more overall but also refer more often to communication itself. In tragedies, the keywords associated with Speakers include verbs such as *pray, tell*, and *say*, as well as second-person – often archaic – forms of address, including *didst, wilt*, and *prithee* (a longer list available in the data repository also contains *thou, read, thank* and *thanks*). In comedies, likewise, Speakers are associated with terms such as *write, word, speak*, and *speaks* (with *read, adieu*, and *thanks* appearing in the extended list), whereas *pray* appear as Connector distinctive in this genre. Furthermore, comic Speakers not only use different forms of address, such as *dear* and *thine*, but also foreground themselves within the communicative act (*us*, *myself*).

Similarly, dialogic markers such as *ha* and *ho* are highly distinctive for both groups of Speakers (also *o* in tragedies). Hamlet frequently exclaims in this manner during moments of intense address or heightened interaction—"Hillo, ho, ho, boy!", "With—ho!—such bugs and goblins in my life," "O villainy! Ho! Let the door be locked"; as does Lear—"ho, are you there with me?" —, but the most striking example is Pandarus' song in *Troilus and Cressida*, which juxtaposes the *ho* of pain with the *ha* of laughter:

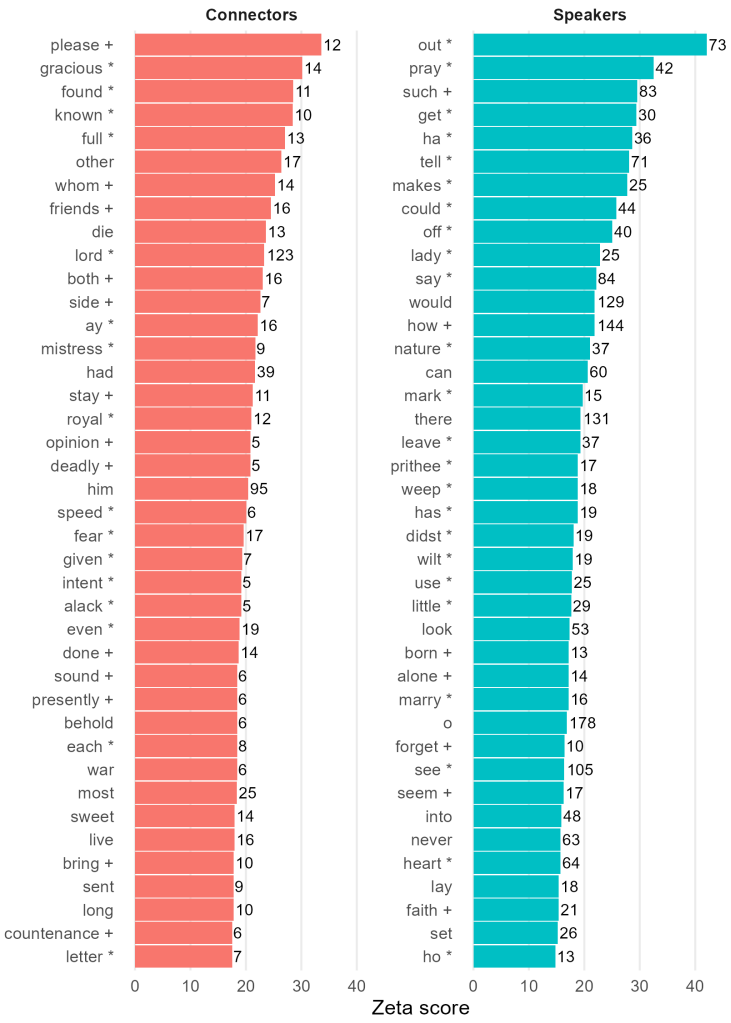
> These lovers cry "O ho!" they die,
> Yet that which seems the wound to kill
> Doth turn "O ho!" to "Ha ha he!"
> So dying love lives still.
> "O ho!" awhile, but "Ha ha ha!"
> "O ho!"groans out for "ha ha ha!"—Hey ho!

These lyrics further illustrate that Speakers display a slight tendency to use emotional and inward-oriented vocabulary. In tragedies and histories, this is reflected in words such as *weep*, *heart*, *alone* and *ha* as a part of *ha-ha*, whereas in comedies it appears in terms such as *love*, *dear*, *ha-ha*, *faith*, *fortune*, and *heaven*. Comparable inward-oriented expressions are largely absent from the keyword lists associated with Connectors.

Using terms of Jakobson 1960, the language of Speakers is not just about the referential

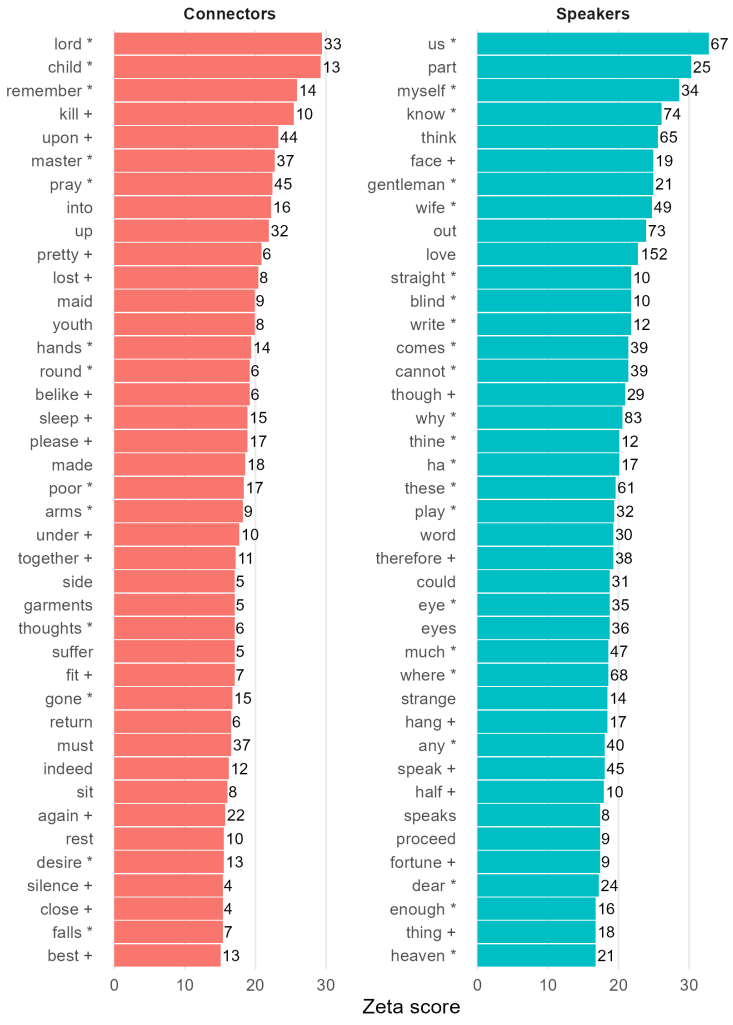
conference version

conference version



(a) Tragedies and histories

conference version



(b) Comedies

Figure 4: Top 40 keywords of Speakers and Connectors with Zeta scores and raw frequencies. Words that also appear among the top 40 keywords of reshuffled texts are marked with *, words in the range of top 41–100 are marked with +.

content, but also uses the metalingual, emotive, or conative functions of communication.[9] Another related characteristic of this group as opposed to Connectors is an emphasis on cognition and reasoning. This tendency is reflected in the frequent use of mental verbs (in comedies: *know, think*, ranked fourth and fifth; in tragedies: *forget* beside *see* and *look*), as well as modal verbs and conjunctions that express contingency and possibility (in comedies: *cannot, could, though*; in tragedies: *can, could, would, seem*). These are complemented by explanatory and interrogative forms that foreground causal and logical relations (in comedies: *therefore, how, why*, with *because* appearing in the extended list; in tragedies: *how*). Beyond explicitly referring to communication, Speakers thus tend to articulate the logical structure of events and to reason aloud, exploring alternatives and conditions rather than presenting their thoughts as compact, impersonal statements.

While the speech of Speakers often highlights the speaker and the addressee, that of Connectors is more consistently oriented toward other characters, though the frequency of the keywords is rather low in this group. In tragedies, this is reflected in terms such as *other* (although the signal disappears in the reshuffled comparison), as well as in references to specific roles or relations, including *lord*, *friends*, *mistress*, and *him* (with *warrant*, *prisoner*, and *lords* appearing in the extended list). Moreover, rather than drawing attention to the producers or receivers of the communicative situation, this group tends to evoke the role of messengers, as indicated by keywords such as *bring*, *sent*, and *letter*. In comedies, a similar pattern emerges in words such as *lord*, *child*, *master*, and *maid*, with *himself* occurring in the extended list. In Jakobson's terms, the language of Connectors primarily relies on the referential function, orienting utterances toward other participants, while the metalingual, conative, and emotive functions remain backgrounded.

This focus on others also evokes a comparatively formal mode of communication grounded in social hierarchy. In tragedies, *lord*, *mistress*, *royal*, and *gracious*, while in comedies *lord*, *master*, and *maid* can be mentioned as examples for Connectors. In both genres, *please* is highly distinctive, further indicating a formalized mode of communication in which requests are conventionally framed through politeness markers.

In summary, the keyword analysis allows for a differentiated characterization of the linguistic profiles associated with the character groups. Within this comparative framework, Speakers' language—being oriented toward address and dialogue—tends to highlight the communicative situation and emotional expression, frequently giving voice to inner thoughts and relational concerns. Connectors, on the other hand, emerge as characters whose speech is primarily oriented toward others in the social world rather than toward sustained self-expression.

4.3 Hypothesis test

To test the validity of these patterns across groups, the frequency of semantic categories was analyzed that capture the observed specificities of language use more comprehensively than individual keywords (see Keyword analysis). The categories include

9. *Metalingual function*: frequent comment on the act of speaking itself; *emotive*: express attitudes, intentions, or emotional states; *conative*: directly address or attempt to influence other characters; *phatic*: speech often serves to establish or maintain interaction; *referential*: language oriented toward external states of affairs; *poetic*: attention to the form and phrasing of the utterance. This typology from communication theory was also used for describing character speeches in drama theory (Pfister 1988, 103-118).

conference version

(1) communication, (2) discourse markers, (3) reference to self versus others (here no embedding was used just first and third person pronouns), (4) cognition, and (5) social hierarchy and politeness (see word lists in Appendix C). The frequency of each category was calculated separately for each character. Mean values were then computed for Speakers and Connectors within each genre (comedies versus non-comedies), and for the whole dataset. In addition, Wilcoxon tests were applied to assess statistical differences between the groups, reporting p-values and effect sizes (r). Table 3 provides an overview of the results.

Category	Direction	ALL Mean (C/S)	ALL p (r)	Comedy p (r)	Other p (r)
Communication	S > C	7.25 / 9.86	**0.016** (0.39)	0.31 (0.26)	**0.016** (0.62)
Discourse	S > C	0.33 / 0.82	**0.013** (0.41)	**0.038** (0.46)	0.24 (0.31)
Self–Other	S > C	12.2 / 25.1	**0.014** (0.4)	0.12 (0.34)	**0.042** (0.52)
Cognition	S > C	5.18 / 6.39	0.09 (0.28)	0.5 (0.15)	**0.016** (0.6)
Hierarchy	C > S	9.43 / 7.01	0.241 (0.2)	0.38 (0.20)	0.47 (0.20)

Table 3: Differences in frequencies of words from category. Mean values show relative frequency in 1000 words. S = Speaker, C = Connector, ALL = whole dataset without genre separation, p value is a result of Wilcoxon test, r value shows effect size.

This follow-up analyses provide partial support for the patterns suggested by the Zeta-based exploration, while also refining them in important ways. Communication-related language shows clear effect: Speakers exhibit higher frequencies than Connectors overall ($p = 0.016$, moderate effect), with the difference becoming particularly pronounced in tragedies and histories (large effect), but not in comedies. On the other hand, discourse markers (e.g., *ha*, *oh*, *ay*) favor Speakers in comedies, pointing to a more expressive interactional style. A similar pattern emerges for self-reference and cognitive language. Speakers tend to use more first-person forms and cognitive verbs, with statistically significant differences in tragedies and histories, while the distinction is attenuated in comedies. These results broadly align with the interpretation of Speakers as more discursively and cognitively oriented characters, though the strength of this association varies by genre. By contrast, terms related to social hierarchy are not distinctive, thus rejecting our hypothesis about Connectors' speech in this respect. The keywords associated with this group may contribute to a more general pattern of referring to others in the social world rather than to themselves.

These results do not simply confirm the Zeta findings but qualify them: the divergence between discursive, cognitive, social dimensions appears to be real but unevenly realized, with the clearest differentiation emerging in tragedies and histories and a tendency toward convergence in comedy. It is also evident that Speakers can be characterized more clearly in positive terms using these categories, whereas Connectors are better described by their relative absence.

5. Conclusion

This study proposes an exploratory approach to the analysis of character prominence in dramatic works. By deliberately allowing different metrics to diverge, the method does not reduce character roles to a single composite model but preserves the traceability of the resulting groups to distinct measures. The measures—the number of words, speech acts, and BC—were chosen for their distinctiveness, selectiveness and interpretability in

conference version

relation to dramas', or even societies' communication system and their relevance to plot function. The resulting character groups reveal multiple ways of prominence. Dominant characters emerge as both the main producers and transmitters in the communication system, most often title characters but sometimes displacing the nominal protagonist; Speakers are verbally prominent characters whose influence is mostly discursive; Connectors mediate between social groups, domains and levels of action, advancing the plot through coordination and mediation rather than reflection.

Linguistic analysis reinforces these functional distinctions. Speakers tend to foreground communication, first-person perspective, emotion, and reasoning; Connectors display a more formal mode of speech oriented towards others. This convergence across different approaches suggests that characters' functional and linguistic characteristics are closely aligned; at the very least, the discursive features of Shakespeare's prominent characters correspond to their structural roles, contributing to their complexity and roundness as characters.

6. Data Availability

The analyzed plays are available at https://github.com/dracor-org/shakedracor/tree/main; the version used in this study corresponds to Git commit 781dd85.

7. Software Availability

Codes and keyword lists can be found here: https://github.com/SzemesBotond/character_groups_and_keywords.

8. Acknowledgements

Funded by the European Union project "The Center for Digital Text Scholarship" under grant agreement ID 101186601.

9. Author Contributions

Botond Szemes: Conceptualization, Methodology, Analysis, Writing – original draft, Writing – review & editing

References

Algee-Hewitt, Mark (2017). "Distributed Character: Quantitative Models of the English Stage, 1550–1900". In: *New Literary History* 48.4, 751–782. 10.1353/nlh.2017.0038.

Archer, Dawn, Jonathan Culpeper, and Paul Rayson (2005). "Love – A Familiar or a Devil? An Exploration of Key Domains in Shakespeare's Comedies and Tragedies". In: *AHRC ICT Methods Network Expert Seminar on Linguistics*. Lancaster University.

conference version

Archer, Dawn and Alison Findlay (2020). "Keywords That Characterise Shakespeare's (Anti)heroes and Villains". In: *Voices Past and Present: Studies of Involved, Speech-related and Spoken Texts*. Ed. by Ewa Jonsson and Tove Larsson. Vol. 97. Studies in Corpus Linguistics. Amsterdam: John Benjamins Publishing Company, 32–46. 10.1075/scl.97.03arc.

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, 1–29. 10.48694/jcls.3670.

Bourgois, Antoine, Jean Barré, Olga Seminck, and Thierry Poibeau (2026). "Toward an ontological representation of fictional characters". In: *Computational Humanities Research* 2, e6. 10.1017/chr.2026.10025.

Burrows, John and Hugh Craig (2012). "Authors and Characters". In: *English Studies* 93.3, 292–309. 10.1080/0013838X.2012.668786.

Burt, Ronald S. (1995). *Structural Holes: The Social Structure of Competition*. Cambridge: Harvard University Press.

Chen, R. H.-G., C.-C. Chen, and C.-M. Chen (2019). "Unsupervised Cluster Analyses of Character Networks in Fiction: Community Structure and Centrality". In: *Knowledge-Based Systems* 163, 800–810. ISSN: 0950-7051. 10.1016/j.knosys.2018.10.005.

Craig, Hugh and Arthur F. Kinney, eds. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Culpeper, Jonathan (2009). "Keyness: Words, Parts-of-Speech and Semantic Categories in the Character-Talk of Shakespeare's *Romeo and Juliet*". In: *International Journal of Corpus Linguistics* 14.1, 29–59. 10.1075/ijcl.14.1.03cul.

Fischer, Frank et al. (2019). "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama". In: *Proceedings of DH2019: "Complexities"*. Utrecht University. 10.5281/zenodo.4284002.

Fischer, Frank, Peer Trilcke, Christopher Kittel, Carsten Milling, and Daniil Skorinkin (2018). "To Catch a Protagonist: Quantitative Dominance Relations in German-Language Drama (1730–1930)". In: *Digital Humanities 2018: Book of Abstracts*. Digital Humanities Conference, 26–29 June 2018. Mexico City.

Goffman, Erving (1959). *The Presentation of Self in Everyday Life*. Garden City, NY: Doubleday.

Granovetter, Mark S. (May 1973). "The Strength of Weak Ties". In: *American Journal of Sociology* 78.6, 1360–1380.

Havrylash, Julia and Christof Schöch (2025). "Exploring Measures of Distinctiveness: An Evaluation Using Synthetic Texts". In: *Journal of Computational Literary Studies* 4.1. 10.48694/jcls.4209.

Hicke, Rebecca M. M. and David Mimno (2024). ""Let Every Word Weigh Heavy of Her Worth": Examining How Women Enact Power in Shakespeare's Comedies through Interactive Speech Pattern Visualizations". In: *Computational Drama Analysis: Reflecting on Methods and Interpretations*. Ed. by Melanie Andresen and Nils Reiter. Berlin and Boston: De Gruyter, 87–106. 10.1515/9783111071824-005.

Hudson, Matthew and James Stiller (2005). "Weak Links and Scene Cliques in Shakespeare's Plays". In: *Journal of Cultural and Evolutionary Psychology* 3.1. 10.1556/JCEP.3.2005.1.4.

Jakobson, Roman (1960). "Linguistics and Poetics". In: *Style in Language*. Ed. by Thomas A. Sebeok. Cambridge, MA: M.I.T. Press, 350–377.

conference version

Jannidis, Fotis (2014). "Character". In: *Handbook of Narratology*. Ed. by Peter Hühn et al. Vol. 1. Berlin and Boston: De Gruyter, 30–45.

Keith, A., A. Rojas Castro, H. Ehrlicher, K. Jung, and Sebastian Padó (2025). "Towards Computational Analysis of Gender Depiction in the Comedias of Calderón de la Barca". In: *Journal of Computational Literary Studies* 4.1. 10.48694/jcls.4055.

Krautter, Benjamin and Janis Pagel (2024). "The Schemer in German Drama: Identification and Quantitative Characterization". In: *Computational Drama Analysis: Reflecting on Methods and Interpretations*. Ed. by Melanie Andresen and Nils Reiter. Berlin and Boston: De Gruyter, 123–148. 10.1515/9783111071824-007.

Krautter, Benjamin, Janis Pagel, Nils Reiter, and Marcus Willand (Nov. 2018). *Eponymous Heroes and Protagonists - Character Classification in German-Language Dramas*. Pamphlet 7. Digital Humanities Cooperation. https://www.digitalhumanitiescooperation.de/wp-content/uploads/2019/06/p07_krautter_et_al_eng-1.pdf.

— (Dec. 31, 2020). "[E]in Vater, dächte ich, ist doch immer ein Vater: Figurentypen im Drama und ihre Operationalisierung". Version 2.0. In: *Zeitschrift für digitale Geisteswissenschaften* 5. Version 2.0 vom 22.05.2023. 10.17175/2020_007_v2.

— (2023). "Properties of Dramatic Characters: Automatically Detecting Gender, Age, and Social Status". In: *Computational Stylistics in Poetry, Prose, and Drama*. Ed. by Anne-Sophie Bories, Petr Plecháč, and Pablo Ruiz Fabo. Berlin and Boston: De Gruyter, 179–202. 10.1515/9783110781502-010.

Masías, Víctor Hugo, Paula Baldwin, Sigifredo Laengle, Augusto Vargas, and Fernando A. Crespo (2017). "Exploring the Prominence of Romeo and Juliet's Characters Using Weighted Centrality Measures". In: *Digital Scholarship in the Humanities* 32.4, 837–858. 10.1093/llc/fqw029.

Mian, Haaris, Melanie Subbiah, Sharon Marcus, Nora Shaalan, and Kathleen McKeown (2026). "Computational Representations of Character Significance in Novels". In: *arXiv preprint*. Preprint. https://arxiv.org/abs/2601.15508.

Moretti, Franco (2011). "Network Theory, Plot Analysis". In: *New Left Review*, 80–102.

Murphy, Sean, Dawn Archer, and Joanne Demmen (2020). "Mapping the Links between Gender, Status and Genre in Shakespeare's Plays". In: *Language and Literature: International Journal of Stylistics* 29.3, 223–245. 10.1177/0963947020949438.

Pfister, Manfred (1988). *The Theory and Analysis of Drama*. European Studies in English Literature. Illustrated, reprint edition. Cambridge: Cambridge University Press, 339. ISBN: 9780521423830.

Reiter, Nils, Benjamin Krautter, Janis Pagel, and Marcus Willand (2018). "Detecting Protagonists in German Plays around 1800 as a Classification Task". In: *Abstracts of EADH: Data in the Digital Humanities*. Galway, Ireland. 10.18419/opus-10162.

Schöch, Christoph (2017). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". In: *Digital Humanities Quarterly* 11.2. 10.1073/pnas.1620741114.

Šeļa, Artjoms, Ben Nagy, Joanna Byszuk, Laura Hernández-Lorenzo, Botond Szemes, and Maciej Eder (2024). "From Stage to Page: Stylistic Variation in Fictional Speech". In: *Computational Drama Analysis: Reflecting on Methods and Interpretations*. Ed. by Melanie Andresen and Nils Reiter. Berlin and Boston: De Gruyter, 149–166. 10.1515/9783111071824-008.

Simmel, Georg (1950). "The Triad". In: *The Sociology of Georg Simmel*. Ed. and trans. by Kurt H. Wolff. Reprint edition. New York: Simon and Schuster, 154–170.

conference version

Szemes, Botond and Bence Vida (2024). "Tragic and Comical Networks: Clustering Dramatic Genres According to Structural Properties". In: *Computational Drama Analysis: Reflecting on Methods and Interpretations*. Ed. by Melanie Andresen and Nils Reiter. Berlin, Boston: De Gruyter, 167–188. 10.1515/9783111071824-009. https://doi.org/10.1515/9783111071824-009.

Ubersfeld, Anne (1999). *Reading Theatre*. Ed. by Paul Perron and Patrick Debbèche. Trans. by Frank Collins. Vol. 3. Toronto Studies in Semiotics and Communication. Illustrated edition. Toronto: University of Toronto Press, 219. ISBN: 9780802082404.

Vishnubhotla, Krishnapriya, Adam Hammond, and Graeme Hirst (2019). "Are Fictional Voices Distinguishable? Classifying Character Voices in Modern Drama". In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Minneapolis, USA: Association for Computational Linguistics, 29–34.

Woloch, Alex (2003). *The One vs. the Many: Minor Characters and the Space of the Protagonist in the Novel*. Princeton, NJ: Princeton University Press.

conference version

A. Appendix

A. Keyword analysis setup

text.slice.length = 700
text.slice.overlap = 0
rare.occurrences.threshold = 3
zeta.filter.threshold = 0.1
oppose.method = "craig.zeta"

B. Keywords of Speakers and Connectors

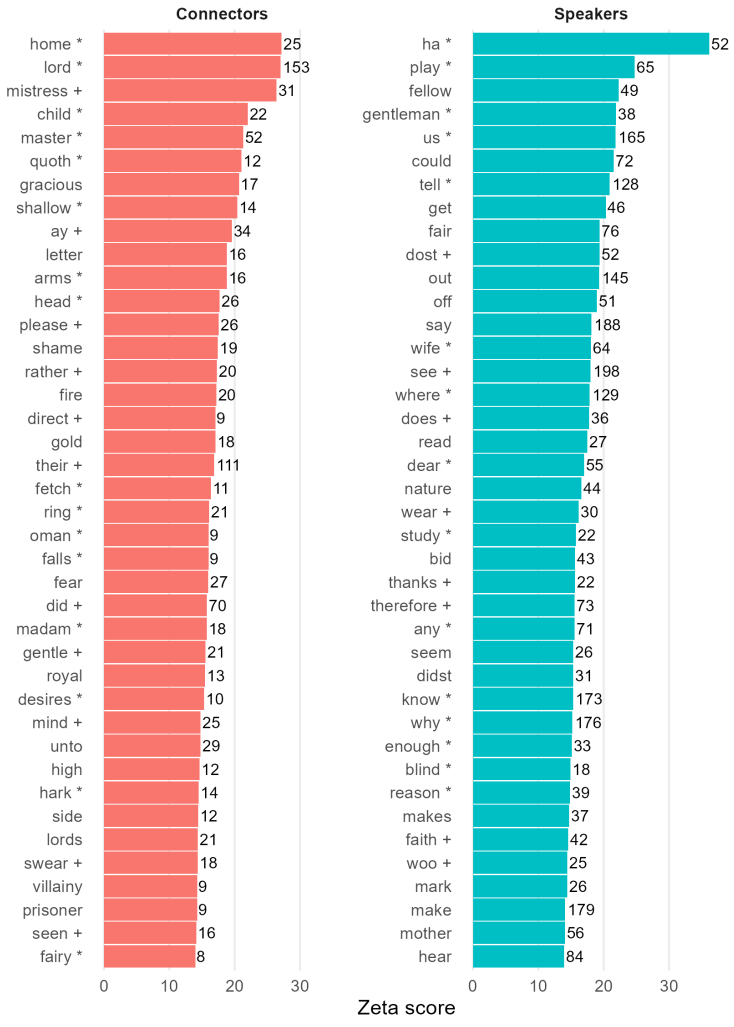


Figure 5: Top 40 keywords of Speakers and Connectors regardless of the genre of the play, with Zeta scores and raw frequencies. Words that also appear among the top 40 keywords of reshuffled texts are marked with *, words in the range of top 41-100 are marked with +.

conference version

C. Lexical Categories: Seed Terms and Word Lists

A. Communication

Seed words: say, tell, speak, ask, answer, pray, write, read

List: say, said, says, tell, told, speak, spoke, ask, answer, hear, call, bid, pray, prithee, beseech, entreat, request, demand, beg, word, words, question, write, written, writing, writ, read

B. Discursive markers

Seed words: ho, ha, oh, ay

List: ha, ho, oh, o, ay, nay, fie, foh, hey, heigh, hark

C. Self vs. others

Positive effect (*first person*): i, me, my, myself, we, us, our, ours, ourselves

Negative effect: he, him, his, himself, she, her, hers, herself, they, them, their, themselves

D. Cognitive

Seed words: think, know, believe, remember, forget

List: think, know, believe, remember, forget, knows, remembered, forgot, perceive, mean, learn, trust, understand, how, why

E. Hierarchy

Seed words: lord, master, sir, madam, king, lady

List: king, queen, prince, princess, noble, royal, lord, lords, lordship, lady, ladyship, sir, madam, master, mistress, liege, grace, highness, majesty, servant, knight, doctor, please, pleaseth, pardon, thank, thanks, humbly, welcome, gramercy

conference version

jcls. Journal of Computational Literary Studies

conference version

Coreference Resolution for Full German Novels using Large Language Models

Agnes Hilger[1]
Anton Ehrmanntraut[1]

1. Institut für Deutsche Philologie, Universität Würzburg, Würzburg, Germany.

OPEN ACCESS

Citation
Agnes Hilger and Anton Ehrmanntraut (2026). "Coreference Resolution for Full German Novels using Large Language Models". In: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7983

Date published 2026-05-05 (preprint)
Date accepted tbc
Date received 2026-01-07

Keywords
coreference resolution, literary corpus, annotation, LLMs, German 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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. The paper introduces the GerFuN dataset, consisting of five German-language novels fully annotated for character coreference, comprising a total of 450,000 tokens. Using a semi-manual pipeline, we first pre-annotated the novels using a LLM, and then manually corrected the annotations in the INCEpTION tool. The annotation guidelines, which build on existing approaches but are made more explicit and refined, are presented in the paper and released alongside the dataset. Finally, we evaluate LLMs on GerFuN, which surpass previous pipelines and exhibit near-human accuracy on prototypical cases of particular interest to literary studies.

1. Introduction

As literary scholars, we are typically concerned with texts in their entirety. Even when we focus on individual parts such as sentences, paragraphs, or chapters, we relate them back to the text as a whole. From this perspective, the existing approaches to coreference resolution in literary texts – the task of identifying when different expressions in a text refer to the same entity – present a challenge. This is because the datasets usually consist of text *excerpts*. This limitation reflects a fundamental trade-off: since coreference annotation is time-consuming and resources are limited, researchers must choose between annotating many short documents or fewer long ones (Bourgois and Poibeau 2025, 2). Pioneering projects have generally opted for the former in order to gain an initial overview of the problem (Krug et al. 2018; Bamman et al. 2020). Consequently, current local pipelines to automatic coreference resolution employed in computational literary studies (CLS), such as BookNLP, are mostly trained and evaluated on excerpts and do not scale reliably to full-text novels (Martinelli et al. 2025). This is illustrated by the qualitative evaluation by Schröder et al. (2021): when applying their model to *The Hound of the Baskervilles*, they obtain 31 distinct clusters for Sherlock Holmes instead of a single one.

More recently, research has started to create datasets with a few novels annotated in full (Bourgois and Poibeau 2025; Martinelli et al. 2025). However, no such dataset exists for German-language literature as yet and, accordingly, no systematic evaluation for it. Against this background, our article makes three main contributions:

1. We present the GerFuN ('German Full Novel') dataset, consisting of five German-language novels fully annotated for character coreference.

2. We develop a semi-manual annotation pipeline and demonstrate its practical viability.

3. We establish and evaluate an automatic annotation pipeline based on Large Language Models (LLMs), which on both the largest German character coreference excerpt dataset (DROC; Krug et al. 2018) and our own dataset outperforms current state-of-the-art methods and approaches human performance.

Unlike some previous work (Roesiger et al. 2018; Bamman et al. 2020), we deliberately restrict our scope to the genre of novels and to the entity type of characters. This choice allows us to control variation across genres and entity types and to develop more detailed annotation guidelines.

Throughout the paper, we adopt the following data model. We distinguish between (i) **mention** spans (or simply *mentions*), which are defined solely by their start and end positions in the text; (ii) **entities** in the discourse, which are associated with attributes such as name, gender, and special-case status (e.g., *generic*); and (iii) **references**, which link precisely one mention to precisely one entity and may carry additional attributes, such as special-case labels (e.g. *figurative* or *part*). Importantly, distinct to other datasets, our model allows a single mention to be linked to more than one entity through multiple references, covering what is commonly referred to in the literature as "plural mention" or "split-antecedent." This data model forms the basis for both the annotation process and the subsequent evaluation presented in this paper.[1]

2. Related Work

2.1 Literary Datasets with Coreference Annotations

For several years now, there has been a growing awareness in the coreference literature that literary texts differ from other types of text in terms of their coreference behavior (Roesiger et al. 2018), as they do in other respects, and that specific datasets are therefore needed for evaluation. Therefore, a number of datasets of literary texts annotated with coreference have been created, differing in document language, genre and whether they use full documents or excerpts (Krug et al. 2018; Bamman et al. 2020; Pagel and Reiter 2020; Bourgois and Poibeau 2025; Martinelli et al. 2025; Han et al. 2021). For an overview, see Table 1.[2]

However, a coreference dataset with fully annotated German-language novels does not exist to date. Krug et al. (2018) annotate German-language novels, but only use excerpts. Pagel and Reiter (2020) annotate German-language texts, some of them in their entirety, but only plays. Bourgois and Poibeau (2025) and Martinelli et al. (2025) annotate entire novels, but in French and English.[3] That is why we are introducing GerFuN.

1. Even though we explicitly model entities and their attributes, our primary task remains *coreference resolution* in the sense of identifying the equivalence between linguistic expressions within the discourse. These entities are not given *a priori* but need to be inferred through the text, which differentiates our setup from the task of *entity linking*, where mentions are linked to entries in a predefined, external knowledge base.
2. For a more comprehensive overview of existing datasets, see Pagel (2024, 39–42) and Bourgois and Poibeau (2025, 3).
3. Martinelli et al. (2025) also annotate Hermann Hesses *Siddhartha*, but apparently in an English translation. In addition, they work with predefined lists of characters based on external resources, which means that not all characters are annotated, but only the main characters and a few minor characters.

conference version

Dataset	Language	Full/Excerpts	Genre	Doc.	Tok.	Avg Tok./Doc.
DROC (Krug et al. 2018)	German	Excerpt	Novels	90	393,000	4,368
LitBank (Bamman et al. 2020)	English	Excerpt	Novels	100	210,532	2,105
GerDraCor-Coref (Pagel and Reiter 2020)	German	Both	Plays	45	298,352	6,630
Long-LitBank-fr (Bourgois and Poibeau 2025)	French	Full	Novels	3	285,176	95,058
$\text{BookCoref}_{\text{gold}}$ (Martinelli et al. 2025)	English	Full	Novels	3	229,000	76,419
GerFuN (ours)	German	Full	Novels	5	459,866	91,973

Table 1: Overview over existing datasets for coreference resolution in literary texts.

2.2 Guidelines for Coreference Resolution in Literary Texts

Together with the datasets mentioned in the previous subsection, the rules specified for annotation were published in varying degrees of detail. However, as far as we know, the complete guidelines as provided to the annotators were not published.

For our context, the rule sets that (also) deal with narrative texts are particularly relevant, namely Roesiger et al. (2018), which first discussed literary-specific phenomena, Krug et al. (2018), which also annotated character coreference in German novels and Bamman et al. (2020). Building strongly on this work, we have developed our own guidelines for GerFuN, which are more explicit than previous rules and will be published together with the paper.

2.3 Automatic Coreference Resolution

Current conventional systems for coreference resolution primarily utilize discriminative Transformer encoder models based on span representation (Lee et al. 2017, 2018; Joshi et al. 2019; Xu and Choi 2020), with the Maverick architecture establishing state-of-the-art performance on English LitBank (Martinelli et al. 2024) and German DROC (Petersen-Frey et al. 2025), yet all of them evaluated only on literary excerpts. Generative approaches based on sequence-to-sequence modeling have recently shown competitive performance (Wu et al. 2020; Bohnet et al. 2023; Zhang et al. 2023; particularly Hicke and Mimno 2024 for literary sentences), but require re-generating the entire input text, making these pipelines impractical for book-level coreference resolution. In a similar regard, compute requirements for discriminative Transformer encoders typically increases quadratically with the input length, which becomes prohibitive when processing long documents.

As such, current pipelines designed for processing full-length (literary) texts must adapt conventional architectures. One strategy involves limiting the antecedent search space, as implemented in the established English BookNLP pipeline[4], which has been evaluated by Martinelli et al. (2025) on their full-novel test set. Similarly, Mélanie-Becquet et al. (2024) introduced the French BookNLP-fr pipeline, which was subsequently improved and evaluated by Bourgois and Poibeau (2025) on their long-document Long-LitBank-fr corpus. Alternatively, models may merge sub-clusters inferred from text segments. For German, Gupta et al. (2024) proposed a hierarchical neural merging approach, reporting results on the novel *Effi Briest*. Similarly, Martinelli et al. (2025) tested their English BookCoref pipeline, which merges clusters via surface-level proper nouns,

4. See https://github.com/booknlp/booknlp.

conference version

Author	Title	Year
Caroline Auguste Fischer	*Gustavs Verirrungen*	1801
Ferdinand Kürnberger	*Der Amerika-Müde*	1855
Wilhelmine Heimburg	*Trudchens Heirat*	1894
Julius Wolff	*Das Wildfangrecht*	1907
Johann Wolfgang von Goethe	*Die Wahlverwandtschaften*	1809

Table 2: Overview over the five annotated novels.

across their annotated English novels. Yet, these evaluations show that the models are struggling with long documents, not exceeding 60 CoNLL F1 points, except for BookCoref, which however constrains its inference and evaluation to a predefined list of (primary) characters.

The emergence of LLMs has introduced a new paradigm, though current results remain mixed. Several studies report promising results using QA-style prompts (Gan et al. 2024; Vadász 2023) and end-to-end annotations (Le and Ritter 2024). However, these experiments often rely on providing the model with gold-standard mention spans. As Le and Ritter (2024) note, and the CRAC 2025 Shared Task demonstrated (Novák et al. 2025), downstream performance is particularly sensitive to high-quality mention detection, while simultaneously LLMs struggle with this subtask. Due to this weakness, we opt for a hybrid setup with a conventional Transformer model for mention detection, and LLMs to perform mention linking.

3. Source Data

We annotate five complete novels, presented in Table 2.

Four of the novels come from a random sample drawn from a larger corpus of 925 German novels published between 1790 and 1915.[5] All four randomly selected novels are not particularly canonical in contemporary German literary studies.[6] Additionally, a fifth novel (Goethe's *Elective Affinities*) was deliberately selected as a canonical example.[7]

It can be assumed that commercial LLMs likely encountered these novels during training. However, we claim that this does not invalidate our subsequent evaluations. First, the models were trained only on raw text, without access to coreference resolution annotations. Second, while secondary material (summaries, lists of characters, etc.) exists online for the canonical *Elective Affinities*, for our other four randomly selected novels, almost no secondary material is available, minimizing the risk of data contamination.

5. The corpus and sample are part of an ongoing PhD project by one of the authors and will be published in the future. Based on the project's research interests, a stratified sample of 14 novels was drawn, of which four are used here.
6. The most canonical of the four is certainly Kürnberger's *Der Amerika-Müde*, which still plays a role in the context of 19th-century German images of America. A search in the scholarly bibliography BDSL yields three and four hits for "Kürnberger Der Amerika-Müde" and "Kürnberger Amerikamüde" respectively, one for "Fischer Gustavs Verirrungen," and none for "Heimburg Trudchens Heirat" and "Wolff Wildfangrecht."
7. The BDSL search for "Johann Wolfgang Goethe" yields 17095 hits, a search for "Goethe Die Wahlverwandtschaften" yields 181 hits.

conference version

conference version

4. Guidelines

We created our guidelines in a semi-cyclical process (Reiter 2020), based on the guidelines by Krug et al. (2018), Bamman et al. (2020) and Roesiger et al. (2018), on narratological theory and our own annotation experience. In the following, we will give a brief introduction.[8]

Fundamental to our guidelines is the conceptual distinction between a 'prototypical case' on the one hand and 'special cases' on the other. By 'prototypical case,' we mean what falls – by our assumption – under the prototypical concept of characters, e.g., Edward, Charlotte, or the gardener in Goethe's *Elective Affinities*. Special cases, on the other hand, are cases that deviate from this prototypical case in some way, for example, generic entities such as 'gardeners in general.' The guidelines focus primarily on the prototypical case, and are less refined when it comes to the special cases.

4.1 Prototypical Case

We focus on the annotation of characters. Therefore, by 'entity,' we always refer to characters, where a character in our understanding is a human or human-like being in a storyworld.[9] We further explicate 'human or human-like' with Jannidis (2004) who draws on conceptions from folk psychology. A being is considered human-like if it has the ability to communicate, has agency, transitory and more stable features, and a difference between an inside and an outside.[10] According to our explication, the horse Caesar from example (1) is not a character, the two lions and the fox from (2) are.

> (1) Caesar, who saw the water jug, instinctively trotted with $[\text{me}]_1$, and it was thanks to his nostrils rather than the $[\text{woman's}]_2$ deathly murmur that $[\text{I}]_1$ found the water source so quickly. (Kürnberger 1856, 294)[11]

> (2) After a poor night's hunt, the $[\text{lion}]_1$ and his young $[\text{son}]_2$ came across a well-fed $[\text{fox}]_3$ at daybreak. The $[\text{fox}]_3$ quickly realized that there was no escape for $[\text{him}]_3$. $[\text{He}]_3$ sighed in $[\text{his}]_3$ heart: "Poor $[\text{thing}]_3$, $[\text{your}]_3$ last hour has come if $[\text{your}]_3$ wits cannot save $[\text{you}]_3$ from danger.["] (Klinger 1810, 78)

Remember that we use the terms 'mention span' or 'mention' to refer to the pure text span, and 'reference' to refer to the link of a mention to an entity. Mentions usually appear in three forms: proper names (e.g., "Charlotte"), common nouns (e.g., "gardener"), and pronouns (e.g., "she"). For proper names, we annotate the full name. For other forms, as Krug et al. (2018), we only annotate the head of the phrase. For nested structures with multiple mentions within a noun phrase, we annotate the corresponding head for

8. For details, please consult the full guidelines: https://github.com/aehrm/llm_literary_coref/blob/main/gerfun_corpus/Guidelines_Coreference_Resolution_(5.0)_(translation).pdf.
9. With this definition, we follow a line of standard conceptions in narratology (Jannidis 2009, 14; Köppe and Kindt 2014, 120; Hillebrandt 2017, 141).
10. See the entry on the 'Basistypus' by Jannidis (2004, 251). Similar criteria are later used by Martínez and Scheffel (2012, 145) in their widely used introduction to narratology and again by Hillebrandt (2017, 141). For a critique of the criterion of 'human likeness,' see, for example, Eder (2008, 56). For an integration of this criticism and extension to other media, see Eder et al. (2010).
11. When using examples from our own or other corpora, we strive to cite a genuine, preferably scholarly edition of the work. If this is not possible, we cite based on the corpus and the text source used therein. For original German-language texts, we cite the original in the appendix. In the body text, we use an existing translation whenever possible; otherwise, we translate the quote ourselves with the help of DeepL.

each mention. Each reference is assigned the ID of the respective entity.

Like Krug et al. (2018) and Bamman et al. (2020), we allow singletons, i.e., entities that are only mentioned once. If entities are mentioned within idioms or other fixed expressions (e.g. *O my god!*), we do not annotate this as a reference. For every entity, we annotate the ID, name and gender.

We understand gender as a fluid social construct. We allow the following values: `f` (female), `m` (male), `u` (unknown or unspecified), `nb` (non-binary), `o` (other) and the option to enter new classes. However, we also point to the fact that the annotated novels stem from 19th century, where gender conceptions were predominantly binary.[12]

The following snippet shows examples for the prototypical case and how we annotate it, as well as an overview of the annotated entities:

> (3) [Trudchen Baumhagen]$_1$ had quickly crossed the quiet church square, opened a gate in the wall opposite, and now stood on family ground. [...] In the large vaulted hallway, [she]$_1$ met [her]$_1$ [brother-in-law]$_2$ standing next to a bicycle. [He]$_2$ was very elegant and dressed in the latest fashion [...] A [servant]$_3$ was busy rubbing the shiny steel of the vehicle with a leather cloth. "Well," asked the young [girl]$_1$ kindly, "are [you]$_2$ going for a ride, [Artur]$_2$?" (Heimburg [1894] 2008)

1. Trudchen Baumhagen, f;
2. Artur Fredrich, m;
3. Diener (*servant*), m.

4.2 Special Cases

Our starting point is always the prototypical case described above. Deviations are primarily considered in terms of the rules for the prototypical case, rather than the rules for deviations. By 'special cases,' we mean a series of frequently occurring cases that deviate from the prototypical case in a certain way.

4.2.1 Mentions with Multiple References and Group Entities

A mention span can refer to more than one character at the same time.

> (4) The first time, for a long while, [Charlotte]$_1$ sate at the head of the table [herself]$_1$ – and it seemed to [Ottilie]$_2$ as if [she]$_1$ was deposed. The two **[ladies]**$_{1,2}$ sate opposite **[each other]**$_{1,2}$[.] (Goethe [1809] 1868, 102)[13]

> (5) A large **[party]**$_1$ had assembled for the occasion. **[They]**$_1$ went first to church, where **[they]**$_1$ found the whole **[congregation]**$_2$ collected together in **[their]**$_2$ holiday dresses. (Goethe [1809] 1868, 56)

We distinguish between two cases here.

12. Since our guidelines are not very detailed in this regard, given the complexity of gender, we recommend adding an additional guideline refining and data correction step if actually using this data for gender-related analyses.
13. Since newer translations are not regularly available in German libraries we use an older translation available online.

conference version

i) Mentions with multiple references: In the first case it is (easily) identifiable which entities are being referred to and/or the referred-to characters are not (frequently) appearing as a unit. Typical examples include plural pronouns, but also the mention span "ladies" from (4): It only refers to two characters at the same time, which are easily identifiable. We treat such cases as a single mention span with multiple references. For 'ladies' from (4) we annotate two references: One to Charlotte, one to Ottilie.

ii) Group Entities: In the second case it is not (easily) identifiable which characters are being referred to and/or it is a group (frequently) appearing as a unit. Prototypical cases for this are the respective groups from example (5). We treat such groups like a single entity with one ID. Additionally, we flag it as a *group* and tag its members, if they or some of them are identifiable.

Our rules for this phenomenon differ in part from previous guidelines for character coreference in literary texts (Roesiger et al. 2018; Krug et al. 2018; Bamman et al. 2020). They treat characters referred to in one mention span always as a group entity with its own ID. Which characters belong to the group is sometimes, but not always recorded in the designation. We decided on this deviation because our annotation provide more information this way: We more systematically record which characters are referred to together, and when. This could be relevant for various questions in literary studies.[14]

However, although we find our approach generally sensible, there are still many issues to be resolved and our data is not yet perfectly annotated in this regard. Member and group relationships can quickly become very complex. Groups appear as members of other groups, the overlap between undefined groups is often unclear, it is not uncommon for the exact membership of a group to remain implicit, and group sizes change over the course of the text, making it difficult to regard the group as an entity.[15]

Therefore, even though we have refined the rules here, a lot of things remain open both in our guidelines and in our annotated data. It has become apparent that the topic is more complex and interesting than it might seem at first glance. Given the qualitative and quantitative relevance of such character groups in literary texts,[16] it is surprising how little attention has been paid to the topic as such in both computational and non-computational literary studies. We believe that there is still much to be done in this area.

4.2.2 Generic Entities

By generic entities, we mean human or human-like beings that are not individuals, but rather a type or class of individuals (Bamman et al. 2020; Reiter and Frank 2010):

14. For example, with our annotations it remains traceable when, in *Elective Affinities*, what units out of the four main characters are formed. At the same time, our annotations can still be converted back to the format of other guidelines automatically: All mentions with multiple references could be identified and converted to a mention with one reference to a group entity.
15. Roesiger et al. (2018) also point out this problem.
16. E.g. in *Elective Affinities*, group composition among the main characters Charlotte, Eduard, Ottilie, and the Hauptmann are genuinely relevant to the meaning of the text. The title *Elective Affinities* describes a theorem from contemporary chemistry that deals with the compositional nature of elements. These elements symbolically represent the protagonists: at the beginning, only Charlotte and Eduard live together, but with the addition of new elements (Ottilie, Hauptmann), a new composition emerges, and a romantic connection develops between Eduard and Ottilie as well as between Charlotte and Hauptmann. With our annotations, it stays traceable when in the text which units are formed.

conference version

(6) IT is a truth universally acknowledged, that a single **[man]**$_1$ in possession of a good fortune, must be in want of a **[wife]**$_2$. (Austen [1813] 1996, 5)

Three main issues arise with respect to generic entities: i) whether they should be annotated at all, ii) how references to generic entities should be treated with regard to coreference, and iii) how to handle mention spans that simultaneously refer to generic and specific entities.

i) Generic entities resemble our prototype of a character in important respects. But they also differ from it to such an extent that treating them simply as characters would be counterintuitive. We therefore annotate generic entities just as entities, but additionally tag them as a special case. This approach is common in related annotation guidelines.[17] Other than for prototypical characters, for generic entities we do not manually assign a name and gender.

ii) We treat references to generic entities as ineligible for coreference.[18] Generic entities are therefore always also singletons in our corpus. This decision was made after several iterations in the guideline creation process, for pragmatic rather than conceptual reasons: It is often only possible with considerable effort to decide whether several mention spans refer to the same generic entity, since the boundaries here are less sharp than with prototypical characters. This problem is exacerbated by the length of the text. In *Elective Affinities*, for example, different characters and the narrator make different statements about 'human beings in general' at very different points in the text. It is counterintuitive to assume that this is the same entity every time, as they are conceptualized differently. Treating each mention as separate, as we currently do, may also be counterintuitive, but it appears to be the right pragmatic solution given our primary research interest.

iii) A single mention span may simultaneously refer to a generic and a specific entity (Bamman et al. 2020, 46; Roesiger et al. 2018, 131). However, from our experience, there are very different degrees of implicitness here. When the dual reference is explicit or only mildly implicit and can be inferred from the immediate context, we annotate multiple references for the same mention span, linking it both to the generic and to the specific entity. In more complex cases, where interpretation depends on broader contextual knowledge or deeper inference, we annotate only the reference to the generic entity.

The opening sentence of *Pride and Prejudice* (6) cited above illustrates such a complex case: although readers may ultimately associate the statement with characters such as Bingley or Darcy, this interpretation requires substantial contextual buildup. We therefore annotate only the generic reference in this instance.

As points ii) and iii) in particular show, our treatment of generic entities deliberately simplifies a highly complex phenomenon. This simplification is justified by our specific research focus on prototypical characters. At the same time, given the literary-specific challenges posed by generic entities and the varying degrees of implicitness involved,

17. In the DROC corpus, generic entities are marked as "abstract and not part of the fictional world" (Krug et al. 2018, 8). We believe that this might conflate two distinct dimensions – genericity and fictional-world status. We therefore follow Krug et al. (2018) in tagging generic entities as such, but adopt the terminology of Roesiger et al. (2018) and Bamman et al. (2020).
18. Therefore, instead of Krug et al. (2018), we are rather following the approach of Bamman et al. (2020) here, but simplifying it even further.

conference version

further research dedicated to this topic appears both necessary and promising. Such work could productively connect to questions in literary studies about the construction of social stereotypes through literary characters or about propositions literary texts make about the world more generally.[19]

4.2.3 Nonfactual Entities

By 'non-factual entities,' we mean humans or human-like beings in the fictional world, who are in some way – in relation to their world – not factual (i.e., possible, negated, etc.):

> (7) "[I]$_1$ want to tell [you]$_2$ something, [Franz]$_2$, joking aside," [he]$_1$ continued, "[you]$_2$ are going to have to get married! And [my]$_1$ advice to [you]$_2$ is to compromise [your]$_2$ ideals a little in this matter. [...] Don't bring [me]$_1$ a poor **[girl]**$_3$, [Franz]$_2$, even if **[she]**$_3$ were the **[pearl]**$_3$ of all the world.[″] (Heimburg [1894] 2008)

We treat such entities as entities, but additionally mark them as *nonfact* special cases.

To our knowledge, this problem has not yet been addressed in guidelines on coreference resolution in literary texts. However, our approach is consistent with CLS guidelines on other aspects of literary texts (Brunner et al. 2020, 16–17; Vauth and Gius 2021, 5–6; Kröncke et al. 2022, 27), as well as with narratological character theory (Margolin 1987, 111–113, 1995, 376–377; Jannidis 2004, 173, 201). As for the other special cases, more fine-grained modeling, especially concerning the relation to generic entities, would be needed here.

4.2.4 Possible Identity

By 'possible identity,' we mean a relationship between two entities: it is possible but not certain that these two entities are actually the same entity. These cases can take many different forms. One example is the deliberate play with, and obscurity of, identity concepts, as found in German Romantic literature by, for example, E.T.A. Hoffmann (e.g., *The Devil's Elixirs*, *The Sandman*). In such cases, it does not seem reasonable to us to assume either one entity or two completely separate entities. We therefore annotate them as two entities with the relationship *possible identity*. Possible identity as we conceptualize it is a case of "true ambiguity," which Roesiger et al. (2018, 133–134) refer to as a specifically literary feature in coreference.

4.2.5 Revelation of Identity and Changing Entities

By 'revelation of identity,' we mean the case in which it only becomes known at a point in the reading that is later than the currently annotated one that a character is actually another character, i.e., that they are identical.[20] By *changing entities*, we mean the case in which entities change so much over the course of the narrative that it becomes

19. Generic Entities could be seen as one way to explicitly establish characters as representatives of larger social groups existing in the real world. For a discussion on such explicit markers see (Jannidis and Lauer 2002, 106). One example from our corpus is the novel *Der Amerika-Müde*, which constantly constructs national and racial stereotypes through connecting characters to larger social groups.
20. This differs from 'possible identity' in that here the relationship is unknown at one point in time but becomes known at a later point, whereas with possible identity, the identity remains open overall.

conference version

questionable whether they are still the same entity. Both phenomena fundamentally concern the question of what knowledge is decisive in annotation.

In our annotation, following Roesiger et al. (2018) and Bamman et al. (2020), the knowledge gained from reading the entire text is seen as decisive, and therefore two entity candidates that are found to be identical at some point in the text are also treated as one entity; changing entities are also treated as one entity. A special case of changing entities are groups of changing size and composition (Roesiger et al. 2018). We assume a new entity when the composition and size of the group changes.

Overall, we are pursuing an approach that could be called static: our annotation is not concerned with capturing the dynamic provision of information by the text, but rather with the static knowledge gained after reading the entire text. However, the fact that we have opted for this approach does not mean that the dynamic approach is unjustified.

4.2.6 Figurative Mentions

By figurative language, we understand an umbrella category encompassing metaphors, metonymy, similes, and other tropes. Mentions can refer to characters in the form of figurative language.[21]

> (8) [T]he young $[\text{man}]_1$ was not really a $[\text{poet}]_1$; but surely he was a $\textbf{[poem]}_1$[.] (metaphor) (Chesterton, *The Man Who Was Thursday*), quoted from (Bamman et al. 2019, 3).

> (9) "[T]he outraged sentiment of the $\textbf{[kitchen]}_1$ was avenged by a bad and hasty dinner. (metonymy) (Oliphant, *Miss Marjoribanks*), quoted from (Bamman et al. 2019, 3).

We treat figurative mentions as such, if they can be identified as such in the immediate context and with a low depth of inference. Additionally we flag them as *figurative*. With that decision we partly follow, partly deviate from Bamman et al. (2020), where metonymic mentions are annotated but metaphor mentions are not.

4.2.7 Character Parts as Mentions

Sometimes a character is only implicitly present in that some part of them is mentioned, but they are not addressed as a whole. As far as we know, this issue is not discussed in existing guidelines. In narratological character theory, it is known as one way of 'indirect character naming' (Jannidis 2004, 123–124). A common case is that characters are initially only heard but not seen, and thus only their voices are mentioned.

> (10) Once, as $[\text{I}]_1$ was indulging in $[\text{my}]_1$ usual daydreams at the theater, $[\text{my}]_1$ attention was drawn to two female $\textbf{[voices]}_{2,3}$ $\textbf{[that]}_{2,3}$ seemed to be coming from the neighboring box. (Fischer [1801] 2012)

Normally, we do not annotate character attributes such as their voice as mentions. An exception to this rule are cases such as (10), where there is no other mention of the character in the immediate context. If such cases were not annotated, passages in which

21. We are only concerned here with cases in which a character is referred to using figurative language, not with cases in which a character itself stands for something else.

conference version

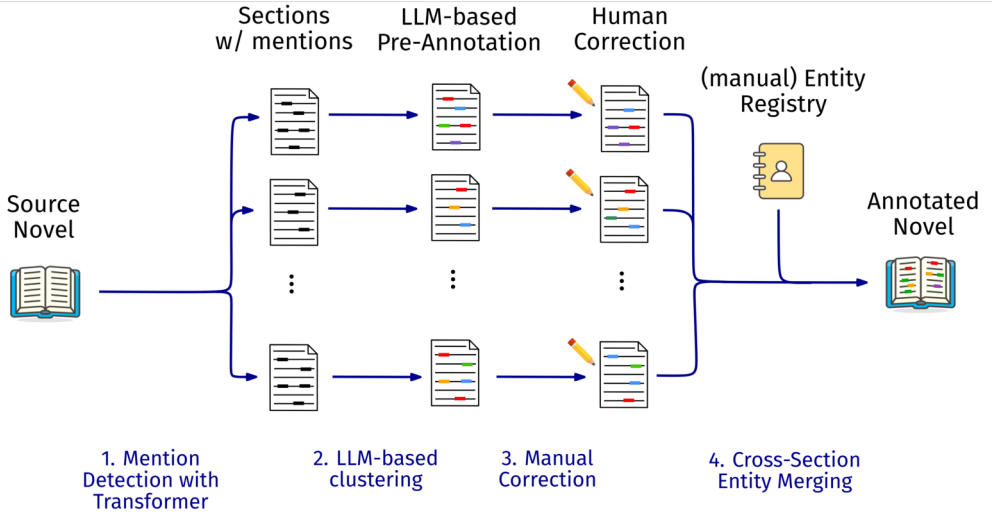


Figure 1: Sketch of our semi-manual annotation pipeline.

the character is present for the reader and other characters, in which attributes are ascribed to them, and in which they appear as speakers, for example, would be lost. Additionally, we flag such 'Character parts as Mentions' as such with the tag *part*.

5. Semi-Manual Annotation Process

Annotating coreference across all five full-length novels from scratch would be highly labor-intensive. Therefore, we adopted a semi-manual, multi-step workflow, using an LLM to create a high-quality draft, which then undergoes intensive manual correction and extension (see Figure 1). The manual correction was done for each novel by one of two trained human annotators, who had read the respective novel in advance during another annotation task.[22] The manual correction was done in the annotation tool INCEpTION.[23]

5.1 Document Partitioning and Mention Detection

Novels were partitioned first by chapter boundaries. Then, for each section, we automatically identified potential mention spans using the incremental coreference model by Schröder et al. (2021), then the state-of-the-art for DROC. We retained only the inferred mention spans, discarding the model's clustering results. (In the fully automatic pipeline in section 8, this was replaced with a ModernGBERT-based detector.)

5.2 LLM-based Pre-annotation for Clustering

After mention spans had been identified, we derived an initial clustering to entities, using Google's *Gemini 2.5 Flash* LLM, guided by a deliberately simple and intuitive 'minimal' prompt modeled around DROC's guidelines.[24] For each section, the predicted mentions were marked inline in the source text with a number, and the LLM assigns in

22. One of the annotators is a student assistant at the end of her undergraduate with experience in different previous CLS annotation projects, the other is one of the authors.
23. See https://inception-project.github.io/.
24. See https://github.com/aehrm/llm_literary_coref/llm_literary_coref/prompts/mention_prompts.py, line 137.

conference version

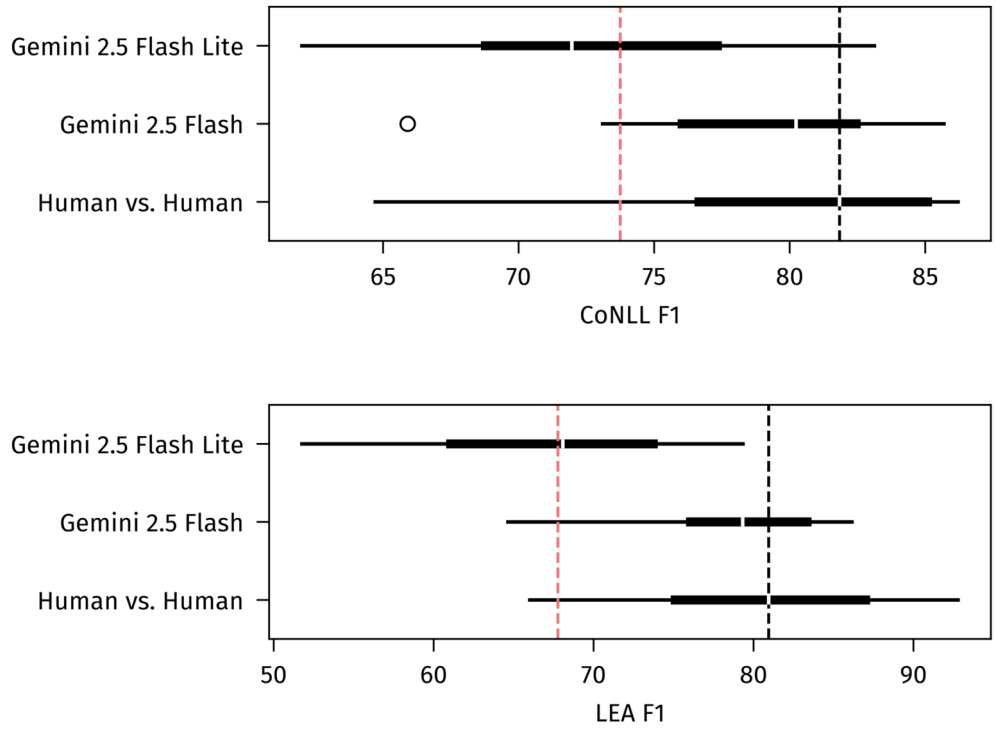


Figure 2: Performance of different LLMs with the 'minimal' prompt on the 12 double-rated documents of the DROC corpus, using gold mentions of the first rater as input, and predictions evaluated against the annotations of the second rater. "Human vs. Human" compares gold annotations of the first rater to the ones of the second rater. Red dashed line are the reported scores for the best-performing Transformer coreference model by Petersen-Frey et al. (2025), though on a different test split.

the output to each numbered mention a human-readable name as ID for the referenced character.[25] We refer to the output of the LLM as the *LLM-based pre-annotation*. Total cost for the automatic LLM inference on all five novels was less than 7 USD.

This model and prompt were selected after experiments on the DROC corpus showed clustering performance comparable to human agreement, outperforming less expensive models and conventional Transformers (Figure 2). Given the similarity in guidelines, this LLM setup provides a reliable baseline for frequent coreference clusters, leaving the complex decisions for the expert human annotators.

5.3 Manual Correction and Enrichment

The LLM-based pre-annotations were imported into INCEpTION as span annotations for intensive manual correction and enrichment. Human annotators refined mention spans, corrected entity IDs, and added special-case attributes like *figurative* or *part*. Mentions with multiple references were realized by layering multiple span annotations over the same span position, each linked to a different entity.

One might assume that the LLM pre-annotation in the data introduces a certain bias in the subsequent evaluation. To estimate this bias, we quantify the extent of the manual

25. While the setup of labeling each mention with the character's name might resemble the task of *entity linking*, we want to stress that this setup is only a practical convenience for the automatic inference of mention clusters, not a linking step to a pre-existing external entity set.

conference version

Metric	Precision	Recall	F1-Score
Mentions	98.10 $^{+0.78}_{-1.53}$	97.75 $^{+0.89}_{-1.28}$	98.00 $^{+0.56}_{-1.56}$
MUC	96.46 $^{+2.53}_{-6.38}$	92.39 $^{+1.98}_{-3.56}$	94.35 $^{+1.72}_{-5.44}$
B^3	92.31 $^{+3.58}_{-8.39}$	74.25 $^{+6.95}_{-6.55}$	80.89 $^{+5.27}_{-5.15}$
$CEAF_e$	57.78 $^{+5.35}_{-6.84}$	56.05 $^{+16.96}_{-16.48}$	54.94 $^{+8.20}_{-9.11}$
CoNLL	—	—	76.11 $^{+4.89}_{-5.13}$
LEA	89.67 $^{+4.57}_{-11.03}$	64.70 $^{+9.53}_{-11.79}$	74.46 $^{+7.23}_{-9.74}$

Table 3: Comparison of the uncorrected LLM-based pre-annotation against the final corrected gold annotation across all sections of all five novels. Scores are reported as median percentage points with IQR. For detailed metric definitions, see Section 6.

revisions. Table 3 compares the initial LLM-based pre-annotation against the final gold annotation. The considerate drop across all metrics highlight that the pre-annotation was intensively updated to obtain the final gold standard.

5.4 Cross-Section Entity Merging

Because annotation was performed section-by-section, a final consolidation step was required to resolve cross-section coreference. Annotators manually created a 'global' entity directory, mapping 'local' cluster IDs to unique global identities. This directory also recorded entity-level attributes, including gender and special cases (e.g., groups, nonfactual, or generic entities). Finally, the source text, corrected annotations, and directory were aggregated into a machine-readable tabular format to form the GerFuN dataset.

6. Metrics

To fully capture all components of the annotated datasets, we will employ three categories of metrics:

1. *Clustering Metrics* (MUC, B^3, $CEAF_e$, LEA) which assess the assignments of mentions to entities;
2. *Reference Attribute Metrics* which compare – on the level of mentions – the presence of reference special cases *figurative* and *part*, and the presence of references to *generic* entities;
3. *Entity Attribute Metrics* which compare entity attributes (gender, presence of entity special cases *nonfact*, *group*) between gold and predicted entities.

6.1 Clustering Metrics

Conventionally, quality of coreference resolution is measured by MUC (Vilain et al. 1995), B^3 (Bagga and Baldwin 1998), and $CEAF_e$ (Luo 2005), and the CoNLL F1-score (arithmetic average of the three respective F1-scores). However, all these metrics possess well-documented flaws (Bagga and Baldwin 1998; Luo 2005; Moosavi and Strube 2016; Duron-Tejedor et al. 2023). We therefore prioritize the LEA metric (Moosavi and Strube 2016), which addresses some of these drawbacks and is increasingly used alongside

conference version

the previous metrics, but also report CoNLL F1-scores for consistency and comparison with previous studies.

Some of these measures cannot handle mentions with multiple references, and thus we follow the generalization by Zhou and Choi (2018) due to their straightforward adjustments and interpretability. Note that, in the conventional setting where all mentions have precisely one reference, all adjusted metrics coincide with the conventional ones.

For discussing these metrics, let $\mathcal{K}$ be the set of entity clusters in the key (gold annotations), and $\mathcal{R}$ be the set of entity clusters in the response (system output). Note that each entity $E \in \mathcal{K} \cup \mathcal{R}$ is modeled as a set of mention spans.

MUC. The MUC metric is a link-based metric and computes the recall based on the minimum number of missing links. MUC has widely and frequently been criticized for not considering singletons and having low discriminative power, and we keep MUC only to be able to derive the commonly used CoNLL score. Since there is no natural way to generalize MUC's partitioning function p to multiple references, we will proceed by just ignoring mentions having multiple references when calculating p.

B^3. The metric B^3 is mention-based. The conventional metric defines the precision of a mention m as the proportion $|K_m \cap R_m|/|R_m|$ where $K_m \in \mathcal{K}$ resp. $R_m \in \mathcal{R}$ is the unique cluster in key resp. response containing mention m. Overall precision is the average precision over all mentions. Recall is derived by swapping the role of key and response. To generalize to mentions referencing multiple entities, we replace K_m resp. R_m with the union of all clusters in the key resp. response referenced by m.

CEAF_e. The metric CEAF_e is an entity-based metric, measuring resolution on the mention-level. CEAF_e recall pairs key entities to response entities, and then evaluates the 'similarity' ϕ of each key–response entity pair, using the Dice–Sørensen coefficient, resp. F1-score, i.e.,

$$\phi(K, R) = \frac{2|K \cap R|}{|K| + |R|}.$$

Then, it uses the Kuhn–Munkres algorithm to find the one-to-one mapping g^* between the key and response entities which maximizes the sum

$$\Phi = \sum_{K \in \mathcal{K}, R \in \mathcal{R}, R = g^*(K)} \phi(K, R)$$

Then, overall recall is $\Phi/|\mathcal{K}|$ and precision is $\Phi/|\mathcal{R}|$, effectively making this metric an *unweighted* average over all entities' resolution ϕ. Since CEAF_e is entity-based, it can immediately handle mentions with multiple references.

LEA. The LEA metric also provides an entity-based evaluation, but measures resolution on the link-level. LEA recall is computed as a weighted average over all key entities, where each entity's contribution is its *resolution*, weighted by its *relevance*, defined as the number of mentions $|E|$. Entity resolution counts the proportion of retrieved links and is defined as

conference version

$$resolution(K) = \begin{cases} \frac{\sum_{R \in \mathcal{R}} links(K \cap R)}{links(K)} & \text{if } |K| > 1, \\ 1 & \text{if } |K| = 1 \text{ and ex. } R \in \mathcal{R} \text{ with } K = R, \\ 0 & \text{otherwise,} \end{cases}$$

where $links(X) = |X|(|X| - 1)/2$.

LEA Precision is defined by swapping the role of key and response. Like CEAF, LEA can immediately handle mentions with multiple references. Note that the weighting distinguishes LEA from the unweighted $CEAF_e$, thus LEA puts more emphasis on larger entities, whereas $CEAF_e$ is more sensitive towards high-frequency small entities such as singletons.

General Evaluation Scenarios. In order to specifically assess some sub-aspects of the coreference resolution, we also calculate above scores for some modified key/response clusters.

1. *Full evaluation*;
2. *Without singletons*, by removing all mentions in the key (resp. response) that reference a singleton entity when computing recall (resp. precision);
3. *Without generics*, like above, but removing mentions to generic entities;
4. *Transformed multi-reference mentions*, by replacing each mention with multiple references by a new pseudo-entity that is identified by the set of referenced entities, thus behaving like conventional coreference resolution datasets and metrics, such as DROC.

6.2 Reference Attribute Metrics

Beyond clustering, we evaluate the classification of *figurative* and *part* reference special cases. To handle mentions with multiple references, we evaluate as a binary classification task on the mention level. Given a special case, we assess whether a mention span in the key and response has *at least one* reference with the given attribute, and then compute precision and recall accordingly. We also evaluate the recognition of *generic* entities in this matter (since generic entities are always singletons), comparing mentions having at least one reference to a generic entity.

6.3 Entity Attribute Metrics

Finally, we also assess a system's ability to correctly classify attributes of the entities. We measure the predictive performance for gender and the special cases *group* and *nonfact*, all treated as separate binary classification tasks.

To compare attributes between key and response entities, we align them by re-using the one-to-one mapping g^* of the $CEAF_e$ metric. This mapping maximizes the mention-level F1-score between entity pairs. The evaluation of attributes is then performed exclusively on the set of successfully aligned entity pairs, disregarding all entities that cannot be matched, in order to isolate the task of attribute classification from upstream clustering errors, thus measuring retrieval of entity attributes, *conditioned* on the entity being successfully identified.

conference version

Title	Tokens	Sentences	Mentions	Entities
Gustavs Verirrungen	24,959	1,380	4,056	223
Trudchens Heirat	64,927	3,782	8,953	380
Amerika-Müde	197,358	9,331	23,638	4,928
Wildfangrecht	78,815	3,203	12,119	463
Wahlverwandtschaften	93,807	3,570	12,541	1,704
Total	459,866	21,266	61,307	7,698
Average	91,973	4,253	12,261	1,540

Table 4: Overview of the annotated corpus.

Mention Type / Attribute	Count	Proportion
Total Mentions	61,307	100.0%
Proper Name Mentions	7,492	12.2%
Common Noun Mentions	14,086	23.0%
Pronominal Mentions	39,729	64.8%
Mentions with multiple references	5,621	9.2%
Mentions with *figurative* reference	352	0.6%
Mentions with *part* reference	78	0.1%

Table 5: Distribution of mention characteristics across the entire corpus.

To refine this analysis, we report attribute scores under several filtering conditions:

1. *Full Evaluation*;
2. *Without Groups*, on the subset of matched entities that are not marked as *group* in either key or response;
3. *Without Groups and Singletons*, like above, additionally excluding singletons.

7. Analysis of the Annotations

7.1 Dataset Statistics

In line with similar corpus publications, we provide in this section some typical descriptive statistics characterizing the GerFuN dataset, also serving as a reference point for future studies regarding long-context literary coreference behavior and evaluation strategies. Our final corpus consists of five novels, totaling over 450,000 tokens, 61,000 mentions and 7,600 entities (Table 4).

Table 5 details the properties of annotated mentions, classified by automatically predicted POS tags into proper names, common noun, and pronominal mentions, with pronominal mentions constituting the vast majority (64.8%). Mentions assigned at least two references account for 9.2% of all mentions, necessitating explicit modeling, whereas mentions with *figurative* or *part* references are rare (0.6%, resp. 0.1%).

We also report in Table 6 the distribution of different entity classes annotated in the corpus. For the purpose of our analysis, we define **Core Entities** as entities being referenced at least twice, and not annotated as group or generic, capturing the prototypical case of entities, constituting 10.1% of all entities. Furthermore, we also consider the category of **Singleton Entities** (entities referenced precisely once) in our investigation,

conference version

Entity Class / Attribute	Count	Proportion
Total Entities	7,698	100.0%
Non-Singleton Entities	1,311	17.0%
Core Entities	774	10.1%
Singleton Entities	6,387	83.0%
Non-Generic	979	12.7%
Generic	5,408	70.3%
Group Entities	1,002	13.0%
Nonfactual Entities	115	1.5%

Table 6: Distribution of entity characteristics across the entire corpus.

	Non-Generic						
Gender	**No Group**		**Group**		**Generic**		**Overall**
f	264	3.43%	103	1.34%	–	–	4.77%
m	892	11.59%	296	3.85%	–	–	15.43%
nb	1	0.01%	–	–	–	–	0.01%
mf	–	–	65	0.84%	–	–	0.84%
u	131	1.70%	538	6.99%	5,408	70.25%	78.94%

Table 7: Distribution of references by gender across different entity classes.

due to its high frequency in GerFuN (83.0%) and their downstream effect on evaluation metrics. Note however, that among the singleton entities, most of them are generic entities (70.3% of all entities).

Regarding the special case entities, we observe that the nonfactual entities are quite rare (1.5%), whereas group entities are relatively common, occurring with a frequency even higher than that of the Core entities. Annotated gender per entity is presented in Table 7. Even though authorial gender in GerFuN is relatively balanced, we see a very high prevalence of male characters compared to other genders.[26]

Table 8 shows a heavily skewed distribution of references: Core entities (10% of all entities) account for 85% of references, while generic entities (70% of entities) contribute only 8% of references. This skew impacts clustering metrics: the weighted entity-based metric LEA favors core entities, while unweighted $CEAF_e$ captures singletons.

Figure 3 visualizes the long-tailed Zipfian distribution for core entity sizes, with the 10

	Entities		References	
Entity Class	**Count**	**Proportion**	**Count**	**Proportion**
Core Entities	774	10.1%	58,017	84.9%
Group Entities	1,002	13.0%	3,977	5.8%
Generic Singletons	5,408	70.3%	5,408	7.9%
Non-Generic Singletons	979	12.7%	979	1.4%
Total	**7,698**	**100.0%**	**68,275**	**100.0%**

Table 8: Distribution of the number of references to a particular class of entities.

26. The single entity marked with nb is the legendary figure Elf Puck mentioned in *Der Amerika-Müde*.

conference version

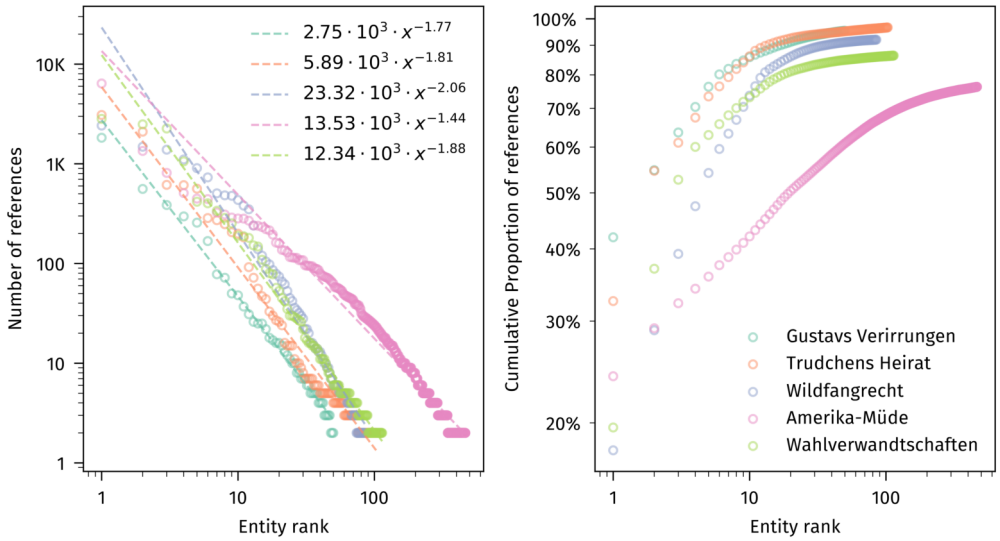


Figure 3: Each circle represents a core entity. Left: Entity Rank (sorted by number of references) versus number of references to that entity. Dashed lines are ordinary least squares regressions in the log-log space. Right: Cumulative proportion of references in the entire novel covered by the top x largest entities. Difference to 100% is accounted by omitted entities (groups, singletons).

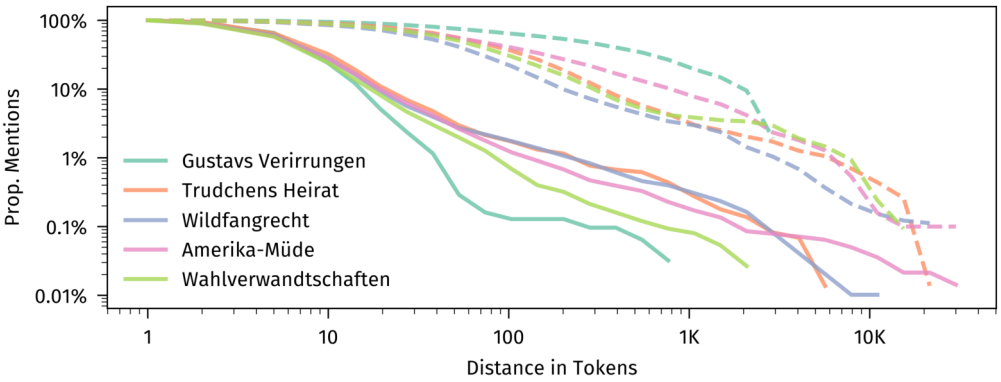


Figure 4: Complementary cumulative distribution function of distance to the nearest antecedent: given a fixed number of tokens n (x-axis), the y-axis indicates the proportion of mentions whose nearest antecedent is *further away* than n tokens. Solid lines represent distance to any mention, dashed lines represent distance to only proper name antecedents.

largest entities accounting for 40%–90% of all references.

We also analyze the locality of coreference by measuring the distance to the nearest antecedent. For each reference on a pronominal or nominal mention, we locate the nearest mention referencing the same entity, and measure the distance between them in unit of tokens. Median distance to the nearest antecedent is 6 tokens, and for 99% of mentions, no more than 137 tokens. The distribution has a very long tail (Figure 4, solid lines), with some antecedents over 10,000 tokens away. This is amplified for proper name antecedents (dashed lines), which are more remote (median 58 tokens; 99% within 5,473 tokens). This long-tail distribution and observations regarding spread imply that, while most references in novel-length texts are local, a significant portion requires document-level understanding beyond typical context windows.

This long-range dependency is further reflected in the entity *spread*, introduced by

conference version

conference version

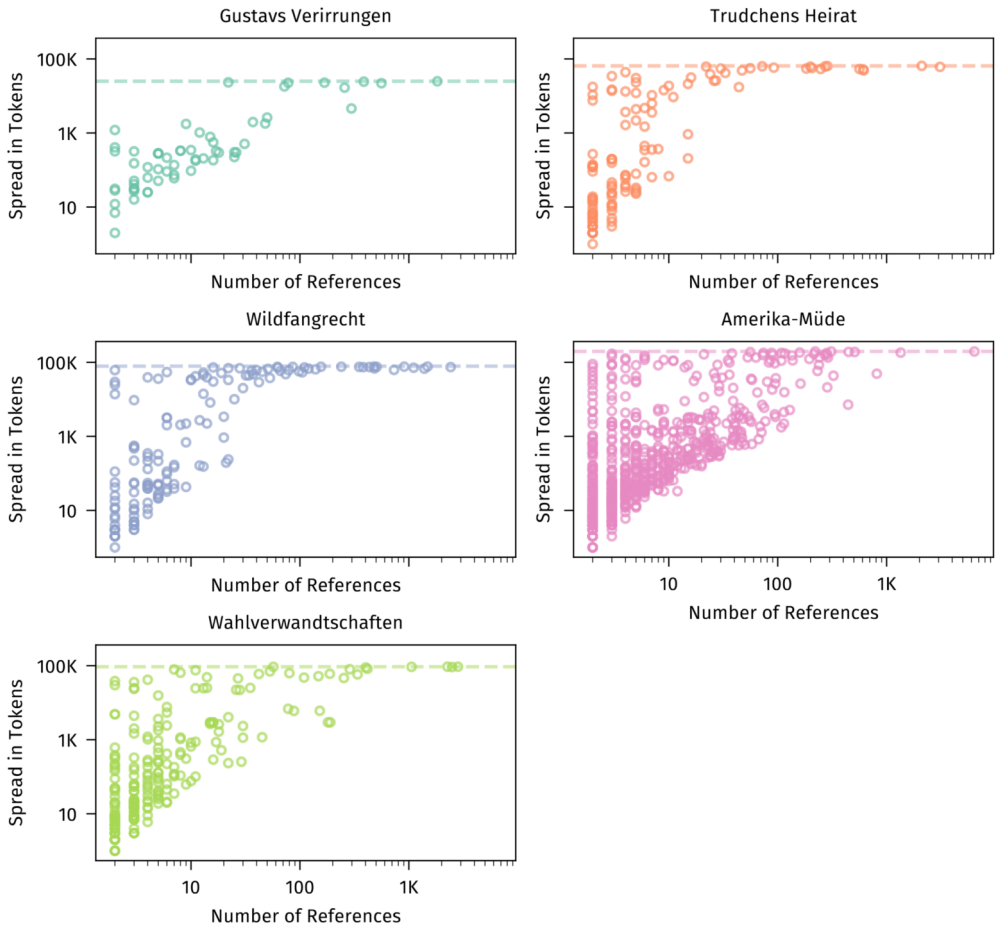


Figure 5: Relationship between entity size and entity spread for non-singleton entities. Each marker represents one entity. Horizontal dashed lines indicate novel length.

Toshniwal et al. (2020) as the token distance between an entity's first and last mention. For non-singleton entities, 50% have a spread of less than 106 tokens, but 10% exceed 47,460 tokens. Entity spread appears independent of size, with many small entities showing document-length spreads (Figure 5).

7.2 Inter-Annotator Agreement

To assess the reliability and consistency of our annotation guidelines, we estimate Inter-Annotator Agreement (IAA). For this, we follow a setup similar to Bamman et al. (2020): each of the two annotators independently annotated the first 2,000 tokens of four novels from our random sample (thus excluding *Wahlverwandtschaften*), totaling 8,000 tokens. Like the full annotation process (section 5), both annotators were given the LLM-based pre-annotation using the minimal prompt and were and tasked to correct and extend them. We then calculated the agreement by treating one annotator's result as the gold standard and the other as the prediction, computing the F1-score for the metrics outlined in the previous section. It is important to note, however, that this excerpt-based assessment cannot fully capture the increased complexity of full-length coreference annotation, hence can only serve as an approximate upper bound for the true IAA.

Table 9 reports the IAA based on our (micro-averaged) clustering metrics, and indicate a high level of agreement (LEA 87.18%), confirming that annotators were largely consis-

Evaluation	Mentions	MUC	B^3	$CEAF_e$	CoNLL	LEA
Full	96.23	94.77	88.99	82.18	88.65	87.18
Transformed Multi-Ref. Mentions	96.23	94.38	90.13	84.05	89.52	88.29
w/o Generics	96.89	95.11	89.80	83.64	89.52	88.44
w/o Singletons	97.21	95.11	89.80	84.13	89.68	88.68

Table 9: Inter-Annotator Agreement (F1-score) on clustering metrics.

		Gender			Special Case	
Evaluation		m	f	u	group	nonfact
Full	F1	96.36	82.76	95.65	86.36	33.33
	Count	55+55	14+15	68+70	20+24	3+3
w/o Groups	F1	97.67	96.00	98.31	–	0
	Count	43+43	12+13	59+59	–	1+2
w/o Groups, Singletons	F1	100.00	100.00	–	–	0
	Count	25+25	10+10	–	–	0+2

Table 10: Inter-Annotator Agreement on entity attributes, showing F1-score and the raw counts from each annotator.

conference version

tent at clustering. As expected, agreement improves slightly when excluding generics and singletons, which often present more ambiguity.

To situate our results, we compare them to the median IAA of the DROC corpus computed in section 5. Our *Transformed multi-reference mentions* scenario is directly comparable to the DROC setup, and here, our IAA scores compare favorably, for instance the IAA of 88.29% LEA F1-score against DROC's 80.95%, suggesting that the annotation task benefited from our explicit guidelines. At the same time, agreement is also notably higher than the machine-to-gold agreement of the LLM-based pre-annotation (Table 3). Taken together, this indicates that, for one, annotators made extensive manual corrections to the LLM-based pre-annotation, and second, among themselves, agreed in these modifications.

The agreement on entity and reference attributes, reported in Table 10 and Table 11, has mixed quality. While agreement is very high for gender assignments (`m`, `f`, `u`),[27] as well as for identifying *group* (86.36%), and satisfactory for *generic* entities (75.91%). For the remaining special case categories, agreement is considerably lower. The F1 scores for *nonfact* entities (33%), resp. *part* (29%) and *figurative* references (25%), are all modest.

This low agreement should not surprise us, since these categories were not at the core of our interest, and we deliberately opted for intuitive and succinct guidelines for these

Attribute	F1-score	Count
part	28.57	5+9
figurative	25.00	3+13
generic	75.91	67+70

Table 11: Inter-Annotator Agreement (F1-score) on mention-level metrics for reference special case *part*, *figurative* and entity special case *generic*.

27. For gender `mf` and `nb` we report no IAA due to the extremely low prevalence in the dataset.

cases.

In summary, the high IAA validates the quality and consistency of GerFuN, particularly for our well-defined prototypical cases, whereas the lower agreement on some special cases provides an assessment of which tasks are straightforward and which are challenging to model for CLS.

8. Using LLMs as Automatic Coreference Annotators

Building on our semi-manual workflow, this section explores a fully automated LLM pipeline for coreference resolution. Extending our DROC findings, this study demonstrates that high-quality resolution on full-length novels is achievable with comparatively little cost. We limit the scope to high-reliability categories, excluding special cases like *part* and *figurative*, as well as complex entity relations like group membership.

8.1 Pipeline

As a natural automation of our manual process, we adopt the same multi-stage workflow. Thus, the automatic pipeline consists of three steps. 1) Mention detection using a conventional local Transformer-based tagger, 2) LLM-based section-level coreference annotation, and 3) LLM-based cross-section entity merging.

Mention Detection. Detection of mention spans is performed using a token classification model based on ModernGBERT (Wunderle et al. 2025), fine-tuned on the mention spans of the DROC corpus. Thus we replace the larger and more complex full coreference model by Schröder et al. (2021) used in the manual pre-annotation phase, in order to minimize the technical barrier of entry. On the (held-out) GerFuN dataset, our model achieves an F1-score of 97%. We utilize the mention spans predicted by this model as the fixed input for the subsequent LLM step.

LLM-Based Section-Level Coreference Annotation. For the core annotation task, we employ the same section-wise processing strategy as in the pre-annotation phase. However, instead of the 'minimal' prompt used to generate the human draft, we utilize an 'extended' prompt adapted to our guidelines.[28] Similarly, the input to the model consists of the raw text of the section with mentions highlighted and numbered inline. Now, instead of only returning, for each mention, a singular human-readable name as ID for the referenced entity, the extended prompt instructs the LLM to populate a structured JSON object for each mention, determining not only the referenced entity ID (or multiple IDs if the mention has multiple references) but also the gender, and applicable special cases.

LLM-Based Cross-Section Entity Merging. Like in the manual pipeline, the local entities need to be merged across sections. Again, similar to the manual entity registry, in our second LLM step, the LLM is given a list of all local entities, with each entity given a set of their mentions' surface words, to provide context. Our merge prompt instructs the

28. See https://github.com/aehrm/llm_literary_coref/prompts/mention_prompts.py, line 17.

conference version

Model	MUC			B^3			$CEAF_e$			CoNLL	LEA		
	P	R	F1	P	R	F1	P	R	F1	F1	P	R	F1
Full Evaluation													
Gemini 2.5 Flash Lite	93.6	85.1	89.1	76.3	57.7	65.4	40.3	63.2	48.7	67.7	70.4	54.0	60.8
Gemini 3 Flash	94.9	92.9	93.9	87.8	83.5	85.6	63.7	67.8	65.4	81.6	85.8	80.4	82.9
w/o Generics													
Gemini 2.5 Flash Lite	93.6	85.8	89.5	75.0	56.8	64.2	37.6	65.2	47.1	67.0	72.6	54.2	61.7
Gemini 3 Flash	94.9	93.9	94.4	87.7	84.9	86.3	61.5	64.9	62.7	81.1	86.7	82.6	84.6
w/o Singletons													
Gemini 2.5 Flash Lite	93.6	85.8	89.5	74.7	56.4	63.8	39.4	61.7	47.7	67.0	73.6	54.0	61.9
Gemini 3 Flash	94.9	93.9	94.4	87.9	85.0	86.4	65.0	69.3	66.8	82.5	87.2	83.0	85.0
Transformed Multi-Ref. Mentions													
Gemini 2.5 Flash Lite	91.5	86.2	88.7	75.4	60.8	67.0	42.6	62.6	50.4	68.7	69.1	57.4	62.5
Gemini 3 Flash	93.8	94.3	94.0	87.0	87.0	87.0	66.8	68.8	67.6	82.9	85.0	84.7	84.8

Table 12: LLM clustering metric performance on GerFuN across different evaluation scenarios. Scores are arithmetic averages in percentage points over all five novels (macro average).

model to assign to each local entity a single, canonical global name to them.[29]

Following the manual pipeline, we evaluate two models from the Google Gemini family: *Gemini 2.5 Flash Lite* and the recent *Gemini 3 Flash* through the OpenRouter platform.[30] Total inference cost was 1.44 USD for *Gemini 2.5 Flash Lite* resp. 9.92 USD for *Gemini 3 Flash*, which translates to 3.13 resp. 21.45 USD per million source tokens.

8.2 Results

Table 12 summarizes the clustering performance of the two evaluated models across different evaluation scenarios. Figure 6 visualizes CoNLL and LEA scores across the five different novels. The performance gap between models confirms that resolving

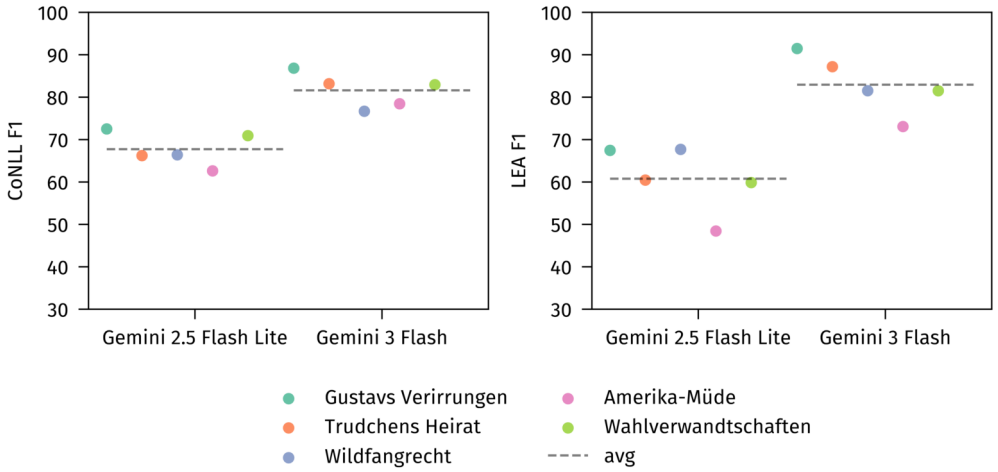


Figure 6: LLM clustering performance on GerFuN per novel (full evaluation scenario).

29. See https://github.com/aehrm/llm_literary_coref/prompts/merge_prompts.py, line 17.
30. We initially intended to include open-source models; however, competitive models available via OpenRouter exhibited non-deterministic behavior, even when controlling for seed, temperature, and provider. Due to resource constraints, we were unable to 1) estimate variance through multiple runs, or 2) host the models locally.

conference version

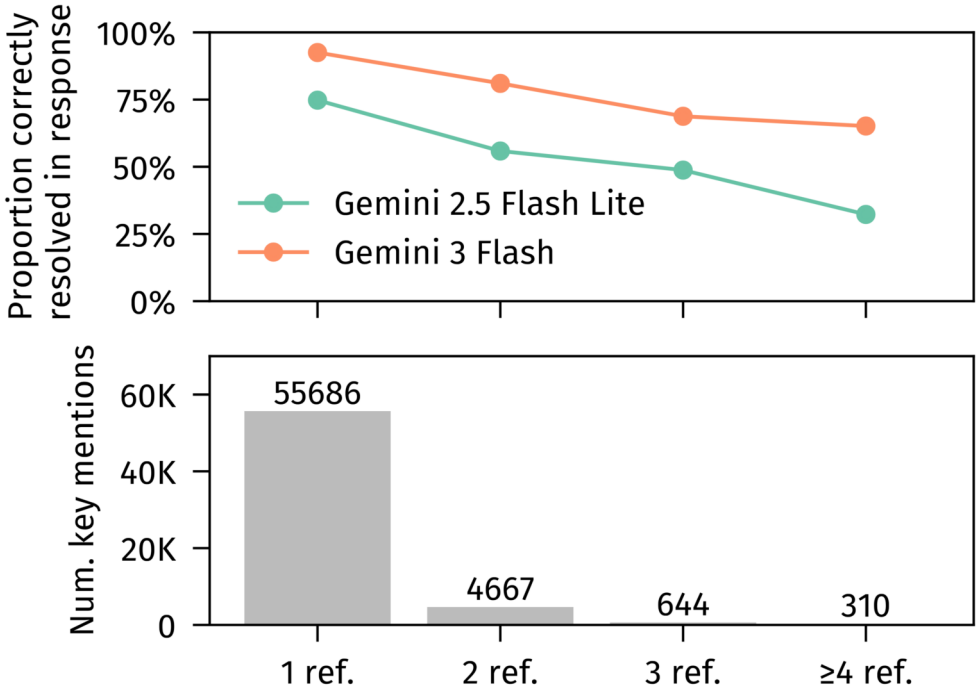


Figure 7: Top: Exact match accuracy of entity assignment for mentions in the response (top) and prevalence in GerFuN (bottom), stratified by the number of referenced entities per mention.

conference version

coreference in long literary texts exceeds the capabilities of smaller, efficient models. In contrast, *Gemini 3 Flash* achieves strong results, with LEA scores exceeding 80% on four of five novels, considerably higher than Transformer-based pipelines on comparable long-novel datasets (subsection 2.3). Measured on section-level, median LEA F1-score for *Gemini 3 Flash* is 83.8% (75.8–88.9% IQR), approaching human agreement levels. *Der Amerika-Müde* presents the lowest scores. This reflects the annotators' experience: more than the other novels, this one requires a wealth of general knowledge even for a basic understanding; the register and level of language pose a greater challenge here, compounded by the novel's extreme length and the large number of minor characters. As indicated by the different evaluation scenarios, both models excel at resolving core entities, with errors concentrated in generic and singleton entities.

GerFuN explicitly models mentions referring to multiple entities. To evaluate this, we analyzed the LLM's accuracy in identifying the exact set of referenced entities (aligned via $CEAF_e$ mapping). As shown in Figure 7, *Gemini 3 Flash* achieves 81% accuracy for mentions with two referents and remains at 65% for those with four or more, demonstrating its capacity to resolve complex character references.

Table 13 details the performance on gender and special case (*group*, *nonfact*) detection.

conference version

	Gender									Special Case					
	m			f			u			group			nonfact		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Full Evaluation															
Gemini 2.5 Flash Lite	72.6	63.3	66.9	80.6	68.0	73.3	80.2	88.6	83.8	66.1	54.7	58.6	20.2	26.5	21.1
Gemini 3 Flash	61.1	89.8	72.6	84.7	89.1	86.7	94.1	81.2	87.1	68.3	80.3	72.8	30.8	69.7	40.4
w/o Groups															
Gemini 2.5 Flash Lite	84.6	73.0	77.6	86.0	84.2	84.8	87.8	94.1	90.7	—	—	—	18.7	33.8	21.1
Gemini 3 Flash	88.5	94.7	91.4	91.5	90.6	91.0	96.7	95.5	96.1	—	—	—	33.2	73.3	42.2
w/o Groups and Singletons															
Gemini 2.5 Flash Lite	97.0	87.7	91.8	97.9	97.1	97.4	11.4	5.0	6.5	—	—	—	24.4	55.0	31.4
Gemini 3 Flash	95.9	97.6	96.6	94.0	97.9	95.8	26.7	11.5	15.8	—	—	—	45.3	83.8	54.5

Table 13: LLM prediction performance on entity attributes across different evaluation scenarios. Reported scores are arithmetic averages in percentage points over all five novels (macro average).

	generic		
Model	P	R	F1
Gemini 2.5 Flash Lite	35.1	56.2	42.4
Gemini 3 Flash	60.6	64.0	61.5

Table 14: LLM prediction performance on mention-level generic special case. Scores are arithmetic averages in percentage points over all five novels (macro average).

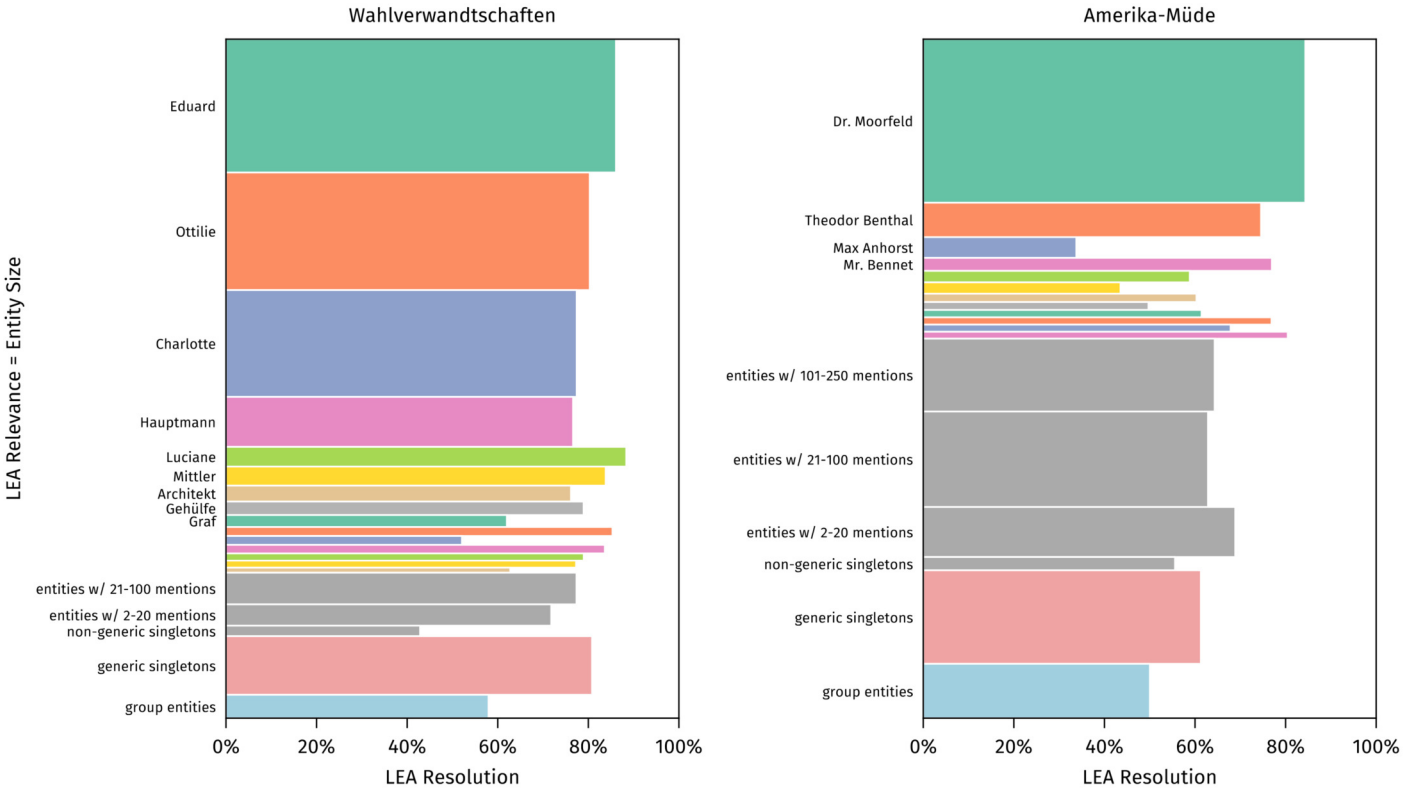


Figure 8: Visualization of the LEA recall of Gemini 3 Flash for the novels *Wahlverwandtschaften* and *Amerika-Müde*. Each horizontal bar represents one key entity, its height is the LEA relevance, and its width is the LEA resolution. For easier readability, some entities have been grouped to a single bar.

The models' ability to infer these entity attributes closely mirrors human performance, particularly for core entities (*w/o Groups and Singletons*). Notably, recall on special cases is considerably higher than precision, indicating a tendency of the LLMs to over-use these attributes.

Table 14 shows the performance on the recognition of generic entities on the mention-level, again showing the difficulty of recognizing these entities. *Gemini 3 Flash* achieves only a F1-score of 61% on the mention-level, behind the IAA of 76%. While there may be potential avenues for improvement by further prompt engineering, these results – together with the relatively low IAA – might hint at the inherent complexity of the phenomenon.

In fact, since LEA is an entity-based metric, we can visualize the error contributions of the different entities resp. entity classes. Figure 8 visualizes the LEA recall, per entity, of *Gemini 3 Flash*'s response for *Wahlverwandtschaften* and *Amerika-Müde*. Under this

conference version

visualization, LEA recall is the proportion of the shaded area on the enclosing rectangle. Thus, the primary error source are entities with few mentions, generic entities, and particularly group entities. However, for main characters, recall is relatively high. All results taken together, this confirms the pipeline's applicability for downstream CLS tasks that analyze the central characters and their attributes in full-length novels.

9. Conclusion

In this paper, we introduced GerFuN, a dataset comprising five full-length German novels fully annotated for coreference resolution. Moving from text excerpts to full-length documents, we employed a semi-manual pipeline using LLM-based pre-annotation. This workflow significantly reduced manual effort while ensuring high data quality through a 'human in the loop,' offering a scalable blueprint for resource creation in CLS.

Our accompanying guidelines introduce a conceptual distinction between prototypical characters and deviating special cases. While inter-annotator agreement was high for prototypical entities we are primarily interested in, the lower agreement on categories like generic entities, group entities, or figurative mentions suggests that future research targeting these phenomena requires more explicit operationalization than intuitively assumed.

Furthermore, we demonstrated that state-of-the-art LLMs like *Gemini 3 Flash* can resolve prototypical coreference with near-human accuracy at a reasonable cost. While our pipeline serves as a successful proof-of-concept for automated full-text annotation, it warrants further investigation, such as evaluating open-source LLMs, the impact on context-length resp. section splitting, a quantitative comparison to other (conventional) coreference resolution pipelines, qualitative error analysis and explore integrating conventional models to further optimize cost efficiency.

Finally, the analysis of the corpus also highlights that current coreference resolution metrics require specific adaptation to the needs of CLS. Technical metrics remain difficult to interpret in a CLS context, where it is often unclear which values are required for sufficient inference quality to make valid interpretation. In future work, we plan to take the quantitative insights from the GerFuN dataset to develop evaluation frameworks that account for the characteristic entity distributions of long novels and are aligned with the specific research interests of CLS.

10. Data Availability

Data, including the full GerFuN dataset, LLM inferences, and annotation guidelines, can be found here: https://github.com/aehrm/llm_literary_coref/.

11. Software Availability

Software can be found here: https://github.com/aehrm/llm_literary_coref/.

conference version

12. Acknowledgements

We warmly thank our student assistant Oana Heckl for her dedicated work as an annotator. We are also grateful to Fotis Jannidis and his Chair of Computational Philology at the University of Würzburg for funding this manual annotation work and for their insightful feedback and support.

13. Author Contributions

Agnes Hilger: Conceptualization, Data curation, Investigation, Resources, Writing – original draft

Anton Ehrmanntraut: Conceptualization, Formal analysis, Investigation, Software, Writing – original draft

References

Austen, Jane [1813] (1996). *Pride and Prejudice*. Ed. by Vivien Jones. Penguin Classics. Penguin Books.

Bagga, Amit and Breck Baldwin (1998). "Entity-Based Cross-Document Coreferencing Using the Vector Space Model". In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*. ACL 1998 (Montreal, Quebec, Canada). Association for Computational Linguistics, 79–85. 10.3115/980845.980859.

Bamman, David, Olivia Lewke, and Anya Mansoor (2020). "An Annotated Dataset of Coreference in English Literature". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020 (Marseille, France). Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association, 44–54. https://aclanthology.org/2020.lrec-1.6/ (visited on 06/25/2025).

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)*. NAACL 2019 (Minneapolis, Minnesota, USA). Association for Computational Linguistics. 10.18653/v1/n19-1220. http://aclweb.org/anthology/N19-1220.

Bohnet, Bernd, Chris Alberti, and Michael Collins (2023). "Coreference Resolution through a seq2seq Transition-Based System". In: *Transactions of the Association for Computational Linguistics* 11, 212–226. 10.1162/tacl_a_00543.

Bourgois, Antoine and Thierry Poibeau (2025). "The Elephant in the Coreference Room: Resolving Coreference in Full-Length French Fiction Works". In: *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*. CRAC 2025 (Suzhou, China). Ed. by Maciej Ogrodniczuk, Michal Novak, Massimo Poesio, Sameer Pradhan, and Vincent Ng. Association for Computational Linguistics, 55–69. 10.18653/v1/2025.crac-1.5.

conference version

Brunner, Annelen, Lukas Weimer, Stefan Engelberg, Fotis Jannidis, and Ngoc Duyen Tanja Tu (2020). *Annotationsrichtlinien des Projekts 'Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse'*. Version 1.2. Zenodo. 10.5281/ZENODO.3759617.

Duron-Tejedor, Ana-Isabel, Pascal Amsili, and Thierry Poibeau (2023). "How to Evaluate Coreference in Literary Texts?" In: *arXiv preprint*. 10.48550/arXiv.2401.00238.

Eder, Jens (2008). *Was sind Figuren? Ein Beitrag zur interdisziplinären Fiktionstheorie*. Mentis. 10.25969/MEDIAREP/18993.

Eder, Jens, Fotis Jannidis, and Ralf Schneider (2010). "Characters in Fictional Worlds. An Introduction". In: *Characters in Fictional Worlds. Understanding Imaginary Beings in Literature, Film, and Other Media*. Ed. by Fotis Jannidis, Jens Eder, and Ralf Schneider. De Gruyter, 3–64. 10.1515/9783110232424.1.3.

Fischer, Karoline Auguste [1801] (2012). *Gustavs Verirrungen*. TextGrid Digitale Bibliothek. https://hdl.handle.net/11858/00-1734-0000-0002-A7D8-4 (visited on 01/02/2026).

Gan, Yujian, Massimo Poesio, and Juntao Yu (2024). "Assessing the Capabilities of Large Language Models in Coreference: An Evaluation". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. LREC-COLING 2024 (Turin, Italy). Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. ELRA and ICCL, 1645–1665. https://aclanthology.org/2024.lrec-main.145 (visited on 11/14/2025).

Goethe, Johann Wolfgang von [1809] (1868). "Elective Affinities". In: *Novels and tales by Goethe*. Trans. by R. Dillon Boylan. Bell & Daldy, 1–246. http://archive.org/details/novelstalesbygoe00goetrich (visited on 12/11/2025).

— [1774] (2006). "Die Wahlverwandtschaften". In: *Die Leiden des jungen Werthers. Die Wahlverwandtschaften. Kleine Prosa. Epen*. Ed. by Waltraud Wiethölter. DKV-Taschenbuch 11. Deutscher Klassiker-Verlag, 269–530.

Gupta, Talika, Hans Ole Hatzel, and Chris Biemann (2024). "Coreference in Long Documents using Hierarchical Entity Merging". In: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. LaTeCH-CLfL 2024 (St. Julians, Malta). Ed. by Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva, and Stan Szpakowicz. Association for Computational Linguistics, 11–17. https://aclanthology.org/2024.latechclfl-1.2/ (visited on 12/31/2025).

Han, Sooyoun, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi (2021). "FantasyCoref: Coreference Resolution on Fantasy Literature Through Omniscient Writer's Point of View". In: *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*. CRAC 2021 (Punta Cana, Dominican Republic). Association for Computational Linguistics, 24–35. 10.18653/v1/2021.crac-1.3.

Heimburg, Wilhelmine [1894] (2008). *Trudchens Heirat*. Projekt Gutenberg. https://www.projekt-gutenberg.org/heimburg/trudchen/chap001.html (visited on 03/03/2022).

Hicke, Rebecca and David Mimno (2024). "[Lions: 1] and [Tigers: 2] and [Bears: 3], Oh My! Literary Coreference Annotation with LLMs". In: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. LaTeCH-CLfL 2024 (St. Julians, Malta). Ed. by Yuri Bizzoni,

conference version

Stefania Degaetano-Ortlieb, Anna Kazantseva, and Stan Szpakowicz. Association for Computational Linguistics, 270–277. https://aclanthology.org/2024.latechclfl-1.27/ (visited on 11/14/2025).

Hillebrandt, Claudia (2017). "Figur". In: *Grundthemen der Literaturwissenschaft: Erzählen*. De Gruyter, 161–173. 10.1515/9783110410747-008.

Jannidis, Fotis (2004). *Figur und Person. Beitrag zu einer historischen Narratologie*. Narratologia 3. De Gruyter. 10.1515/9783110201697.

— (2009). "Character". In: *Handbook of Narratology*. Ed. by Peter Hühn, Jan-Christoph Meister, John Pier, and Wolf Schmid. De Gruyter, 14–29. 10.1515/9783110217445.14.

Jannidis, Fotis and Gerhard Lauer (2002). "'Bei meinem alten Baruch ist der Pferdefuß rausgekommen' – Antisemitismus und Figurenzeichnung in *Der Stechlin*". In: *Fontane und die Fremde, Fontane und Europa*. Ed. by Konrad Ehlich. Königshausen & Neumann, 103–119.

Joshi, Mandar, Omer Levy, Luke Zettlemoyer, and Daniel Weld (2019). "BERT for Coreference Resolution: Baselines and Analysis". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. EMNLP-IJCNLP 2019 (Hong Kong, China). Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Association for Computational Linguistics, 5803–5808. 10.18653/v1/D19-1588.

Klinger, Friedrich Maximilian von (1810). *Der Faust der Morgenländer oder Wanderungen Ben Hafis, Erzählers der Reisen von Sündfluth*. https://www.digitale-sammlungen.de/view/bsb11341825 (visited on 01/02/2026).

Köppe, Tilmann and Tom Kindt (2014). *Erzähltheorie: eine Einführung*. Reclams Universal-Bibliothek Reclam-Sachbuch Nr. 17683. Reclam.

Kröncke, Merten, Fotis Jannidis, Leonard Konle, and Simone Winko (2022). *Annotationsrichtlinien Emotionsmarker und Emotionen*. Version 1.1. Zenodo. 10.5281/ZENODO.6021152.

Krug, Markus, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, and Fotis Jannidis (2018). *Description of a Corpus of Character References in German Novels – DROC [Deutsches ROman Corpus]*. DARIAH-DE Working Papers 27. Georg-August-Universität Göttingen. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2018-2-9 (visited on 11/15/2025).

Kürnberger, Ferdinand (1856). *Der Amerika-Müde*. Vol. 8. Deutsche Bibliothek: Sammlung auserlesener Original-Romane. Verlag von Meidinger Sohn & Cie. https://www.deutschestextarchiv.de/book/show/kuernberger_amerikamuede_1855 (visited on 01/02/2026).

Le, Nghia T. and Alan Ritter (2024). "Are Language Models Robust Coreference Resolvers?" In: *First Conference on Language Modeling*. COLM 2024 (Philadelphia, Pennsylvania, USA). https://openreview.net/forum?id=MmBQSNHKUl (visited on 11/14/2025).

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (2017). "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2017 (Copenhagen, Denmark). Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Association for Computational Linguistics, 188–197. 10.18653/v1/D17-1018.

Lee, Kenton, Luheng He, and Luke Zettlemoyer (2018). "Higher-Order Coreference Resolution with Coarse-to-Fine Inference". In: *Proceedings of the 2018 Conference of the*

conference version

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). NAACL 2018 (New Orleans, Louisiana, USA). Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. Association for Computational Linguistics, 687–692. 10.18653/v1/N18-2108.

Luo, Xiaoqiang (2005). "On Coreference Resolution Performance Metrics". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. HLT-EMNLP 2005 (Vancouver, British Columbia, Canada). Ed. by Raymond Mooney, Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff. Association for Computational Linguistics, 25–32. https://aclanthology.org/H05-1004/ (visited on 09/18/2025).

Margolin, Uri (1987). "Introducing and Sustaining Characters in Literary Narrative: A Set of Conditions". In: *Style* 21.1, 107–124. http://www.jstor.org/stable/42945634 (visited on 01/02/2026).

— (1995). "Characters in Literary Narrative. Representation and Signification". In: *Semiotica* 106.3–4, 373–392.

Martinelli, Giuliano, Edoardo Barba, and Roberto Navigli (2024). "Maverick: Efficient and Accurate Coreference Resolution Defying Recent Trends". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2024 (Bangkok, Thailand). Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Association for Computational Linguistics, 13380–13394. 10.18653/v1/2024.acl-long.722.

Martinelli, Giuliano, Tommaso Bonomo, Pere-Lluís Huguet Cabot, and Roberto Navigli (2025). "BOOKCOREF: Coreference Resolution at Book Scale". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2025 (Vienna, Austria). Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Association for Computational Linguistics, 24526–24544. 10.18653/v1/2025.acl-long.1197.

Martínez, Matías and Michael Scheffel (2012). *Einführung in die Erzähltheorie*. 9th ed. C.H. Beck.

Mélanie-Becquet, Frédérique, Jean Barré, Olga Seminck, Clément Plancq, Marco Naguib, Martial Pastor, and Thierry Poibeau (2024). "BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature". In: *Journal of Computational Literary Studies* 3.1. 10.48694/jcls.3924.

Moosavi, Nafise Sadat and Michael Strube (2016). "Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2016 (Berlin, Germany). Association for Computational Linguistics, 632–642. 10.18653/v1/P16-1060.

Novák, Michal, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman (2025). "Findings of the Fourth Shared Task on Multilingual Coreference Resolution: Can LLMs Dethrone Traditional Approaches?" In: *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*. CRAC 2025 (Suzhou, China). Ed. by Maciej Ogrodniczuk, Michal Novak, Massimo Poesio, Sameer Pradhan, and Vincent Ng. Association for Computational Linguistics, 95–118. https://aclanthology.org/2025.crac-1.9/ (visited on 11/14/2025).

conference version

Pagel, Janis (2024). "Enhancing Character Type Detection using Coreference Information: Experiments on Dramatic Texts". PhD thesis. http://nbn-resolving.de/urn:nbn:de:bsz:93-opus-ds-150190 (visited on 12/20/2025).

Pagel, Janis and Nils Reiter (2020). "GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020 (Marseille, France). Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association, 55–64. https://aclanthology.org/2020.lrec-1.7/ (visited on 01/02/2026).

Petersen-Frey, Fynn, Hans Ole Hatzel, and Chris Biemann (2025). "Efficient and Effective Coreference Resolution for German". In: *Proceedings of the 21st Conference on Natural Language Processing: Long and Short Papers*. KONVENS 2025 (Hannover, Germany). Ed. by Christian Wartena and Ulrich Heid. HsH Applied Academics, 128–137. https://aclanthology.org/2025.konvens-1.12/ (visited on 09/16/2025).

Reiter, Nils (2020). "Anleitung zur Erstellung von Annotationsrichtlinien". In: *Reflektierte algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. De Gruyter, 193–202. 10.1515/9783110693973-009.

Reiter, Nils and Anette Frank (2010). "Identifying Generic Noun Phrases". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL 2010 (Uppsala, Sweden). Ed. by Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre. Association for Computational Linguistics, 40–49. https://aclanthology.org/P10-1005/ (visited on 07/16/2025).

Roesiger, Ina, Sarah Schulz, and Nils Reiter (2018). "Towards Coreference for Literary Text: Analyzing Domain-Specific Phenomena". In: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. LaTeCH-CLfL 2018 (Santa Fe, New Mexico). Ed. by Beatrice Alex, Stefania Degaetano-Ortlieb, Anna Feldman, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz. Association for Computational Linguistics, 129–138. https://aclanthology.org/W18-4515/ (visited on 01/02/2026).

Schröder, Fynn, Hans Ole Hatzel, and Chris Biemann (2021). "Neural End-to-end Coreference Resolution for German in Different Domains". In: *Proceedings of the 17th Conference on Natural Language Processing*. KONVENS 2021 (Düsseldorf, Germany). Ed. by Kilian Evang, Laura Kallmeyer, Rainer Osswald, Jakub Waszczuk, and Torsten Zesch, 170–181. https://aclanthology.org/2021.konvens-1.15 (visited on 04/15/2024).

Toshniwal, Shubham, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel (2020). "Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2020 (Online). Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics, 8519–8526. 10.18653/v1/2020.emnlp-main.685.

Vadász, Noémi (2023). "Resolving Hungarian Anaphora with ChatGPT". In: *Text, Speech, and Dialogue: 26th International Conference, Proceedings*. TSD 2023 (Pilsen, Czech Republic). Ed. by Kamil Ekštein, František Pártl, and Miloslav Konopík. Springer Nature, 45–57. 10.1007/978-3-031-40498-6_5.

conference version

Vauth, Michael and Evelyn Gius (2021). *Richtlinien für die Annotation narratologischer Ereigniskonzepte*. Zenodo. 10.5281/zenodo.5078175.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman (1995). "A Model-Theoretic Coreference Scoring Scheme". In: *Sixth Message Understanding Conference*. MUC-6 (Columbia, Maryland, USA). Morgan Kaufmann, 45–55. https://aclanthology.org/M95-1005/ (visited on 12/30/2025).

Wu, Wei, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li (2020). "CorefQA: Coreference Resolution as Query-based Span Prediction". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020 (Online). Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Association for Computational Linguistics, 6953–6963. 10.18653/v1/2020.acl-main.622.

Wunderle, Julia, Anton Ehrmanntraut, Jan Pfister, Fotis Jannidis, and Andreas Hotho (2025). "New Encoders for German Trained from Scratch: Comparing ModernGBERT with Converted LLM2Vec Models". In: *arXiv preprint*. 10.48550/arXiv.2505.13136.

Xu, Liyan and Jinho D. Choi (2020). "Revealing the Myth of Higher-Order Inference in Coreference Resolution". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2020 (Online). Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics, 8527–8533. 10.18653/v1/2020.emnlp-main.686.

Zhang, Wenzheng, Sam Wiseman, and Karl Stratos (2023). "Seq2seq is All You Need for Coreference Resolution". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2023 (Singapore). Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Association for Computational Linguistics, 11493–11504. 10.18653/v1/2023.emnlp-main.704.

Zhou, Ethan and Jinho D. Choi (2018). "They Exist! Introducing Plural Mentions to Coreference Resolution and Entity Linking". In: *Proceedings of the 27th International Conference on Computational Linguistics*. COLING 2018 (Santa Fe, New Mexico, USA). Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Association for Computational Linguistics, 24–34. https://aclanthology.org/C18-1003/ (visited on 11/14/2025).

conference version

conference version

A. Original quotes

(1) Cäsar, der den Wasserkrug sah, trabte instinctmäßig mit mir, und seinen Nüstern mehr als dem todesmatten Gemurmel der Frau verdankt' ich das directe Auffinden der Wasserquelle. (Kürnberger 1856, 294)

(2) Nach einer schlechten, nächtlichen Jagd stieß der Löwe bei anbrechendem Tage mit seinem jungen Sohne auf einen wohlgenährten Fuchs. Schnell sah der Fuchs, für ihn sei keine Rettung mehr. Er seufzte in seinem Herzen: „Armer, deine letzte Stunde ist nun gekommen, wenn dir dein Verstand nicht aus der Gefahr hilft.[“] (Klinger 1810, 78)

(3) Trudchen Baumhagen war rasch über den stillen Kirchplatz geschritten, hatte in der gegenüberliegenden Mauer eine Pforte geöffnet und stand nun auf väterlichem Boden. Ziemlich eilig ging sie durch die mit Buchs eingefaßten Wege des im altfranzösischen Stile angelegten Gartens und über einen stillen geräumigen Hof in das Haus. Auf dem großen gewölbten Flur traf sie ihren Schwager neben einem Fahrrad stehend. Er war sehr elegant und nach neuester Mode gekleidet[...] Ein Diener war beschäftigt, den glänzenden Stahl des Vehikels mit einem Lederlappen abzureiben. „Nun“, fragte das junge Mädchen freundlich, „willst du ausreiten, Artur?“„Ausreißen, meinst du, Trudchen? Ja, ja, was soll man anfangen!“gab er verdrießlich zur Antwort. „Jenny hat ja heute ausnahmsweise wieder einmal einen Damentee arrangiert – da bin ich überflüssig. Ich fahre mit Karl Röben nach Bodenstedt[.]“(Heimburg [1894] 2008)

(4) Eduard und der Hauptmann fehlten, Charlotte hatte seit langer Zeit zum erstenmal den Tisch selbst angeordnet, und es wollte Ottilien scheinen, als wenn sie abgesetzt wäre. Die beiden Frauen saßen gegen einander über; (Goethe [1774] 2006, 378)

(5) Es hatte sich diesen Tag viel Gesellschaft eingefunden. Man ging zur Kirche, wo man die Gemeinde im festlichen Schmuck versammelt antraf. (Goethe [1774] 2006, 330)

(7) „Ich will dir etwas sagen, Franz, Scherz in die Ecke“, fuhr er fort, „du wirst heiraten müssen! Und da gebe ich dir den Rat, tue bei dieser Angelegenheit deinen Idealen ein wenig Zwang an. Sieh ab von elfengleichem Wuchs, sinnigen Augen und holdester Weiblichkeit – zugunsten eines anderen Vorzuges, der durch nichts zu ersetzen ist in unserem prosaischen Leben. Bringe mir kein armes Mädchen, Franz, und wäre sie die Perle aller Weltteile.[“] (Heimburg [1894] 2008)

(10) Einst da ich mich im Schauspielhause meinen gewöhnlichen Träumereyen überließ; ward meine Aufmerksamkeit durch zwey weibliche Stimmen angezogen, welche aus der benachbarten Loge zu kommen schienen. (Fischer [1801] 2012)

jcls. Journal of Computational Literary Studies

conference version

Character Development in Oral Testimonies

Computational Modeling of Religiosity in Holocaust Narratives

OPEN ACCESS

Esther Shizgal[1]
Eitan Wagner[1]
Omri Abend[1]
Renana Keydar[2]

1. Computer Science, The Hebrew University of Jerusalem, Jerusalem, Israel.
2. Law and Digital Humanities, The Hebrew University of Jerusalem, Jerusalem, Israel.

Citation
Esther Shizgal, Eitan Wagner, Omri Abend, and Renana Keydar (2026). "Character Development in Oral Testimonies. Computational Modeling of Religiosity in Holocaust Narratives". In: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7997

Date published 2026-05-05 (preprint)
Date accepted tbc
Date received 2026-01-08

Keywords
character development, testimonial narratives, holocaust, oral testimonies, religious trajectories

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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. Character is a core feature of oral testimonies, inviting readers into the protagonist's process of identity formation. This work explores character trajectories in 1,000 Holocaust testimonies from a religious development perspective, focusing on religious practice and belief. We extract religious trajectories by fine-tuning large language models and then perform a quantitative analysis across the narratives. The trajectories are clustered to identify common structures of religious evolution, yielding a taxonomy of five structural types. The analysis is validated automatically, manually, and through close reading of selected examples. We find that the most common structure for religious practice involves oscillations, whereas a constant-positive pattern is most prevalent for religious belief. Finally, we demonstrate how the proposed method can be used to select similar testimonies. While some results support existing theories, others offer new insights into the complexity of survivors' lived religious experiences, highlighting the potential of computational methods for testimony analysis.

1. Introduction

Oral testimonies are unique in the study of narratives. Shaped by vision and memory and guided by interviewer prompts, their synthesis of personal experiences within a relatively confined domain of topics and locations distinguishes them from many fictional narratives (Eaglestone 2002). These first-person narratives offer a window into the witness's identity formation and development, where the portrayal of the self serves as a central element. Thus, Holocaust testimonies, which trace the deformation and shattering of individual life patterns, constitute a rich narrative space for analyzing character development (Assmann 2006; Felman and Laub 1992; Howarth 1974; McAdams 2001; Sultana et al. 2022).

Reconstructing a life story through testimonial narration is one way that narrative serves as a bridge for understanding human behavior and belief. By transforming fragmented memories into a self-portrait, both the witnesses and audiences can interpret their world, derive meaning, and motivate their actions and those of others (Bruner 1991; Piper et al. 2021). As the narrative unfolds, characters evolve, manifesting changes in personality, motivation, and relationships (Chaturvedi et al. 2017). Such evolution

conference version

drives the development of narratives, conveying themes, and engages readers with novel forms of social experience. As a result, character analysis has been widely studied across disciplines (Piper et al. 2017; Yang and Pianzola 2025).

Religious consciousness provides an example of an evolving character dimension. For religious individuals, ritual observance and belief in God are not merely matters of doctrinal change but constitute a core layer of identity, which has been shown to evolve over the life course (Fowler 1981; Fowler and Keen 1978; Ingersoll-Dayton et al. 2002; Meddin 1998; Melia 1999). Theoretical perspectives on religious development suggest that this evolution is often shaped by an individual's social background and personality characteristics (Reich 1992). These shifting patterns of faith are often described as religious trajectories (Ingersoll-Dayton et al. 2002).

The influence of trauma on an individual's religiosity is a widely debated topic within the psychology and sociology of religion. One perspective emphasizes the role of faith as a source of resilience in the face of crisis (Berkovits 1973; Dein 2022; Leo et al. 2021). In contrast, other scholars contend that trauma is often followed by common processes of religious transformation manifesting as either a loss of faith or an intensification of it (Ben-Ezra et al. 2010). Montell (2001) illustrates how adversity can influence religious belief, describing how belief continues even as the catastrophe fades from immediate consciousness: "The most salient examples are in the strengthening of religious beliefs occurring in the aftermath of the Holocaust and of subsequent mass murders."

The question of religious trajectories among Holocaust survivors is commonly studied across disciplines. The extreme experiences described by survivors likely shape their belief systems, while the wartime conditions imposed constraints on maintaining religious practice. Indeed, the subject of religious life during the Holocaust is distinct because the Nazis directly attacked Orthodox Judaism. The Germans often timed Aktions and transportations to coincide with Jewish holidays to degrade observant individuals (Michman 1993). This degradation included physical assaults, such as German forces violently tearing off beards. Salomon Gutter, a survivor from Chrzanów, Poland, vividly describes one such act on the holy day of Yom Kippur in 1939, a couple of weeks after the German invasion into Poland: "They came into our house, and they found my great-grandfather in bed... They cut off his beard in the bed. It was a great tragedy... And they made them dig potatoes a whole day. And they left from them—it's your holiday today" (Gutter 1996).[1]

Following the unparalleled trauma, many scholars of Jewish history and theologians have fundamentally confronted the question of post-Holocaust theology (Dein 2022). However, less attention has been given to the lived religious experiences of the survivors themselves, a gap that oral testimonies uniquely address. Our work builds on this by examining how these religious shifts are articulated as a continuous trajectory across the testimony narrative timeline.

Recent advances in NLP and large language models have opened avenues for computational narratology, allowing for sophisticated representations of character that support systematic comparison across narratives, while remaining sensitive to character complexity and narrative evolution (Algee-Hewitt 2017; Papoudakis et al. 2024; Piper et al.

1. Salomon Gutter's full interview transcript is included in the supplementary material.

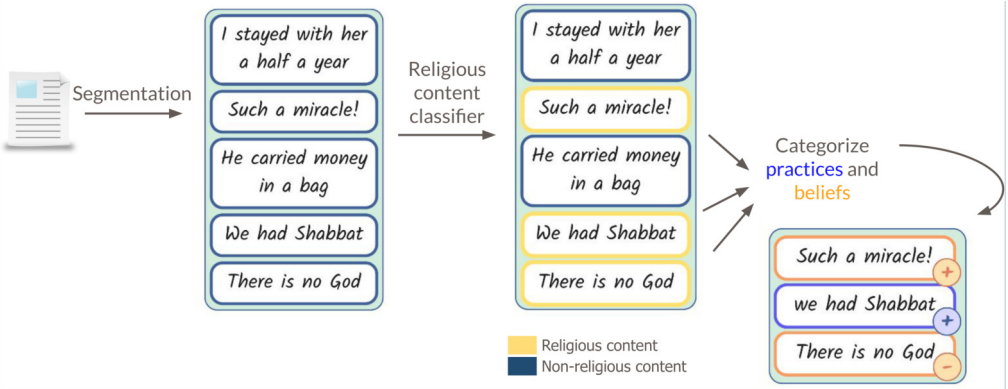


Figure 1: Fabricated example of extracting a single trajectory.

2017). Such computational frameworks provide tools to analyze dimensions of identity development at scale.

This work examines how religious character evolution is woven into the narratives of Holocaust survivor testimonies. Using a computational approach, we systematically analyze a large collection of testimonies to identify the predominant structures of religious trajectories as narrated by the survivors.

Given a testimony transcript, the process of extracting the trajectory follows: (1) segmenting the text into question and answer pairs, (2) filtering the segments that reference religion, and (3) classifying those segments with respect to their valence of practice and belief (Figure 1). This is implemented by fine-tuning and prompting large language models. To provide a consistent representation across all testimonies, the trajectories are represented as a series of annotated segments across time, allowing comparison of each.[2]

Our contributions are twofold. First, we propose a computational method for analyzing thematic character trajectories in oral testimonies. Second, this approach enables scholars to identify comparable religious journeys across testimonies, supporting close reading, thematic grouping, and case-based historical analysis, while revealing patterns that validate the method as historically and sociologically sensitive and open avenues for further, more nuanced interpretation.

2. Character Development

Computational character analysis has been applied across diverse narrative dimensions that enable the systematic modeling and comparison of characters across cultures. Prior work has examined character attributes such as personas, archetypes, and relationships in narrative text (Chaturvedi et al. 2017; Radak et al. 2024; Rafaeli et al. 2025; Wilkens et al. 2024; Yang and Pianzola 2025). Within NLP, scholars have highlighted the challenges of modeling culturally salient attributes like gender, ethnicity, and religion, due to the biases embedded in narrative data and models (Piper et al. 2021; Plaza-del-Arco et al. 2024). Tasks such as modeling protagonists' emotional trajectories (Brahman and Chaturvedi

2. We note that the technical and algorithmic part of the paper was adopted with minimal changes from Shizgal et al. (2025).

conference version

2020) and generating character descriptions (Brahman et al. 2021; Papoudakis et al. 2024) further emphasize the need for better models of narrative comprehension.

Computational methods offer particular advantages for analyzing the entire collection of testimonies, offering a data-driven perspective that mitigates the selection bias inherent in manual qualitative research. This approach allows for a scale of analysis far beyond human capacity (Mohr et al. 2020; Nelson 2020). As a result, computational approaches have increasingly become an essential component of the Holocaust research toolkit.

The USC Shoah Foundation (SF) archive itself consists of over 120,000 hours of testimony, illustrating this necessity. It is thus beyond the capacity of any research team to manually review the entire collection. Presner et al. (2024) advocate for the urgent need to embrace these technologies to derive insights from the vast archival material, noting that it would take a single person approximately thirty years to view the entire collection. At the same time, scholars have stressed the unique historical value of first-hand testimony: Young (1997) argues that no document is more historically authentic than one that captures victims' contemporaneous understanding of events, and Langer (1991) emphasizes that they reveal layers of knowledge that are inaccessible from any other source.

Accordingly, computational models have opened interdisciplinary pathways for tracing patterns of belief, resilience, identity, and trauma in Holocaust testimonies. Prior studies demonstrate the utility of these methods through use cases including topical segmentation, location mapping, the analyses of gender differences among Auschwitz prisoners, and sentiment analysis of Holocaust memories (Blanke et al. 2019; Ifergan et al. 2024; Tóth et al. 2022; Wagner et al. 2025, 2022).

Trajectory analysis is a well-established methodological framework utilized across the social and natural sciences to model longitudinal changes and life-course developments over time (Gabadinho et al. 2011). In computational domains, Alqahtani et al. (2021) present a thorough review of time-series data analysis, focusing on deep time-series clustering (DTSC), proposing a clustering approach using deep convolutional auto-encoders (DCAE). Furthermore, Chang et al. (2023) discuss methods that include traditional metrics for trajectories, and deep learning approaches that embed trajectories for similarity measurement. However, these methods typically require labeled data for supervised learning.

Within the humanities, and particularly Holocaust studies, such methods have been used on archival material to trace the individual, familial, and community trajectories of both survivors and victims (Chopard 2020; Mariot and Zalc 2017). This approach has recently been expanded to computational methodologies scaling the analysis of mapping events and locations (Wagner et al. 2023, 2025). Moving beyond geographical locations, systematic approaches have been applied to articulate emotional transitions along survivor trajectories. By applying representational models to subjective experiences, these methods focus on exploring the relationships between people, place, events, and emotions (Cole and Giordano 2025).

Religion, as a layer of identity, is constantly evolving as a character undergoes life experiences. Building on prior efforts, this study models how religious development is conveyed in Holocaust testimonies. Ingersoll-Dayton et al. (2002) conducted a retrospective quantitative analysis of more than one hundred interviews with older adults

conference version

identifying as Christian, to study changes in religiosity from a life-course perspective. Their findings highlight multiple social forces that shape religious evolution, some encouraging increased religiosity, while others are associated with decreases in religious intensity. Adverse life experiences emerged in both categories and were often described by the participants as turning points in their religious faith and commitment.

To quantify these patterns, methods from modern statistics have been employed to examine religious development. Brennan and Mroczek (2003) applied growth curve models and McCullough et al. (2005) demonstrated how growth mixture models overcome the limitations of the former approach, applied to a dataset of 1,151 American adults, between ages 24 and 80, across multiple religious affiliations.

In the context of the religious evolution of Holocaust survivors, extensive research has been conducted through the manual selection of testimonies from various religious affiliations and backgrounds. For instance, Lassley (2015) concentrates on a selection of testimonies focusing on the loss of faith. Similarly, Giuffra Darbyshire (2019) studied religious rituals and belief in Auschwitz, reviewing survivor testimonies and memoirs, attempting to gain first-hand insight into survivors' Jewish activity and their faith. The author cites Michman (1993), arguing for the need to systematically explore how individuals and small groups dealt with the dissonance between their experiences and the tenets of their religious traditions. Indeed, the frequent discussion of God and prayer has been observed to be a characteristic feature of survivor memoirs (Patterson 1998).

Brenner (1980) conducted the first systematic research on faith among individual Shoah survivors, interviewing 708 survivors living in Israel. The essence of this survey was the question of how the Holocaust affected their beliefs. While thoroughly comprehensive, this work differs from the present study in key aspects. First, the survivors surveyed were questioned specifically about religion, while our study analyzes testimonies covering many topics, and we extract religious descriptions whether the topic was prompted or arose spontaneously. In addition, Brenner interviewed survivors living in Israel, most of whom originated from Eastern Europe and were highly observant before the war. In contrast, our selection consists of testimonies filmed in English-speaking countries, of survivors from diverse backgrounds.

We build on these foundations by extending computational character modeling to first-person Holocaust survivor narratives. Here the protagonist's character development in Holocaust testimonies is examined through a focus on described religious practices and expressed religious belief systems as evolving aspects of character. To the best of our knowledge, this is the first large-scale, systematic analysis of accounts of religious practices and beliefs embedded within the survivors' own narratives, leveraging the power of computational methods to explore a vast corpus.

3. Data

This work analyzes a corpus of 1,000 Holocaust survivor testimony transcripts, documented and archived by the SF.[3]

3. The collection of testimonies was received with permission to use from the SF, and we do not have further details on the selection process.

conference version

conference version

The testimonies were recorded on video in English between 1996 and 2015, each typically lasting several hours. Consequently, the narratives provide a retrospective lens on the survivors' memories, reflecting on experiences that occurred decades prior. The survivors represented in the dataset come from 31 countries, with a significant representation from Poland (34%) and Germany (19%), and the dataset maintains a balanced gender distribution of 531 males (53%) and 469 females (47%). The volunteers documenting the testimonies conducted them as interviews, following strict guidelines regarding their structure and themes. [4]

As an overview exploration of pre- and post-war religious backgrounds, we utilized Gemini-3[5] to extract and categorize the primary identities described by the survivors in the dataset. The analysis of pre-war backgrounds revealed a diverse range of affiliations: Orthodox (39%), Secular (11%), Assimilated (4%), and Christian (1%). Additionally, 14% of survivors described some degree of observance (Reform, Conservative, or Traditional), while the remaining 31% did not specify a particular denomination.

Post-war identities draw a remarkably different image, with 68% of survivors no longer affiliating with any particular religious movement (grouping general Jewish identity with diverse unclassified responses), 15% identifying as Orthodox, 11% Secular, 6% maintaining observance at some level, and less than 1% identifying as assimilated Jews or Christians. This extraction was performed via zero-shot prompting, followed by manual validation of two samples from each background to serve as an initial heuristic, and alignment with the visual history archive (VHA) metadata.[6] Prompt details are available in the supplementary material.

Reviewing the data manually to validate this analysis revealed several identities, covering traditional Jews and witnesses who were baptized or grew up in a non-Jewish environment. In those cases, accounts of practicing Christianity and mentions of violating Jewish rituals appear as inactive statements within their religious trajectories (discussed further in section 4).

It is important to note that the generated identity titles must be interpreted with attention to the cultural biases embedded in pre-trained language models, as well as to geographical-historical differences, when discussing religious nuances in Europe and in the countries where the survivors later lived (Plaza-del-Arco et al. 2024).

3.1 Testimony Structure

To identify common patterns across the collection rather than within individual testimonies, we require a way to compare narratives to one another. This raises the question of how the chronological order of events aligns thematically along the narrative sequence. Although personal narratives, by nature, each have their own structure, previous research on this dataset suggests that religiosity is discussed with higher frequency at certain points in the narrative (Ifergan et al. 2024). This is likely influenced by the structured interview guidelines of the SF, which direct interviewers to discuss pre-war life and religious practice early on and to discuss the survivor's present views toward

4. http://tiny.cc/sf-interview-guidelines
5. https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview
6. See: https://vha.usc.edu/.

the end.

Consistent with this, Wagner et al. (2022) mapped SF thesaurus keywords[7] onto four temporal categories, and found their average narrative positions to suggest a rough but consistent thematic structure across testimonies, supporting the alignment of religious trajectories along the shared narrative timeline (Figure 2, and in the supplementary material).

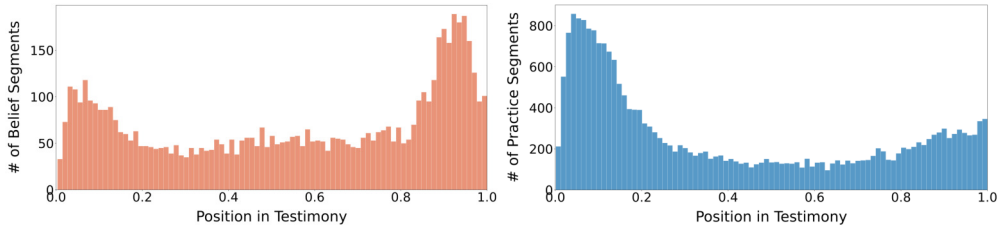


Figure 2: Distributions of religious content along the narrative timeline: the left plot isolates the three belief classes, and the right panel illustrates religious practice. For evaluating trajectories against proxies, the distributions used to define the baselines are computed separately for each label.

conference version

4. Annotations

The trajectories are formed using all religious content segments from the protagonist's perspective. Although Judaism encompasses many dimensions, this work broadly distinguishes between two aspects of religiosity: practice and belief. This distinction arises naturally, as it reflects incongruities often observed among religious observers. As Chaves (2010) notes, "Ideas and practices exist as bits and pieces that come and go as situations change, producing many inconsistencies... Religious ideas, values, and practices generally are not congruent".

A key step in extracting the trajectories lies in defining labels for annotating each potential segment. With a team of domain experts, we collected annotations in two phases.

First, the annotators were directed to identify all segments that express the narrator's Jewish religious observance and belief or indicate their absence. Examples of interest include "He was able to survive... a miracle. It was simply a miracle. Because, by rights, he was dying. He was on death's door" and "I got my presents under the Christmas tree, and then my girlfriend came to Hanukkah to my place. She got a present". For consistency, we do not treat Jewish identity as equivalent to religion; therefore, mentions such as "If you ask him, what's your religion, I am a Jew" are excluded. Similarly, Zionism descriptions are out of scope, since the Zionist movement saw the land of Israel and the Hebrew language as national heritage, and not necessarily tied to religious significance. Since the majority of segments contain no religious significance, this step serves as initial filtering before the detailed annotation. The complete guidelines are provided in the supplementary materials.

The second stage categorizes the segments identified as applicable in the first phase, meaning those that are highly likely to contain descriptions of the protagonist practicing religion or reflections on their belief. After performing the first annotation phase, we

7. sfi.usc.edu/content/keyword-thesaurus

conference version

fine-tune a classifier on this data (section 5) and then apply it to identify segments of interest, sampling 1,000 positive predictions to build the dataset for this task.

Each sample is ranked according to the intensity of the practice and belief reflected, or marked as a false positive. Practices are defined as rituals, actions, or activities motivated by the Jewish religion. Descriptions reflecting the absence of Jewish religious practices are also included in this category. We note that the practice label records the objective occurrence or absence of a ritual, regardless of whether the absence was due to personal choice or external wartime constraints; mentions of the individual's internal willingness or intention are evaluated separately by the belief dimension. Religious belief, in contrast, is defined by inner-life descriptions, including ideas, thoughts, philosophy, and feelings related to God or religion. This approach is broadly supported in the sociology of religion literature (Leo et al. 2021; McIntosh 1995; Spilka et al. 1985).

The classification schema distinguishes between the valences of expressed religious practice and belief. A segment describing a practice is annotated with one of the following classes: *active*, *inactive*, or *other*. The first two classes identify instances of participation in or abstention from religious practices. In contrast, *other* is used when the description does not clearly meet the criteria of either category. This includes, for instance, references to religious holidays without sufficient context to determine whether they were observed, as well as cases in which a single segment simultaneously contains elements of both participation and non-participation.

Belief is annotated using a parallel schema consisting of the classes *positive*, *negative*, and *other*. Examples include sincere praying (*positive*)—"What did you do? ... Nothing. I just prayed. I just prayed."—and rejection of God (*negative*)—"Why didn't he help us? I don't believe... There is no God". The *other* category captures ambivalent or complex positions, where belief is expressed with uncertainty, conditionality, or internal tension rather than clear affirmation or denial (See example in Figure 3, and the annotation guidelines in the supplementary materials for further examples of all classes).

Importantly, practice and belief are annotated independently, allowing for combinations that reflect the nuanced ways individuals relate to religion. For example (practice-*other*, belief-*positive*): "I believe in God... I kiss the mezuzah... I don't go to the synagogue... every morning... I pray my own way." Figure 3 illustrates the annotation process.

This complex annotation task involves knowledge of history and the Jewish religion. Many segments can be ambiguous or context-dependent. Therefore, there may be a lack of consensus among annotators on the classification. These factors contribute to the complexity of the task.

The annotation process was carried out by three annotators, collecting a total of 4,000 randomly sampled testimony units in the binary classification stage and 1,165 in the second stage. We evaluate the inter-annotator agreement (IAA) with Krippendorff's Alpha Krippendorff (1970), which accounts for agreement beyond chance, on a subset of overlapping samples annotated by at least two annotators (833 for the first phase, 540 for the second phase). Alpha values range from -1 to 1, with higher values indicating stronger agreement. The score reflects both the quality of the labels and the complexity of the task. The first task attains an average Alpha of 0.74, indicating a substantial level of agreement, and 0.48 for the second, which is more complex, indicating mod-

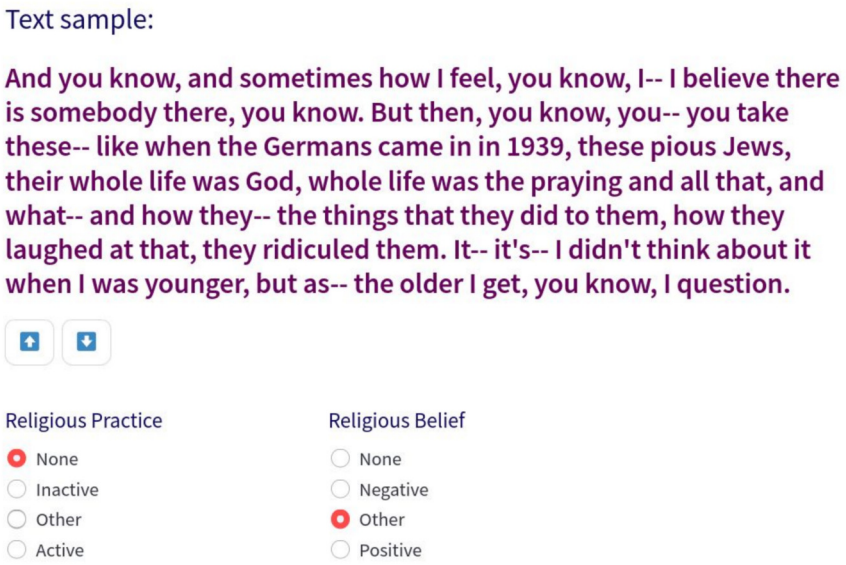


Figure 3: An annotation example from the platform we provided the annotators with to identify the survivor's valence of religious practice and belief in each segment.

conference version

erate agreement, and suggesting that this classification task is challenging even for human annotators, which is reasonable given the increased complexity and the task's subjectivity.

The Belief dataset contains 122 *Positive* samples, 68 *Negative*, and 44 *Other*. For Practice predictions, the class distribution is 343 *Active*, 77 *Inactive*, and 49 *Other*. We divide the data into three splits with equal proportions of each class. sampled entirely from the overlapping set to ensure the reliability of our results.

5. Experiments

For a single testimony, the pipeline for extracting its trajectory comprises segmenting the transcript into smaller text units, identifying and filtering religious content, and categorizing the segments that mention religion by practice, belief, and their respective intensities (Figure 1). Formally, each trajectory is represented as a sequence of 3-dimensional vectors of the form (t, b, p), where $0 < t < 1$ denotes the segment's position along the normalized narrative timeline, b and p indicate the belief and practice labels, respectively (as defined in section 4).

The interview question-answer structure is leveraged to segment the testimonies into text units, each of which is assumed to correspond to a single point in the survivor's trajectory sequence. To construct these units, we use the speaker annotations to divide the transcript into question-answer pairs, then merge segments containing fewer than 10 words and split those exceeding 100. This process creates segments typically between 50-100 words long, mostly keeping separate ideas distinct. Yet, due to the nature of rule-based methods, some samples may contain contradicting religious signals. For example: "Saturday I would go, and I was encouraged to go to the synagogue. And Sunday, I was encouraged to go to church."

Filtering is performed using a RoBERTa classifier (Liu et al. 2019) fine-tuned on the

conference version

curated annotated data. This is a balanced dataset of 1,348 samples (674 positive samples with religious content), divided into three splits: train (0.8), validation (0.1), and test (0.1).[8] To enhance the reliability of the model and reduce individual bias in the evaluation, the entire test split was randomly sampled from the overlap data. Evaluation of the model on the test set yields accuracy and recall scores of 97% and 94%, respectively.

A total of 45,717 segments were identified in the religious content filtering phase, representing an average of 11% of the data across samples. These segments were then prompted to identify practice and belief valence. The average trajectory length constitutes 11% of a testimony's total segments, exhibiting meaningful variance across the testimonies: the bottom 10% of testimonies (the "least religious") contain religious content in an average of 4% of their segments, whereas the top 10% (the "most religious") cover an average of 28%.

The subsequent classification of belief and practice is carried out by prompting and fine-tuning large language models. Specifically, we fine-tune Mistral-7B-instruct (Jiang et al. 2023), on data from the second annotation task. Before reaching this configuration, we compared multiple setups to the annotated test-set, including prompting a series of GPT-4.1 models[9] on multiple settings. For training, we use Mistralai's mistral-finetune repository[10] with their default hyperparameters and test multiple seeds.[11] The best performing model across all practice and belief classes is GPT4.1 with an average F1 of 0.55 for practice and 0.64 for belief. The fine-tuned Mistral-7B scores 0.43 and 0.45 for practice and belief, respectively.[12]

6. Validation

To validate the correctness of the resulting trajectories, we draw on prior work conducted on the same subset of testimonies to generate silver-standard reference trajectories using two complementary methods, one based on topic modeling and the other on the SF thesaurus (subsection 3.1).

Ifergan et al. (2024) applied topic modeling on this corpus, uncovering multiple topics that correspond to religious practices and beliefs. Using this prior knowledge, we align each relevant topic with the corresponding practice or belief label, determine its position along the narrative timeline and treat the resulting sequences as silver standard data.

The testimony transcripts from the SF were divided into segments of one minute each, indexed with a subset of keywords from the SF's detailed thesaurus of 8,000 terms.[13] The second method to approximate the trajectories is by utilizing the thesaurus's terms related to religious practices and beliefs, and map them to practice/belief labels to extract

8. Hyperparameters: 4 epochs, batch size=4, learning rate=1e-5, seed=5
9. https://platform.openai.com/docs/models/gpt-4.1, version: 2025-04-14
10. https://github.com/mistralai/mistral-finetune
11. Hyperparameters: Lora Rank=64, sequence length=16,384, batch size=1, learning rate=1e-4, number of micro-batches=1. Number of training steps=100, weight decay=0.1, pct-start=0.05. Seed for belief=1; for practice=5
12. Owing to limitations with large-scale API access, we use Mistral-7B to generate the full dataset of trajectories. The prompts we run follow the format of Self-Consistency Wang et al. (2023) and are carefully designed with guidance from Anthropic's prompt generator (https://console.anthropic.com/login). Considering the complicated annotation process and IAA, along with the models' bias toward stereotypes in the context of religion Plaza-del-Arco et al. (2024), these moderate F1 scores are unsurprising.
13. sfi.usc.edu/content/keyword-thesaurus

reference trajectories. The full list of addressed terms and the induced topics that we compare against is provided in the supplementary material.

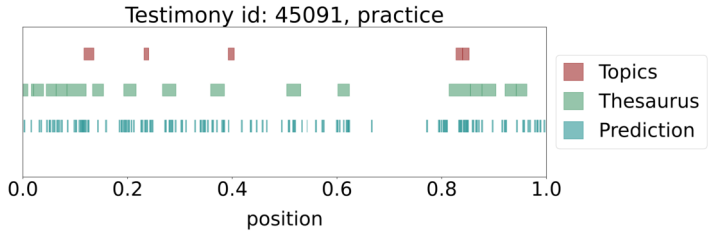


Figure 4: Alignment of predicted trajectories with reference trajectories for testimony ID: 45091, illustrating the differences and overlaps between them. The colored rectangle widths correspond to the segment lengths. The x-axis represents the normalized position within the testimony transcript.

Both methods have several limitations, which make the results suitable only for comparison. Since the segments that were used for topic modeling and the thesaurus are larger than the segments we use, the resulting sequences are expected to be sparser in relation to our method. For the topics, our analysis of the topic model selects a single topic for each segment, rather than all the top relevant topics. For the thesaurus, despite the large number of terms, many segments lack labels for their matching terms. This may be a result of the exceedingly large set of keywords, which may have resulted in recall issues in the SF annotation. In addition, the label set of the references is partial, meaning that the majority of the topics and terms we address do not have a specific valence. Considering these points, the false positive rate of the predictions is unknown, making the references suitable only for measuring recall. An example illustrating the alignment of the references and predictions is shown in Figure 4.

The fine-tuned model is used to produce 1,000 trajectories. In this setting, relying on our fine-tuned classifier, only religious content predictions are prompted, and not every single one of the original dataset segments. For each of the reference and predicted sequences, we quantify their distance. For a given non-empty trajectory T and a reference path R:

$$min\_sum\_dist(T, R) = \sum_{r \in R} min_{t \in T}(|t - r|)$$

If T is empty, return the number of segments in R, and if R is empty, return 0.

We compare the sum of the distances to baseline trajectories. These baselines are artificial sequences designed by using our prior knowledge about the distribution (Figure 2) of religious content throughout the dataset; each is the same length as the predicted sequence. The different baselines for a given class are defined in Table 1. Results of this evaluation are presented in Table 2, validating the predictions' recall as almost every reference point aligns with a predicted one.

7. Findings

The extracted trajectories offer insights into religious experiences as narrated by the survivors. They allow us to examine the intensity of religious practice across the narrative

conference version

Baseline	Definition
Equal	Scatter the points evenly between 0 and 1.
Original	Randomly select points from the original distribution of the predicted label.
Edges and middle	Randomly select points from three equal splits of $(0, 1)$; each third contains the same portion of points that it has in the predicted distribution of the label.
G-Edges and middle	Same as Edges&middle except we sample the point from the normal distribution.
2-Gaussian	All of the points sampled normally from the first half, same for the second half.

Table 1: The different baseline definitions for evaluating the predicted trajectories.

	Topic			Thesaurus				
	B	**P**	**P$^-$**	**B**	**P**	**P$^+$**	**B$^+$**	**B$^-$**
Predicted	**5**	**10**	**47**	**5**	**8**	**11**	**9**	**90**
Original scatter	33	40	**47**	29	34	50	11	105
Edges & middle	19	29	59	19	28	29	16	99
Equal scatter	21	20	60	19	17	21	**9**	142
Normal-original	16	57	52	18	47	53	17	104
# Reference paths	335	905	187	282	787	761	98	217
# predicted paths	954	998	742	954	998	990	796	480
# Reference points	456	2,768	301	439	2,434	2,214	171	253
# predicted points	6,051	23,274	1,987	6,051	23,274	20,086	3,204	917

Table 2: Sum of *min_sum_dist* for fine-tuned Mistral-7B predictions and baseline trajectories on the full dataset. **Original:** Sample from the distribution of the predictions. **Edges&middle:** Random sample from three equal splits of (0, 1); according to the predicted distribution of the label. **Equal:** Even scatter. **Normal-original:** Sample from the Gaussian with the variance and mean values of the predictions.

timeline and compare patterns across individuals and groups.

7.1 Individual Level Analysis

Both gender and geography serve as a proof of concept for this methodology, with the findings inviting scholars of history and religion to engage with the results and offer their own interpretations. We conduct six primary chi-square tests examining the association between demographic variables—gender, age group, and geographic origin—and two outcome variables—belief and practice. To account for multiple comparisons, we apply a Bonferroni correction across these tests; all reported p-values reflect this adjustment.

Analyzing practice-label frequencies by gender reveals notable differences in the narration of religious observance. Male testimony segments contain a higher proportion of Active labels than female ones, while the opposite holds for Inactive labels. A chi-square test indicates that female segments are 1.4 times more likely to express the absence of religious practice than its presence relative to male segments, a statistically significant effect ($p < 10^{-11}$).

One interpretation of these findings is that, before the Holocaust, observant Jewish women in Europe had fewer formal rituals available to them. They did not attend Heder or Yeshiva (religious school systems primarily for boys) and did not take on roles in

conference version

synagogue communal prayers. Surprisingly, the distribution of Inactive labels along the narrative timeline does not show that these gender differences diminish in the middle portion of the timeline. This period corresponds to descriptions of religious resilience in the camps, an experience common to both men and women, though manifested differently. For instance, Schmolling (1984) describes prisoners creating a Sabbath candle (a practice rooted in Jewish tradition associated with women) by hollowing out a potato, placing a rag as a wick, and filling it with margarine. While male expressions of religious resilience often involved smuggling tefillin or forming prayer groups, rituals more commonly associated with men Berkovits (1979).

A similar trend emerges when examining geographic origins in a subset of testimonies. We compared 510 testimonies of survivors originating in Eastern countries and 301 testimonies from Western countries. Segments drawn from testimonies of survivors originating in Western countries are 1.7 times more likely to contain negative belief expressions than those from Eastern countries, aligning with Banik (2016) and statistically significant ($p < 10^{-8}$).

Examining age group differences did not reveal consistent patterns; we therefore do not report these results here, though they may warrant further investigation in future work.

7.2 Clustering Analysis

To capture an overview of the entire trajectory set distribution, we cluster the testimonies in two methods: (1) by a Structure-based taxonomy, and (2) using hierarchical clustering algorithms.

Taxonomy

The trajectories are mapped by converting the belief and practice labels into numerical values based on structure and valence (*constant positive/negative, ascending/descending, oscillating*), and on the degree of coverage relative to the testimony storyline, $t_{max} - t_{min}$ (Table 3). This is a simplified taxonomy designed to reduce complexity as much as possible, ignoring the density or frequency of tags and segments labeled as *other*. While this abstraction collapses temporal frequency and density, which the hierarchical clustering addresses, the coverage attribute ensures the trajectory's temporal span remains a core factor in the classification. For this analysis, each sequence is simplified by removing any neutral values it contains and then collapsing consecutive identical values into a single instance. For example, [-1,1,0,1,1] becomes [-1,1].

Structure	**Filtered&shrunk series**
Constant-Negative/Inactive	[-1]
Constant-Positive/Active	[1]
Ascending	[-1, 1]
Descending	[1, -1]
Oscillating	$\{1, -1\}^n$
Coverage Level Definitions	
Low	$t_{max} - t_{min} \leq 0.33$
Medium	$0.33 \leq t_{max} - t_{min} \leq 0.67$
High	$0.67 < t_{max} - t_{min}$

Table 3: Definitions of the structures and coverage levels for the predefined taxonomy.

conference version

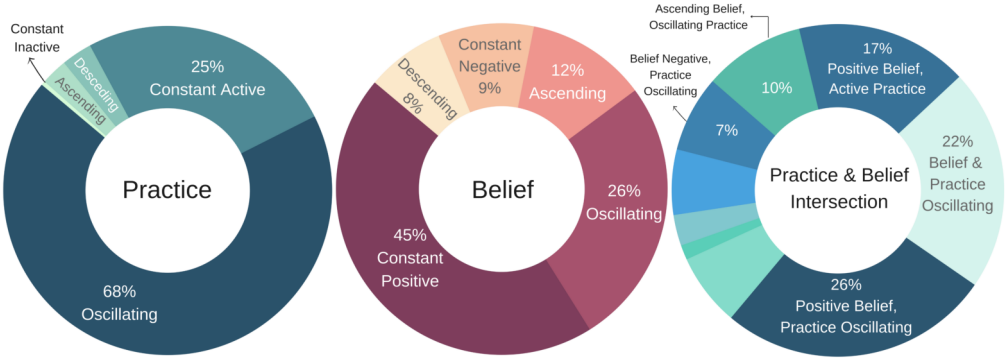


Figure 5: Religious Trajectory structure distributions, from left to right: 68% of the practice trajectories have an oscillating structure and 25% are constant-active. The belief trajectories have a similar distribution, with the largest cluster sharing a constant-positive structure (45%) and the rest distributed among oscillating (26%), ascending (12%), constant-negative (9%), and descending (8%). For the intersection of the two aspects, the two large groups cover 44% of the intersection, all have an oscillating practice structure, while the belief valence distributes evenly between oscillating and constant-positive structures.

The taxonomy mapping reveals which trajectory structures are dominant within the corpus, providing an overview of the findings. As hypothesized, the results show a notable divergence between religious practice and belief, supporting the premise that religious belief and practice are not always congruent (section 4). Among the extracted trajectories, 83% contain at least two points for both practice and belief.

For religious practice, the vast majority of trajectories (68%) are Oscillating, with the second largest group being Constant-Active (25%). The Oscillating trajectory captures non-linear accounts of observance, characterized by fluctuations in the level of practice over time. For example, the survivor Harry Thalheimer recalls some degree of observance during childhood, despite occasionally eating non-Kosher meat (Thalheimer 1997).

These patterns may be understood as a lived form of traditional Judaism. As argued by Banik (2016), traditional Jewish life in the interwar period often involved a complex identity where ritual observance was not a binary state but a series of situational negotiations. Within this framework, a non-linear trajectory is not necessarily a sign of religious crisis or environmental constraints, but rather a reflection of the fluid ways individuals practiced Judaism.

For religious belief, Constant-Positive (45%) and Oscillating (26%) are the most dominant. The remaining trajectories are split between Ascending (12%), Constant-Negative (9%) and Descending (8%). Examples of expressions along a Constant-Positive trajectory include persistent statements of faith, such as "he instilled in us to be thankful to HaShem" , "God will help. And we will be saved", and "I think he was there all the time. Because I only knew one thing – that God created us."

To investigate the religious observance and belief inconsistencies in the current data, we analyzed the most common combinations of belief and practice trajectories within the same testimony. The findings reveal three major trends, which together account for 65% of the intersectional data: 26% of testimonies pair a Positive Belief trajectory with an Oscillating Practice trajectory, 22% exhibit Oscillations in both belief and practice, and 17% of testimonies show positive valence for both dimensions (i.e., Positive Belief and

conference version

Active Practice).

Salomon Gutter's testimony (Gutter 1996) serves as an example for positive Belief and Oscillating Practice trajectories. Beginning with active practice of learning with the 'best rebbes', followed by the Inactive descriptions of not 'being Bar Mitzvahed' due to the war constraints, and having to violate the Jewish holidays: "They degrade you... you work on Yom Kippur (the Day of Atonement) ... you run away Rosh Ha-Shana (the Jewish New Year's)." These low points are then offset by positive signals of religious activity, which include secretly continuing Jewish religious life in a hidden attic as an act of spiritual resistance, thus creating an oscillating sequence.

Gutter's Positive Belief trajectory remains consistent and explicitly stated throughout the testimony. Pre-war, his belief is evident in the description of the holy atmosphere of the Sabbath. Crucially, he interprets multiple events during the Holocaust as the hand of God. He viewed his family's escape from deportation as two miracles in one day, attributing the first stage of salvation to his mother's resourcefulness, acting as a conduit for Divine Providence. Even the horrific reality of the camps is framed through this lens, as he recounts leaving Graditz camp: "We saw God's hand... he took us away from there", because shortly after the prisoners were evacuated, a typhus epidemic broke out in Graditz.

This consistent valence is finalized in Gutter's post-liberation and post-war reflections. He explicitly declared, "God has decreed that we should already be saved", and when asked about staying religious after the war, he emphasizes the human limitations of understanding God's actions: "There is [sic] many things we don't understand.... We are only human beings here for a short period of time."

An example for the second most dominant combination is the survivor Frank Dobia's testimony (Dobia 1996), accompanied by many recollections and reflections on religion, which create oscillations in both practice and belief. His practice oscillates between recalling his father's devout weekly Kiddush (*Active*) and his own later avoidance of the synagogue (*Inactive*). Tightly after liberation, he did the Jewish mourning custom of "Kaddish", and went to synagogue on the New Year holiday (*Active*), though, when asked by the interviewer to repeat a traditional prayer he once knew by heart, he notes, "it's a long time since I've done it" (*Inactive*).

Dobia's belief trajectory begins positively, describing a moment of uplifting during his father's Yom-Kippur prayers: "Another scene I'll never forget – the tears and the cries about his son – that we didn't know where he was – very moving and touching scene." Later, reflecting on an internal conflict. While in a labor camp, he states, "There was [sic] a lot of nonbelievers. I myself had become a nonbeliever", yet recounts in the same period, "On my bed..., I would remember about something, and I would pray quietly to myself... about my family... my brothers. What's happened? Where are they? When will we meet? When will we see them again? When will we meet up again?"

This testimony, moving fluidly between active and inactive, positive and negative beliefs, serves as a quintessential example of a non-linear religious journey that our computational model identifies as a key, recurring pattern within the corpus.

conference version

Hierarchical Clustering

The hierarchical clustering process aims to identify data-driven potential patterns across testimonies. This analysis includes neutral labels. Attempting to quantify religious evaluation similarity, considering sequences in various lengths and structures, we adapt Dynamic Time Warping (Berndt and Clifford 1994) with multiple window sizes to generate a distance matrix, and then run Agglomerative (Zepeda-Mendoza and Resendis-Antonio 2013) and HDBSCAN clustering (McInnes et al. 2017) to group similar trajectories.

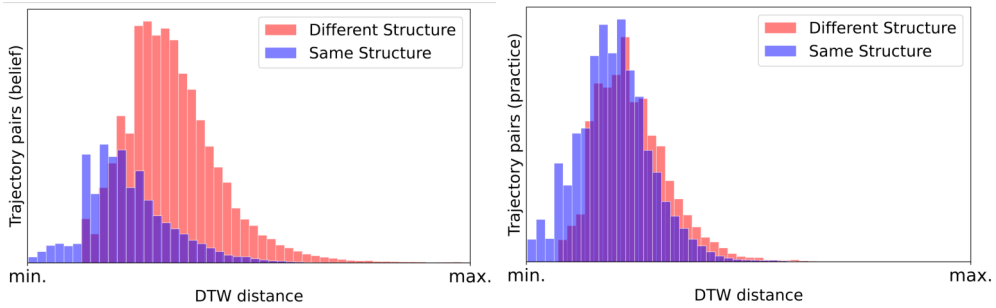


Figure 6: DTW distance distributions of trajectory pairs: the left plot represents belief and the right practice. Blue indicates pairs with the same structures, and red indicates pairs with different structures. The distribution for identical structures is shifted toward smaller DTW distance values.

Our results show that the DTW method is useful for selecting similar trajectories and that the hierarchical cluster found validates the taxonomy analysis, as we noticed that trajectories with the same structure tend to show smaller DTW distances (Figure 6).

The testimonies of Wanda Mehr (Mehr 1997) and Esia Shor (Shor 1998) appear to have closely related trajectories, with normalized DTW values of 0.14 and 0.21 for belief and practice, respectively. Their trajectories diverge in the domain of practice. Shor identifies with the Conservative movement at the time of the recording, whereas Mehr describes no longer maintaining any ritual observance.

Both recount experiences of desperate solitary prayer during periods of hiding. Shor testifies: "I hid in the field... all the time my mind was with prayer. And if I didn't say a hundred times Shema Yisrael I would be lying to you... that kept me sane... it, like, felt like I'm talking with God". Mehr's prayers, however, are not rooted in a specifically Jewish belief, though they share the emphasis on loneliness.

Each survivor also experienced a period of being hidden by righteous Christians, and after the war, each helped bring a Jewish child back to Judaism. In Mehr's case, it was her own daughter, born in 1939, who was taught by the woman who rescued her to hate Jews. She testifies: "I had to battle with her a long time later to convince her that it's not so at all." For Shor, it was her three-year-old cousin, raised in a Christian home from the age of nine months; she took her to church one Sunday and gently explained to her that this was not their religion.

A final point of convergence lies in their current, complex attitudes toward belief. Mehr states, "I don't know to believe or not to believe", and, "I don't believe in miracles. My life there during the war was all miracles—if there are some miracles." Shor similarly testifies to questioning religion in the post-Holocaust period.

conference version

This comparison illustrates the value of trajectory-based similarity for identifying testimonies that are structurally aligned yet diverge in meaningful ways, inviting close, historically grounded interpretation rather than superficial grouping. Their shared patterns highlight belief as a source of resilience in conditions of extreme isolation, while their divergent post-war practices show that such belief can persist independently of religious observance. Together, these similarities and disparities underscore the plurality of religious evolution and affirm the distinction between practice and belief.

7.3 Clustering Evaluation

For manual evaluation of the clusters, we selected 121 trajectories with at least four points and coverage greater than 0.75. Annotators classified each trajectory according to the taxonomy and identified the most similar pair within a set of triplets. IAA for the classification task was substantial for belief ($\alpha = 0.66$) and moderate for practice ($\alpha = 0.56$). Agreement was lower for the triplet similarity task, reaching moderate levels ($\alpha = 0.58$ for belief, $\alpha = 0.32$ for practice), emphasizing the task's complexity.

We found that DTW distances correlate with the similarity annotations, supporting their use for identifying similar trajectories. They also align with the taxonomy, with a t-test showing that the average normalized distance between trajectory pairs with the same structure is significantly smaller than that for pairs with different structures ($p < 10{-}12$), indicating that the distance captures fundamental structural characteristics of trajectories.

Comparing the automatic and manual classifications yields F1 scores of 0.40 for belief and 0.38 for practice. Some discrepancies are a result of the assumption that the religious trajectories follow the narrative timeline, which does not hold in all cases.

The evaluation confirms that while computational models cannot replace interpretation, they provide a reliable scaffold for large-scale analysis and hypothesis generation.

8. Discussion

This study systematically addressed the question of character development from a perspective of religion as a layer of identity. We analyzed religious trajectories extracted from a corpus of 1,000 Holocaust survivor testimony transcripts. This included conducting a detailed annotation process, leveraging large language models to build a pipeline for automatic extraction, and defining clustering strategies to identify common patterns across the narratives.

Interpreting Narrative Religious Trajectories. We examined trajectories of religious practices and beliefs along the narrative timeline, relying on the assumption that the testimonies in the corpus share a rough thematic structure that often aligns chronologically. While human memory, and consequently personal testimony, is inherently non-linear, the strict interview guidelines provide a sufficient chronological macrostructure, allowing for the treatment of the narrative sequence as a developmental trajectory.

Regarding the structure of belief trajectories, particularly the interpretation of oscilla-

conference version

conference version

tions, belief descriptions can convey meaning about how beliefs are articulated rather than provide evidence of the survivor's development. Many belief segments reflect past beliefs, for example, "Every night I would take my prayer book... and cry". However, there is an inherent tension between one's personal life as expressed in present-day narration and their historically lived experience (Rosenthal 2006), this analysis focusing on the former. Examining the relationship between narrated belief and lived experience shows that the two are closely aligned, though not always temporally synchronized.

Specifically, the grammatical tense of a belief statement does not always correspond to its position in narrative time. For instance, the statement "I don't know [what] to believe..." (subsection 7.2) appears at the beginning of the testimony, although it reflects the speaker's present religious views.

Performing a structure-based taxonomy revealed results that align with theoretical research on post-Holocaust religious responses. Examining the common belief trajectory reveals that the most frequent structure is Constant-Positive (45%). This is consistent with the argument that faith can persist after unimaginable trauma only through a theodicy framework capable of containing such evil. This framework serves as a source of resilience by reconciling the existence of a benevolent God with the reality of suffering (Dein 2022). In parallel, 46% of the trajectories describe transformations of religious views (Oscillating, Ascending, and Descending), Schweid (1988)'s theory of a post-Holocaust crisis of theodicy in Jewish and Christian religious thought, arguing that it necessitated substantial changes in religious norms. On the other hand, about 5% of the survivors interviewed were changed from atheists to believers, wearas the the taxonomy analysis we performed found a non-negligible portion (12%) of Ascending trajectories.

When turning to the findings regarding practice trajectories, 68% display an oscillating pattern while 25% show a constant active disposition of religious practices. Brenner (1980) reports a significant decline in religious observance in the immediate post-war period; however, when comparing participants' observance levels before the war with those at the time of the study, 71% reported no overall change. This finding supports the commonality of oscillating practice trajectories, reflecting an initial disruption followed by a return to prior levels of religious observance. While this pattern closely aligns with one common form of oscillation observed in our data, it does not capture the full diversity of oscillating trajectories, which include multiple non-linear configurations of change in religious practice.

Limitations. While this work demonstrates the potential of computational character analysis, its limitations must be acknowledged. The testimonies, recorded decades after the war and therefore shaped by hindsight, may not serve as precise mirrors of historical events (Felman and Laub 1992). A blindspot is further present due to reporter and survivor biases, capturing only the stories of those who survived and chose to speak, which complicates efforts to quantify the full spectrum of faith across the broader victim population (Dein 2022).

Additionally, a geographical selection bias, of English-speaking countries limits the generalizability of our findings, and broader secularization trends likely influenced individual trajectories rather than trauma alone (Banik 2016).

Finally, the large language models employed embed cultural and religious biases (Plaza-del-Arco et al. 2024). Despite fine-tuning and careful prompting, the models frequently assigned positive valence to acts such as praying to Jesus or attending church, suggesting a bias toward Christianity as the default religious framework. Overall, these limitations highlight the need for cautious interpretation and point to directions for future work to mitigate these biases.

9. Conclusion

To conclude, this study demonstrates a proof of concept for leveraging computational methods to track character development in survivor narratives. Our proposed pipeline, comprising fine-grained annotation, large language model deployment, and quantitative analysis, reveals interpretable patterns of religious belief and practice across testimonies. More broadly, this approach suggests how computational models can be used to enhance the ability to scale interpretive questions and enable systematic engagement with large testimonial corpora, opening multiple interdisciplinary pathways.

10. Ethics Statement

Any personal details regarding the examples from the SF testimonies were cited in accordance with the Visual History Archive (VHA) guidelines. The code for the experiments and evaluation will be released upon request; however, it will not contain any private data from the archive. The data and trained models used in our work will not be shared with third parties without the archive's consent. Permission to browse and research the testimonies can be requested from the SF archive.

11. Data Availability

The code for training the classifier and the hierarchical clustering, as well as supplementary materials, including annotation guidelines and prompts, can be found here: https://doi.org/10.5281/zenodo.19327829.

12. Acknowledgements

The authors acknowledge the USC Shoah Foundation – The Institute for Visual History and Education for supporting this research. We thank Dr. Naama Seri-Levi, and Dr. Mali Eisenberg for their valuable insights, and our team of annotators for their research assistance. Grants from the Israeli Ministry of Science and Technology, the Israeli Council for Higher Education, and the Alfred Landecker Foundation supported this research.

13. Author Contributions

Esther Shizgal: Conceptualization, Implementation, Writing – original draft, Writing – review and editing

Eitan Wagner: Conceptualization, Supervision, Writing – review and editing

conference version

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

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

References

Algee-Hewitt, Mark (2017). "Distributed Character: Quantitative Models of the English Stage, 1550–1900". In: *New Literary History* 48.4, 751–782. 10.1353/nlh.2017.0038.

Alqahtani, Ali, Mohammed Ali, Xianghua Xie, and Mark W. Jones (2021). "Deep Time-Series Clustering: A Review". In: *Electronics*. 10.3390/electronics10233001.

Assmann, Aleida (2006). "History, Memory, and the Genre of Testimony". In: *Poetics Today* 27.2, 261–273. 10.1215/03335372-2005-003.

Banik, Vibeke Kieding (2016). "The Faith of the Fathers, the Future of the Youth". In: *Scripta Instituti Donneriani Aboensis* 27, 153–172. 10.30674/scripta.66573.

Ben-Ezra, Menachem, Yuval Palgi, Dina Sternberg, Dina Berkley, Hadar Eldar, Yael Glidai, Liron Moshe, and Amit Shrira (2010). "Losing my religion: A preliminary study of changes in belief pattern after sexual assault". In: *Traumatology* 16.2, 7–13. 10.1177/1534765609358465.

Berkovits, Eliezer (1973). *Faith after the Holocaust*. NY: Ktav Publishing House. ISBN: 0870681931, 9780870681936.

— (1979). *With God in Hell: Judaism in the Ghettos and Deathcamps*. Sanhedrin Press. ISBN: 9780884829379.

Berndt, Donald J. and James Clifford (1994). "Using Dynamic Time Warping to Find Patterns in Time Series". In: *KDD Workshop*. AAAI Press, 359–370. 10.5555/3000850.3000887.

Blanke, Tobias, Michael Bryant, and Mark Hedges (2019). "Understanding Memories of the Holocaust—A New Approach to Neural Networks in the Digital Humanities". In: *Digital Scholarship in the Humanities* 35.1, 17–33. 10.1093/llc/fqy082.

Brahman, Faeze and Snigdha Chaturvedi (2020). "Modeling Protagonist Emotions for Emotion-Aware Storytelling". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics, 5277–5294. 10.18653/v1/2020.emnlp-main.426.

Brahman, Faeze, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi (2021). ""Let Your Characters Tell Their Story": A Dataset for Character-Centric Narrative Understanding". In: *Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 1734–1752. 10.18653/v1/2021.findings-emnlp.150.

Brennan, Mark and Daniel K. Mroczek (2003). "Examining Spirituality Over Time: Latent Growth Curve and Individual Growth Curve Analyses". In: *Journal of Religious Gerontology* 14.1, 11–29. 10.1300/J078v14n01_02.

Brenner, Reeve Robert (1980). *The Faith and Doubt of Holocaust Survivors*. Free Press. https://archive.org/details/faithdoubtofholo0000bren.

Bruner, Jerome (1991). "The Narrative Construction of Reality". In: *Critical Inquiry* 18.1, 1–21. 10.1086/448619.

conference version

Chang, Yanchuan, Egemen Tanin, Gao Cong, Christian S. Jensen, and Jianzhong Qi (2023). "Trajectory Similarity Measurement: An Efficiency Perspective". In: *Proc. VLDB Endow.* 17, 2293–2306. 10.48550/arXiv.2311.00960.

Chaturvedi, Snigdha, Haoruo Peng, and Dan Roth (2017). "Story Comprehension for Predicting What Happens Next". In: *Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Association for Computational Linguistics, 1603–1614. 10.18653/v1/D17-1168.

Chaves, Mark (2010). "Rain Dances in the Dry Season: Overcoming the Religious Congruence Fallacy". In: *Journal for the Scientific Study of Religion* 49.1, 1–14. 10.1111/j.1468-5906.2009.01489.x.

Chopard, Thomas (2020). "Post-Holocaust Migrations from Poland to America: An Exercise in Microhistory". In: *S.I.M.O.N. Shoah: Intervention, Methods, Documentation* 7.1. 10.23777/SN.0120.

Cole, Tim and Alberto Giordano (2025). "Mapping the Emotional Landscapes of the Holocaust: Visualizing Space and Place in Survivor Trajectories". In: *Holocaust and Genocide Studies*. 10.1093/hgs/dcaf044.

Dein, Simon (2022). "Trauma, theodicy and faith: maintaining religious beliefs in the Holocaust". In: *Mental Health, Religion & Culture* 25.3, 388–400. 10.1080/13674676.2022.2027900.

Dobia, Frank (1996). *Interview 20367*. Interview by Christian Froelicher. USC Shoah Foundation Visual History Archive. https://vha.usc.edu/testimony/20367.

Eaglestone, Robert (2002). "Identification and the Genre of Testimony". In: *Immigrants & Minorities* 21.1-2, 117–140. 10.1080/02619288.2002.9975035.

Felman, Shoshana and Dori Laub (1992). *Testimony: Crises of Witnessing in Literature, Psychoanalysis, and History*. NY: Taylor & Francis/Routledge.

Fowler, James W. (1981). *Stages of Faith: The Psychology of Human Development and the Quest for Meaning*. San Francisco: Harper & Row.

Fowler, James W. and Sam Keen (1978). *Life Maps: Conversations on the Journey of Faith*. Ed. by Jerome Berryman. Waco, TX: Word Books.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas S Müller, and Matthias Studer (2011). "Analyzing and Visualizing State Sequences in R with TraMineR". In: *Journal of Statistical Software* 40.4, 1–37. 10.18637/jss.v040.i04.

Giuffra Darbyshire, Flavia (2019). "Was God Behind the Barbed Wire? An Inquiry into Jewish Faith and Practice in Auschwitz". Master's Thesis. Faculteit der Geesteswetenschappen (FGw), University of Amsterdam.

Gutter, Salomon (1996). *Interview 22999*. Interview by Miriam Cofsky. USC Shoah Foundation Visual History Archive. https://vha.usc.edu/testimony/22999.

Howarth, William L. (1974). "Some Principles of Autobiography". In: *New Literary History* 5.2, 363–381. 10.2307/468400.

Ifergan, Maxim, Omri Abend, Renana Keydar, and Amit Pinchevski (2024). "Identifying narrative patterns and outliers in holocaust testimonies using topic modeling". In: *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*. Torino, Italia: ELRA and ICCL, 44–52. https://aclanthology.org/2024.htres-1.7.pdf.

Ingersoll-Dayton, Berit, Neal Krause, and David Morgan (2002). "Religious Trajectories and Transitions Over the Life Course". In: *The International Journal of Aging and Human Development* 55.1, 51–70. 10.2190/297Q-MRMV-27TE-VLFK.

conference version

Jiang, Dongsheng, Yuchen Liu, Songlin Liu, Zhao, Jin'e, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong (2023). "From CLIP to DINO: Visual Encoders Shout in Multi-Modal Large Language Models". In: *arXiv preprint*. 10.48550/arXiv.2310.08825.

Krippendorff, Klaus (1970). "Bivariate Agreement Coefficients for Reliability of Data". In: *Sociological Methodology* 2, 139–150. 10.2307/270787.

Langer, Lawrence L. (1991). *Holocaust Testimonies: The Ruins of Memory*. New Haven: Yale University Press. 10.1017/S0364009400004505.

Lassley, Jennifer (2015). "A Defective Covenant: Abandonment of Faith among Jewish Survivors of the Holocaust". In: *International Social Science Review* 90.2, 1–17. https://www.jstor.org/stable/intesociscierevi.90.2.03.

Leo, Darius, Zahra Izadikhah, Erich C. Fein, and Sayedhabibollah Ahmadi Forooshani (2021). "The Effect of Trauma on Religious Beliefs: A Structured Literature Review and Meta-Analysis". In: *Trauma, Violence & Abuse* 22.1, 161–175. 10.1177/1524838019834076.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692*. 10.48550/arXiv.1907.11692.

Mariot, Nicolas and Claire Zalc (2017). "Reconstructing Trajectories of Persecution: Reflections on a Prosopography of Holocaust Victims". In: *Microhistories of the Holocaust*. Ed. by Claire Zalc and Tal Bruttmann. Vol. 24. War and Genocide. New York: Berghahn Books, 85–112.

McAdams, Dan P. (2001). "The Psychology of Life Stories". In: *Review of General Psychology* 5.2, 100–122. 10.1037/1089-2680.5.2.100.

McCullough, Michael E., Craig K. Enders, Sharon Brion, and Andrea R. Jain (2005). "The varieties of religious development in adulthood: a longitudinal investigation of religion and rational choice". In: *Journal of Personality and Social Psychology* 89.1, 78–89. 10.1037/0022-3514.89.1.78.

McInnes, Leland, John Healy, and Steve Astels (2017). "hdbscan: Hierarchical Density Based Clustering". In: *Journal of Open Source Software* 2.11, 205. 10.21105/joss.00205.

McIntosh, Daniel N. (1995). "Religion-as-Schema, with Implications for the Relation between Religion and Coping". In: *International Journal for the Psychology of Religion* 5.1, 1–16. 10.1207/s15327582ijpr0501_1.

Meddin, Jacob Robert (1998). "Dimensions of Spiritual Meaning and Well-Being in the Lives of Ten Older Australians". In: *The International Journal of Aging and Human Development* 47.3, 163–175. 10.2190/1LXA-K5TN-BGY4-FAXV.

Mehr, Wanda (1997). *Interview 26609*. Interview by Richard Reisner. USC Shoah Foundation Visual History Archive. https://vha.usc.edu/testimony/26609.

Melia, Susan Perschbacher (1999). "Continuity in the Lives of Elder Catholic Women Religious". In: *The International Journal of Aging and Human Development* 48.3, 175–189. 10.2190/X0LY-TERK-XQCC-CXAV.

Michman, Dan (1993). "Jewish religious life under Nazi domination: Nazi attitudes and Jewish problems". In: *Studies in Religion / Sciences Religieuses* 22.2, 147–165.

Mohr, John W., Christopher A. Bail, Margaret Frye, Jennifer C. Lena, Omar Lizardo, Terence E. McDonnell, Ann Mische, Iddo Tavory, and Frederick F. Wherry (2020). *Measuring Culture*. NY: Columbia University Press. 10.7312/mohr18028.

conference version

Montell, Conrad (2001). "Speculations on a privileged state of cognitive dissonance". In: *Journal for the Theory of Social Behaviour* 31.2, 119–137. 10.1111/1468-5914.00151.

Nelson, Laura K. (2020). "Computational Grounded Theory: A Methodological Framework". In: *Sociological Methods & Research* 49.1, 3–42. 10.1177/0049124117729703.

Papoudakis, Argyrios, Mirella Lapata, and Frank Keller (2024). "BookWorm: A Dataset for Character Description and Analysis". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Association for Computational Linguistics, 4471–4500. 10.18653/v1/2024.findings-emnlp.258.

Patterson, David (1998). *Sun Turned to Darkness: Memory and Recovery in the Holocaust Memoir*. NY: Syracuse University Press.

Piper, Andrew, Mark Algee-Hewitt, Koustuv Sinha, Derek Ruths, and Hardik Vala (2017). "Studying Literary Characters and Character Networks". In: *Digital Humanities 2017*. Alliance of Digital Humanities Organizations. https://dh2017.adho.org/abstracts/103/103.pdf.

Piper, Andrew, Richard Jean So, and David Bamman (2021). "Narrative Theory for Computational Narrative Understanding". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 298–311. 10.18653/v1/2021.emnlp-main.26.

Plaza-del-Arco, Flor Miriam, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, and Dirk Hovy (2024). "Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Association for Computational Linguistics, 4346–4366. 10.18653/v1/2024.findings-emnlp.251.

Presner, Todd, Anna Bonazzi, Rachel Deblinger, Lizhou Fan, Michelle Lee, Kyle Rosen, and Campbell Yamane (2024). *Ethics of the Algorithm: Digital Humanities and Holocaust Memory*. Princeton: Princeton University Press. 10.1515/9780691258980.

Radak, Tamara, Lou Burnard, Pieter Francois, Agnes Hilger, Fotis Jannidis, Gábor Palkó, Roxana Patras, Michael Preminger, Diana Santos, and Christof Schöch (2024). "Towards a Computational History of Modernism in European Literary History: Mapping the Inner Lives of Characters in the European Novel (1840–1920)". In: *Open Research Europe* 3, 128. 10.12688/openreseurope.16290.2.

Rafaeli, Omri, Noam Krohn-Borojovich, Itay Marienberg-Milikowsky, and Dan Vilenchik (2025). "Mind the Gap: Word-Embedding and Multi-Layered Literary Networks". In: *Digital Scholarship in the Humanities*, fqaf112. 10.1093/llc/fqaf112.

Reich, K. Helmut (1992). "Religious development across the life span: Conventional and cognitive developmental approaches". In: *Religion and Mental Health*. Ed. by John F. Schumaker. Hillsdale, NJ: Erlbaum, 145–188. 10.4324/9781315807508.

Rosenthal, Gabriele (2006). "The narrated life story: On the interrelation between experience, memory and narration". In: *Narrative, Memory & Knowledge: Representations, Aesthetics, Contexts*. Ed. by Kate Milnes, Christine Horrocks, Nancy Kelly, Brian Roberts, and David Robinson. Huddersfield: University of Huddersfield, 1–16. https://eprints.hud.ac.uk/id/eprint/4894/2/Chapter_1_-_Gabriele_Rosenthal.pdf.

Schmolling, Paul (1984). "Human reactions to the Nazi concentration camps: a summing up". In: *Journal of Human Stress* 10.3, 108–120. 10.1080/0097840X.1984.9934964.

conference version

Schweid, Eliezer (1988). "'Faith, Ethics and the Holocaust': The Justification of Religion in the Crisis of the Holocaust". In: *Holocaust and genocide studies* 3.4, 395–412. https://academic.oup.com/hgs/article/3/4/395/758021.

Shizgal, Esther, Eitan Wagner, Renana Keydar, and Omri Abend (2025). "Computational Analysis of Character Development in Holocaust Testimonies". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng. Association for Computational Linguistics, 22721–22745. 10.18653/v1/2025.emnlp-main.1156.

Shor, Esia (1998). *Interview 41035*. Interview by Ruth Meyer. USC Shoah Foundation Visual History Archive. https://vha.usc.edu/testimony/41035.

Spilka, Bernard, Phillip Shaver, and Lee A. Kirkpatrick (1985). "A General Attribution Theory for the Psychology of Religion". In: *Journal for the Scientific Study of Religion* 24.1, 1–20. 10.2307/1386272.

Sultana, Sharifa, Renwen Zhang, Hajin Lim, and Maria Antoniak (2022). "Narrative Datasets through the Lenses of NLP and HCI". In: *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Ed. by Su Lin Blodgett, Hal Daumé III, Michael Madaio, Ani Nenkova, Brendan O'Connor, Hanna Wallach, and Qian Yang. Association for Computational Linguistics, 47–54. 10.18653/v1/2022.hcinlp-1.7.

Thalheimer, Harry (1997). *Interview 27080*. Interview by Anita Hecht. USC Shoah Foundation Visual History Archive. https://vha.usc.edu/testimony/27080.

Tóth, Gábor Mihály, Tim Hempel, Krishna Somandepalli, and Shri Narayanan (2022). "Studying Large-Scale Behavioral Differences in Auschwitz-Birkenau with Simulation of Gendered Narratives". In: *Digital Humanities Quarterly* 16.3. http://www.digitalhumanities.org/dhq/vol/16/3/000622/000622.html.

Wagner, Eitan, Renana Keydar, and Omri Abend (2023). "Event-Location Tracking in Narratives: A Case Study on Holocaust Testimonies". 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, 8789–8805. 10.18653/v1/2023.emnlp-main.544.

— (2025). "Unsupervised Location Mapping for Narrative Corpora". In: *arXiv preprint*. 10.48550/arXiv.2504.05954.

Wagner, Eitan, Renana Keydar, Amit Pinchevski, and Omri Abend (2022). "Topical Segmentation of Spoken Narratives: A Test Case on Holocaust Survivor Testimonies". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Association for Computational Linguistics, 6809–6821. 10.18653/v1/2022.emnlp-main.457.

Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai-hsin Chi, and Denny Zhou (2023). "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: *Proceedings of the Eleventh International Conference on Learning Representations*. International Conference on Learning Representations. https://openreview.net/forum?id=1PL1NIMMrw.

Wilkens, Matthew, Elizabeth F. Evans, Sandeep Soni, David Bamman, and Andrew Piper (2024). "Small Worlds: Measuring the Mobility of Characters in English-Language Fiction". In: *Journal of Computational Literary Studies* 3.1, 1–16. 10.48694/jcls.3917.

conference version

Yang, Xiaoyan and Federico Pianzola (2025). "Fans Reconstruct Heroes: Modeling Fictional Characters in Participatory Culture". In: *Semantic Web: Interoperability, Usability, Applicability*. in press. https://pure.rug.nl/ws/portalfiles/portal/1390729016/swj3885.pdf.

Young, James E. (1997). "Between History and Memory: The Uncanny Voices of Historian and Survivor". In: *History and Memory* 9.1-2, 47–58. 10.2979/HIS.1997.9.1-2.47.

Zepeda-Mendoza, Marie Lisandra and Osbaldo Resendis-Antonio (2013). "Hierarchical Agglomerative Clustering". In: *Encyclopedia of Systems Biology*. Springer New York, 886–887. 10.1007/978-1-4419-9863-7_1371.

conference version

jcls. Journal of Computational Literary Studies

conference version

Echoes of Emotion

Linking Narrative and Reader Response of Web Novels in Chinese and English

OPEN ACCESS

Ze Yu[1]
Lanping Zhang[2]
Federico Pianzola[1]

1. Center for Language and Cognition, University of Groningen, Groningen, The Netherlands.
2. Department of Digital Humanities, King's College London, London, United Kingdom.

Citation
Ze Yu, Federico Pianzola, and Lanping Zhang (2026). "Echoes of Emotion. Linking Narrative and Reader Response of Web Novels in Chinese and English". In: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7994

Date published 2026-05-05 (preprint)
Date accepted tbc
Date received 2026-01-08

Keywords
computational literary studies, reader response, web novels, sentiment analysis, cross-cultural comparison

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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. This article examines how fictional narratives' emotional valence affects reader response across languages and platforms. We reproduce findings from research on Wattpad, which shows stories' emotional valence influences reader reactions, with sentiment alignment varying by genre. We extend the analysis to two additional platforms (*Qidian* and *WebNovel*) with Chinese and English content, and employ Generalized Additive Models to also investigate nonlinear relationships between narrative and comment sentiment. Results confirm that narrative emotional valence significantly shapes reader response, with stronger correlations in Chinese narratives than English. Platform characteristics emerge as significant moderating factors in this relationship. Our multilingual corpus analysis reveals that while original findings are reproducible, cross-cultural dimensions add complexity to the interplay between narrative emotion and reader engagement in web novel communities.

1. Introduction

Online reading platforms have transformed literary consumption into a social, interactive experience, generating unprecedented data about reader engagement and emotional responses to stories (Pianzola 2025). These developments resonate with reception theory (Willis 2017), specifically, how the intertwined relationship between texts and readers shapes meaning. These platforms create new opportunities for empirical literary studies while facilitating cross-cultural research. Consistent with this perspective, recent computational research has explored the interaction between narratives and reading engagement (Koolen et al. 2022; Kuijpers et al. 2024). One central method to analyze this interaction is sentiment analysis, which enables examination of emotional patterns in both narrative content and reader responses, revealing engagement dynamics (Bizzoni and Feldkamp 2023).

Building on this methodological foundation, Pianzola et al. 2020 conducted an analysis of reader response on the online reading platform Wattpad. The study revealed

that there is a statistically significant positive association between the narrative sentiment and the sentiment of readers' comments, at the paragraph level. This work established methodological and theoretical contributions to computational literary studies and provided large-scale empirical evidence to reader-response theory. Additionally, it showed how online social reading platforms can be useful corpora for reception studies.

These findings align with media psychology frameworks that distinguish between emotions embedded in text and those experienced by audiences (Schmidt et al. 2023). At the textual level, emotions are expressed through story events, character experiences, and narrative elements, while at the audience level, emotions encompass multiple types including narrative emotions that emerge from story engagement, aesthetic emotions related to quality judgments (Menninghaus et al. 2020, 2019; Schmidt et al. 2023), and relived emotions triggered by personal associations (Mar et al. 2011). Shifts between emotions of the same or different valence can be a proxy for how readers experience the narrative (Nabi and Green 2015), with these responses likely mirroring the emotional patterns presented in the text (Appel et al. 2019; Tilmatine et al. 2024). However, empirical evidence suggests the strength of this correspondence varies across different narrative moments and structural features (Schmidt et al. 2023).

While Pianzola et al.'s study provided valuable insights into narrative-reader emotional dynamics, several factors limit the generalizability of these findings to other digital reading contexts. First, the original study was limited to English-language stories on Wattpad with specific genres (Teen Fiction and Classics), leaving questions about whether these relationships extend to other genres and platforms. Second, the lexicon-based sentiment analysis methods they used can overlook idiosyncratic language use or unusual sentence structure, which often occur in social media data. Third, linear statistical modelling may have obscured more complex nonlinear relationships between narrative progression and reader engagement.

Our research addresses these limitations by analyzing a parallel multilingual dataset of web novels from two reading platforms, namely Chinese stories translated into English (Yu et al. 2025), and by refining the methods for sentiment analysis. To this end, we first attempt to reproduce the original results using the same statistical method but apply a different sentiment analysis procedure based on a fine-tuned transformer optimized for online readers' comments. Secondly, we test the same hypothesis with Generalized Additive Models (GAMs) to also investigate the nonlinear relationship between the progression of narrative sentiment and the associated comment sentiment. While we do not differentiate by genre in this reproduction, we hope to use this composite genre of corpus to test the generality of the relationship between narrative and reader response.

conference version

2. Methods

2.1 Corpus

This study employs the Qidian-WebNovel Corpus (Yu et al. 2025), a parallel dataset of 110 Chinese web novels from Qidian.com (original Chinese edition) and Webnovel.com (translated English edition), including readers' comments at the paragraph level (thus, the granularity of the data is the same as that of Pianzola et al. 2020) (see Table 1).

Source	Genres	Tags	Total Comments	Replies	Primary Lang.	Comments Lang. Distribution	Replies Lang. Distribution
Qidian	10	27	2,791,837	855,577	CN	CN: 95.7%, EN: 0.1%	CN: 97.2%, EN: 0.05%
WebNovel	8	40	327,988	96,250	EN	EN: 72.7%, Other: 27.3%	EN: 68.2%, Other: 31.8%

Table 1: Metadata for stories on Qidian and WebNovel

The broad categorization of the selected corpus is *Male Lead*[1] and the stories span the following genres[2]: Sci-Fi, Games, Eastern, Fantasy, Urban (see Table 2). This selection broadly reflects the genre distribution observed in Chinese web novel platforms. All selected works were complete novels; however, due to copyright restrictions, our analysis focuses exclusively on publicly available chapters on both platforms. These free chapters typically encompass the opening narrative, including world-building, character introductions, and the plot's first major turning point. This part of the story represents the most strategically crafted portion of the web novel, designed to engage readers sufficiently to encourage continued reading through paid content (Shao 2024; 光明网 2023). Recent research highlights the crucial role of setting in establishing the fictional world at the start of a narrative, which helps readers orient themselves within the spatial and sensory framework of the story (Rohrbacher 2025). Therefore, the available chapters of each story offer an important perspective for analyzing how readers respond to the narrative openings. For each story, available data include the full text of the freely accessible chapters and the respective comments left by readers at the paragraph level.[3] Each selected book underwent manual verification to ensure proper chapter alignment and content consistency between platforms.

For each story, available data include the full text of the freely accessible chapters and

1. Male Lead Stories (Nan Pin Xiao Shuo) often focus on the male protagonist's individual growth, adventure, conquest, and success, and feature harem elements (one male protagonist in romantic relationships with multiple female characters), while Female Lead stories (Nv Pin Xiao Shuo) are primarily written by women and reflect female aesthetic preferences and desires, emphasising interpersonal relationships. Though both categories share similar genres (fantasy, cultivation, science fiction, mystery, historical fiction, romance, etc.), male lead stories tends to feature more historical, Games, Eastern Fantasy, Horror, Urban stories, while female lead stories includes more urban romance, danmei (boys' love), and romance narratives where romantic plots serve as the primary storyline (China Writers Network 2021).

2. Based on the genre information provided by Web Novel for the stories. Differences in categorization and translation may exist for the Qidian Platform, as shown in the right side of the table.

3. Reader emotional response in this study refers only to readers' comments left in direct response to paragraphs, excluding replies to other readers' comments.

conference version

Web Novel Genre	Count	Qidian Genre	Count
Eastern[4]	46	Xuan Huan	37
Urban	18	Game	17
Fantasy	16	Xian Xia	14
Games	15	Sci-fi	11
Sci-fi	12	Urban	9
Action	1	Fantasy (Infinity)	7
Horror	1	Romance (Ancient/Modern/Xuan Huan)	6
War	1	Light novel	6
-	-	Wu Xia	2
-	-	Mystery	1
Total	110		110

Table 2: Platform-specific Genre Labels for the Corpus

the respective comments left by readers at the paragraph level.[5] The Chinese and English comments display both similarities and notable differences. The English dataset contains 268,619 comments, while the Chinese dataset includes over 2 million. We investigated the length of the comments in both languages. Figures 1 show the distribution of token [6] length of the comments in each language. Both datasets have a median token count of 7.0, with Chinese comments having a mean of 10.0 and English comments a mean of 11.0. However, the dispersion differs markedly: Chinese comments exhibit a standard deviation of 11.0, whereas English comments show an extremely high standard deviation of 126.0. We examined several long English comments and found out that they generally include both extensive expressions of opinions and occurrences of repeated words. Since the comments are conveying a meaningful message from the reader, they are not being removed from the dataset. This choice ensures that the dataset reflects the full diversity of commenting behavior, including both short, spontaneous reactions and longer, more elaborated discussions.

2.2 Data Processing

In the Qidian dataset, chapters do not contain recognizable paragraph delimiters. Chapter metadata includes partial information through *segmentId*, indicating the number of paragraphs per chapter, but no individual paragraph indicators. To address this, we manually examined the paragraph structuring of each story pair on two platforms. We confirmed that the number of paragraphs in each chapter is the same for stories on Qidian and their translation on WebNovel. This allowed us to determine the number of paragraphs for each chapter in all Qidian stories. Based on this, we segmented the chapter content based on the number of paragraphs in each chapter. We are aware that the method of paragraph segmentation in Chinese text presents challenges and may introduce inconsistencies. However, our subsequent

5. Reader emotional response in this study refers only to readers' comments left in direct response to paragraphs, excluding replies to other readers' comments.

6. Here, token means an individual English word or a Chinese character

conference version

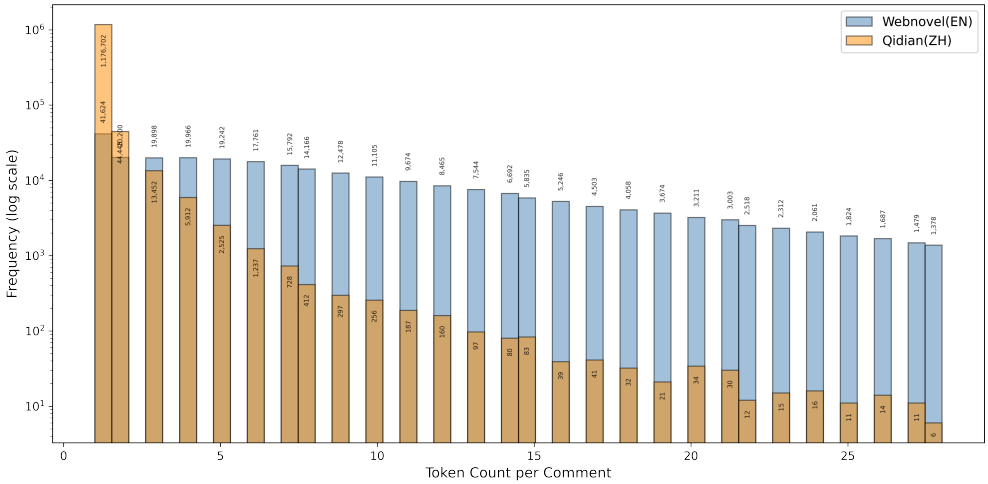


Figure 1: Length Distribution of Comments in both platforms
Note: The distribution of comment lengths is highly left-skewed. To preserve visibility of the full distribution without discarding information, the y-axis is displayed on a logarithmic scale.

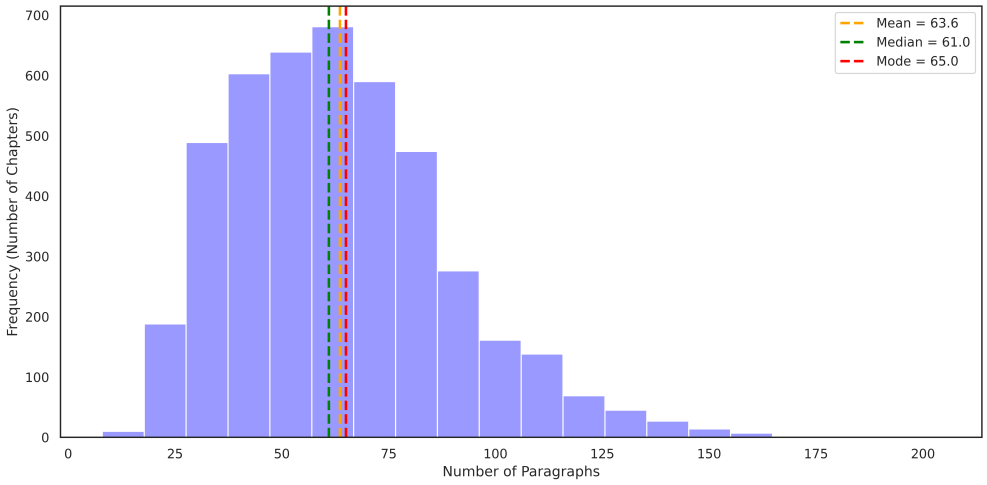


Figure 2: Number of paragraphs per chapter

sentiment analysis uses a sliding window that mitigates this problem by reducing sensitivity to individual paragraph boundary errors. Figure 2 shows the paragraph distribution per chapter for the stories.

2.3 Multilingual Sentiment Analysis

2.3.1 Manual Annotation

We adapted our annotation framework based on the semantic-based approach proposed by Mohammad 2016. This framework was refined to address the unique pragmatic characteristics of web novel reader comments, such as domain-specific vocabulary, and the conversational nature of online social reading platforms. The system utilizes a two-tier labeling structure: a primary tier for sentiment labels (Positive, Neutral, and Negative) and a secondary tier to capture more complex emotions, including *Playful/humorous/joking*, *Sarcasm/irony/mockery*, and *Role-play*.

conference version

The full annotation guideline is available at the project's GitHub repository.

The implementation followed a double-blind procedure involving two independent annotators fluent in both English and Chinese. The annotators independently annotated a sample of the dataset, strictly adhering to the annotation guidelines. The inter-annotator agreement was measured using Cohen's Kappa across two levels of granularity. For the main sentiment labels, the agreement reached 0.57 for the Chinese dataset and 0.53 for the English dataset as summarized in Table 3. When considering the combined main and sub-labels, the scores were 0.49 and 0.50, respectively. All obtained scores signify moderate agreement according to standard interpretation scales. The agreement level across both languages reflects the inherent linguistic challenges of web novel reading comments, where subtle pragmatic cues often complicate the definitive categorization of sentiment.

Dataset	**Total Sampled** (N)	**Annotated** (n)	**Cohen's κ (Main Label)**	**Cohen's κ (Main + Sub-label)**
Chinese (Qidian)	999	999	0.57	0.49
English (WebNovel)	999	981	0.53	0.50

Table 3: Summary of Manual Annotation Dataset and Inter-annotator Agreement (Cohen's κ). Note: The English dataset initially comprised 999 samples, but 18 comments were excluded as they consisted solely of a single emoji. This decision was based on the semantic instability of isolated emojis, which are prone to significant misinterpretation across platforms and cultures.

Several unique linguistic and pragmatic challenges emerged during the annotation process. One primary challenge was the pervasive use of domain-specific or culture-specific slang and memes. For instance, expressions such as " 老起誓人了" (literally means "veteran swearer" in Chinese) refer to someone's recurring habit of making empty promises. Without context knowledge, such phrases remain ambiguous or may be misclassified. The interpretation of emoticons and emojis also presented a challenge due to their inherent semantic instability. Miller et al. 2016 demonstrated that emoji renderings vary across platforms and cultural contexts, leading to disagreement rates as high as 25%. For the annotation dataset, we excluded 18 English comments consisting solely of a single emoji. This approach ensures that the dataset relies on verifiable linguistic context rather than isolated, ambiguous cues prone to subjective bias.

Furthermore, identifying sarcasm and irony proved demanding, as readers frequently employed a rather harsh joking manner to convey positive emotions. We also observe unique expression as readers writing a comment adopt character personas. For example, a reader wrote, "*Klein: If I knew this then, I would just leave you there*", and *Klein* here is the name of the main character. In such cases, our guidelines prioritized the sentiment conveyed in the role-played speech itself to maintain annotation consistency.

Based on the unique and difficult cases that two annotators identified in the first

conference version

round of annotation, we refined our annotation guidelines and conducted the second round of annotation, where comments with different labels were discussed, and a finalized gold annotation dataset was established.

2.3.2 Synthetic Data Generation

Due to the substantial time and resources demanded for manual annotation, our gold annotation dataset is limited in size. To address this limitation, we generated synthetic data via Large Language Models (LLMs) to eventually finetune a better performance transformer-based sentiment classifier. Synthetic data can augment real-world datasets by introducing controlled variations and increasing the diversity of training samples (Chang et al. 2024; Jaipuria et al. 2020).

Building on prior synthetic data generation frameworks, we propose a refined pipeline that improves the accuracy and computational efficiency of synthetic data generation. As the basis for generating synthetic comments, we randomly selected 100 paragraphs from story beginnings (approximately 10 per story) and 50 corresponding comments, which we divided into five groups of 10 comments each. We rotated these comment groups across prompts as examples to avoid biased comment generation patterns. The LLM was then instructed to read the story paragraphs and emulate the observed reader comment styles to generate synthetic comments. The prompts requested a balanced distribution of positive, negative, and neutral comments. Sublabels were not specified in the prompt, as they were not required for the generation task.

We conducted a preliminary qualitative evaluation of multiple LLMs, including both open-weight and closed-source options. Based on initial performance, we shortlisted `Kimi2`, `Gemma 3 27B`, `Stepfun 3.5`, `nemotron-nano-9b-v2`, and `Mistral 3`. A second qualitative evaluation round assessed generation quality and refined prompts, ultimately leading to the selection of `Gemma 3 27B` as the best model for generating synthetic English comments and `Stepfun 3.5` for Chinese.

Then, a two-tier quality control process is being set. The first tier is a post-generation filter where we asked the LLM to "double check" its generation and remove any data if the text-label mismatch. Subsequently, human annotators validated 100 randomly sampled comments per language to verify generation quality. The inter-annotator agreement scores between synthetic data and annotators were 0.99 for Chinese and 0.95 for English. Comments failing to meet quality criteria were discarded. The final augmented dataset includes 2,717 Chinese generated comments (positive: 954; neutral: 823; negative: 940) and 2,792 English generated comments (positive: 944; neutral: 859; negative: 989).

We have noted several limitations in the generation of synthetic comments. First, a distribution shift emerged between the original and synthetic data since we requested the model to generate balanced labels. Table 4 presents the sentiment distribution comparison: the generated data exhibits a notable divergence, particu-

conference version

larly in the Chinese sentiment distribution on negative (35% vs. 50% in the original data) and positive (35% vs. 20% in the original data). Second, over-reliance on prompt-specific patterns was observed; for instance, one sample comment in the prompt contained an emoji, and most generated comments incorporated an emoji, suggesting the model's sensitivity to surface-level features in few-shot examples. Lexical diversity metrics further indicated that the model prioritized memorization over generalization. Type-token ratios (TTR) for synthetic comments were substantially lower than human-authored texts (English: 0.10 vs. 0.25; Chinese: 0.034 vs. 0.088), revealing marked repetition of vocabulary and sentence structural patterns.

Dataset	Positive	Neutral	Negative	Total
Chinese Original	20%	30%	50%	
Chinese Synthetic	954 (35%)	823 (30%)	940 (35%)	2,717
English Original	26%	28%	46%	
English Synthetic	944 (34%)	859 (30%)	989 (36%)	2,792

Table 4: Sentiment distribution of human-annotated data and synthetic data

conference version

2.3.3 Finetuning and Evaluation

As discussed in the annotation section, we have noticed the unique pragmatic characteristics of web novel reader comments. To better conduct the sentiment analysis, we evaluated five multilingual pre-trained transformer models, specifically selected for their cross-lingual sentiment analysis capabilities. These models included: `bert-base-multilingual-uncased` (Devlin et al. 2019), `xlm-roberta-base` (Conneau et al. 2020), `roberta-base-multilingual-sentiment` (clapAI 2025a), `roberta-large-multilingual-sentiment` (clapAI 2025b), and `twitter-xlm-roberta-base-sentiment-multilingual` (Barbieri et al. 2022). This selection encompasses both pre-trained base models and fine-tuned models that were trained on a social media dataset. The baseline benchmark was conducted using 10% of the manually labeled dataset ($n = 100$ for Chinese and $n = 98$ for English). This subset serves as a high-quality Gold Standard to validate model performance prior to full-scale fine-tuning and evaluation.

`twitter-xlm-roberta-base-sentiment-multilingual` achieved the highest English zero-shot scores (Macro-F1 0.71, Kappa 0.57), while `roberta-large-multilingual-sentiment` led the Chinese results (Macro-F1 0.65, Kappa 0.46) (Table 5). Based on the evaluation result, we fine-tuned both `twitter-xlm-roberta-base-sentiment-multilingual` and `xlm-roberta-base`. The former was selected because of the superior zero-shot performance prior to fine-tuning, whereas the latter is to avoid interference from prior sentiment task biases (Phang et al. 2018). Further evaluation indicates that `twitter-xlm-roberta-base-sentiment-multilingual` achieves better performance on the fine-tuning task for this specific dataset. Our model achieved an overall macro F1 score of 0.73, with language specific scores of 0.75 for English and 0.71 for Chinese.

Language	Model	Acc.	Recall	Macro F1	Cohen's Kappa
English	`xlm-roberta-base`	0.48	0.33	0.22	0.00
	`bert-base-multilingual-uncased`	0.27	0.33	0.14	0.00
	`roberta-base-multilingual-sentiment`	0.53	0.54	0.53	0.28
	`roberta-large-multilingual-sentiment`	0.53	0.53	0.53	0.28
	`twitter-xlm-roberta-base-sentiment-multilingual`	0.72	0.73	0.71	0.57
	`qidian-webnovel-twitter-xlm-roberta-base-sentiment-multilingual` (ours)	**0.76**	**0.75**	**0.75**	**0.58**
Chinese	`xlm-roberta-base`	0.30	0.33	0.15	0.00
	`bert-base-multilingual-uncased`	0.30	0.33	0.15	0.00
	`roberta-base-multilingual-sentiment`	0.52	0.48	0.50	0.22
	`roberta-large-multilingual-sentiment`	0.67	0.63	0.65	0.46
	`twitter-xlm-roberta-base-sentiment-multilingual`	0.64	0.59	0.60	0.39
	`qidian-webnovel-twitter-xlm-roberta-base-sentiment-multilingual` (ours)	**0.72**	**0.72**	**0.71**	**0.634**

Table 5: Performance of 3-class sentiment classification by multilingual models on Qidian and WebNovel comments. Best results are in bold, second best are underlined.

conference version

2.3.4 Sliding Window

To construct sentiment arcs, we applied a rolling-mean smoothing approach (like Pianzola et al. 2020), choosing a window size of 51 paragraphs for the statistical tests (25 paragraphs before and after the target paragraph). This smoothing method was particularly important given that the stories in the corpus have editions in two languages: while translation should preserve paragraph structure and alignment, word choice differences related to language and cultural needs can alter emotional intensity or shift sentiment positioning relative to the original text. The rolling-mean approach mitigates these translation-induced variations and improves cross-language alignment by reducing localized inconsistencies. We experimented with multiple window sizes and found that small windows introduced excessive noise, whereas large windows over-smoothed the trajectory and obscured local variation. The window size of our choice (51 paragraphs) provided the best balance between preserving the overall arc shape and retaining meaningful local emotional fluctuations.

2.3.5 Dynamic Time Warping

To confirm our assumption that translation variation of the same story won't hinder our cross-platform comparison. We used Dynamic Time Warping (DTW) to align the sentiment sequences of stories. Dynamic time warping is a well-established technique for finding optimal alignment between two time-dependent sequences under specific constraints (Müller 2007; Wevers et al. 2021). Our analysis focused on the morphological characteristics of sentiment arcs, specifically the positioning of emotional peaks and valleys. This approach enables evaluation of whether critical emotional turning points in the original narrative are maintained through translation. In this implementation, we computed the DTW distance over the entire sentiment sequence without additional smoothing. The resulting normalized distance and corresponding similarity score capture the global structural alignment

of emotional progression across translations. We visualized the DTW warping paths by chapter progression to qualitatively assess alignment patterns and deviations.

2.4 Statistical Testing

2.4.1 Direct Reproduction

In the original study (Pianzola et al. 2020), linear regression has been used to examine the relationship between narrative sentiment and reader response. We first attempt a direct reproduction of the original findings using the same statistical test.

2.4.2 Generalized Additive Models

In addition to linear regression, we also test the main hypothesis using Generalized Additive Models (GAM). GAMs are a flexible regression technique that can capture non-linear relationships, meaning the model can detect patterns that change direction or strength across the story, rather than assuming a constant effect. This is especially valuable in modeling complex behavioral or linguistic patterns that may show complex associations (Winter and Wieling 2016). This methodological adaptation enables us to test the robustness of the original findings under less restrictive assumptions about the associations of variables and to gain additional insight into the narrative parts that are significantly associated with more positive or more negative reader response.

We fitted GAMs, incrementally adding predictors to explain variation in comments' sentiment as stories progress. We allowed patterns of associations between variables to differ between Qidian and WebNovel, and accounted for the fact that some books tend to attract more positive or negative comments overall, regardless of paragraph content.

3. Results

3.1 Cross-Platform Sentiment Arcs

3.1.1 Dynamic Time Warping Results

We calculated the dynamic time warping (DTW) distance between the chapter-level sentiment arcs of each Qidian-Webnovel book pair and compared these values against the mean sentiment difference computed for aligned chapters. As shown in Figure 3, the y-axis captures the similarity of the emotional progression across chapters, whereas the x-axis captures systematic shifts in overall emotional tone between versions. In an ideal situation, the original and translation pair should share the same sentiment arc and have no difference in mean sentiment.

Since no domain-specific benchmark exists for sentiment arc alignment, DTW

conference version

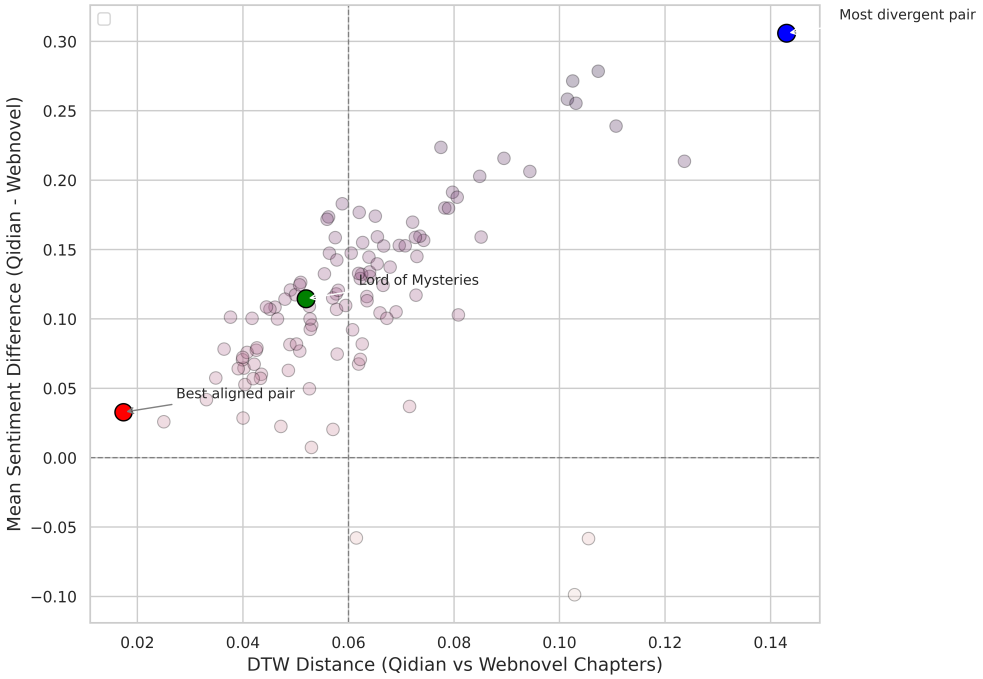


Figure 3: 2D map of sentiment arcs' structure similarity and mean sentiment score at chapter level.

distances are interpreted relative to the empirical distribution of this corpus. The majority of books cluster within a normalized DTW distance of 0.04 to 0.08, and mean sentiment differences between 0 and 0.25, suggesting that the sentiment arcs of story pairs are relatively consistent across platforms.

Lord of Mysteries serves as a particularly informative reference case. As a widely acclaimed novel with an active reader community, its Webnovel translation represents a high-effort rendering of the original Chinese text. The DTW score and the mean sentiment difference of this story are intermediate. Comparing the best aligned pair (DTW $\approx$ 0.015, mean sentiment difference $\approx$ 0.03) and the most divergent pair (DTW $\approx$ 0.15, mean sentiment difference $\approx$ 0.3), all three cases receive reader-assigned translation quality ratings of 4 out of 5 stars. Given that chapter segmentation and story content are held the same across platforms, this divergence in DTW distance is most plausibly attributable to differences in translatorial choices. The same narrative content is rendered through word selection, phrasing, and stylistic conventions, which cumulatively shift the emotional arcs slightly, which is detectable but falls below the perception of readers.

3.1.2 Sentiment Arcs of narrative and comments

Figure 4 shows the sentiment arcs of the narrative by paragraph and the corresponding comments (the mean sentiment score of all comments on a paragraph) with a sliding window of 25 for the story *Lord of Mysteries*. We can observe that the sentiment of the narrative progression in both platforms relatively keeps the same peak and valley, while the sentiment of the comments varies.

conference version

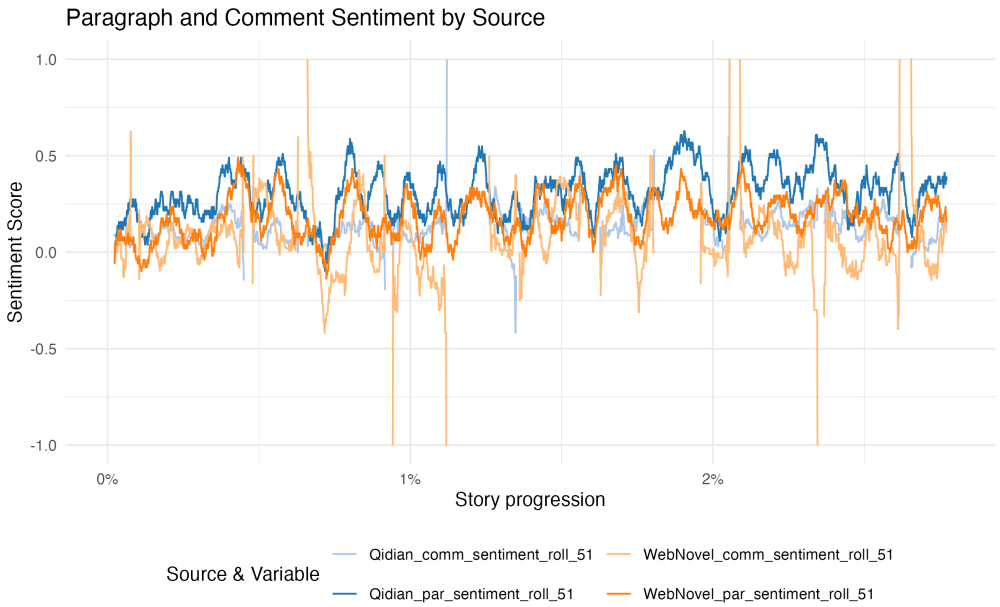


Figure 4: Example of sentiment arcs of the story *Lord of Mysteries* and its respective comments. The *x-axis* shows the narrative progression expressed as a percentage of the whole story.

3.2 Hypothesis Testing

3.2.1 Direct Reproduction

We started with a simple linear model, with *paragraph sentiment* as a predictor and *comment sentiment* as outcome. The association between variables is significant: $\beta = 0.05$, SE $= 0.002$, $p < .001$, $R^2 = 0.001$. This is a successful reproduction of the results in Pianzola et al. 2020, using a larger corpus, coming from two different digital social reading platforms, and in two different languages. There is a direct positive relationship between the variation of the sentiment of a paragraph and of the sentiment of the comments on it. However, it is important to note the very small R^2, meaning that the paragraph sentiment can only explain 0.01% of the variation in comment sentiment (under the assumption that their association is linear).

3.2.2 Generalized Additive Models

We built the GAM model incrementally, adding one component at a time — story progression, the interaction between sentiment and progression, platform differences, and book-level variation — to test whether each addition meaningfully improved the model's ability to explain comment sentiment. Each successive model component significantly improved fit (all $p < .001$), with the final model explaining 14.2% of the variation in comment sentiment (R^2).

The results confirm that comment sentiment broadly mirrors paragraph sentiment: sections with more positive paragraph sentiment elicit more positive comments,

6. We smoothed the arcs using a rolling mean of 51 paragraphs to reduce the noise in the visualization, instead of the more fine-grained 11-paragraph window used for the statistical tests.

conference version

and vice versa (Figure 5, Panel B). This pattern holds across both Qidian and WebNovel, though GAM reveals that the strength of the association varies over story progression, as indicated by the significant interaction terms. The observed data (Figure 5, Panel A) corroborate this pattern, although with considerably more noise due to the smaller sample sizes within each bin. Notably, the overall effect is modest in magnitude, with predicted comment sentiment ranging approximately between -0.15 and 0.10, suggesting that paragraph content is only one of many factors shaping reader responses.

An important limitation is that this analysis is restricted to heavily commented story sections, as the rolling average requires a minimum number of non-missing comment scores within each window. Approximately 77% of the paragraphs get excluded due to sparse commenting (less than 10 comments over a window of 51 paragraphs), and the distribution of paragraph sentiment differed significantly between included and excluded observations in most book – source combinations, though the absolute differences were small (median = 0.036). The GAM results therefore, allow to present a more nuanced description of the paragraph – comment sentiment relationship, showing how this is driven by sections where readers actively engage. Accordingly, the results may not generalize to less-commented portions of the narratives, which are nevertheless relevant to shape reader response.

3.3 Qualitative Analysis

3.3.1 High Engagement Chapters

As mentioned in the previous section, our GAM results characterize the non-linear relationship between narrative progression and high reader engagement. To better understand the reason for reader response eliciting narratives, we conducted a qualitative analysis of the five most commented chapters (chapters 1, 22, 27, 29, and 31) of a story sample. Close reading of these chapters revealed that they generally contain key narrative events, including world-building, character introductions, interpersonal conflicts, and plot twists commonly referred to as "face slapping" scenes[7] in web novel. Across both platforms, we observed that readers actively engage with strategically designed events, such as character confrontations, the appearance of female characters (remember these are all Male Lead stories), and plot twists. However, differences emerged in responses to non-critical narrative events. For instance, a description of a librarian generated substantial comments on Qidian because readers recognized an important Chinese figure who started his career as a librarian[8], sparking discussion, while receiving little to no response on WebNovel. Conversely, a citation of a classical Chinese idiom triggered substantial discussion on Qidian but drew no comments on WebNovel (examples shown in

7. In the web novel context, "Face Slapping" refers to the discovery that a plot event is the opposite of what one would expect. For example, when readers think the protagonist cannot achieve something, the protagonist suddenly demonstrate ability and strength to prove that it is achievable.

8. The important Chinese figure here refers to Mao Zedong

conference version

conference version

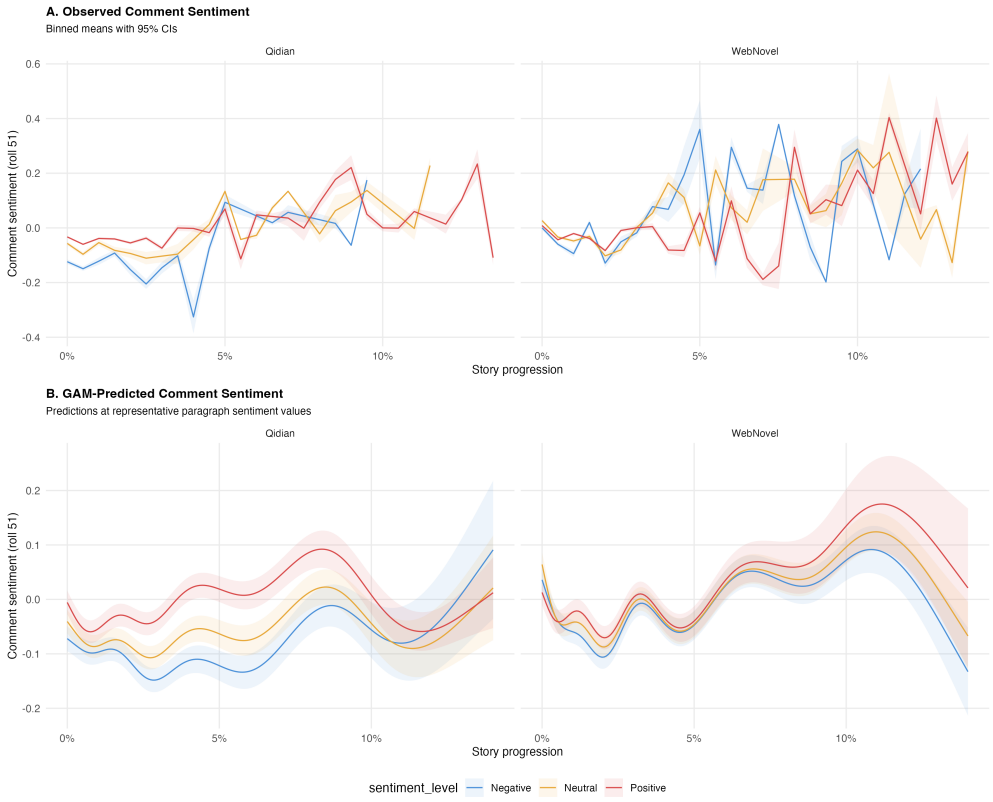


Figure 5: Observed (A) and GAM-predicted (B) comment sentiment across story progression, grouped by paragraph sentiment level (Negative = -0.1, Neutral = 0.0, Positive = 0.3). Observed values are binned means at 0.5% intervals. Predicted values are derived from the final GAM model, excluding book-level random effects. Shaded areas represent 95% confidence intervals.

Appendix A). These preliminary findings suggest that while readers actively engage with triggers of the plot progression, they demonstrate varied attention to other non-major events. This pattern may indicate that cultural or platform specific factors shape how readers emotionally engage with different narrative events.

4. Conclusion

This study successfully reproduced and extended previous work on narrative emotional valence and reader response in digital social reading environments (Pianzola et al. 2020), providing robust evidence for the cross-cultural validity of computational literary study methods. Through analysis of the beginning part (initial 3-10%) of 109 web novels across Chinese (Qidian) and English (WebNovel) platforms, we have shown that the narrative's emotion has an effect on reader response. Our implementation of Generalized Additive Models represents a methodological advancement over previous linear approaches, revealing more complex relationships that vary across narrative progression and platform contexts. The GAM analysis confirmed that the sentiment of narrative progression significantly influences reader response sentiment, explaining 14.2% of its variance, with the Chinese corpus exhibiting stronger correlations than the English, especially for the first 3-8% narrative progression. Our results are consistent with the notion that reader response likely mirror emotional patterns in the text (Appel et al. 2019; Tilmatine et al. 2024), however, the strength of this correspondence varies across narrative moments (Schmidt et al. 2023).

Our qualitative analyses of the chapters with high levels of engagement explain this correlation, as readers actively engage with content that is designed to elicit narrative tension and emotional reactions. The original findings in Pianzola et al. 2020 showed that Teen Fiction stories exhibited higher correlation and more harmonious trends. While our corpus contains different genres, they all fall under the broader category of "male lead stories" and have less genre distance between them compared to the original study's contrast between Teen Fiction and Classics. For this reason, we did not do any further genre categorization, in order not to decrease the statistical power of the fitted models. Our close reading of the narrative and the comments suggests that main narrative events like interpersonal conflicts and plot twists are associated with more active reader response.

5. Limitations and Future Steps

Several limitations should be taken into account when interpreting these results. First, our analysis was constrained to publicly available chapters, potentially limiting insights into reader engagement patterns across complete narrative arcs. Free opening chapters may strategically differ from subsequent paid content in their emotional intensity, pacing, and narrative techniques, as authors and platforms

conference version

optimize these chapters to attract and convert readers into paying readers (光明网 2023). Additionally, readers who engage with free chapters may represent a distinct subset of the platform's reader base whose commenting behavior differs from committed, paying readers. Free-chapter commenters may include casual browsers exploring multiple stories, readers sampling content before deciding whether to purchase, or readers facing economic constraints that prevent access to paid chapters. These distinct reader motivations may systematically influence how readers engage with narratives and express their responses in comments. This selection bias limits our ability to generalize findings about the relationship between narrative sentiment and reader response to complete works or to the full reader population. However, our focus on story beginning does enable systematic analysis of a critical narrative juncture where reader engagement is established, aligning with research on the importance of structural moments in narrative reception (Schmidt et al. 2023).

Second, the systematic sentiment differences between languages may reflect model bias rather than genuine linguistic variations, showing the necessity of language and context specific sentiment models with better performance. Future steps would involve expanding our annotation dataset to improve both the generation and evaluation of synthetic data, which will in turn enhance model performance. Third, we adopt *source platform* as the grouping variable in our GAMs due to the fact that we are analyzing only one (major) language per platform in our dataset. Consequently, the platform variable may represent a conflation of multiple underlying factors that require disaggregation in future investigations. The platform effect encompasses cultural context embedded within narratives, reader demographic characteristics, interface design features, and social interaction mechanisms, among other variables.

6. Data Availability

Data can be found here: `https://github.com/GOLEM-lab/Qidian-Webnovel-sentiment`.

7. Software Availability

Software can be found here: `https://github.com/GOLEM-lab/Qidian-Webnovel-sentiment`.

8. Author Contributions

Ze Yu: Conceptualization, Methodology, Formal analysis, Visualization, Writing – original draft, Writing - review and editing

Lanping Zhang: Methodology, Writing - original draft

conference version

Federico Pianzola: Conceptualization, Methodology, Formal analysis, Visualization, Writing – original draft, Writing - review and editing, Supervision, Funding acquisition

References

Appel, M., C. Schreiner, M. B. Haffmans, and T. Richter (2019). "The mediating role of event-congruent emotions in narrative persuasion". In: *Poetics* 77, 101385. 10.1016/j.poetic.2019.101385.

Barbieri, Francesco, Luis Espinosa Anke, and Jose Camacho-Collados (2022). "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual.

Bizzoni, Yuri and Feldkamp (2023). "Comparing Transformer and Dictionary-Based Sentiment Models for Literary Texts: Hemingway as a Case-Study". In: *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*. Association for Computational Linguistics, 219–228. https://aclanthology.org/2023.nlp4dh-1.25/.

Chang, Hsin-Yu, Pei-Yu Chen, Tun-Hsiang Chou, Chang-Sheng Kao, Hsuan-Yun Yu, Yen-Ting Lin, and Yun-Nung Chen (2024). *A Survey of Data Synthesis Approaches*. arXiv: 2407.03672 [cs.LG]. https://arxiv.org/abs/2407.03672.

China Writers Network (Mar. 2021). *From "golden finger" to sweet romance: What are the differences between male- and female-oriented fiction?* Accessed: 2025-07-13. https://www.chinawriter.com.cn/n1/2021/0317/c404027-32053332.html.

clapAI (2025a). *roberta-base-multilingual-sentiment: A Multilingual Sentiment Classification Model*. https://huggingface.co/clapAI/roberta-base-multilingual-sentiment.

— (2025b). *roberta-large-multilingual-sentiment: A Multilingual Sentiment Classification Model*. https://huggingface.co/clapAI/roberta-large-multilingual-sentiment.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://huggingface.co/FacebookAI/xlm-roberta-base.

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*. https://huggingface.co/google-bert/bert-base-multilingual-uncased.

conference version

Jaipuria, Nikita, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N. Murali (2020). *Deflating Dataset Bias Using Synthetic Data Augmentation*. arXiv: 2004.13866 [cs.CV]. https://arxiv.org/abs/2004.13866.

Koolen, Marijn, Julia Neugarten, and Peter Boot (2022). "'This Book Makes Me Happy and Sad and I Love It' : A Rule-Based Model for Extracting Reading Impact from English Book Reviews". In: *Journal of Computational Literary Studies* 1.1. 10.48694/jcls.104.

Kuijpers, Moniek M., Matteo Lusetti, Piroska Lendvai, and Simone Rebora (2024). "Validation of the Story World Absorption Scale through Annotation of Online Book Reviews". In: *Journal of Cultural Analytics* 9.1. 10.22148/001c.92531.

Mar, R. A., K. Oatley, M. Djikic, and J. Mullin (2011). "Emotion and narrative fiction: Interactive influences before, during, and after reading". In: *Cognition & Emotion* 25.5, 818–833. 10.1111/j.1745-6924.2008.00073.x.

Menninghaus, W., I. Schindler, V. Wagner, E. Wassiliwizky, J. Hanich, T. Jacobsen, and S. Koelsch (2020). "Aesthetic Emotions are a Key Factor in Aesthetic Evaluation: Reply to Skov and Nadal (2020)". In: *Psychological Review* 127.4, 650–654. 10.1037/rev0000213.

Menninghaus, W., V. Wagner, E. Wassiliwizky, I. Schindler, J. Hanich, T. Jacobsen, and S. Koelsch (2019). "What are Aesthetic Emotions?" In: *Psychological Review* 126.2, 171–195. 10.1037/rev0000135.

Miller, Hannah, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht (2016). ""Blissfully happy" or "ready to fight": Varying Interpretations of Emoji". In: *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM)*.

Mohammad, Saif M. (2016). "A Practical Guide to Sentiment Annotation: Challenges and Solutions". In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. San Diego, California: Association for Computational Linguistics, 174–179.

Müller, Meinard (2007). *Dynamic Time Warping. In: Information Retrieval for Music and Motion*. Springer, 69–84.

Nabi, R. L. and M. C. Green (2015). "The role of a narrative's emotional flow in promoting persuasive outcomes". In: *Media Psychology* 18.2, 137–162. 10.1080/15213269.2014.912585.

Phang, Jason, Thibault Févry, and Samuel R. Bowman (2018). "Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks". In: *arXiv preprint arXiv:1811.01088*.

Pianzola, Federico (2025). *Digital Social Reading: Sharing Fiction in the Twenty-First Century*. Paperback. The MIT Press. ISBN: 9780262550918.

Pianzola, Federico, Simone Rebora, and Gerhard Lauer (2020). "Wattpad as a Resource for Literary Studies: Quantitative and Qualitative Examples of the Importance of Digital Social Reading and Readers' Comments in the Margins". In: *PLOS ONE* 15.1, e0226708. 10.1371/journal.pone.0226708.

conference version

Rohrbacher, Katrin (2025). "Opening Worlds: Narrative Beginnings and the Role of Setting". In: *CCLS2025 Conference Preprints* 4.1. 10.26083/tuprints-00030149.

Schmidt, M.-L. C. R., J. R. Winkler, M. Appel, and T. Richter (2023). "Tracking emotional shifts during story reception: The relationship between narrative structure and affective responses". In: *Scientific Study of Literature* 12.1, 17–39. 10.61645/ssol.177.

Shao, JunRu (2024). "数字时代下对青少年网络小说付费阅读行为的研究". In: *新闻传播科学* 12.4, 1183–1188. 10.12677/jc.2024.124181.

Tilmatine, Mesian, Jana Lüdtke, and Arthur M. Jacobs (2024). "Predicting subjective ratings of affect and comprehensibility with text features: a reader response study of narrative poetry". In: *Frontiers in Psychology* 15. ISSN: 1664-1078. 10.3389/fpsyg.2024.1431764.

Wevers, Melvin, Jan Kostkan, and Kristoffer L. Nielbo (2021). "Event Flow: How Events Shaped the Flow of the News, 1950–1995". In: *Proceedings of the Conference on Computational Humanities Research 2021*, 62–76. http://ceur-ws.org/Vol-2989/long_paper16.pdf.

Willis, Ika (2017). *Reception*. 1st ed. Taylor and Francis.

Winter, Bodo and Martijn Wieling (Feb. 2016). "How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling". In: *Journal of Language Evolution* 1.1, 7–18. ISSN: 2058-4571. 10.1093/jole/lzv003.

Yu, Ze, Federico Pianzola, and Emin Tatar (Dec. 2025). "Qidian-Webnovel Corpus: A Dataset of Chinese Web Novels with Multilingual Reader Response". In: *Journal of Open Humanities Data*. 10.5334/johd.368.

光明网 (July 16, 2023). 网络小说和短视频的"长"与"短". https://news.gmw.cn/2023-07/16/content_36698429.htm (visited on 07/09/2025).

conference version

A. Appendix - Sample Chapter with High Engagement

Example 1:

English: "The callous heaven regard all beings as nothing more than straw dogs..."; Comment count: 3

Chinese: "天地不仁，以万物为刍狗....."; Comment count: 195

The meaning of this old saying has two prevailing interpretations. Qidian comments featured discussions about these interpretations, while WebNovel showed no discussion even when footnotes explained the meaning of the sayings.

Example 2:

English: "Could this be the gift pack for transcenders? A library? Damn it, I was also a librarian my previous life. Am I to continue on with this occupation in this life as well?"; Comment count: 11

Chinese: "这难道是穿越众的大礼包？图书馆？尼玛，我上辈子是图书管理员，不会到了这个世界，还是一样吧！"; Comment count: 53

conference version

jcls. Journal of Computational Literary Studies

conference version

The Anatomy of the Online Book Review

Marijn Koolen [1]
Joris J. Van Zundert [1,2]
Peter Boot [2]
Silvia Lilli [4]
Katja Tereshko [3]

1. DHLab, Humanities Cluster, Royal Netherlands Academy of Arts and Sciences (KNAW), Amsterdam, The Netherlands.
2. Computational Literary Research, Huygens Insitute, Amsterdam, The Netherlands.
3. Vrije Universiteit Amsterdam, Amsterdam, The Netherlands.
4. Università di Roma Tor Vergata, Rome, Italy.

OPEN ACCESS

Citation
Marijn Koolen, Joris J. van Zundert, Peter Boot, Silvia Lilli, and Katja Tereshko (2026). "The Anatomy of the Online Book Review". In: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7995

Date published 2026-05-05 (preprint)
Date accepted tbc
Date received 2025-11-26

Keywords
computational literary studies, online book reviews, reading impact, reader response, content 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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. With the availability of massive amounts of data, online reviews have become a promising source for detecting readers' engagement with fiction. Yet identifying expressions of response to reading (such as emotions, evaluations, and feelings) remains challenging for automated analysis, given the unstructured nature of reviews. This study proposes a method for categorizing review content in order to distinguish between sentences referring to readers' experiences and those addressing other book-related aspects, such as plot, characters, or author. We first designed an annotation schema, after assessing previous classifications. Then, we manually annotated according to our schema the sentences from 1,400 Dutch fiction reviews. Subsequently, we trained a robBERT2023 classifier to automate review sentence classification and applied it to a corpus of 670,751 reviews. Our analysis allows us to draw conclusions on the composition of online reviews and correlations among their components.

1. Introduction

For the field of literary studies, online book reviews offer unprecedented access to information on how readers experience books (Rebora et al. 2021). Even though there are important concerns with respect to their authenticity, representativeness, and immediacy (Hu et al. 2024; Koolen et al. 2020), the volume and variation of reviews make it feasible to empirically study reading 'in the wild'. Accordingly, researchers have tried to come to grips with the varying content of online reviews, just like they have studied the content and processes related to more conventional reviews in e.g. newspapers (Chong 2020).

A large number of papers has looked at online book reviews in order to answer research questions in areas as diverse as literary studies (Koolen et al. 2020; Walsh and Antoniak 2021), reader studies (Pianzola 2025, Ch.3), computer science (Bartl et al. 2024) and marketing (Chevalier and Mayzlin 2006). In this paper, our focus is on the reviews themselves. We believe that the field of computational literary studies needs a better understanding of the online review as a genre in its own right before we can draw conclusions from what happens in the reviews and before fruitful computational processing is possible. If we know what type of content we are likely to encounter in an online

book review, how this content depends on, for instance, the genre of the book reviewed and the platform on which the review is published, and if we know which sentences in the review can be assigned to which type of content, then we can better understand the review as a cultural product, closely linked to the reception of the fictional text and the mode of experiencing reading. It will also help us build more sophisticated computational tools to analyse the reviews. We see our contribution as a step towards this better understanding of the review. We have no radically new findings to report, but point out various relationships that must be taken into account when analysing reviews.

An example of a tool that could benefit from a deeper understanding of review content is the Review Impact model (Boot and Koolen 2020). This model sometimes incorrectly finds reading impact in sentences that do not contain any reader response. It may, for instance, find evidence of aesthetic impact because the reviewer quoted a particular emotional outcry by a character as part of a description of the narrative. Obviously, this emotion should not readily be attributed to the reader-reviewer. If we can better distinguish reader response from other content, such as a description of the narrative, we can better extract different types of reading impact.[1]

So our main question is whether we can decompose online book reviews into statements related to various aspects of the reading experience, such as book content, reader response and comparison with other works or authors. The paper addresses the following, more specific, research questions:

- RQ1: Is it possible to define a succinct set of content categories to meaningfully capture different aspects mentioned in online book reviews?
- RQ2: Can human annotators reliably annotate these different aspects of book reviews?
- RQ3: Can we train automated taggers to accurately annotate fiction reviews?
- RQ4: How are the types of review statement distributed within and across reviews?

In order to answer these RQs,[2] we set up a process of manual annotation of sentences in online book reviews, which served as the basis for training an automatic model for this task. This work was done in the context of the Impact & Fiction project, in which we analyse the expression of reading impact in Dutch-language reviews and relate them to the textual characteristics of novels in Dutch (both originally written in Dutch and translated to Dutch). For this reason, we select reviews for annotation from this corpus of Dutch-language reviews. We note that any specific language selection limits generic insights and lessons we can draw from the analysis, as languages not only influence how thoughts are expressed, but also introduce specificities from the culture of their speakers. This paper reports on the annotation effort and on what we have learned from it. We focus on online reviews of fiction novels and thus ignore reviews of non-fiction. Although the term 'book' is broader than 'novel', we use the terms interchangeably. In

1. Of course, all content of the review is written from the reader's perspective, and the description of the narrative will necessarily also show aspects of the reader's way of processing the story. Theoretically, we cannot 'factor out' the reviewer. That doesn't mean that every sentence expresses the reviewer's opinion.
2. A fifth research question we set aside for a future article: Are there differences in the composition of reviews between different fiction genres and in the reviews posted on different review platforms?

conference version

2 we discuss earlier annotation efforts and why we chose to develop our own set of categories.

We developed this annotation schema within the context of the Impact and Fiction project [3]

2. Related Work

For a fuller overview of studies into online book reviews and other forms of digital social reading, we refer to Rebora et al. (2021) and Pianzola (2025). Our aim is to develop a schema that is *comprehensive* in the sense that it can be used to categorise all content of a review, *relevant* to literary studies (so no categories for aspects of e.g. book purchase and shipment as found in reviews on websites of online booksellers, see Dimitrov et al. 2015; Koolen et al. 2020). Furthermore, we want the schema to be *succinct*. Very fine-grained schemas have categories that are very rare or highly subjective, which are hard to annotate and agree on, thereby multiplying the effort required. Below we present a short summary of studies that introduce annotation schemas of book review content which identify the most commonly distinguished categories. A more extensive overview can be found in Appendix A.

2.1 Literature review

A comprehensive descriptive annotation scheme for online book reviews was developed by Kutzner et al. (2021), in which they annotated 282 book reviews in German posted on Amazon to study the types of content, using a schema consisting of 65 categories. Another comprehensive schema is represented by Mehling et al. (2018). Their annotation scheme focuses on what happens in an online review. It contains 153 categories and they annotate 507 Amazon reviews.

These schemas show a number of annotation categories relevant for our purposes, which overlap with our final annotation schema. They make a distinction between descriptions of a novel's *content* (differentiating between plot, characters and themes), aspects of *style* (language, structure), statements about the *author*, references to other works, statements of *evaluation*, *emotion*, *identification* and of *recommendation*.

A more succinct schema was developed by Milota (2014). Her schema also distinguishes between *content* aspects (*plot*, *setting*, *characters*, *moral/message/theme*) from aspects of response (*evaluation*, *emotional response*, *personal tie-in*) as well as comments about the *author* and references to other works (*intertextual references*).

Álvarez-López et al. (2017) provided a dataset of 300 reviews with 2977 sentences labelled for various categories of aspect based sentiment mining. They annotated sentences of 40 randomly selected novels from the Amazon/LibraryThing corpus provided by the Social Book Search Lab (Koolen et al. 2016). They annotated polarity (positive, negative, or neutral) of aspects ("targets") in 13 categories (characters, plot, general, author, genre, title, audience, quality, structure, period of story, period of publishing, length, and price).

3. Impact & Fiction is a three-year research project funded by the Netherlands eScience Center (see acknowledgements for more details). See more info at https://impactandfiction.huygens.knaw.nl/about-page/.

conference version

conference version

Op de Beek (2014) and Linders (2014) conducted an analysis of the aspects in 734 fiction reviews of professional critics published in five Dutch newspapers. They based their analysis on a method developed by Praamstra (1984), who distinguishes between statements about the book (and the author) and statements about different matters. The former can be categorised in three groups: descriptive (e.g. the novel contains twelve chapters), interpretive (e.g. this novel is about hope) and evaluative. The descriptive category includes aspects of narrative and characters as well as of a novel's organisational structure (e.g. number of chapters).

Rebora and Vezzani (2024) annotated 100 reviews from Goodreads of which 89 eventually were used for actual analysis. Their interest is specifically in evaluation: whether a sentence is evaluative, what criteria are used in the evaluation and what the evaluation is based on. They build their tagset on von Heydebrand and Winko's theory (Von Heydebrand and Winko 1996, 2008) and interpret online reviews as acts of "linguistic evaluation" which require an explicit "standard of value" as well as "certain categorizing assumptions" . In their tag set they devoted special attention to criteria based on a reader's individual thoughts, feelings, and experiences. Sentences are labelled as mentioning "personal experience", "cognitive processes", and "emotions". Next to this they classify sentences into the categories of "aesthetic", "social", and "generic evaluation".

Specifically interested in emotional engagement, M. Kuijpers et al. (2023) provided a dataset of 493 curated reviews from Goodreads totalling 2000 statement level annotations. They particularly aimed at identifying absorption related statements in five main categories (attention, emotional engagement, mental imagery, transportation (of self), and impact) and several subcategories.

Finally, our own Reading Impact models for Dutch (Boot and Koolen 2020) and English (Koolen et al. 2022), which are rule-based models to identify expressions of reading impact in sentences, have categories for general affect and for narrative and stylistic impact. In these models, the focus is on identifying reader response, and in the case of narrative and aesthetic feelings, the expressed response has an explicitly target of either narrative elements (plot, characters, events) or stylistics elements (word choice, use of metaphor, text structure).

There are a few common denominators in these schemas. One is the distinction between (1) the novel's intrinsic characteristics (2) that of the reader's response to it, (3) comments about the author, (4) references to other works, (5) statements of classification (placing a novel in the broader context of genres, period and literary traditions), and (6) other statements. Among the intrinsic characteristics, several schemas distinguish between aspects of the story (e.g. plot/storyline, characters) and aspects of style such as language and structure (Álvarez-López et al. 2017; Boot and Koolen 2020; Kutzner et al. 2021; Mehling et al. 2018; Milota 2014). Among reader's response characteristics, several schemas distinguish between statements of evaluation, identification, immersion and reflection (M. Kuijpers et al. 2023; Kutzner et al. 2021; Mehling et al. 2018; Milota 2014). Finally, several schemas include a separate category for recommendation (Kutzner et al. 2021; Mehling et al. 2018).

conference version

3. A Model for Annotating Review Aspects

We first describe how we developed the review categories and the coding scheme, then the annotation process.

3.1 Coding scheme

We iteratively developed the coding scheme over a period of six months, between April and September 2024. We started with a comparison of several existing coding schemes (see Section 2 and Appendix A) with a coding scheme we developed for an earlier, unpublished experiment. With the additional coding schemes in mind, we refined our coding scheme using multiple iterations of annotating review sentences and gathering examples per category, in which three of the authors individually annotated a small set of reviews in different languages, including Dutch, English and Italian. In plenary sessions we discussed the categories and examples and added, deleted, grouped and modified categories until we reached agreement on the relevance and interpretation of each category.[4]

We decided to annotate at the level of the sentence. While it would clearly be possible, in many cases, to assign one type of content to one part of a sentence and another type of content to another part of the sentence, we felt this would complicate further processing and be of limited value. We also decided that each sentence could be assigned to any number of categories. We initially made a distinction between the 'main' category of a sentence and 'additional' categories, but decided against it as we found that it proved to be a highly subjective decision, unnecessary for our purposes and in many cases difficult or impossible to determine.

Main (top-level) categories We were interested in a simple descriptive schema to identify the elements that could be found in the reviews, distinguishing three main aspects: book, author, and reader related. The schema curation followed an inductive process, combining abstraction from close reading of reviews, and then again a verification of the efficacy of the classification on the set of testing reviews. The scheme we developed was hierarchical, to keep granularity without losing precision.

Regarding book aspects, we derived three main categories: metadata (bibliographic information and comments regarding the book as a physical object), classification (literary, topographical and chronological context) and content (explicit and implicit contents). For authorial aspects, we considered two main categories: author (as individual) and style (writing aspects). For reader's aspects we considered two main categories: reader response (context of reading and all kinds of reaction to the reading act) and recommendations (general or specific to some group).

Subcategories Within each of these seven main categories, we then listed all the possible subcategories we could derive (e.g., for the first category 'metadata': title, author, translator, publisher, edition, price, source, reading practice), finally obtaining 50 subcategories in total. Because of the high number of alternative categories, and the

4. The full coding scheme is available at https://github.com/impact-and-fiction/JCLS-2026-review-composition/blob/main/Annotation-categories-book_reviews.pdf.

frequent overlap of many of them, we reduced this to 21 subcategories distributed in the seven main groups. Our main interest was not in a granular classification of reviews' contents, but rather in a clear separation between what refers to book content and what refers to reader's experience, especially when dealing with expression of feelings and emotions. However, we were also interested in a flexible and broader schema, so that the annotation model could serve for other analyses not specifically related to reading impacts.

We adopted a hierarchical schema. For reader response aspects, we defined a set of 6 subcategories (*evaluation of quality*, *feelings*, *identification and immersion*, *reflection*, *reading context*, *reception*). We also kept the subdivisions for content (*narrative*, *quote*, *theme*, *other*) and style (*stylistic features*, *context*, *structure*), as we considered these elements to be particularly meaningful in determining reading impact. Finally, we further refined the list, dropping the metadata category (which contained mostly redundant information derivable from ISBN), merging mentions of other characteristics such as length, cover etc. in the 'content: other' category, and separating 'author' and 'other works'. This last category includes references to other works and adaptations by the same or a different author, for filtering evaluation and response referring to these other works.

As we developed the categories, we added scope notes and examples to illustrate positive and negative examples. Many example sentences fit multiple categories, so it was clear that a single sentence could capture multiple aspects from different categories.

We refer to appendix B for the full coding scheme. Table 1 summarizes the categories.

Top level category	Lower level category
Author	-
Classification	-
Content	Narrative
	Other
	Quote
	Theme
Other works	-
Reader response	Evaluation of quality
	Feelings
	Identification and immersion
	Reading Context
	Reception
	Reflection
Recommendations	-
Style	Context
	Structure
	Stylistic features

Table 1: The annotation scheme.

3.2 Review sampling

For the annotation phase we wanted a diverse set of reviews that would allow us to study the anatomy of reviews from different platforms and for books of different genres. The reviews are sampled from a large dataset of almost 700,000 reviews written in Dutch (Koolen et al. 2024). We excluded all reviews of non-fiction books. We used the genre

conference version

classification introduced by Zundert et al. (2022), which is a mapping of NUR codes to 11 genres (see D for more details).

Readers post reviews on various platforms and the platform may affect what readers write in their reviews (Dimitrov et al. 2015; Hu et al. 2024; Koolen et al. 2020, 2023). Therefore, we include reviews from three different platforms: Goodreads (international) and Hebban (Dutch) both focus on book cataloguing, rating, reviewing and discussion, while Bol (Dutch) is an online webshop selling books and many other products.

We also included reviews from NBD Biblion,[5] a service providing acquisition information for Dutch public libraries. Their reviews are written by professional (paid) reviewers and adhere to a (for us unknown) set a criteria, including length (all reviews are between 100 and 200 words).

Rebora and Vezzani (2024) use a lower bound of 200 words to remove "extremely short" reviews. The advantage of focusing on longer reviews is that fewer reviews are needed to capture a broad range of evaluative and non-evaluative statements. However, we note that in a set of 15 million Goodreads reviews crawled without length restriction only 23% of all reviews are 200 words or more (Wan and McAuley 2018; Wan et al. 2019). In a similar set of 29 million Amazon reviews (Hou et al. 2024), only 11% of reviews are longer than 200 words. The consequence of using a minimum length of 200 words is that this excludes the majority of reviews (77% and 89% of Goodreads and Amazon reviews respectively) and possibly a large fraction of reviewers who never write more than 200 words. To be able to say something about the composition of online book reviews in general, we need a more representative sample, and thus to include shorter reviews.

We use a minimum review length of 10 words, which resulted in excluding just over 8% of all reviews. Reviews shorter than 10 words may contain evaluative statement, but often just mention a star rating ("I give this book 4 stars." or just "4 stars.") or an incomplete or incomprehensible statement.

For the full list of criteria and sampling steps, see Appendix D.

The resulting selection of reviews has imbalances in the relation between platform, genre and rating. The reason is that in the corpus of reviews these dimensions are associated with each other, e.g. fiction genre is related to platform. Goodreads has a relatively high percentage of literary fiction reviews (64%) compared to Bol (41%), Hebban (37%) and NBD Biblion (40%) while Goodreads has virtually no reviews of regional fiction (0.09%). The other platforms have still low, but much higher percentages. On NBD Biblion the fraction (0.8%) is an order of magnitude higher. Up-sampling the number of reviews of regional fiction will introduce the same association between genre and platform in the selection. If we were to create a completely balanced review selection, each combination of platform, genre and rating, we would have to limit it to the combination with the lowest number of reviews, which would be one or just a few reviews. As a consequence, in analysing the resulting annotations, we need to consider this imbalance.

5. https://www.nbdbiblion.nl

conference version

3.3 Annotation

Research question 2 asks whether our annotation categories can be reliably applied on a corpus of reviews. To address this question, we hired three annotators to annotate the sampled reviews using the coding scheme introduced in the previous section. We used the sentence as the unit of analysis. A single sentence can discuss multiple aspects of a book or the reading experience or something else. Therefore, a single sentence can be assigned to multiple categories. For example, the sentence *Het verhaal is moeilijk te volgen door de vele personages die allemaal een eigen agenda hebben.* (EN: "The story is hard to follow because of the many character who all have their own agendas.") is categorised as both *Content – Narrative* and *Reader response – Evaluation of quality*.

This is different from Rebora and Vezzani (2024) who focused on whether a sentence is evaluative or not, and if so, what type of evaluation it expresses. They assigned each sentence to exactly one class to be able to train a classifier, but noticed that for some evaluative sentences it is hard to reach agreement, as multiple aspects of evaluation are expressed.

We use the Trankit (Nguyen et al. 2021) sentence tokenizer to split the text of all reviews into sentences. For annotation we used INCEpTION (Klie et al. 2018),[6] providing each review as a document consisting of a list of sentences.

All review sentences were annotated by three trained annotators between December 2024 and March 2025. The annotators were students of programs related to Literary Studies who were hired and compensated for their work. They first participated in an introductory session in which the annotation categories, the nature of the task, and the annotation tool INCEpTION were explained and discussed using several example reviews.

Based on the available budget and the estimated annotation speed (≈4 minutes per review), we calculated that three annotators could annotate 1400 reviews, amounting to 278 hours of annotation in total (93.33 hours per annotator, 120 hours in total including breaks).

The annotators then completed a training phase of 150 reviews (December 2024) to familiarize themselves with the material, the annotation schema, and the tool. After the training phase, questions and disagreements were discussed with the authors, and the annotation guidelines were refined.

After that, the annotators worked on the target set of 1250 reviews (January–March 2025). To monitor reliability and resolve difficult cases, three further meetings were held at roughly monthly intervals. At the first meeting, questions mainly concerned long quotations, book annotations, or reviews shifting from book discussion to personal reflection in form of stories barely related to the book any more. Later meetings contained fewer such issues. As a result of these meetings, the annotators could change their annotations. Annotators could in addition raise questions directly with the authors, though they rarely needed to do so. In some cases, short reviews that consisted mainly of links or references remained unannotated. This process resulted in a ground-truth dataset of annotated review sentences. Unfortunately, we did not think to ask the

6. Version 23.2, see https://github.com/inception-project/inception/releases/tag/inception-23.2.

conference version

annotators to adjudicate all their disagreements until our budget for annotation work was used up, so they did not revisit and resolve disagreements. As a consequence, we do not have a curated set of annotations, which in turn has consequences for the machine learning phase (see Section 5).

4. Agreement

Category	Freq.	Kappa
Author	1,328	0.75
Classification	343	0.61
Content	6,197	0.70
Narrative	5,339	0.69
Other	259	0.56
Quote	338	0.88
Theme	172	0.40
Other works	612	0.60
Reader response	5,798	0.73
Evaluation of quality	3,182	0.64
Feelings	778	0.25
Identification and immersion	320	0.53
Reading Context	603	0.56
Reception	83	0.66
Reflection	997	0.42
Recommendations	243	0.73
Style	1,110	0.58
Context	11	0.08
Structure	58	0.35
Stylistic features	962	0.58

Table 2: Inter-Rater Reliability of agreement in terms of Light's variant of Cohen's Kappa, per category. The frequency is in number of sentences, based on majority vote.

Because all three annotators annotated all sentences, we have a fully-crossed design (Hallgren 2012), which makes Fleiss' Kappa (Fleiss 1971) inappropriate, as it assumes each item to be rated by a random sample of n raters from a larger sample of raters. Following Hallgren (2012), we use Light's variant of Cohen's Kappa (Light 1971), where the overall Kappa score is the arithmetic mean of the Cohen's Kappa scores between each pair of raters. We report the results in Table 2.

Following Landis and Koch (1977) for interpreting these values, we find agreement is substantial for the top level categories, except for Style, where agreement is moderate. For the lower level categories, the results vary. For the category of Quotes, agreement is almost perfect, for the category of Context (under Style), agreement is slight.

We note that the agreement score for evaluation of quality ($\kappa = 0.64$) is the same as in Rebora and Vezzani (2024).[7] Although they annotated reviews written in English, the similar agreement scores suggests a similar level of intersubjectivity of identifying evaluation in review sentences in both languages.

Overall, the agreement scores suggest that the main categories can be reliably annotated

7. Although Rebora and Vezzani uses the inverse category "no evaluation", it still comes down the same level of agreement.

conference version

by a group of annotators, but that care is needed for the lower-level categories.

Finally, we note that it could be useful to test whether an LLM can act as an additional annotator to augment the dataset by annotating additional reviews. If an LLM can reliably annotate review sentences, it could lower the effort needed to create ground truth annotations for different languages.

5. Training a Classifier

Our next research question (RQ3) is whether we can train a classifier to automate review sentence classification and investigate to what extent each (sub-)category can be learned. Since we did not ask the annotators to curate their annotations, we need some way to merge the annotations of the individual annotators into a single decision per sentence and per (sub-)category. This can be done in different ways, e.g. mapping the decisions of the three annotators into a binary decision that a sentence is assigned to a category (1) or not (0), or we could represent it as a degree to which the sentence belongs to the category, using the number of annotators who assign it to the category as the degree, ranging from 0 to 3.

For ease of analysis, we turn the annotations into binary decision using majority vote, so that a sentence is assigned to a category if at least two annotators assigned it to the category, otherwise it is not. The number of sentences assigned to each category is shown in the second column in Table 2.

For training a classifier, we treated the multiple annotations per sentences as a multiple classification task. Although multi-classification can be trained as a single task, we chose to train individual classifiers per (sub-)category and train them in a multi-task training framework (Caruana 1993, 1997) where a single model is trained on multiple tasks simultaneously and model parameters are shared between tasks. For this we used MaChAmp (Goot et al. 2021), which is a flexible toolkit for multi-task training built around the AllenNLP Python library (Gardner et al. 2018) based on PyTorch (Paszke et al. 2019).

With only 11 occurrences in the entire data, we excluded *Style - Context*, as there are not enough positive examples for proper training, validation and testing.

We use a train/dev/test split of 60/20/20 to make sure that the less common sub-categories still have enough sentences in the development and test sets to reliably measure performance. We used both the MaChAmp-default model BERT multilingual base (mBERT) (Devlin et al. 2018)[8], the 2023 version of robBERT (Delobelle and Remy 2023)[9] and the larger XLM-RoBERTa model (Conneau et al. 2019)[10] for training to compare their performance. We include robBERT-2023 because it is one of the most recent and best performing models for Dutch, and XLM-RoBERTa because it has been shown to be highly effective for many tasks on many different languages. Both mBERT and XLM-RoBERTa are widely used.[11]

8. https://huggingface.co/google-bert/bert-base-multilingual-cased
9. https://huggingface.co/DTAI-KULeuven/robbert-2023-dutch-large
10. https://huggingface.co/FacebookAI/xlm-roberta-large
11. At the time of writing, the HuggingFace website reports over 10 million downloads per month for BERT multilingual base and over 4 million for XLM-RoBERTa.

conference version

Category	Precision	Recall	F_1	Support
Author	**0.91**	**0.91**	**0.91**	282
Classification	**0.88**	**0.87**	**0.87**	62
Content	**0.89**	**0.89**	**0.89**	1224
Narrative	**0.88**	**0.88**	**0.88**	1064
Other content	**0.94**	**0.83**	**0.87**	61
Quote	**0.92**	**0.78**	**0.84**	62
Theme	**0.78**	**0.74**	**0.76**	25
Other works	**0.87**	**0.83**	**0.85**	120
Reader response	**0.88**	**0.88**	**0.88**	1142
Evaluation of quality	**0.87**	**0.87**	**0.87**	661
Feelings	**0.76**	**0.74**	**0.75**	147
Identification and immersion	**0.86**	**0.83**	**0.84**	53
Reading Context	**0.90**	**0.82**	**0.85**	128
Reception	**0.95**	**0.84**	**0.89**	25
Reflection	**0.72**	**0.66**	**0.68**	187
Recommendations	**0.90**	**0.93**	**0.91**	46
Style	**0.89**	**0.85**	**0.87**	232
Structure	**0.75**	**0.73**	**0.74**	15
Stylistic features	**0.89**	**0.86**	**0.87**	196
Weighted average	**0.88**	**0.87**	**0.87**	5732
Macro average	**0.87**	**0.83**	**0.84**	5732

Table 3: Evaluation results of RobBERT-2023 for the (sub-)categories. Support is the number of positive examples in the test dataset.

For all three models, we use the default MaChAmp parameters settings, including a learning rate of 0.0001, a batch size of 32 and train them for 20 epochs. The evaluation measure used for optimising between epochs and for testing is F_1. Although accuracy is a common measure for classification, the task in our case is predicting the presence of a category (1 for presence, 0 otherwise), leading to unbalanced datasets for most categories, with the majority of sentences assigned to the zero class. Accuracy pays attention to both true positives and true negatives, which would put most of the weight on the zero class, whereas we mostly care about the few times that the category is present. Instead, F_1 combines both precision and recall, which focus on the true positives and false negatives.

For reasons of space, we only show here the evaluation results for the best performing model, robBERT-2023, in Table 3. A table with scores for all three models is in Appendix E. Per category we report precision, recall, F_1 and support, which is the number of positive example of that category in the ground truth test data. The model performs well on almost all categories, and even on the most difficult categories, *Theme*, *Feelings*, *Reflection* and *Structure*, performance is around 0.7. This means that all categories are learnable to some extent.

For many categories, the F_1 score is substantially higher than the Cohen's κ agreement. What this means is that, although human annotators disagree regularly, the majority vote corresponds to a stable signal that can be learned by a BERT-based model. It is

conference version

possible that for these categories, two of the three annotators tend to agree with each other, and the third annotator deviates, leading to low overall agreement, but the model learns to follow the behaviour of the two agreeing annotators, leading to a high F_1. An alternative explanation is that the cases where at least two out of three annotators agree are the clearest cases of the occurrence of a category, and the majority vote binarisation retains only these clear examples, giving the model a clear and learnable target. An analysis of the agreement per pair of annotators shows that for most categories, the highest agreement between pairs is still lower than the F_1 score, so we speculate that it is more likely that the majority vote represents the clearest cases that form an easier target to learn.

For all six main categories, the F_1 score for robBERT-2023 is at least 0.85, suggesting the main categories can be automatically annotated with close to human-level performance.

Of the twelve sub-categories, there are four for which performance is below 0.8, namely *Theme*, *Feelings*, *Reflection* and *Structure*. We speculate that this is partly because these are infrequent categories and partly because they are more abstract than e.g. direct quotes, statements of evaluation and descriptions of specific textual and non-textual features.

However, there is no direct relation between support and F_1. There are some rare categories with a high F_1 and thus are easily learnable, such as *Quotes*, *Other* content features, *Classifications* and *Recommendations*. We think this is because they are more concrete, such that human annotators and LLMs find it easier to identify them, which is also reflected by the higher agreement on these categories (see Table 2).

Rebora and Vezzani (2024) trained three different BERT-based models and found that for classifying sentences as either containing evaluation or not, all three models reach an $\mathbf{F_1}$ score in the range of 0.82 and 0.86, which is very similar to the scores we found for mBERT, robBERT-2023 and XLM-RoBERTa. Taking this together with the similar agreement scores, this suggests that the evaluation categories in both sets of annotations are comparable in terms of reliability.

6. Review Composition Analysis

In this section we analyse the composition of reviews using the annotations (RQ4).

Presence of categories We first look at the percentage of individual sentences and whole reviews that contain at least one assignment to one of the main categories. For whole reviews, that means that at least one sentence is assigned to a main category. The percentages are shown in Figure 1. The percentages per sentence, in blue, show that *Reader response* and *Content* are roughly equally common, with around 50% of sentences assigned to each of them. *Style* and *Author* each cover around 10% of sentence, while the other three categories are rare. At the level of entire reviews, the percentages are of course higher. Notably, 98% of reviews have at least one *Reader response* statement, and 77% have at least one sentence describing the *Content* of a novel.

The total number of assigned categories per reviews is related to the length of the review. Since most sentences are assigned to one or two categories, reviews with a single

conference version

conference version

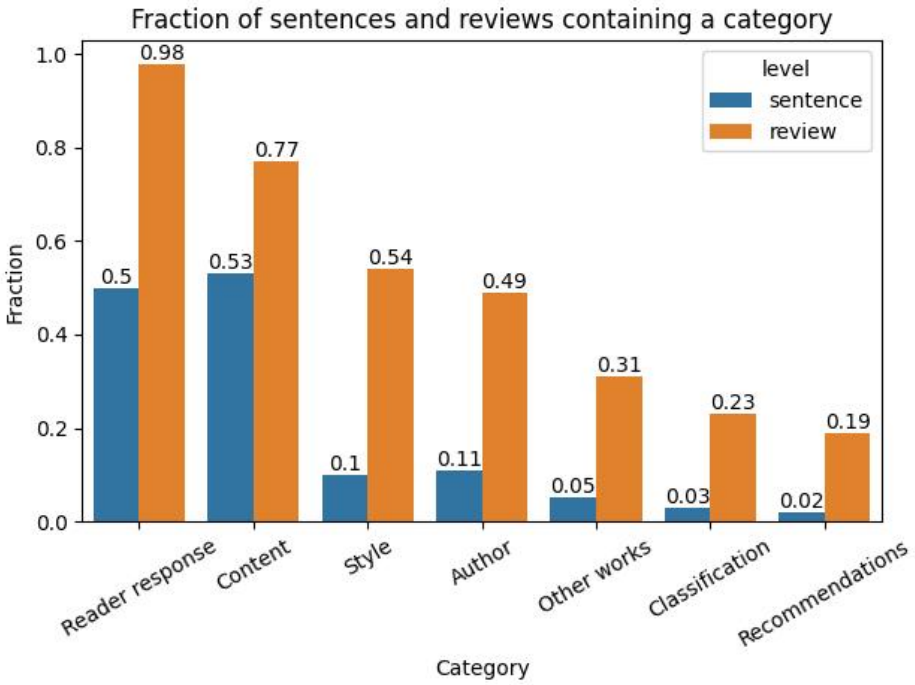


Figure 1: The fraction of sentences and reviews containing at least one statement assign with a given main category label. For reviews, this is aggregated over all sentences in the review.

sentence can have a maximum of six possible assignments, but tend have only one or two. Reviews with more sentences are thus likely have more category assignments.

Our findings seem to confirm some of the results obtained in the annotation process described by Kutzner et al. 2021, where the vast majority of text segments are assigned to 10 out of 65 categories, mostly related to reader response (positive/negative assessment, emotions) and content (summary, storyline, etc.). Their category system, however, distinguishes review content at the token level, with segments consisting of as little as a two-word phrase (e.g., "the book" for "view of the artefact as a whole"), and is therefore only partly comparable with ours. Notably, they do not report the distribution of categories across reviews. The skewed distribution of category occurrence observed both by Kutzner et al. and by us is probably due to the bottom-up nature in which both coding schemes were developed.

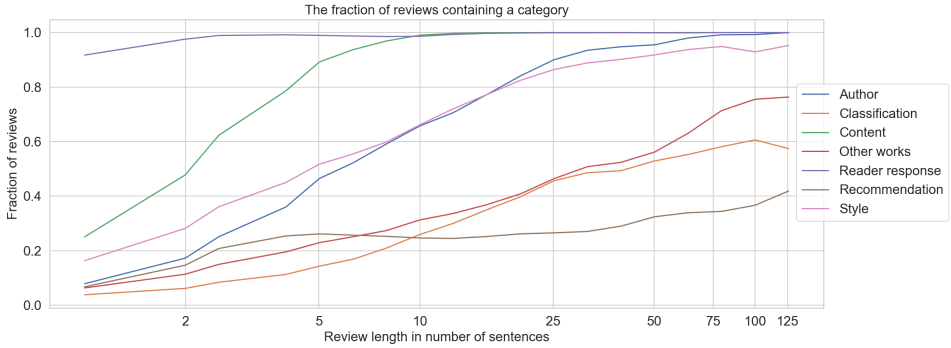


Figure 2: The relation between review length (on a logarithmic scale) and the fraction of reviews containing at least one statement assigned to a main category.

conference version

Composition and review length In reviews with many sentences there is more space to mention a diverse set of aspects than reviews with only one or two sentences (unless the individual sentences are long and convoluted). Therefore, there may be a relationship between number of sentences in a review and the presence of categories and their frequency of occurrence.

The probability that a review of a certain length has at least one sentence assigned to one of the main categories is shown in Figure 2, with the X-axis representing the number of sentences in a review, shown on a logarithmic scale. We used the annotations generated by our trained robBERT-2023 model on all 670,751 reviews in our corpus, so that the curves are smooth and clear. The same plot based only on the human annotations is available in Appendix F, which shows roughly the same trend, but because the limited number of reviews, the curves have very few data points. The presence of *Reader response* is not related to review length. Virtually all reviews of any length have at least one statement of reader response.

For the other categories, there is a clear length effect. Among short reviews of a single sentence, around 40% contain a description of *Content* and 25% mention *Style*. Looking at increasingly longer reviews, the probability that they describe *Content* rises quicker than other categories. Reviews of around 10 sentences and longer almost always contain some content description. The next two categories that rise in probability as length increases are *Style* and *Author*, reaching a fraction of close to 1.0 for reviews with close to 100 sentences.

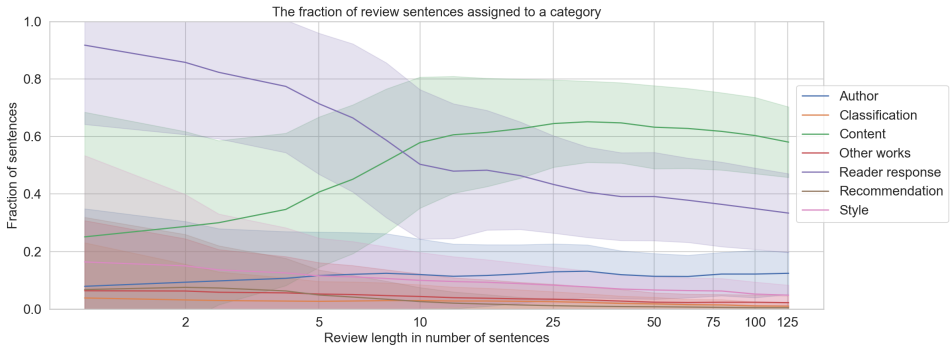


Figure 3: The relation between review length (on a logarithmic scale) and the fraction of sentences in a review containing a statement assigned to a main category. The lines represent the mean and the shaded areas are the standard deviation.

Another insightful analysis is the relationship between the total number of sentences in a review and the number of sentences that mention some category. We plot the distribution of the fraction of sentences in a review that are assigned to each of the six main categories in Figure 3.[12] For reviews of a single sentence, the fraction of sentences containing a *Reader response* statement is close to 100%. This is no surprise, given that almost all reviews with a single sentence have at least one reader response statement (Figure 2). As reviews get longer, this percentage slowly drops, to around 20% for reviews with a few hundred sentences. They have a high absolute number of sentences mentioning *Reader response*, but a larger fraction of sentences mention something else.

12. Again, this is based on annotation on all 670,751 reviews. The same plot based on human-only annotations is available in Appendix F.

In contrast, statements of *Content* make up only around 30% of sentences in reviews with one or two sentences, but this percentage goes up with review length to around 60% for reviews with 10 sentences or more. In other words, as reviews get longer, a larger fraction consists of content description.

For the other main categories, the mean fraction of sentences mentioning them is low at all review lengths, indicating these categories are mentioned at a more or less fixed rate.

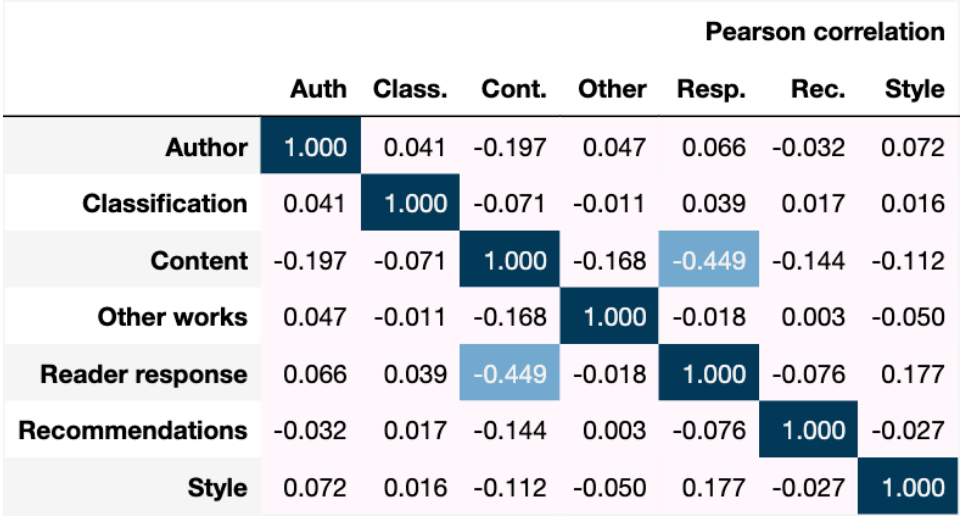


Figure 4: Pearson correlation between main categories at the sentence level.

Category overlap All categories can co-occur with each other in a single sentence, but some category pairs may be more likely to co-occur than others.

# main cat.	# sent.	%
0	526	0.05
1	7419	0.64
2	2941	0.25
3	664	0.06
4	82	0.01
5	2	0.00
Total	11,634	1.00

Table 4: The distribution of number of main categories assigned per sentence.

The number of main categories assigned per sentence is shown in Table 4. Of the 11,634 sentences, 7,419 are assigned to a single category (64%), while 526 are not assigned to any category (5%) and roughly one in three (31%) are assigned to multiple main categories. It is thus fairly common that a sentence covers aspects of multiple main categories.

This prompts the question which categories tend to be mentioned together in a sentence. For this we compute the Pearson correlations between pairs of main categories, which are shown in Figure 4. Most correlations are close to zero. Based on the standard interpretation that correlations $\rho \in (-0.2, 0.2)$ signal no correlation, this means that for most pairs of categories there is no association between the presence of one category and the presence of another. The only clear exception is the moderate negative correlation

conference version

between *Content* and *Reader response* ($\rho = -0.449$). In other words, reviewers tend to describe content and reading experience in separate sentences.

For the sub-categories there are only two pairs with a weak correlation, namely *Content - Narrative* and *Reader response - Reflection* ($\rho = -0.207$) and *Style - Stylistic Features* and *Reader response - Evaluation of Quality* ($\rho = 0.299$). The full set of percentages for all pairs is shown in Appendix F. The two sub-categories that are by far the most frequent are *Narrative* and *Evaluation of quality*, and as a consequence, most co-occurrences of sub-categories are with these two. Sentences that mention *Identification and immersion* are more likely to also mention *Narrative* aspects (45%) than *Evaluation of quality* (24%).

Mention of *Stylistic features* more often co-occurs (72%) with *Evaluation of quality* than with *Narrative* (30%).

We can repeat the same analysis at the review level, that is, whether a (sub)category occurs in any sentence of a review, and how often two (sub)categories co-occur in a single review. Notable co-occurrences of subcategories are *quote*s co-occurring with either statements about the *author* (71%) and about *stylistic features* (75%). This could mean that quotes are mostly used to illustrate the writing style of the book or the author.

7. Conclusions

In the introduction to this paper (1) we formulated four research questions:

- RQ1: Is it possible to define a succinct set of content categories to meaningfully capture different aspects mentioned in online book reviews?
- RQ2: Can human annotators reliably annotate these different aspects of book reviews?
- RQ3: Can we train automated taggers to accurately annotate fiction reviews?
- RQ4: How are the types of review statement distributed within and across reviews?

As to RQ1 and RQ2, we defined a hierarchical set of content categories that also mostly satisfies the other criteria we formulated in section 2.1 (relevance, clarity, comprehensiveness, succinctness). Our set strikes a balance between smaller sets of categories, designed for specific research projects, and some of the (for our purposes) overly detailed schemes defined in other projects, as we discussed in section 2. The clarity criterion is the most problematic, as appears from the Inter-Rater Reliability discussed in section 4. Some of our definitions were obviously not precise enough for our annotators to agree on the presence of, e.g., *feelings*, *theme* or *structure*. However, for our top-level categories the agreement is substantial, except for the moderate agreement for *style*, where we should have included a definition rather than have assumed that the concept was intuitively clear.

As for RQ3, we showed in section 5 that most categories are sufficiently, perhaps even surprisingly, learnable to a close to human level of performance. This shows the power of the multi-task training framework, where model parameters are shared between tasks. This approach makes knowledge of the texts acquired in learning one category accessible for learning the others. It should be especially helpful for learning categories

conference version

that occur less frequently.

We summarize the most important findings w.r.t. RQ 4 as follows:

- There are virtually no reviews without some form of reader response; three-quarters of the reviews also devote space to book content. Style and author are addressed in about half of the reviews.
- At the level of sentences, response as well as content are discussed in about half of them. The other main categories occur much less frequently.
- For all categories except response, the frequency at the level of reviews increases with the length of the review. Reviews longer than 75 sentences almost always refer to response, content, style, and author.

With the questions we discussed here, however, the research potential of this dataset is far from exhausted. One question that we have postponed for further work is how book genre and platform influence the content categories in the reviews. Dimitrov et al. (2015) showed differences between Goodreads and Amazon reviews in one genre (biography); the same was shown for Dutch Hebban (comparable to Goodreads) and Bol (comparable to Amazon) by Daniëls (2016). Antoniak et al. (2021) showed that reviews for different genres are tagged differently and use different topics. We are likely to find corresponding differences between the content categories. There may also be an interaction between platform and genre in this respect.

We already alluded to another issue for further work: to use the review sentence classification in order to refine the Reading Impact Model from Boot and Koolen (2020). The Impact Model is used to identify the book's impact on the reader. Once we know that a sentence is about the content of the narrative, we can assume that the sentiment that it expresses is not the reader's sentiment. This should increase the correctness of the model's predictions.

Another issue that we would like to explore is the question of the rhetoric of the review. The online review itself is a genre (Domsch 2009) and that should bring with it some form of structure. Accordingly, it should make sense to investigate the order in which the various categories occur in the review.

For researchers in computational literary studies, the online book review is obviously a welcome source of data. But it is also clear that we still have much to learn about how the online book review functions, and, therefore, about the potentials and pitfalls of its use. The online reviews that we analyse are always 'capta', not 'data' in the simple sense of the word (Drucker 2011). We know that countless factors —– among which, the behaviours that the sites encourage, the expectations of fellow readers, the changing fashions of social media, and our downloading choices —– all influence what we find in the 'data'. This article is an attempt to take that knowledge seriously.

8. Data Availability

Data can be found here: https://github.com/impact-and-fiction/JCLS-2026-review-composition.

conference version

9. Software Availability

Software can be found here: https://github.com/impact-and-fiction/JCLS-2026-review-composition.

10. Author Contributions

Marijn Koolen: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Writing – original draft

Joris J. Van Zundert: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Visualization, Writing – review & editing

Peter Boot: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Visualization, Writing – review & editing

Silvia Lilli: Methodology, Software, Validation, Writing – review & editing

Katja Tereshko: Methodology, Project administration, Writing – original draft, Writing – review & editing

References

Álvarez-López, Tamara, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, Jonathan Juncal-Martínez, Silvia García-Méndez, and Patrice Bellot (Nov. 2017). "A Book Reviews Dataset for Aspect Based Sentiment Analysis". In: *8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, 49–53. https://ltc.amu.edu.pl/book2017/papers/SANA-1.pdf (visited on 08/12/2025).

Antoniak, Maria, Melanie Walsh, and David Mimno (Apr. 2021). "Tags, Borders, and Catalogs: Social Re-Working of Genre on LibraryThing". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1. 10.1145/3449103. https://doi-org.proxy.uba.uva.nl/10.1145/3449103.

Bartl, Christoph, Sharwin Rezagholi, and Mareike Schumacher (2024). "A quantitative study of gender representation and authors' gender in a large-market print medium". In: *Proceedings of the Computational Humanities Research Conference*. Ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Vol. 3834. CEUR Workshop Proceedings, 1037–1052.

Boot, Peter and Marijn Koolen (2020). "Captivating, splendid or instructive? Assessing the impact of reading in online book reviews". In: *Scientific Study of Literature* 10.1, 35–63.

Caruana, Rich (1993). "Multitask learning: A knowledge-based source of inductive bias". In: *Proceedings of the Tenth International Conference on Machine Learning*, 41–48.

— (1997). "Multitask learning". In: *Machine learning* 28, 41–75.

conference version

conference version

Chevalier, Judith A and Dina Mayzlin (2006). "The Effect of Word of Mouth on Sales: Online book reviews". In: *Journal of marketing research* 43.3. Publisher: SAGE Publications Sage CA: Los Angeles, CA, 345–354.

Chong, Phillipa K (2020). *Inside the critics' circle: Book reviewing in uncertain times*. Princeton University Press.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2019). "Unsupervised Cross-lingual Representation Learning at Scale". In: *CoRR* abs/1911.02116. arXiv: 1911.02116. http://arxiv.org/abs/1911.02116.

Daniëls, Lianne (2016). "'Als ik realisme wil ga ik wel een uur uit het raam staan kijken'". MA thesis. Radboud University Nijmegen. https://theses.ubn.ru.nl/server/api/core/bitstreams/ea6fcc28-5e58-4321-936a-3a18149d553e/content.

Delobelle, P and F Remy (Sept. 2023). *RobBERT-2023: Keeping Dutch Language Models Up-To-Date at a Lower Cost Thanks to Model Conversion*. Antwerp, Belgium. https://clin33.uantwerpen.be/abstract/robbert-2023-keeping-dutch-language-models-up-to-date-at-a-lower-cost-thanks-to-model-conversion/.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. http://arxiv.org/abs/1810.04805.

Dimitrov, Stefan, Faiyaz Zamal, Andrew Piper, and Derek Ruths (2015). "Goodreads versus Amazon: the effect of decoupling book reviewing and book selling". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 9. 1, 602–605.

Domsch, Sebastian (2009). "Critical genres: Generic changes of literary criticism in computer-mediated communication". In: *Genres in the Internet*. John Benjamins Publishing Company, 221–238.

Drewry, John Eldridge (1974). *Writing book reviews*. Greenwood Press.

Drucker, Johanna (2011). "Humanities approaches to graphical display". In: *Digital Humanities Quarterly* 5 (1).

Fleiss, Joseph L (1971). "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5, 378.

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew E Peters, Michael Schmitz, and Luke Zettlemoyer (2018). "AllenNLP: A Deep Semantic Natural Language Processing Platform". In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 1–6.

Goot, Rob van der, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank (Apr. 2021). "Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, 176–197. 10.18653/v1/2021.eacl-demos.22. https://aclanthology.org/2021.eacl-demos.22.

Graf, Guido, Ralf Knackstedt, and Kristina Petzold (2022). *Rezensiv-Online-Rezensionen und Kulturelle Bildung*. transcript Verlag.

Hallgren, Kevin A. (2012). "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial". In: *Tutorials in Quantitative Methods for Psychology* 8.1, 23–34. 10.20982/tqmp.08.1.p023. http://www.tqmp.org/RegularArticles/vol08-1/p023/p023.pdf.

conference version

Hou, Yupeng, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley (2024). "Bridging Language and Items for Retrieval and Recommendation". In: *arXiv preprint arXiv:2403.03952*.

Hu, Yuerong, Zoe LeBlanc, Jana Diesner, Ted Underwood, Glen Layne-Worthey, and J Stephen Downie (2024). "Complexities of leveraging user-generated book reviews for scholarly research: transiency, power dynamics, and cultural dependency". In: *International Journal on Digital Libraries* 25.2, 317–340.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych (June 2018). "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation". In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018). Santa Fe, USA: Association for Computational Linguistics, 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/.

Koolen, Marijn, Toine Bogers, Maria Gäde, Mark Hall, Iris Hendrickx, Hugo Huurdeman, Jaap Kamps, Mette Skov, Suzan Verberne, and David Walsh (2016). "Overview of the CLEF 2016 Social Book Search Lab". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro. Cham: Springer International Publishing, 351–370. ISBN: 978-3-319-44564-9.

Koolen, Marijn, Peter Boot, and Joris J. van Zundert (2020). "Online Book Reviews and the Computational Modelling of Reading Impact". In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18-20, 2020*. Ed. by Folgert Karsdorp, Barbara McGillivray, Adina Nerghes, and Melvin Wevers. Vol. 2723. CEUR Workshop Proceedings. CEUR-WS.org, 149–169. https://ceur-ws.org/Vol-2723/long13.pdf.

Koolen, Marijn, Olivia Fialho, Julia Neugarten, Joris J van Zundert, Willem van Hage, and Ole Mussmann (2023). "How Can Online Book Reviews Validate Empirical In-depth Fiction Reading Typologies?" In: *IGEL 2023: Rhythm, Speed, Path: Spatiotemporal Experiences in Narrative, Poetry, and Drama*.

Koolen, Marijn, Julia Neugarten, and Peter Boot (2022). "'This book makes me happy and sad and I love it'. A Rule-based Model for Extracting Reading Impact from English Book Reviews". In: *Journal of Computational Literary Studies* 1.1.

Koolen, Marijn, Joris van Zundert, Eva Viviani, Carsten Schnober, Willem van Hage, Katja Tereshko, and Joris J van Zundert (2024). "From Review to Genre to Novel and Back. An Attempt To Relate Reader Impact to Phenomena of Novel Text". In: *Journal of Computational Literary Studies* 3.1.

Kuijpers, Moniek, Piroska Lendvai, Massimo Lusetti, Simone Rebora, Lina Ruh, Jonathan Tadres, Tina Ternes, and Johanna Vogelsanger (Sept. 2023). "Absorption in Online Reviews of Books: Presenting the English-Language AbsORB Metadata Corpus and Annotation Guidelines". In: *Journal of Open Humanities Data*. 10.5334/johd.116.

Kutzner, Kristin, Thorsten Schoormann, and Ralf Knackstedt (2021). "Exploring the content composition of online book reviews". In: *INFORMATIK 2020*. Gesellschaft für Informatik, Bonn, 1335–1344.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, 159–174.

conference version

Light, Richard J (1971). "Measures of response agreement for qualitative data: some generalizations and alternatives." In: *Psychological bulletin* 76.5, 365.

Linders, Yvette Francisca Maria (2014). "Met waardering gelezen. Een nieuw analyse-instrument en een kwantitatieve analyse van evaluaties in Nederlandse literaire dagbladkritiek, 1955-2005". PhD thesis. Radboud University, Faculty of Arts.

Mehling, Gabriele, Axel Kellermann, Holger Kellermann, and Martin Rehfeldt (2018). *Leserrezensionen auf amazon. de: Eine teilautomatisierte inhaltsanalytische Studie*. Vol. 7. Bamberger Beiträge zur Kommunikationswissenschaft. University of Bamberg Press.

Milota, Megan (2014). "From "compelling and mystical" to "makes you want to commit suicide": Quantifying the spectrum of online reader responses". In: *Scientific Study of Literature* 4.2, 178–195.

Mohammad, Saif (June 2016). "A Practical Guide to Sentiment Annotation: Challenges and Solutions". In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Ed. by Alexandra Balahur, Erik van der Goot, Piek Vossen, and Andres Montoyo. San Diego, California: Association for Computational Linguistics, 174–179. 10.18653/v1/W16-0429. https://aclanthology.org/W16-0429/.

Nguyen, Minh Van, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen (2021). "Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Op de Beek, EA (2014). "'Een literair fenomeen van de eerste orde'. Evaluaties in de Nederlandse literaire dagbladkritiek, 1955-2005: een kwantitatieve en kwalitatieve analyse". PhD thesis. Radboud University Nijmegen.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems* 32.

Pianzola, Federico (Jan. 2025). *Digital Social Reading: Sharing Fiction in the Twenty-First Century*. The MIT Press. ISBN: 9780262381352. 10.7551/mitpress/14588.001.0001. eprint: https://direct.mit.edu/book-pdf/2498089/book_9780262381352.pdf. https://doi.org/10.7551/mitpress/14588.001.0001.

Praamstra, Olf (1984). "De analyse van kritieken". In: *Voortgang. Jaarboek voor de neerlandistiek* 5, 241–264.

Rebora, Simone, Peter Boot, Federico Pianzola, Brigitte Gasser, J Berenike Herrmann, Maria Kraxenberger, Moniek M Kuijpers, Gerhard Lauer, Piroska Lendvai, Thomas C Messerli, et al. (2021). "Digital humanities and digital social reading". In: *Digital Scholarship in the Humanities* 36.Supplement_2, ii230–ii250.

Rebora, Simone and Gabriele Vezzani (2024). "Models of Literary Evaluation and Web 2.0. An Annotation Experiment with Goodreads Reviews". In: *Proceedings of the Computational Humanities Research Conference*. Ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Vol. 3834. CEUR Workshop Proceedings, 1175–1200.

Rosch, Eleanor H. (1973). "Natural categories". In: *Cognitive Psychology* 4.3, 328–350. ISSN: 0010-0285. https://doi.org/10.1016/0010-0285(73)90017-0. https://www.sciencedirect.com/science/article/pii/0010028573900170.

conference version

Sich, Christy (2017). "A Comparison of Traditional Book Reviews and Amazon. com Book Reviews of Fiction Using a Content Analysis Approach". In: *Evidence Based Library and Information Practice* 12.1, 85–96.

Spiteri, Louise F and Jen Pecoskie (2016). "Affective taxomonies of the reading experience: Using user-generated reviews for readers' advisory". In: *Proceedings of the Association for Information Science and Technology* 53.1, 1–9.

Von Heydebrand, Renate and Simone Winko (1996). *Einführung in die Wertung von Literatur. Systematik–Geschichte–Legitimation.* Paderborn : F. Schöningh.

— (2008). "The qualities of literatures: A concept of literary evaluation in pluralistic societies". In: *The Quality of Literature: Linguistic studies in literary evaluation*. John Benjamins Publishing Company, 223–239.

Wahlberg, Camilla (2019). "Features of Crossover Literature-Analysis of Reader Responses". PhD thesis. University of Vaasa, School of Marketing and Communication.

Walsh, Melanie and Maria Antoniak (Apr. 2021). "The Goodreads "Classics": A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism". In: *Journal of Cultural Analytics* 6.2. 10.22148/001c.22221.

Wan, Mengting and Julian J. McAuley (2018). "Item recommendation on monotonic behavior chains". In: *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*. Ed. by Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan. ACM, 86–94. 10.1145/3240323.3240369. https://doi.org/10.1145/3240323.3240369.

Wan, Mengting, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley (2019). "Fine-Grained Spoiler Detection from Large-Scale Review Corpora". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, 2605–2610. 10.18653/V1/P19-1248. https://doi.org/10.18653/v1/p19-1248.

Willemsen, Lotte M, Peter C Neijens, Fred Bronner, and Jan A De Ridder (2011). ""Highly recommended!" The content characteristics and perceived usefulness of online consumer reviews". In: *Journal of computer-mediated communication* 17.1.

Zundert, Joris J. van, Marijn Koolen, Julia Neugarten, Peter Boot, Willem Van Hage, and Ole Mussmann (2022). "What Do We Talk About When We Talk About Topic?" In: *Proceedings of the Computational Humanities Research Conference 2022, CHR 2022, Antwerp, Belgium, December 12-14, 2022*. Ed. by Folgert Karsdorp and Kristoffer L. Nielbo. Vol. 3290. CEUR Workshop Proceedings. CEUR-WS.org, 398–410. https://ceur-ws.org/Vol-3290/short%5C_paper5533.pdf.

conference version

A. Existing Review Annotation Schemas

For a fuller overview of studies into online book reviews and other forms of digital social reading, we refer to Rebora et al. (2021) and Pianzola (2025). Below we present a non-exhaustive but representative meta-review of studies that have relied on some form of annotation of book review content. Some studies annotate reviews with a specific research question in mind, while others have annotated for the sake of analysing the content of the (online) review as a genre.

Besides presenting previous research related to our investigation, our interest in these studies is equally driven by the possibility to use (part of) their annotation scheme for our purposes. Criteria that we apply in judging these annotation schemes include:

- Comprehensiveness: Does the scheme cover all major content in the reviews?
- Relevance: are the categories relevant to current research in computational literary studies?
- Clarity: Are definitions clear and do they clearly demarcate the categories from each other?
- Succinctness: Is the number of categories sufficiently small to be manageable? Aren't the categories too small for automatic processing?
- Hierarchy: Hierarchy makes grouping of items possible and results in more robust categories.

A.1 Comprehensive schemas

We firstly referred to works which propose a comprehensive descriptive annotation scheme for online book reviews. Among those, Kutzner et al. (2021) annotated 282 book reviews in German posted on Amazon to study the types of content, using a schema consisting of 65 categories. However, Amazon is an online shop, and the content of such reviews tends to differ from reviews on book discussion platforms (Dimitrov et al. 2015; Koolen et al. 2023). Therefore, their annotation schema may not be a good fit for reviews on other platforms. The research of Kutzner et al. (2021) was carried out in a larger project reported in Graf et al. (2022). Though their interest is rather specific (what do reviews show about the development of literary competence), the complete annotation scheme is impressively comprehensive and distinguishes 158 categories.

Another attempt at comprehensive annotation of what happens in an online review is represented by Mehling et al. (2018). Their annotation scheme contains 153 categories and they annotate 507 Amazon reviews. Their aim is not just to describe the content of the review, but also aspects of the reviewer (e.g. gender), as well as the evaluation. Given Amazon as the source, the same caveats that pertain to Graf et al. (2022) apply here as well.

These schemas show a number of annotation categories relevant for our purposes, which overlap with our final annotation schema (e.g. making a distinction between descriptions of the *content*, statements of *evaluation* and of *recommendation*). Nevertheless, adopting these extensive schemas in full would be infeasible given their breadth and

our available annotation budget, and would go beyond our scope. Our schema does not aim at a detailed classification of all kinds of book review content, but at improving the automatic distinction between references to book content, to the author or other works, and to the reader's experience, as stated in the introduction. Still, identifying the overlap as well as the mismatch that exists between our categorisation and the labels used in Graf et al. (2022) could be useful for downstream purposes.

A.2 Focused schemas

More directly comparable with the schema we finally developed (although we opted for different labels), Milota (2014) annotated reviews of Marilynne Robinson's debut novel *Housekeeping* from Amazon (196), Goodreads (100), and Librarything (75).

Her schema also distinguishes between content-related aspects (*plot*, *setting*, *characters*, *moral/message/theme*) from aspects of reader response (*evaluation*, *emotional response*, *personal tie-in*) as well as comments about the *author* and *intertextual references*.

Op de Beek (2014) and Linders (2014) conducted an analysis of the aspects in 734 fiction reviews of professional critics published in five Dutch newspapers. They based their analysis on a method developed by Praamstra (1984), who distinguishes between statements about the book (and the author) and statements about different matters. The former can be categorised in three groups: descriptive (e.g. the novel contains twelve chapters), interpretive (e.g. this novel is about hope) and evaluative. This schema provides the useful distinction between descriptive statements about the *structure* of the book (which we placed under *Style*) and interpretation of aspects of the *Content* such as *Narrative* and *Theme*.

Álvarez-López et al. (2017) provided a dataset of 300 reviews with 2977 sentences labelled for various categories of aspect based sentiment mining. They annotated sentences of 40 randomly selected novels from the Amazon/LibraryThing corpus provided by the Social Book Search Lab (Koolen et al. 2016). They annotated polarity (positive, negative, or neutral) of aspects ("targets") in 13 categories (characters, plot, general, author, genre, title, audience, quality, structure, period of story, period of publishing, length, and price). We recognised these aspects from our familiarity with Dutch reviews, but this is more fine-grained than we need for our purposes. We group *characters*, *plot* under *Content - Narrative*, and the aspects *title*, *audience*, *period of publishing*, *length* and *price* under *Content - Other*.

Sharing our interest in reader response, Rebora and Vezzani (2024) annotated 100 reviews from Goodreads of which 89 eventually were used for actual analysis. Their interest is specifically in evaluation: whether a sentence is evaluative, what criteria are used in the evaluation and what the evaluation is based on. They build their tagset on von Heydebrand and Winko's theory (Von Heydebrand and Winko 1996, 2008) and interpret online reviews as acts of "linguistic evaluation" which require an explicit "standard of value" as well as "certain categorizing assumptions" . They turned to the practice of "interpretative markup" to form a tagset bottom up by "recording [...] observations and conjectures in an open-ended way [as] argued by Gius and Jake" (Rebora and Vezzani 2024, p.1176). In their tag set they devoted special attention to criteria based on a reader's individual thoughts, feelings, and experiences. For this

conference version

sentences are labelled as mentioning "personal experience", "cognitive processes", and "emotions". Next to this they classify sentences into the categories of "aesthetic", "social", and "generic evaluation". Four rounds of annotation followed by discussion on (dis)agreement according to kappa score led to a convergent use of labels.

For our purposes the final tagset used by Rebora and Vezzani (2024) was published after we had finished our annotation scheme and started the annotation phase. We were familiar with an early version of this tagset and ran a test to annotate Dutch review sentences with them, but could not reach agreement on some of the specific evaluation categories, and did not use some other categories. We did however adopt the distinction between *emotions*, *personal experience* and *cognitive processes* (corresponding in our schema to *feelings*, *immersion and identification* and *reflection* under the main category *Reader response*). We decided to keep a distinction between such statements and statements of *evaluation* because we came across many example sentences that were either only statements of *evaluation* or statements expressing emotion, personal experience and/or cognitive processes. In other words, the latter three seem to be distinct from pure evaluative statements such as "I loved it."

Specifically interested in emotional engagement, M. Kuijpers et al. (2023) provided a dataset of 493 curated reviews from Goodreads totalling 2000 statement level annotations. They particularly aimed at identifying absorption related statements in five main categories (attention, emotional engagement, mental imagery, transportation (of self), and impact) and several subcategories.

Similarly interested in emotional aspects of reviews, but with the specific aim of providing reading recommendations (Spiteri and Pecoskie 2016) annotated affect in 536 user-generated book reviews obtained from the Canadian Public Library. They annotated for emotion, tone ("dramatic", "humorous", "cerebral"), and association (events, places, etc.). Their annotations were categorized into 44 unique emotions (as a subdivision of 9 basic emotion categories from Rosch's Prototype Theory, Rosch 1973), 141 grounded theory inferred unique tones in 11 main tone categories, and 31 unique associations in 7 categories.

Most of their taxonomy could be mapped to our categories of *Content - Narrative* and *Reader response - Feelings*. However their schema appeared too granular for any feasible annotation effort at scale in our project.

Finally, our own Reading Impact models for Dutch (Boot and Koolen 2020) and English (Koolen et al. 2022), which are rule-based models to identify expressions of reading impact in sentences, have categories for general affect and for narrative and stylistic impact. In these models, the focus is on identifying reader response, and in the case of narrative and aesthetic feelings, the expressed response has an explicit target of either narrative elements (plot, characters, events) or stylistics elements (word choice, use of metaphor, text structure).

A.3 Other schemas

In the area of library science, Sich (2017) focused on the helpfulness of traditional reviews versus that of Amazon reviews as a source of information for making library purchase decisions. For her annotation scheme, she takes the elements described in a

conference version

guide for writers of traditional reviews (Drewry 1974) that may be part of a "helpful" review.

The scheme focuses on characters, plot, theme, setting, style, and judgement, but ignores elements that frequently occur in online reviews, such as discussion of the author and details about the reader's response.

In a similar vein, Willemsen et al. (2011) used content analysis to investigate what makes consumer reviews useful as consumer information. The authors annotated 400 Amazon reviews across nine different product categories using the NET method (Willemsen et al. 2011, p.25) which divides a text into core statements that describe the relations between objects in the form of triples that capture positive, neutral or negative meaning. The annotations were related to the usefulness ratings attributed to reviews by other consumers. The authors focus on the usefulness of reviews for fellow consumers, which is different from our interest in the impact of reading fiction. This makes their annotation scheme less useful for our purposes.

However, the findings of Willemsen et al. (2011) remain useful in that they both draw our attention to, and confirm the so-called "negativity effect" for experience goods "such as recreational services, [which] are dominated by intangible attributes that cannot be known until purchase, and for which performance evaluations can be verified only by [...] experience or consumption". This effect, that may pertain to books as well, predicts that negative reviews warrant more perceived usefulness as customers are more averse to possible negative properties of an experience product.

Not directly aligned with our research, Wahlberg (2019) annotates reviews with a view to the question what drove the (adult) reviewer to read and review a book written for children. The annotation scheme consists of six thematic categories (narrative aspects, religious aspects, moral aspects, reading as a child vs. reading as an adult, plot timelines, and movie adaptions) into which readers' remarks are divided.

Some of the annotations may map well to various labels we intend to use, but the categorisation is decidedly different. Unfortunately, the annotation data for Wahlberg do not seem to be available.

Tangentially related to our work is the approach taken by Antoniak et al. (2021), who use tags generated by the LibraryThing community. The tags are used to elucidate the genre conceptions of the users and not, as in more conventional approaches, to normalize genre categories or enforce a hierarchy.

Similarly, Mohammad (2016) report some hard problems that pertain in general to (sentiment) annotation, such as sarcasm, quotations, and identifying the target of opinion.

Not comparable to our research questions, Hu et al. (2024) annotated 300 sponsored book reviews for keywords that might identify sponsorship of reviews to train a classifier. They found that classifiers are not able to identify sponsored reviews (if no sponsorship statement is given) and that there is no typical or significant skew in sponsored book reviews as compared to non-sponsored ones.

conference version

A.4 Commonalities among schemas

Most of the schemas contain relevant categories for our purposes, although some provide too many fine-grained categories (i.e. not succinct), others too few, and yet others too many categories outside the scope of our purposes (i.e. not relevant), or only categories to cover some parts of the review content (i.e. not comprehensive). Several schemas distinguish between on the one hand evaluation and reader response, and on the other, aspects of the books content and style. In our schema we retain these broad distinctions. Other broad categories that only some schemas have are statements about the *author*, about *other works* and statements of classification *e.g. classifying a book as belonging to a particular genre, period or group of authors*. There are several more fine-grained elements that are shared by several schemas in some form or other. Below is a rough mapping between our categories and the schemas referenced above:

- Content:
  - Narrative aspects: plot/storyline, setting, characters (Álvarez-López et al. 2017; Kutzner et al. 2021; Mehling et al. 2018; Milota 2014; Sich 2017; Spiteri and Pecoskie 2016; Wahlberg 2019)
  - Thematic aspects: moral, theme (Mehling et al. 2018; Milota 2014; Sich 2017; Wahlberg 2019)
  - Quote: citation (Kutzner et al. 2021; Mehling et al. 2018)
  - Other: length, illustrations (Álvarez-López et al. 2017; Mehling et al. 2018)
- Reader Response:
  - Evaluation (Álvarez-López et al. 2017; Kutzner et al. 2021; Mehling et al. 2018; Rebora and Vezzani 2024; Sich 2017)
  - Identification and immersion (M. Kuijpers et al. 2023; Kutzner et al. 2021; Milota 2014)
  - Feelings (Kutzner et al. 2021; Milota 2014; Rebora and Vezzani 2024; Spiteri and Pecoskie 2016)
- Author (Álvarez-López et al. 2017; Kutzner et al. 2021; Milota 2014)
- Classification: genre, literary-historical epoch (Álvarez-López et al. 2017; Kutzner et al. 2021; Mehling et al. 2018)
- Other works: relation to other artefacts (Kutzner et al. 2021; Mehling et al. 2018; Milota 2014)
- Recommendation (Kutzner et al. 2021; Mehling et al. 2018)
- Style
  - Context
  - Structure: chapter structure (Álvarez-López et al. 2017; Praamstra 1984)
  - Stylistic features: language, style (Kutzner et al. 2021; Mehling et al. 2018; Milota 2014; Sich 2017)

conference version

B. The Full Coding Scheme

For the accompanying positive and negative examples that were used to clarify the definitions, see our GitHub repository.[13]

- **Author** General discussion about the author (also addressing the author), biography and contextualization of the author.
- **Classification** Descriptions of the genre of the book (Suspense, Literary novel, Fantasy etc.), canonisation (e.g. classics, real thriller) and references to genre standards, or to period/literary epoch.
- **Content**
  - **Narrative** Narrative features such as plot, characters, emotion of characters, setting, etc.
  - **Quote** All and only direct quotes from the reviewed book
  - **Theme** Topic/subject of the book (when not referring explicitly to the plot), also when containing personal interpretation, or referring to facts of reality or specific period/place/culture.
  - **Other** Other non-linguistic features of the book, e.g. illustrations, length (even if number of words is textual/linguistic aspect), cover page etc.
- **Other works** Reflection on other works of the author (including mentions of book's series) or other authors or adaptations.
- **Reader response**
  - **Evaluation of quality** Any explicit positive or negative evaluation of the book. It may be general (good, bad, interesting) or referring also to emotional response (boring, exciting).
  - **Feelings** Any expression of feelings and emotions related to the book, both when referring to the content or referring to personal feelings and desires including the desire to read on.
  - **Identification and immersion** Any process of a reader's identification with the book, in terms of immersion in the story, identification with the characters, connection of the content of the book with the reader's biography.
  - **Reflection** Reflections about (own) life and society, including changes in the reader's perceptions or appreciations.
  - **Reading context** Any mention of time, place, medium of reading, as well as reading plans, order of reading, read/not read and prepublication reading. Expectations for the book caused by the other parts of a series or other books of the author are also in this category. Including the context in which they bought the book.

13. https://github.com/impact-and-fiction/JCLS-2026-review-composition/blob/main/Annotation-categories-book_reviews.pdf.

conference version

NUR code	NUR label	Genre label
280	Children's Fiction general	Children's fiction
281	Children's fiction 4 - 6 years	Children's fiction
282	Children's fiction 7 - 9 years	Children's fiction
283	Children's fiction 10 - 12 years	Children's fiction
284	Children's fiction 13 - 15 years	Young adult
285	Children's fiction 15+	Young adult
300	Literary fiction general	Literary fiction
301	Literary fiction Dutch	Literary fiction
302	Literary fiction translated	Literary fiction
305	Literary thriller	Literary thriller
312	Pockets popular fiction	Literary fiction
313	Pockets suspense	Suspense
330	Suspense general	Suspense
331	Detective	Suspense
332	Thriller	Suspense
334	Fantasy	Fantasy fiction
339	True crime	Suspense
342	Historical novel (popular)	Historical fiction
343	Romance	Romance
344	Regional- and family novel	Regional fiction

Table 5: The selected NUR codes of novels in our dataset of 18,885 novels, and their mapping to genres.

conference version

  - **Reception** Reception by others, official reception, prizes.

- **Recommendation**

- **Style**

  - **Stylistic features** Any reference to stylistic features such as tone, formality, complexity, narrative perspective, focalization, etc. Humor is also related to style.

  - **Context** describing the style in the context of the author, genre, translation or a specific period.

  - **Structure** Formal organization of the book's contents.

C. Mapping NUR Codes to Genre Labels

The complete mapping from NUR codes to genre labels is shown in Table 5.

We mapped the remaining codes to *other fiction* and *non-fiction*:

- *Other fiction* All other NUR codes in the ranges 286-350, which includes poetry, essays, biographies.

- *Non-fiction* All NUR codes in the ranges 1-279 and 351-999.

conference version

D. Review Sampling

We had several goals for the review sentence annotations, which informed our sampling approach.

There are number of dimensions along which we want to be able to comparatively analyse reviews with different values along each dimension.

The dimensions are:

- Review source: reviews posted on different platforms are slightly different in nature and characteristics Dimitrov et al. 2015; Hu et al. 2024; Koolen et al. 2020, and we want to understand the differences between reviews from different platforms in terms of review composition. Therefore, there should be enough reviews from each platform to be able to make a useful comparison.
- Ratings: we want to be able to compare reviews with a low rating, medium and high rating,
- Genre: the reviews in the corpus are labelled with the genre of the book that is reviewed. There are 12 distinct genre labels. The genres are, in order the number of available reviews from high to low: *Literary fiction*, *non-fiction*, *unknown*, *literary thriller*, *suspense*, *other fiction*, *young adult*, *children's fiction*, *fantasy fiction*, *romance*, *historical fiction* and *regional fiction*. We want to focus on fiction, therefore, removed reviews of *non-fiction* and with an *unknown* genre.

To sample reviews for annotation, we used multiple criteria. From a trial in which we tested our developed annotation schema, we calculated how long it would take an annotator to annotate a review on average. From our available budget and the cost of hiring three student assistants to do the annotations, we calculated that we could have 1400 reviews annotated. This was our target number of reviews.

First, we selected only reviews from the four selected sources (Bol, Goodreads, Hebban, NBD Biblion) that contain at least 10 words and that have a genre label from one of the ten included genres.

Second, we randomly sampled 100 reviews for each of the ten included genres. This resulted in a set of 1000 reviews in which Goodreads and NBD Biblion were under-represented (fewer than 100 reviews each), and with few 0, 1 and 2 star reviews (5, 21 and 44 respectively).

Third, we randomly sampled additional reviews from Goodreads and NBD Biblion until we had 100 reviews from each of these platforms, resluting in a set of 1081 reviews.

Fourth, we randomly sampled from reviews with a rating below 3 until we had a set of 1400 reviews.

E. Evaluation

The full evaluation scores of all three models, mBERT, XLM-RoBERTa and robBERT 2023 are shown in Figure 7.

Genre	# Reviews	Rating	# Reviews	Source	# Reviews
Literary fiction	297	5	394	Hebban	632
Suspense	169	4	365	Bol	504
Literary thriller	164	2	259	Goodreads	164
Other fiction	130	3	148	NBD Biblion	100
Young adult	116	1	123		
Children's fiction	109	1	10		
Fantasy fiction	107				
Historical fiction	105				
Romance	102				
Regional fiction	101				
Total	1400		1400		1400

Table 6: Distribution of the sampled reviews over the included genres, ratings and source platforms.

F. Review composition

Here we show the relation between review length and (1) the fraction of reviews that contain at least one sentence assigned to a category (Figure 5) and (2) the fraction of sentences in a review that contain a category (Figure 6), but only for the reviews that were annotated by the human annotators, using the majority vote of their annotations.

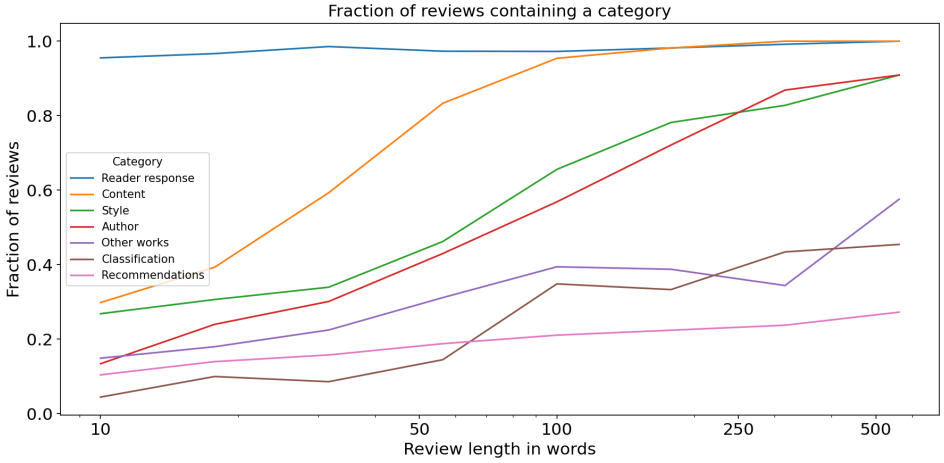


Figure 5: The relation between review length (on a logarithmic scale) and the fraction of reviews containing at least one statement assigned to a main category.

We also show the Pearson correlation of all categories overlapping with each other at the sentence level (Figure 7).

conference version

conference version

Category	Model	Precision	Recall	F_1	Support
Author	mBERT	0.89	**0.91**	0.90	282
	RobBERT-2023	**0.91**	**0.91**	**0.91**	282
	XLM-RoBERTa	0.90	**0.91**	**0.91**	282
Classification	mBERT	**0.88**	0.83	0.85	62
	RobBERT-2023	0.88	0.87	0.87	62
	XLM-RoBERTa	**0.89**	**0.92**	**0.90**	62
Content	mBERT	0.87	0.87	0.87	1224
	RobBERT-2023	**0.89**	**0.89**	**0.89**	1224
	XLM-RoBERTa	**0.89**	0.88	0.88	1224
Narrative	mBERT	0.86	0.86	0.86	1064
	RobBERT-2023	**0.88**	**0.88**	**0.88**	1064
	XLM-RoBERTa	**0.88**	**0.88**	**0.88**	1064
Other content	mBERT	0.89	0.73	0.79	61
	RobBERT-2023	**0.94**	**0.83**	**0.87**	61
	XLM-RoBERTa	**0.94**	0.82	0.87	61
Quote	mBERT	0.86	0.76	0.80	62
	RobBERT-2023	0.92	**0.78**	**0.84**	62
	XLM-RoBERTa	**0.93**	0.76	0.82	62
Theme	mBERT	0.71	0.70	0.71	25
	RobBERT-2023	**0.78**	0.74	**0.76**	25
	XLM-RoBERTa	0.73	**0.76**	0.74	25
Other works	mBERT	0.81	0.79	0.80	120
	RobBERT-2023	0.87	0.83	0.85	120
	XLM-RoBERTa	**0.88**	**0.85**	**0.87**	120
Reader response	mBERT	0.86	0.86	0.85	1142
	RobBERT-2023	**0.88**	**0.88**	**0.88**	1142
	XLM-RoBERTa	**0.88**	**0.88**	**0.88**	1142
Evaluation of quality	mBERT	0.82	0.85	0.83	661
	RobBERT-2023	**0.87**	**0.87**	**0.87**	661
	XLM-RoBERTa	0.85	0.86	0.86	661
Feelings	mBERT	0.74	0.67	0.70	147
	RobBERT-2023	**0.76**	0.74	**0.75**	147
	XLM-RoBERTa	0.74	**0.76**	0.75	147
Identification and immersion	mBERT	0.78	0.66	0.70	53
	RobBERT-2023	**0.86**	**0.83**	**0.84**	53
	XLM-RoBERTa	0.84	0.79	0.81	53
Reading Context	mBERT	0.87	0.76	0.81	128
	RobBERT-2023	0.90	0.82	0.85	128
	XLM-RoBERTa	**0.92**	**0.84**	**0.87**	128
Reception	mBERT	0.91	0.70	0.77	25
	RobBERT-2023	0.95	**0.84**	**0.89**	25
	XLM-RoBERTa	**0.96**	0.76	0.83	25
Reflection	mBERT	0.70	0.62	0.65	187
	RobBERT-2023	**0.72**	0.66	0.68	187
	XLM-RoBERTa	0.72	**0.68**	**0.70**	187
Recommendations	mBERT	0.84	0.92	0.88	46
	RobBERT-2023	**0.90**	0.93	**0.91**	46
	XLM-RoBERTa	0.88	**0.95**	**0.91**	46
Style	mBERT	0.88	0.82	0.85	232
	RobBERT-2023	**0.89**	0.85	**0.87**	232
	XLM-RoBERTa	0.88	**0.86**	**0.87**	232
Structure	mBERT	0.73	0.67	0.69	15
	RobBERT-2023	**0.75**	**0.73**	**0.74**	15
	XLM-RoBERTa	0.64	0.66	0.65	15
Stylistic features	mBERT	0.88	0.83	0.85	196
	RobBERT-2023	**0.89**	**0.86**	**0.87**	196
	XLM-RoBERTa	**0.89**	**0.86**	**0.87**	196
Weighted average	mBERT	0.85	0.84	0.84	5732
	RobBERT-2023	**0.88**	**0.87**	**0.87**	5732
	XLM-RoBERTa	0.87	0.86	0.87	5732
Macro average	mBERT	0.83	0.78	0.80	5732
	RobBERT-2023	**0.87**	**0.83**	**0.84**	5732
	XLM-RoBERTa	0.85	**0.83**	0.84	5732

Table 7: Evaluation results of mBERT, RobBERT-2023 and XLM-RoBERTa for the (sub-)categories. Support is the number of positive examples in the test dataset.

conference version

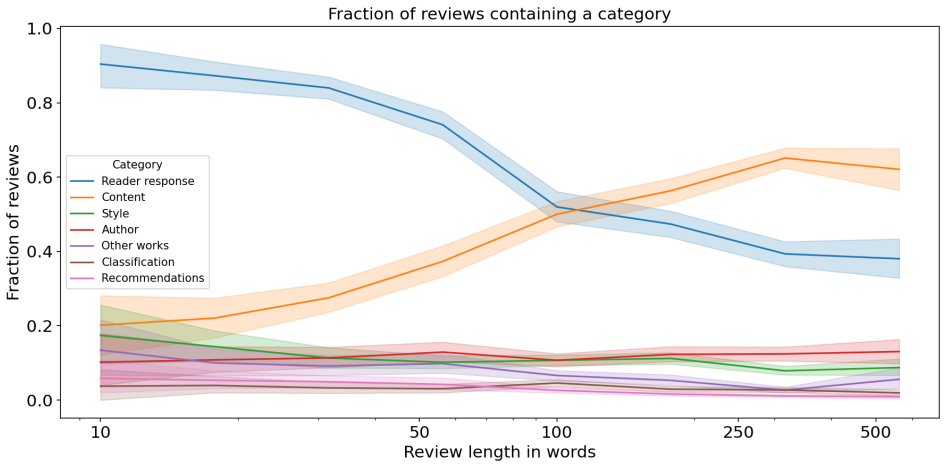


Figure 6: The relation between review length (on a logarithmic scale) and the fraction of sentences in a review containing a statement assigned to a main category. The lines represent the mean and the shaded areas are the standard deviation.

conference version

Figure 7: Overlap between sub-categories at the sentence level. The overlap fractions are with respect to the category in the rows.

jcls. Journal of Computational Literary Studies

conference version

Grace as a Formal Turning Point

Computational Detection in Flannery O'Connor's Short Fiction

Arthur F. Ramos[1]

1. Microsoft, Tampa, United States.

OPEN ACCESS

Citation
Arthur F. Ramos (2026). "Grace as a Formal Turning Point. Computational Detection in Flannery O'Connor's Short Fiction". In: *CCLS Conference Preprints* 5 (1). 10.26083/tuda-7982

Date published 2026-05-05 (preprint)
Date accepted tbc
Date received 2026-01-05

Keywords
Flannery O'Connor, turning points, computational narratology, close reading, grace, digital humanities

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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. This paper operationalizes Flannery O'Connor's theological concept of "the moment of grace" as a narratologically detectable turning point. We propose a formal definition of grace-typed turning points as discourse spans exhibiting both epistemic rupture and action-alternative visibility, drawing on O'Connor's own critical writings, later criticism, and narratological theory. Working with twelve short stories, we develop an annotation protocol at the paragraph level and evaluate computational detection methods. Our cross-validated grace-aware model incorporating position weighting, enhanced vocabulary detection, cross-story phrase cues, and window smoothing achieves 75% Hit@1 in leave-one-out evaluation; an exploratory story-specific phrase condition reaches 100% on the full corpus. The cross-validated success demonstrates that O'Connor's grace moments share a transferable linguistic signature: recognition, violence, body/incarnation, and spiritual vocabulary combine in characteristic ways across her corpus. We argue this success illustrates a productive workflow for digital humanities: build interpretable baselines, analyze failures, incorporate domain knowledge, and validate through cross-validation.

1. Introduction

In her essay "On Her Own Work," Flannery O'Connor identifies a precise moment in "A Good Man Is Hard to Find" as the story's crucial pivot:

> I often ask myself what makes a story work, and what makes it hold up as a story, and I have decided that it is probably some action, some gesture of a character that is unlike any other in the story, one which indicates where the real heart of the story lies. This would have to be an action or a gesture which was both totally right and totally unexpected. (O'Connor 1969, 111)

This "gesture" – the grandmother's reaching out to The Misfit and calling him "one of my own children" – functions as what narrative theorists would call a nucleus (Barthes 1966) or turning point (Ouyang and McKeown 2015): a structurally privileged moment whose removal or displacement would render the story fundamentally different. For O'Connor, such moments are not merely structural but theological; they represent the "action of grace in territory largely held by the devil" (O'Connor 1969, 118).

This paper asks: **Can we detect grace as a formal phenomenon?** More precisely, can computational methods identify the turning points that O'Connor identifies as

moments of grace, and what do the successes and failures of such detection reveal about the relationship between narrative structure and theological meaning?

The question matters for several reasons. First, O'Connor's concept of grace is both theologically specific and formally consequential – she argues that grace moments structure her narratives, producing the "totally right and totally unexpected" gestures that define her stories. If grace has formal correlates, computational methods should detect them. Second, O'Connor's critical prose provides unusually explicit commentary on where she locates grace in her own fiction. We do not treat that commentary as definitive authority; rather, we use it as one especially strong interpretive witness, triangulated with later scholarship, against which to test computational methods. Third, the challenge of operationalizing a theological concept for computational analysis illuminates broader questions about the relationship between humanistic interpretation and algorithmic detection in digital humanities.

Our approach synthesizes three methodological traditions:

1. **Computational narratology**, which models narrative structure through formal features and algorithmic detection (Otake et al. 2020; Ouyang and McKeown 2015; Piper 2018)
2. **O'Connor scholarship**, which has long attended to the formal dimensions of her theological vision (Brinkmeyer 1989; Desmond 1987; Wood 2004)
3. **Digital humanities hermeneutics**, which theorizes the relationship between computational analysis and interpretive reading (Ramsay 2011; Rockwell and Sinclair 2016; Underwood 2019)

1.1 Contributions

This paper makes four contributions:

1. **Conceptual**: A portable definition of "grace-typed turning point" grounded in the dual criteria of epistemic rupture and action-alternative visibility. By connecting theology-inflected interpretation to narratological structure, the definition remains applicable beyond O'Connor's specific theological framework.
2. **Methodological**: An explicit annotation protocol at the paragraph level that makes the concept of grace-as-turning-point auditable, reusable, and amenable to computational analysis, validated through leave-one-out cross-validation. The protocol includes five decision rules that resolve common annotation ambiguities.
3. **Computational**: Development of a grace-aware detection model achieving 75% Hit@1 in leave-one-out cross-validation (0% → 75% improvement over baselines). The model encodes literary-critical knowledge through position weighting, enhanced vocabulary features (including body/incarnation and perception terms derived from O'Connor scholarship), cross-story phrase cues, and window smoothing to capture multi-paragraph grace moments. A separate exploratory story-specific phrase condition reaches 100% on the full corpus and serves as an interpretive upper bound rather than the paper's central claim.

conference version

4. **Interpretive**: Close readings of "A Good Man Is Hard to Find" and "Revelation" that use computational output as a scaffold for, rather than replacement of, literary interpretation – demonstrating how scholarship-informed computation extends close reading by making implicit interpretive assumptions explicit and testable.

1.2 Paper Structure

Section 2 reviews background in narratology, O'Connor criticism, and computational turning-point detection. Section 3 describes our corpus selection, annotation protocol, and the characteristics of annotated grace moments. Section 4 presents our computational methods, with particular attention to operationalization and the derivation of features from literary scholarship. Section 5 reports evaluation results including ablation analysis. Section 6 provides extended close readings of two central stories, demonstrating how computational detection informs interpretation. Section 7 discusses implications, limitations, and future directions.

2. Background

2.1 Turning Points in Narrative Theory

The concept of the turning point has roots in Aristotle's *Poetics*, where *peripeteia* (reversal) and *anagnorisis* (recognition) mark pivotal narrative moments. Aristotle defines peripeteia as "a change from one state of affairs to its opposite" that occurs "in accordance with probability or necessity" (1452a22-23). Recognition, the "change from ignorance to knowledge," often coincides with reversal in complex plots. These concepts establish that narratives have structurally privileged moments – points where the direction of action changes and where characters come to understand their situations differently.

In structuralist narratology, Barthes distinguishes between *nuclei* (cardinal functions that open alternatives and advance the narrative) and *catalyzers* (supplementary functions that fill narrative space without altering trajectory). The nucleus is defined by its structural indispensability: "a nucleus cannot be deleted without altering the story" (Barthes 1966, 93). This deletion test provides an operational criterion for identifying turning points: if removing a passage would require fundamental revision to the story's logic, that passage is structurally privileged.

Contemporary narratology has refined this concept through attention to what Herman calls the "storyworld" – the mental model readers construct of the narrative's characters, events, and causal relations (D. Herman 2002). A turning point, in this view, is a discourse span that forces significant revision to the storyworld model, particularly regarding what Ryan terms the "modal structure" of narrative: the configuration of knowledge, obligation, and possibility that determines what characters can know, must do, and might choose (Ryan 2006). Ryan's possible-worlds approach is particularly relevant here: a turning point reconfigures the space of narrative possibilities, closing some paths and opening others.

conference version

2.2 O'Connor on Grace and Gesture

O'Connor's fiction repeatedly stages what she calls "the action of grace" through violent or shocking pivotal moments. Her critical essays in *Mystery and Manners* provide unusually explicit commentary on both the theology and the craft behind these moments. In "The Fiction Writer and His Country," she explains:

> The novelist with Christian concerns will find in modern life distortions which are repugnant to him, and his problem will be to make these appear as distortions to an audience which is used to seeing them as natural; and he may well be forced to take ever more violent means to get his vision across to this hostile audience. (O'Connor 1969, 33–34)

Violence in O'Connor's work functions not as gratuitous spectacle but as the necessary shock that opens characters (and readers) to recognition. As she explains in "On Her Own Work," discussing the grandmother's gesture in "A Good Man Is Hard to Find":

> I think myself that [the grandmother's gesture] is a moment of grace, the most significant position in the story ...Her head clears for an instant and she realizes, even in her limited way, that she is responsible for the man before her and joined to him by ties of kinship which have their roots deep in the mystery she has been merely prattling about so far. (O'Connor 1969, 111–112)

This passage is crucial for our operationalization. O'Connor identifies the "head clearing" as the core of the grace moment – a sudden epistemic shift in which the grandmother recognizes both her responsibility and her kinship with The Misfit. The gesture (reaching out, calling him "one of my children") is the visible manifestation of this internal change. Grace, then, has both cognitive and behavioral correlates: a change in understanding accompanied by a meaningful action.

O'Connor scholarship has elaborated on these themes. Wood emphasizes the "incarnational" dimension of O'Connor's theology: grace operates through the body, not despite it (Wood 2004). Desmond traces how O'Connor's characters are "risen" through violence into new understanding (Desmond 1987). Brinkmeyer examines the dialogical structure of O'Connor's fiction, where grace moments often involve characters speaking across profound difference (Brinkmeyer 1989). These scholarly emphases inform our feature design: the body vocabulary, the violence vocabulary, and the recognition vocabulary all derive from established themes in O'Connor criticism.

2.3 Grace Outcomes: Accepted, Refused, and Ambiguous

O'Connor distinguishes between grace offered and grace received. In some stories, characters accept the moment of recognition and are transformed (even if they die immediately after, as the grandmother does). In others, characters refuse grace, remaining in their delusions. In still others, the outcome is ambiguous – the story ends before we see whether the character will integrate the recognition.

This tripartite distinction structures our corpus. We classify grace moments as *accepted* (clear transformation, as in "A Good Man Is Hard to Find"), *refused* (explicit rejection,

conference version

as in "Good Country People" where Hulga's recognition leads nowhere), or *ambiguous* (uncertain outcome, as in "Revelation" where Mrs. Turpin's response to her vision remains undeveloped). This classification does not affect our detection task – we aim to identify where grace moments occur, not how characters respond – but it enriches interpretation of results.

2.4 Computational Approaches to Narrative Structure

Computational narratology has developed several approaches to detecting structurally significant moments in narrative:

Event salience models (Otake et al. 2020) operationalize Barthes' cardinal functions by measuring how much narrative coherence degrades when events are removed. Using neural language models, they identify events whose removal most damages the probability of subsequent text, operationalizing the deletion test computationally.

Turning point detection (Ouyang and McKeown 2015) frames narrative pivots as "reportable events" – moments that license narrative retelling. Their model for personal narratives uses features including temporal expressions, quotations, and verb tense patterns to identify the moment a narrator would choose to highlight.

Movie plot analysis (Papalampidi et al. 2019) extends turning point detection to film synopses, identifying five canonical turning points (opportunity, change of plans, point of no return, major setback, climax) using supervised learning.

Narrative arc extraction (Reagan et al. 2016) applies sentiment analysis to trace emotional trajectories, identifying pivot points as significant direction changes in sentiment curves. Their analysis of thousands of novels finds recurring arc shapes, suggesting that narrative structure has statistical regularities.

Change-point detection more broadly (Truong et al. 2020) provides statistical methods for identifying moments of regime change in sequential data. Applied to narrative features (vocabulary, sentiment, topic), these methods can identify structural discontinuities without supervision.

Long before recent narrative-turning-point models, computational literary studies and empirical stylistics were already pursuing domain-attuned literary questions through carefully designed features. Herman, Hogenraad, and van Mierlo's content analysis of *Gravity's Rainbow* is an early example of using quantitative signals to test interpretive claims specific to a single literary work rather than to solve a generic NLP task (L. Herman et al. 2003). Jockers likewise argues for literary models that are built around historically and formally specific research questions rather than around off-the-shelf textual features alone (Jockers 2013). Our project belongs to this longer tradition of hypothesis-driven operationalization: the goal is not a universal turning-point detector but an interpretable model of one author's distinctive formal theology.

Prior computational work on O'Connor is limited, but Hardy's study of ergative and reflexive voice patterns demonstrates the value of computational linguistics for O'Connor analysis (Hardy 2007). Hardy shows that O'Connor uses ergative constructions ("the door opened" vs. "she opened the door") to create effects of characters being acted upon by forces beyond their control – a linguistic correlate of her theological vision. Just as

conference version

importantly for the present study, Hardy's attention to bodily diction and linguistic voice suggests that apparently local stylistic patterns can register O'Connor's incarnational theology in a systematic way. Our body/incarnation vocabulary extends that insight from sentence-level agency to paragraph-level turning-point detection: where Hardy tracks how grammar and bodily reference shape the felt texture of experience, we ask whether comparable cues help mark the structurally decisive moments toward which those experiences move.

Our work differs from prior computational narratology in the specificity of its literary target. Rather than seeking generic turning-point features, we encode the characteristics of O'Connor's grace moments as identified in literary scholarship. This scholarship-informed approach is necessary because grace moments are not generic narrative pivots; they have distinctive theological and formal properties that generic methods cannot capture.

3. Corpus and Annotation

3.1 Corpus Selection

We work with twelve short stories from O'Connor's collected fiction (O'Connor 1971), selected to represent the range of grace-moment types in her work. Selection criteria included: (1) clear consensus in O'Connor scholarship that the story contains a grace moment; (2) sufficient length for paragraph-level analysis (at least 90 paragraphs); and (3) diversity in grace outcome (accepted, refused, ambiguous) and trigger type (violence, vision, illness, humiliation).

Table 1 presents the corpus with annotations. Stories range from 96 paragraphs ("Everything That Rises Must Converge") to 305 paragraphs ("The Lame Shall Enter First"). Gold turning points occur between 95.9% and 100% of story length, confirming that O'Connor's grace moments cluster at story endings.

Table 1: Corpus of twelve O'Connor stories with gold turning point annotations. "Type" indicates grace outcome; "Trigger" indicates the proximate cause of the grace moment.

Story	Paras	Gold TP	Type	Trigger
A Good Man Is Hard to Find	123	118	Accepted	Violence
Revelation	154	148–150	Ambiguous	Vision
Good Country People	129	126	Refused	Humiliation
Greenleaf	151	149–150	Ambiguous	Violence
The Enduring Chill	171	171	Ambiguous	Illness
Everything That Rises Must Converge	96	95–96	Accepted	Humiliation
The River	123	121–123	Accepted	Violence
The Displaced Person	250	248–249	Ambiguous	Violence
The Artificial [N-word]	130	128–129	Accepted	Vision
A View of the Woods	135	134–135	Refused	Violence
The Lame Shall Enter First	305	304–305	Accepted	Violence
Parker's Back	132	130–132	Refused	Humiliation

3.2 Annotation Unit

We annotate at the **paragraph level** for three reasons. First, *stability*: paragraph boundaries are consistent across editions, unlike sentence boundaries which may vary with

conference version

editorial decisions about punctuation. Second, *interpretability*: paragraphs correspond to interpretively meaningful discourse units, typically organized around a single action, description, or thought. Third, *computational tractability*: paragraph-level features balance granularity with statistical reliability – sentences are too short to extract stable vocabulary features, while larger units lose precision.

Paragraphs are defined structurally as text blocks separated by blank lines in the source. We use PDF extraction with manual verification, following the paragraph structure of the FSG 1971 edition.

3.3 Annotation Protocol

Our annotation protocol operationalizes O'Connor's concept of grace through narratological criteria:

Turning Point (TP): A discourse span whose removal would change the story's downstream trajectory in a way readers recognize as "a different story." This operationalizes Barthes' deletion test.

Grace-Typed Turning Point (GTP): A TP exhibiting both:

- *Epistemic rupture*: The focal character's belief state changes abruptly – what O'Connor calls the "head clearing." This may involve recognition of kinship, collapse of self-deception, or sudden understanding.
- *Action-alternative*: A meaningful choice becomes visible. The character could act differently; they face what O'Connor calls "the moment of decision."

3.4 Decision Rules

Five decision rules resolve common annotation ambiguities:

1. **Leverage test**: If the span is removed, does the later narrative lose coherence? If removing a paragraph leaves the story logically intact, it is not a turning point.
2. **Discontinuity preference**: Prefer spans where perception shifts abruptly over gradual developments. Grace in O'Connor is sudden, not gradual.
3. **Choice visibility**: A GTP must expose an action-alternative. Pure shock without choice does not qualify.
4. **Minimal span**: Use the smallest span containing the decisive transition, typically 1–3 paragraphs. Multi-paragraph spans are appropriate when the grace moment unfolds across adjacent paragraphs.
5. **Shock vs. turn distinction**: Violence or shock may trigger the grace moment but is not identical to it. The grandmother's murder is not the turning point; her reaching out is.

3.5 Annotation Process

Annotations were developed through iterative reading informed by O'Connor scholarship. For each story, we identified the conventional critical interpretation of the grace

conference version

moment, then located the specific paragraph(s) where that moment occurs. Where O'Connor herself comments on the story (as with "A Good Man Is Hard to Find"), her remarks were treated as historically important evidence rather than final authority. Where scholarship diverges, we followed the interpretation best supported by the textual passage in conjunction with the dominant line of critical discussion.

This annotation approach is interpretive rather than empirical. We do not claim inter-annotator reliability in the psychometric sense; rather, we claim that our annotations represent defensible interpretations grounded in O'Connor's critical prose, later scholarship, and the stories themselves. The leave-one-out cross-validation tests whether features learned from eleven stories detect turning points in a held-out twelfth, providing indirect validation of the annotation scheme without requiring us to posit a single unchallengeable reading.

4. Computational Methods

Our computational approach prioritizes interpretability over predictive power. We extract transparent features and evaluate simple baselines that produce analyzable outputs. This design philosophy follows the hermeneutic tradition in digital humanities (Rockwell and Sinclair 2016): computation should enhance interpretation, not replace it.

4.1 Operationalization

A key challenge in computational literary studies is *operationalization*: translating humanistic concepts into computable features (Moretti 2013). Operationalization requires making interpretive assumptions explicit, which can feel reductive. However, as Underwood argues, explicit operationalization is preferable to implicit operationalization, where assumptions remain hidden in intuitive judgments (Underwood 2019).

Our operationalization of "grace moment" proceeds through four levels, each grounded in O'Connor scholarship:

1. **Structural**: Grace moments are turning points occurring in the final 5–10% of stories. O'Connor's stories build toward terminal grace moments; the moment of recognition occurs near the end. This is not a universal narrative property but a specific feature of O'Connor's technique.

2. **Lexical**: Grace moments deploy distinctive vocabulary. Recognition vocabulary ("clear," "see," "realize") signals epistemic rupture. Violence vocabulary ("kill," "blood," "die") signals the shocking triggers. Spiritual vocabulary ("soul," "grace," "heaven") signals theological content. Body vocabulary ("head," "eyes," "face") signals incarnational theology.

3. **Phrasal**: Grace moments contain characteristic phrases identified in O'Connor's prose and scholarship. "Head cleared" echoes O'Connor's own description. "One of my children" appears verbatim in "A Good Man Is Hard to Find." "Action of mercy" appears in "The Artificial [N-word]."

conference version

4. **Sequential**: Grace moments span multiple paragraphs. The recognition may begin in one paragraph and culminate in another. Window smoothing captures this sequential structure.

4.2 Feature Categories

Position Weight: We apply exponential weighting favoring late-story paragraphs. Position ratios and weights are:

- Final 5% (≥ 0.95): weight 5.0
- 90–95%: weight 3.5
- 85–90%: weight 2.0
- 80–85%: weight 1.0
- 70–80%: weight 0.3
- First 70%: weight 0.05

This strong position prior reflects O'Connor's technique: grace moments do not occur in the middle of her stories. The low weight (0.05) for early paragraphs effectively filters noise without eliminating early paragraphs from consideration.

Generic Grace Vocabulary: A base vocabulary of approximately 70 terms capturing recognition, violence, familial, spiritual, and emotional language:

- *Recognition*: clear, cleared, recognize, saw, see, vision, understand, realize, truth
- *Violence*: kill, killed, shot, blood, pierce, die, dying, death, wound, bull, horn
- *Familial*: child, children, mother, own, baby
- *Spiritual*: grace, soul, heaven, light, fire, holy, ghost, spirit, mercy
- *Emotional*: terror, horror, love, terrible, beautiful, guilt

Enhanced Vocabulary: Three domain-specific sets derived from O'Connor scholarship, not from gold paragraph inspection:

Body/incarnation terms: head, face, eyes, eye, body, flesh, skin, hands, feet, mouth, lips, throat, chest, back. O'Connor's theology emphasizes grace through the body – what Wood calls "incarnational realism" (Wood 2004). The word "head" appears in O'Connor's description of the grandmother's grace moment ("her head clears"), suggesting body vocabulary as a signal.

Perception terms: looked, looking, stared, staring, watched, watching, gazed, gazing, appeared, visible, blind, sight, seen. Grace involves "seeing clearly" – the epistemic rupture is often figured as a change in vision.

Stillness terms: still, stood, standing, stopped, frozen, paralyzed, silent, silence, quiet, motionless, unable. Grace moments involve stopping – the character pauses before the moment of choice.

conference version

conference version

Phrase Score: A bonus for matching key recognition phrases. We use two phrase settings. The cross-validated setting combines nine generic phrases with the 2–3 short phrase patterns contributed by each of the other eleven stories, yielding 37–38 test-time patterns that do *not* come from the held-out story. A separate exploratory full-model setting additionally permits the held-out story's own 2–3 phrases. These phrases were manually curated from O'Connor's prose and established criticism rather than automatically mined as frequent n-grams; each match receives a uniform 50-point bonus so that phrase evidence is strong but not learned on a per-phrase basis. Because story-specific phrases encode interpretation directly, we treat the full-model phrase condition as an upper-bound demonstration rather than as the main evidentiary result. Examples include:

- "head clear" / "head cleared"
- "one of my ...children"
- "action of mercy"
- "believing in nothing"
- "souls climbing"

Window Smoothing: Scores are averaged over adjacent paragraphs using a sliding window of size 5 (i.e., ± 2 paragraphs). This captures multi-paragraph grace moments and reduces noise from paragraph-level variation. The window size was selected based on domain knowledge (grace moments span 1–3 paragraphs) and validated empirically (performance degrades at both smaller and larger windows).

4.3 Scoring Function

For paragraph i in a story of n paragraphs, we compute:

$$S_i = \left(\frac{1}{|W_i|} \sum_{j \in W_i} (G_j \times 40 + P_j \times 2) \right) \times \text{pos}(i)$$

Where:

- G_j is the grace vocabulary density (percentage of tokens in vocabulary) for paragraph j
- P_j is the phrase score for paragraph j (50 points for each matched phrase)
- W_i is the set of paragraphs in the window around i
- $\text{pos}(i)$ is the position weight for paragraph i

The paragraph with the highest score S_i is predicted as the turning point.

4.4 Implementation

Listing 1 shows the core scoring implementation in Python. The vocabulary sets encode the domain knowledge described above; the `score_paragraphs` function implements window smoothing and position weighting.

conference version

```python
# Enhanced vocabulary derived from O'Connor scholarship
BODY_VOCAB = {'head', 'face', 'eyes', 'body', 'hands', ...}
PERCEPTION_VOCAB = {'looked', 'stared', 'watched', 'saw', ...}
STILLNESS_VOCAB = {'still', 'stood', 'frozen', 'silent', ...}
GRACE_VOCAB = GENERIC | BODY_VOCAB | PERCEPTION_VOCAB | STILLNESS_VOCAB

def position_weight(index, total):
    """Exponential weighting for late-story paragraphs."""
    ratio = index / total
    if ratio >= 0.95: return 5.0
    elif ratio >= 0.90: return 3.5
    elif ratio >= 0.85: return 2.0
    elif ratio >= 0.80: return 1.0
    else: return 0.05

def score_paragraphs(paragraphs, vocab, phrases, window=5):
    """Score paragraphs with window smoothing."""
    n = len(paragraphs)
    # First pass: raw scores
    raw = [grace_density(p, vocab)*40 + phrase_score(p, phrases)*2
           for p in paragraphs]
    # Second pass: window smoothing * position weight
    scores = []
    for i in range(n):
        start, end = max(0, i-2), min(n, i+3)
        smoothed = sum(raw[start:end]) / (end - start)
        scores.append(smoothed * position_weight(i+1, n))
    return scores
```

Listing 1: Core scoring function (simplified)

Full source code is available at the repository listed in Software Availability.

4.5 Evaluation Metrics

We evaluate using:

- **Hit@1**: Whether the top prediction falls within the gold span. This is our primary metric.
- **Average Distance**: Mean paragraphs between prediction and nearest gold boundary. Captures "near misses."

4.6 Leave-One-Out Cross-Validation

To test whether features generalize across stories, we use leave-one-out (LOO) cross-validation, a procedure in which each story serves once as held-out test data while the other eleven stories supply the phrase inventory. For each story:

1. Train: Collect phrases from the other 11 stories
2. Test: Predict turning point in the held-out story using enhanced vocabulary plus cross-story phrases

This setting is the paper's primary evaluation because it withholds the test story's own phrase list. High LOO performance indicates transferable signal; low LOO performance suggests overfitting to story-specific vocabulary.

5. Results

5.1 Main Results

Table 2 presents results across evaluation conditions, showing incremental improvement as features are added:

Table 2: Evaluation results on 12-story corpus. Each row adds features to the previous configuration.

Method	Hit@1	Avg Dist
Position weighting only	0/12 (0%)	5.2
+ Generic vocabulary	4/12 (33%)	3.0
+ Enhanced vocabulary	5/12 (42%)	1.6
+ Window smoothing (LOO main result)	9/12 (75%)	0.25
+ Story-specific phrases (exploratory)	12/12 (100%)	0.0

Several patterns emerge. First, **position weighting alone fails completely**. Although grace moments cluster in the final 5% of stories, position alone cannot distinguish the grace paragraph from adjacent paragraphs. This baseline failure confirms that domain-specific features are necessary.

Second, **vocabulary features add substantial signal**. Generic grace vocabulary achieves 33% Hit@1; enhanced vocabulary (body/perception/stillness) improves to 42%. The enhanced vocabulary was derived from O'Connor scholarship, not from inspecting gold

conference version

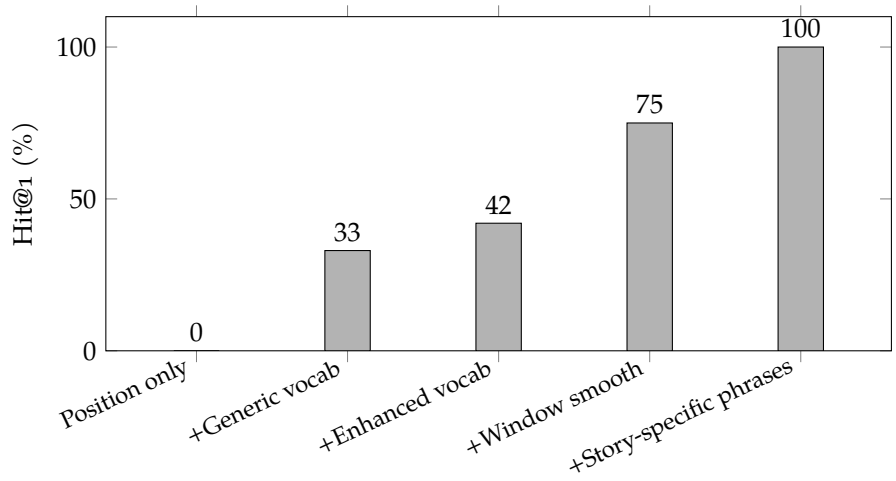


Figure 1: Incremental improvement in Hit@1 as features are added. The leave-one-out cross-validated model (+Window smooth) is the paper's main result at 75%; the story-specific phrase condition is reported separately as an exploratory upper bound. Figure by the author, CC BY 4.0.

conference version

paragraphs, supporting the claim that literary-critical knowledge can be operationalized computationally.

Third, **window smoothing dramatically improves LOO performance**. The jump from 42% to 75% reflects the multi-paragraph nature of grace moments. Optimal window size (5 paragraphs) was determined empirically: performance peaks at window=5 and declines at both smaller (window=1: 42%) and larger (window=7: 50%) sizes.

Fourth, **the exploratory story-specific phrase condition reaches perfect detection**. When the model is allowed to use the held-out story's own curated phrases, Hit@1 reaches 100%. We report this setting as an interpretive ceiling rather than as confirmatory evidence of generalization, precisely because those phrases are not withheld at test time.

5.2 Cross-Validation Analysis

The 75% LOO Hit@1 is our central result. It confirms that O'Connor's grace moments share recognizable patterns that transfer across her corpus. This is not trivial: one might expect each story to use distinctive vocabulary that does not transfer. Instead, the enhanced vocabulary and cross-story phrases capture a shared linguistic signature.

Critically, all three stories not achieving exact Hit@1 miss by only one paragraph:

- **Revelation**: Predicted paragraph 147, gold 148–150 (distance 1)
- **Greenleaf**: Predicted paragraph 151, gold 149–150 (distance 1)
- **The Artificial [N-word]**: Predicted paragraph 130, gold 128–129 (distance 1)

The average distance of 0.25 paragraphs means the model identifies the correct region even when it misses the exact paragraph. These near-misses occur because the immediately adjacent paragraph also contains grace vocabulary, creating scoring ties resolved by small differences.

5.3 Ablation Analysis

Ablation analysis estimates the contribution of each feature by removing one component at a time from an otherwise fixed model and then re-running the evaluation. In Table 3, the Δ column therefore reports the percentage-point drop relative to the full 75% LOO model:

Table 3: Ablation analysis: LOO Hit@1 when each feature category is removed

Configuration	Hit@1	Δ
Full LOO model	9/12 (75%)	–
− Position weighting	3/12 (25%)	−50%
− Body vocabulary	7/12 (58%)	−17%
− Perception vocabulary	8/12 (67%)	−8%
− Stillness vocabulary	8/12 (67%)	−8%
− Window smoothing	5/12 (42%)	−33%

Position weighting is most critical: without it, Hit@1 drops to 25%. This confirms that O'Connor's end-placement of grace moments is a structural regularity the model must capture. Window smoothing is second most important (−33%), validating the multi-paragraph nature of grace moments. Enhanced vocabulary categories contribute more modestly but still measurably: body vocabulary contributes 17%, perception and stillness each contribute 8%. Put differently, the model's success depends less on any single magic token than on the interaction of late placement, embodied/perceptual diction, and local sequential context.

5.4 Per-Story Analysis

Table 4 presents per-story results under LOO cross-validation:

Table 4: Per-story LOO results. "Pred" is predicted paragraph; "Gold" is gold span; "Dist" is distance to nearest gold boundary.

Story	Pred	Gold	Result
A Good Man Is Hard to Find	118	118	Hit
Revelation	147	148–150	Miss (d=1)
Good Country People	126	126	Hit
Greenleaf	151	149–150	Miss (d=1)
The Enduring Chill	171	171	Hit
Everything That Rises Must Converge	96	95–96	Hit
The River	123	121–123	Hit
The Displaced Person	248	248–249	Hit
The Artificial [N-word]	130	128–129	Miss (d=1)
A View of the Woods	134	134–135	Hit
The Lame Shall Enter First	305	304–305	Hit
Parker's Back	131	130–132	Hit

The model succeeds on stories with diverse grace outcomes (accepted, refused, ambiguous) and trigger types (violence, vision, illness, humiliation). No systematic pattern distinguishes successful from unsuccessful predictions.

conference version

6. Close Readings

This section demonstrates how computational detection informs close reading. We examine two stories in detail: "A Good Man Is Hard to Find," where O'Connor's retrospective commentary and later criticism provide an unusually strong interpretive anchor (Desmond 1987; O'Connor 1969; Wood 2004), and "Revelation," where the model distinguishes between violence and vision as turning points (Brown Smith 2012; Desmond 1987; Tolomeo 1978).

6.1 "A Good Man Is Hard to Find"

6.1.1 Computational Detection

The model correctly identifies paragraph 118 as the turning point. This paragraph contains the grandmother's decisive gesture:

> She saw the man's face twisted close to her own as if he were going to cry and she murmured, "Why you're one of my babies. You're one of my own children!" She reached out and touched him on the shoulder.

The paragraph scores highest on multiple features:

- **Grace vocabulary density**: 8.7% of tokens match the vocabulary ("saw," "own," "children," "reached," "touched")
- **Phrase match**: "one of my ...children" matches a story-specific phrase (+50 points); in LOO validation, generic cues such as "she saw" still contribute
- **Position**: 95.9% of story length (weight: 5.0)
- **Body vocabulary**: "face," "shoulder" signal incarnational moment
- **Perception vocabulary**: "saw" signals epistemic shift

6.1.2 Interpretive Implications

The computational success aligns with O'Connor's own account and with later critical readings that treat the grandmother's gesture as the story's decisive recognition. O'Connor identifies this as "a moment of grace" where the grandmother's "head clears for an instant"; Desmond and Wood likewise read the scene as a brief breakthrough into kinship and responsibility rather than merely as the prelude to murder (O'Connor 1969, 111–112); see also (Desmond 1987; Wood 2004). The model detects exactly what those readings emphasize: recognition ("saw"), kinship ("my own children"), and gesture ("reached out and touched").

More interestingly, the model correctly distinguishes this paragraph from the surrounding violence. The Misfit shoots the grandmother in paragraph 119, but paragraph 118 – the gesture, not the violence – scores highest. This validates our decision rule 5: shock may trigger the grace moment but is not identical to it, a distinction that also structures critical accounts of O'Connor's violent climaxes (Brinkmeyer 1989; Desmond 1987).

The feature breakdown also illuminates the paragraph's construction. O'Connor packs

conference version

recognition, familial, and contact vocabulary into three short sentences. The density is strategic: this is the story's most linguistically concentrated moment. Computational analysis makes this density visible and measurable.

6.1.3 The Near-Miss at Paragraph 117

Paragraph 117, immediately preceding the grace moment, also scores highly:

> "Jesus!" the old lady cried. "You've got good blood! I know you wouldn't shoot a lady! I know you come from nice people!"

This paragraph contains spiritual vocabulary ("Jesus") and familial hints ("good blood," "nice people"), but lacks the recognition and gesture that define the grace moment. The grandmother is still manipulating, not recognizing. The model's slight preference for paragraph 118 reflects the difference between manipulation (117) and genuine recognition (118) – a distinction encoded in vocabulary density rather than explicit annotation.

6.2 "Revelation"

6.2.1 The Challenge of Two Candidates

"Revelation" presents a detection challenge because it contains two dramatic moments: Mary Grace's assault on Mrs. Turpin in the doctor's waiting room (paragraphs 60–62) and Mrs. Turpin's vision at the pig parlor (paragraphs 148–150). Both involve shock and transformation, and criticism has plausibly treated each as central to the story's spiritual drama (Brown Smith 2012; Desmond 1987; Tolomeo 1978). Which is the turning point?

6.2.2 Computational Detection

The model identifies paragraphs 148–150 as the turning point, corresponding to Mrs. Turpin's vision:

> She saw the streak as a vast swinging bridge extending upward from the earth through a field of living fire. Upon it a vast horde of souls were rumbling toward heaven ...And bringing up the end of the procession was a tribe of people whom she recognized at once as those who, like herself and Claud, had always had a little of everything ...Yet she could see by their shocked and altered faces that even their virtues were being burned away.

Key features include:

- **Spiritual vocabulary density**: "souls," "heaven," "fire"
- **Perception vocabulary**: "saw," "recognized," "could see"
- **Position**: 96.1% of story length

The earlier assault (paragraphs 60–62) scores lower despite its violence because it occurs at only 39% of story length (weight: 0.05) and contains different vocabulary (anger, attack) rather than grace vocabulary (recognition, vision).

conference version

conference version

6.2.3 Interpretive Implications

The model's preference for the vision over the assault reflects O'Connor's narrative structure. The assault is the *trigger* – it poses the question that troubles Mrs. Turpin for the rest of the story ("Who do you think you are?"). But the turning point is the *answer* – the vision that reconfigures Mrs. Turpin's understanding of herself and her relationship to others.

This distinction matters for interpretation. Readings that foreground the waiting-room assault emphasize violence as the means by which complacency is broken, in line with Desmond's broader account of shocking grace in O'Connor. Readings that foreground the final vision instead stress the story's symbolism of sight, judgment, and revelation, as in Tolomeo's biblical framing and Smith's account of ocular symbolism (Brown Smith 2012; Desmond 1987; Tolomeo 1978). The model sides with the latter cluster of readings, and does so on formal grounds: the vocabulary and position of the vision paragraph mark it as structurally privileged.

The model's detection also illuminates the vision's construction. O'Connor builds the paragraph around perception verbs ("saw," "recognized," "could see") and spiritual nouns ("souls," "heaven," "fire"). The vocabulary density is strategic – this is the story's most spiritually concentrated moment. As with "A Good Man Is Hard to Find," computational analysis makes the strategic density visible.

6.2.4 The Ambiguity of Grace Outcome

We classify "Revelation" as having an *ambiguous* grace outcome. The story ends with Mrs. Turpin "hearing the voices of the souls climbing upward into the starry field and shouting hallelujah." Has she accepted the vision? Will she be transformed? The text does not say.

This ambiguity does not affect detection – the model identifies *where* grace occurs regardless of outcome – but it affects interpretation. The computational detection locates the moment of possibility; close reading explores what that possibility means.

7. Discussion

7.1 What the Model Learns

The grace-aware model encodes four key insights from O'Connor scholarship. First, grace moments are terminally placed: the strong position prior captures O'Connor's habit of building toward late recognition. Second, grace is incarnational: body vocabulary such as "head," "face," "eyes," and "hands" matters because O'Connor repeatedly stages spiritual recognition through bodily perception and gesture. Third, grace is phenomenologically marked by stillness and altered perception: characters stop, stare, freeze, or suddenly see differently. Fourth, grace often unfolds across adjacent paragraphs rather than within a single isolated sentence, which is why window smoothing contributes so much to performance (Desmond 1987; O'Connor 1969; Wood 2004).

These features derive from O'Connor scholarship, not from inspecting gold paragraphs. The enhanced vocabulary categories (body, perception, stillness) operationalize critical

concepts: Wood's "incarnational realism," Desmond's emphasis on violence and recognition, and Hardy's demonstration that bodily and grammatical patterning in O'Connor can carry theological weight at the level of style (Desmond 1987; Hardy 2007; Wood 2004). The success of these features validates the operationalization: literary-critical knowledge, when made explicit, improves computational detection.

7.2 Why Generic Methods Fail

Position weighting alone achieves 0% Hit@1. This failure illuminates what makes grace moments distinctive. Generic turning-point detection often assumes that pivots are marked by abrupt change: sentiment reversal, topic shift, or lexical discontinuity (Ouyang and McKeown 2015; Reagan et al. 2016; Truong et al. 2020). But O'Connor's grace moments are not primarily moments of *change*; they are moments of *recognition*. The grandmother does not become a different person; she *sees* The Misfit differently. This seeing has specific vocabulary correlates (recognition, kinship, gesture) that generic change detection misses.

The failure of generic methods also reflects O'Connor's technique. Her stories build steadily toward the grace moment; there is no abrupt topic shift, no sudden sentiment reversal. The narrative arrives at grace through careful preparation, not rupture. Change-point detection looks for ruptures; O'Connor provides culminations. Scholarship-informed features are necessary because O'Connor's grace moments are not generic narrative pivots but historically and theologically specific formal phenomena.

7.3 Cross-Validated Signal

The 75% LOO Hit@1 represents genuine signal rather than overfitting. Several features of the result matter here. The enhanced vocabulary derives from O'Connor scholarship, not from inspection of the gold paragraphs; the body, perception, and stillness categories were hypothesized from critical concepts before validation. The phrase inventory also transfers across stories: each held-out run uses only generic phrases plus phrases borrowed from the other eleven stories, so success requires cross-story rather than purely local lexical cues. Just as importantly, all three LOO failures miss by only one paragraph, suggesting boundary ambiguity rather than wholesale misrecognition of the relevant region. Finally, the ablations degrade performance gradually rather than catastrophically, which indicates that the model is not riding on a single overfit trick but on the interaction of several independently motivated features.

7.4 Limitations

Several limitations constrain our findings. The most obvious is corpus size: twelve stories limit statistical power and make leave-one-out the only viable validation regime. A larger corpus would allow stronger significance testing and more stable estimation of feature importance. A second limitation is author specificity. The features that work here encode O'Connor-specific regularities, especially terminal placement and incarnational vocabulary, and they may not transfer intact to writers whose moments of insight are structured differently.

A third limitation is circularity risk. Although the enhanced vocabulary derives from

conference version

scholarship rather than gold-paragraph inspection, the story-specific phrase lists are interpretation-laden and informed by close reading of the relevant scenes. Leave-one-out cross-validation mitigates this risk for the paper's main result, but it does not eliminate it entirely, which is why the story-specific phrase condition is reported only as exploratory. Finally, the annotations themselves remain interpretive judgments, and our preference for interpretable features may miss signals that richer contextual models would capture. We accept that trade-off because interpretability is part of the argument, not merely a constraint.

7.5 Implications for Computational Literary Studies

Our results support a practical workflow for computational literary studies: begin with interpretable baselines, examine their failures, operationalize the missing domain knowledge, and then validate the resulting model on held-out texts. That sequence is closer to the interpretive and hypothesis-driven traditions described by Moretti, Ramsay, and Rockwell and Sinclair than to benchmark-oriented machine learning (Moretti 2013; Ramsay 2011; Rockwell and Sinclair 2016; Underwood 2019). In this workflow, feature engineering is not an embarrassment to be minimized but a way of making literary assumptions explicit and testable.

The 75% LOO Hit@1 is therefore not just a performance metric. It is evidence that O'Connor's grace moments have transferable formal properties and that scholarship-specific operationalization can recover them. The larger implication is methodological: when the object of study is historically and formally distinctive, interpretable, domain-specific models may reveal more than generic models optimized for portability alone.

7.6 Future Directions

Several extensions are possible. The most immediate is corpus expansion: O'Connor's complete short fiction includes 31 stories, and enlarging the dataset would strengthen validation and make more nuanced comparisons across outcome types possible. A second extension is comparative. Grace-like moments in other religious writers, such as Graham Greene or Walker Percy, would provide a strong test of which features are specifically O'Connor's and which belong to a broader poetics of religious recognition.

There is also a clear modeling extension. Contextual embeddings and other neural representations might capture cues that handcrafted vocabularies miss, but they would need to be compared against interpretable features rather than substituted for them uncritically. Finally, a multi-annotator version of the dataset would sharpen the distinction between stable critical consensus and genuinely contested cases.

8. Reproducibility

All annotation data (paragraph indices and offsets, not copyrighted text) and Python code for all experiments are publicly available at `https://github.com/Arthur742Ramos/ConnorGrace`. The repository includes the gold annotations for the twelve stories, vocabulary lists, generic and story-level phrase patterns, evaluation scripts for both the

conference version

exploratory full model and the leave-one-out main result, and the feature extraction code used to produce the reported tables and figures.

The experiments use text extracted from lawfully accessed copies of *The Complete Stories* (O'Connor 1971). Because O'Connor's fiction remains under copyright, the repository does not redistribute the primary texts or provide third-party download links. Researchers must secure lawful access independently and consult the legal regime that governs their use. In the United States, 37 C.F.R. § 201.40 contains specified exemptions that may be relevant to some research uses of literary works (*37 CFR §201.40 - Exemptions to prohibition against circumvention* 2026). In the European Union, Directive (EU) 2019/790 establishes a scientific-research text-and-data-mining exception for research organizations and cultural heritage institutions with lawful access (*Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market* 2019). The public package therefore distributes only derived data and code, not the copyrighted stories themselves.

8.1 Computational Environment

Experiments require Python 3.8+ with PyMuPDF for PDF text extraction. No GPU is required; all experiments complete in under one minute on a standard laptop.

9. Data Availability

Annotation data, guidelines, and paragraph offsets are available at: `https://github.com/Arthur742Ramos/ConnorGrace/tree/master/data`.

10. Software Availability

Python code and evaluation scripts are available at: `https://github.com/Arthur742Ramos/ConnorGrace`.

11. Acknowledgements

The author thanks the reviewers for their careful reading and constructive suggestions, which materially improved the manuscript.

12. Author Contributions

Arthur F. Ramos: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing

References

37 CFR §201.40 - Exemptions to prohibition against circumvention (2026). Legal Information Institute. `https://www.law.cornell.edu/cfr/text/37/201.40` (visited on 03/21/2026).

conference version

Barthes, Roland (1966). "Introduction to the Structural Analysis of Narratives". In: *Communications* 8.1. Translated in *Image–Music–Text*, 1977, 1–27. 10.3406/comm.1966.1113.

Brinkmeyer, Robert H. (1989). *The Art and Vision of Flannery O'Connor*. Baton Rouge: Louisiana State University Press.

Brown Smith, Julie (2012). "Eye Symbolism in Flannery O'Connor's REVELATION". In: *The Explicator* 70.3, 231–233. 10.1080/00144940.2012.703707.

Desmond, John F. (1987). *Risen Sons: Flannery O'Connor's Vision of History*. Athens: University of Georgia Press.

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market (2019). European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790 (visited on 03/21/2026).

Hardy, Donald E. (2007). *The Body in Flannery O'Connor's Fiction: Computational Technique and Linguistic Voice*. Columbia: University of South Carolina Press.

Herman, David (2002). *Story Logic: Problems and Possibilities of Narrative*. Lincoln: University of Nebraska Press.

Herman, Luc, Robert Hogenraad, and Wim van Mierlo (2003). "Pynchon, Postmodernism and Quantification: An Empirical Content Analysis of Thomas Pynchon's *Gravity's Rainbow*". In: *Language and Literature: International Journal of Stylistics* 12.1, 27–41. 10.1177/096394700301200102.

Jockers, Matthew L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Moretti, Franco (2013). ""Operationalizing": Or, the Function of Measurement in Literary Theory". In: *New Left Review* 84, 103–119.

O'Connor, Flannery (1969). *Mystery and Manners: Occasional Prose*. Ed. by Sally Fitzgerald and Robert Fitzgerald. New York: Farrar, Straus and Giroux.

— (1971). *The Complete Stories*. New York: Farrar, Straus and Giroux.

Otake, Takaki, Sho Yokoi, Naoya Inoue, Ryo Takahashi, Tatsuki Kuribayashi, and Kentaro Inui (2020). "Modeling Event Salience in Narratives via Barthes' Cardinal Functions". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 1784–1794. 10.18653/v1/2020.coling-main.160.

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*. Lisbon, Portugal: Association for Computational Linguistics, 2149–2158. 10.18653/v1/D15-1257.

Papalampidi, Pinelopi, Frank Keller, and Mirella Lapata (2019). "Movie Plot Analysis via Turning Point Identification". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 1707–1717. 10.18653/v1/D19-1180.

Piper, Andrew (2018). *Enumerations: Data and Literary Study*. Chicago: University of Chicago Press.

Ramsay, Stephen (2011). *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press.

conference version

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, 31. 10.1140/epjds/s13688-016-0093-1.

Rockwell, Geoffrey and Stéfan Sinclair (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge, MA: MIT Press.

Ryan, Marie-Laure (2006). *Avatars of Story*. Minneapolis: University of Minnesota Press.

Tolomeo, Diane (1978). "Flannery O'Connor's "Revelation" and the Book of Job". In: *Renascence* 30.2, 78–90. 10.5840/renascence197830223.

Truong, Charles, Laurent Oudre, and Nicolas Vayatis (2020). "Selective Review of Offline Change Point Detection Methods". In: *Signal Processing* 167, 107299. 10.1016/j.sigpro.2019.107299.

Underwood, Ted (2019). *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press.

Wood, Ralph C. (2004). *Flannery O'Connor and the Christ-Haunted South*. Grand Rapids: Eerdmans.

conference version

jcls. Journal of Computational Literary Studies

conference version

Much Ado about Meaning: Shakespeare in Localization

Ella Montgomery[1]
Alexandra Montgomery[2]

1. School of Information, University of California, Berkeley, United States of America.
2. Independent Scholar & Practicing Psychologist,, United States of America.

OPEN ACCESS

Citation
Ella Montgomery and Alexandra Montgomery (2026). "Much Ado about Meaning: Shakespeare in Localization". In: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7998

Date published 2026-05-05 (preprint)
Date accepted tbc
Date received 2026-01-09

Keywords
intralingual translation, semantic drift, Shakespeare, modernization, localization

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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

conference version

Abstract. This study operationalizes the modernization of Shakespeare as a measurable act of localization. By analyzing parallel corpora of *Hamlet* and *A Midsummer Night's Dream* using a novel Triangulation protocol, this study analyzes the semantic drift inherent in contemporary educational modernizations of the bard. The results reveal a significant genre gap. While the logical propositions of tragedy demonstrate high semantic reversibility, the specific imagery of comedy suffers from modernization choices that resolve ambiguity and flatten aesthetic form.

1. Introduction

The many modernizations of Shakespeare's early modern English into contemporary parlance are striking cases of localization. They adapt linguistic form, cultural reference, and rhetorical density to meet the expectations of a specific, contemporary readership. As Mityagina and Volkova (2019) argue, localization involves not only a linguistic transfer, but a deliberate cultural adaptation that foregrounds the interpretive assumptions of the editor. In both the classroom and the market, these strategies are shaped by a pedagogical imperative to render opaque texts accessible, often privileging readability over historical nuance.

This tension between accessibility and authenticity recalls Friedrich Schleiermacher's classic distinction in translation theory: the translator must choose between bringing the reader to the text (foreignization) or the text to the reader (domestication). While originally formulated in the context of interlingual translation, this framework provides a useful heuristic for understanding the interpretive stakes of modernization. From this perspective, modernization risks not only altering linguistic form but also reshaping the conditions under which meaning is encountered, potentially narrowing the interpretive horizon available to first-time readers. Across pedagogical adaptations, recent anglophone editions, and performance-oriented rewritings of Shakespeare intended for contemporary audiences, there is a marked tendency toward domestication, opting for linguistic and rhetorical accessibility at the expense of historical density. While literary scholars such as Albright (2019) have critiqued this resulting "flattening" of rhetorical density, and Cross (2017) has historicized the recurring impulse to "re-speak" Shakespeare, these critiques have largely remained qualitative in nature. From a computational perspective, however, the availability of side-by-side original and modern texts offers a means of operationalizing localization not as an assumed tendency but as

a measurable textual transformation.

This paper investigates the linguistic and interpretive consequences of this localization by constructing a parallel corpus of original and translated lines for both *Hamlet* and *A Midsummer Night's Dream*. By comparing a Tragedy and a Comedy, this study seeks to quantify how editorial choices encode assumptions about the audience's interpretive capacity. The methodology moves beyond pairwise comparison by introducing a novel triangulation technique, utilizing the original Early Modern text, the Modernized translation, and a "Re-Shakespeareanized" control version (generated using an online Shakespeare translator).

This study does not advocate for modernization as a substitute for the original, but rather analyzes the 'Shakescleare' corpus as a pedagogical artifact that shapes a student's first encounter with the Bard. Modernization is treated as a pedagogical intervention that may improve immediate accessibility while also risking the production of a hermeneutical gap between readers and the linguistic-historical texture of the original text. Extending Schleiermacher's hermeneutic framework, interpretive labor is not merely a barrier to understanding but part of the process through which understanding is developed; when this labor is systematically pre-empted by modernization, readers may be less frequently required to engage in the inferential and reconstructive work that supports interpretive skill development.

Using semantic similarity metrics (SBERT) and this triangulation framework, this study seeks to identify to what extent modernization preserves the semantic core of Shakespeare's iconic lines, and where it drifts into simplification. Preliminary evidence suggests that modernization is highly selective, often preserving iconic lines verbatim while aggressively localizing complex metaphors, and that the reversibility of this translation varies significantly by genre. By quantifying these shifts, we can render visible the trade-offs between pedagogical clarity and the loss of metaphysical ambiguity, demonstrating how the modern Shakespeare is constructed through specific, measurable acts of deletion and reframing.

2. Literature Review

While this study draws on translation theory, it does not treat Shakespearean modernization as equivalent to interlingual translation. Instead, modernization is understood here as an intralingual editorial practice that selectively rearticulates Early Modern English for contemporary readers. Unlike translation across languages, modernization operates within a shared linguistic system but across significant historical and rhetorical distance, raising distinct theoretical and pedagogical concerns.

Traditional translation theory has long grappled with the ethical responsibility of the translator. In his seminal 1813 lecture, Schleiermacher posited that a translator operates within a strict binary: they must either demand that the reader labor to understand the foreign author, or they must scrub the text of its foreignness to accommodate the reader Woodstein (2024). This dilemma is framed by Schleiermacher as an "utterly foolish undertaking," not because translation should be abandoned, but because any attempt to transmit meaning across languages presupposes an arduous hermeneutic labor that can

conference version

never be complete. Ideally, Schleiermacher argued, the translator should leave the author in peace and move the reader toward him Woodstein (2024). In practice, the market for educational aids invariably chooses the opposite path, bringing the author to the reader, flattening linguistic and cultural difference to favor immediate comprehension. This alternative path would be interpreted by Schleiermacher as a conceptual fiction, insofar as it assumes that thought can be cleanly detached from the language in which it is formed Hermans (2019).

In the context of Early Modern English, this binary is particularly fraught. While "foreignization" would require the reader to grapple with the historical otherness of Shakespeare's syntax, commercial modernizations tend to adopt a strategy of extreme domestication. This preference is driven by the unique position Shakespeare occupies as both a mandatory pedagogical subject and a pillar of the cultural canon. Because the plays are ubiquitous in Western secondary education curricula, there is a powerful commercial incentive to render them immediately germane to the modern student. Where Schleiermacher treats the difficulty of understanding as the necessary condition of interpretation, pedagogical markets and publishers effectively promise to bypass the difficulty and historical friction of the text altogether. Mityagina and Volkova (2019) argue that this goes beyond mere translation; it is an act of localization. Originally a term found by this research most often in translating shows and games—adapting a product to a specific locale—literary localization involves aggressively filtering cultural and rhetorical markers to fit the immediate expectations of a target market. In this view, a text like No Fear Shakespeare (widely used educational editions of Shakespeare's works that presents Early Modern English lines alongside sentence-by-sentence paraphrases in contemporary prose) is not a "translation" in the literary sense, but a localized product designed to minimize friction for a specific demographic: the contemporary student.

Read through the lens of hermeneutical injustice, this turn toward pedagogical localization has consequences that extend beyond fidelity or stylistic preference. Following Fricker (2007), hermeneutical injustice arises when structural conditions deprive individuals of the shared interpretive resources needed to make sense of their experiences. In the context of Shakespeare pedagogy, commercial modernizations increasingly substitute interpretive labor with preemptive explanation, reshaping institutional expectations about what it means to understand a text. Historical linguistic difficulty is no longer positioned as an object of analysis and is instead a barrier to understanding to be removed. This restructuring operates at both the institutional and cognitive levels: institutionally by normalizing curricular reliance on pre-digested paraphrase, and cognitively by limiting students' opportunities to develop competence in navigating diachronic register variation, pragmatic inference, and rhetorical ambiguity. The results suggest a potential narrowing of the interpretive field, rendering readers increasingly positioned as receivers of meaning rather than as active constructors of interpretation.

Within Shakespeare studies, modernization is often treated with skepticism. Scholars of textual editing such as Tanselle (1989) argue that linguistic form is not merely a vehicle for meaning but a constitutive element of it, such that alterations to diction, syntax, or rhetorical structure necessarily reshape interpretation. Similarly, editorial theorists such as McGann (1983) emphasize that all acts of textual transmission involve interpretive intervention, with modernization representing a particularly visible and contested form

conference version

of such mediation. Within this framework, modernized editions are frequently viewed as compromising interpretive integrity by collapsing historical linguistic difference into contemporary norms. This creates a disciplinary tension: while modernization is pedagogically widespread, it remains theoretically contested within literary scholarship. This study engages that tension directly by examining modernization not as a replacement for the original text, but as a pedagogical instrument that shapes how readers first encounter it.

The impetus for this localization is what Cross (2017) terms "Shakesfear"—the cultural anxiety that Shakespeare is inherently unintelligible. Cross argues that this anxiety drives a pedagogical imperative to prioritize plot extraction over poetic form, effectively treating the text as a data delivery system rather than an aesthetic object. Albright (2019) provides a granular critique of this phenomenon, noting that modernizations often reduce the "forms, figures, and shapes" of Early Modern English into utilitarian prose. Crucially, Albright observes that this process strips the text of its polysemy. Where a Shakespearean line might hold three simultaneous meanings, the modernization collapses it into a single, unambiguous statement. This flattening suggests that localization does not just clarify the text; it narrows the interpretive field, effectively deciding for the student what the line means and discarding the surplus ambiguity. However, not all modernizations simplify in the same way. Hill-Madsen (2019) introduces the concept of intralingual heterogeneity, positing that rewritings vary wildly based on their functional purpose. A modernization aimed at actors might preserve the iambic rhythm while updating the vocabulary (to aid memorization), whereas a study guide will be unattuned to the rhythm in order to maximize syntactic clarity. This study focuses on the latter—specifically the high-resource "pedagogical localization" found in LitCharts. By treating these texts as a cohesive corpus, we can isolate specific patterns of simplification that are distinct from other forms of adaptation, such as performance or film.

3. Methodology

To operationalize the study of localization in Shakespearean modernization, this project employs a computational pipeline combining parallel corpus construction, vector semantic analysis, and a novel triangulation technique for iconic lines. Annotating the dataset for iconicity (operationalized here as a proxy for cultural salience) allows for an initial test of whether widely circulated lines are treated differently in modernization. This approach complements anecdotal close reading by quantifying systematic shifts in linguistic complexity and semantic fidelity across genre boundaries.

3.1 On the Corpus

The primary dataset consists of side-by-side translations of *Hamlet* Florman (2014b) and *A Midsummer Night's Dream* Florman (2014a). While numerous modernizations exist, notably SparkNotes' No Fear Shakespeare, this study utilizes the Shakescleare translations provided by LitCharts.[1]

LitCharts was selected as the source text for three reasons. First, it is widely used

1. Both companies were made by the same people, Florman and Kestler (n.d.)–the Litcharts webpage description is "From the creators of Sparknotes, something better."

conference version

by students and educators alike; according to usage statistics reported by LitCharts, more than 50 million students, teachers and parents have engaged with its guides, with companies such as Course Hero reporting that over 36 percent of LitCharts subscribers are classroom educators Mascarenhas (2021). Second, unlike No Fear Shakespeare, which often prioritizes colloquial slang for relatability, LitCharts aims for a register of "explanatory prose," making it a cleaner dataset for measuring semantic explication rather than stylistic slang. Where published No Fear Shakespeare volumes present the original scenes in their totality and follow them with the translations, Shakescleare entries present the original and localized lines side by side. Finally, the digital format of LitCharts (PDFs) retains rigorous spatial alignment, facilitating the extraction of parallel text blocks. The selection of *Hamlet* and *A Midsummer Night's Dream* controls for the variable of genre. *Hamlet*, defined by high rhetoric and blank verse, presents a translation challenge rooted in philosophical ambiguity. In contrast, *Midsummer* relies heavily on rhyme, song, and high-fantasy imagery. Comparing a Tragedy and a Comedy allows this study to test if localization is genre-dependent. Specifically, we hypothesize that the prose-heavy modernizations of LitCharts struggle more to capture the formal constraints of comedic rhyme than the rhetorical arguments of tragedy.

To evaluate the selectivity of modernization, the dataset was annotated for "Iconicity." A line was designated as iconic if it met two criteria: Lexicographical Canonization, inclusion in standard reference texts (specifically Bartlett's Familiar Quotations or The Oxford Dictionary of Quotations) and Recurrence, evidence of the quote circulating in wider culture as a headline or title online. This operationalization simplifies what is in reality a complex phenomenon of cultural circulation. Large-scale studies, such as the JSTOR Labs 'Understanding Shakespeare' project, demonstrate that the frequency and distribution of Shakespearean quotations can be quantified across extensive corpora of scholarly and public texts. While this study does not replicate that scale, it adopts a constrained proxy for cultural salience sufficient for a pilot analysis. Annotating the dataset for iconicity is critical because iconic lines are likely to be treated differently in localization. The iconic lines can resist full paraphrase, be selectively preserved, or receive disproportionate attention in explanatory prose. By distinguishing iconic from non-iconic lines, this paper can assess whether modernization systematically privileges salient lines, revealing how editorial decisions shape the localization process.

3.2 Data Processing

Text extraction was performed using a custom Python pipeline using pdfplumber. To preserve the integrity of the dialogue, the script employed a dynamic gutter detection algorithm. Rather than assuming a fixed page split, the algorithm calculated the vertical whitespace density for each page to accurately separate the Original (Left) and Modern (Right) columns. The extraction unit was defined not as the line, but as the full dialogue utterance. Shakespearean syntax frequently enjambs across line breaks; segmenting by line often results in fragmented, nonsensical semantic vectors. By aggregating before embedding, the analysis captures the complete semantic thought of the character.

conference version

conference version

3.3 Analytical Metrics

To quantify the vectors of localization, this study departs from standard pairwise comparison to employ a triangulation protocol. While traditional machine translation relies on n-gram overlap metrics, these are ill-suited for intralingual tasks where the objective is to alter vocabulary while preserving meaning. The core analysis of this study utilizes Sentence-BERT (SBERT), a modification of the BERT network that uses siamese structures to derive semantically meaningful sentence embeddings Reimers and Gurevych (2019). We establish a three-node relationship for each speech block: The Original (O), the Modern translation (M), and a synthetic Re-Shakespeareanized Control (C). A pairwise comparison between original and modernized texts (O–M) provides a measure of surface-level semantic divergence, but it cannot distinguish whether observed differences arise from systematic simplification, selective preservation of salient phrases, or asymmetrical paraphrastic compression. As a result, O–M comparisons conflate multiple transformation processes into a single scalar estimate of similarity, limiting interpretability. To resolve this, this study introduces a triangulation method. The third node, the Re-Shakespeareanized Control[2] (C), was generated applying a style transfer procedure that maps localized prose back toward Early Modern English[3]. The inclusion of a Re-Shakespeareanized control condition (C) is intended not as a reconstruction of historical Shakespearean language, but as a probe of semantic reversibility under stylistic re-expansion. By subjecting modernized text (M) to a constrained stylistic transformation, we introduce a third reference point that allows us to assess whether semantic information lost during modernization is recoverable through generative re-expression. Without a third condition, it is not possible to determine whether semantic compression observed in modernization is reversible or structurally irreversible. The O–M–C triangulation therefore enables the separation of two analytically distinct phenomena: (1) semantic drift induced by modernization, and (2) the extent to which that drift can be recovered or re-expressed under a generative re-stylization process. This creates a triangular relationship of similarity scores:

1. **Translation Fidelity (O↔M)**: How far did the translation drift? This metric measures the cosine similarity between the Original and Modernization. In line with the findings of Zhang et al. (2019) on paraphrase detection. A lower SBERT score indicates a high degree of semantic drift.

2. **Control Fidelity (M↔C)**: Did the style transfer accurately reflect the modern text? This measures the similarity between the Modern text and the generated Control. A high score suggests the model has successfully mapped the content of the modern sentence into the target style without introducing noise.

3. **Restoration Gain (O↔C)**: If the Control (C) is closer to the Original (O) than the Modern (M), we observe a "positive restoration," suggesting the semantic core was preserved. If the Control drifts further away, it indicates a fundamental loss of meaning in the localization process.

2. It is common for costume directors to be faced with a problem: a costume, pristine and obviously new, appears in the text as ragged, worn, or otherwise obviously lived-in. To achieve this effect they will "weather" the garments, adding wear at the elbows with gritted sandpaper and the impression of dirt/ fluids/other adulterants with paint. The term weathering would describe the process of the backwards translation to antiquated phrasing nicely, but is not used in this analysis.
3. Aptly titled Shakespearean-English-Translator: https://shakespearean-english-translator.com/

Restoration gain serves as a quantitative proxy for reversibility. When the similarity between O and C exceeds that between O and M, the result indicates a positive restoration suggesting that modernization preserved a semantic core that could be partially recovered through transformation. Conversely, when the control diverges further from the original modernization itself, this indicates that localization involves a loss of semantic content that can not be readily reconstructed. This triangulation provides a quantitative proxy for reversibility, serving as a validity check on the modernization itself. Rather than assuming that localization preserves meaning, it empirically tests whether meaning survives transformation in a form robust enough to support reverse mapping. By treating reversibility as an observable property rather than a theoretical assumption, the method allows for comparison of the effect of localization across genre (tragedy vs comedy), rhetorical features (poetic or syntactically complex lines vs straightforward prose), cultural salience (iconic quotations vs less well known lines), and line function (soliloquy vs dialogue). Statistical significance for the difference in Restoration Gain between the tragic and comedic corpora was determined using Welch's t-test for independent samples.

4. Results

The computational analysis revealed distinct patterns of semantic drift across the two plays, challenging the assumption that modernization is a uniform process of simplification. To quantify this, the study first employs a dual-metric approach comparing Base BERT against SBERT. SBERT is explicitly optimized to capture deep semantic equivalence; Base BERT is used, due to its anisotropic embedding, to measure lexical overlap, and, when contrasted against SBERT, gives an indication of distance between maintaining original vocabulary and maintenance of semantic preservation. This is called drift.

Figure 1 displays the distribution of similarity scores. A significant "Drift Gap" emerged between the Base BERT and SBERT scores across both corpora. The average BERT score (0.91) remained high, indicating that the modernized syntax roughly mirrors the original sentence structure. However, the SBERT scores fluctuated wildly, dropping as low as 0.34 for lines like "The lady doth protest too much, Meethinks" (Modernized: "The lady's promising a bit much, I think."). This specific delta (BERT - SBERT), when narrow, suggests a more literal translation. The SBERT Triangulation Analysis provided the most significant insight into genre differences. As shown in **Figure 2**, the "Restoration Gain" (the degree to which the Re-Shakespeareanized control resembled the original text) varied starkly by genre. *Hamlet*, the tragedy, demonstrated a positive restoration gain (+0.14 average). Lines such as "To be, or not to be" saw a massive restoration (+0.50), suggesting that the modern translation ("To live, or to die") retained the semantic core required for the AI to reconstruct the original phrasing. *A Midsummer Night's Dream* (Comedy) demonstrated a neutral-to-negative restoration gain (-0.01 average). In many cases, the "Re-Shakespeareanized" version drifted further away from the original than the modern translation did, suggesting that the modernization of comedy involves a loss of specific imagery that is difficult for algorithmic style-transfer to recover. To assess the statistical validity of the genre gap, an independent samples t-test was conducted assuming unequal variances. While the initial magnitude of semantic drift

conference version

conference version

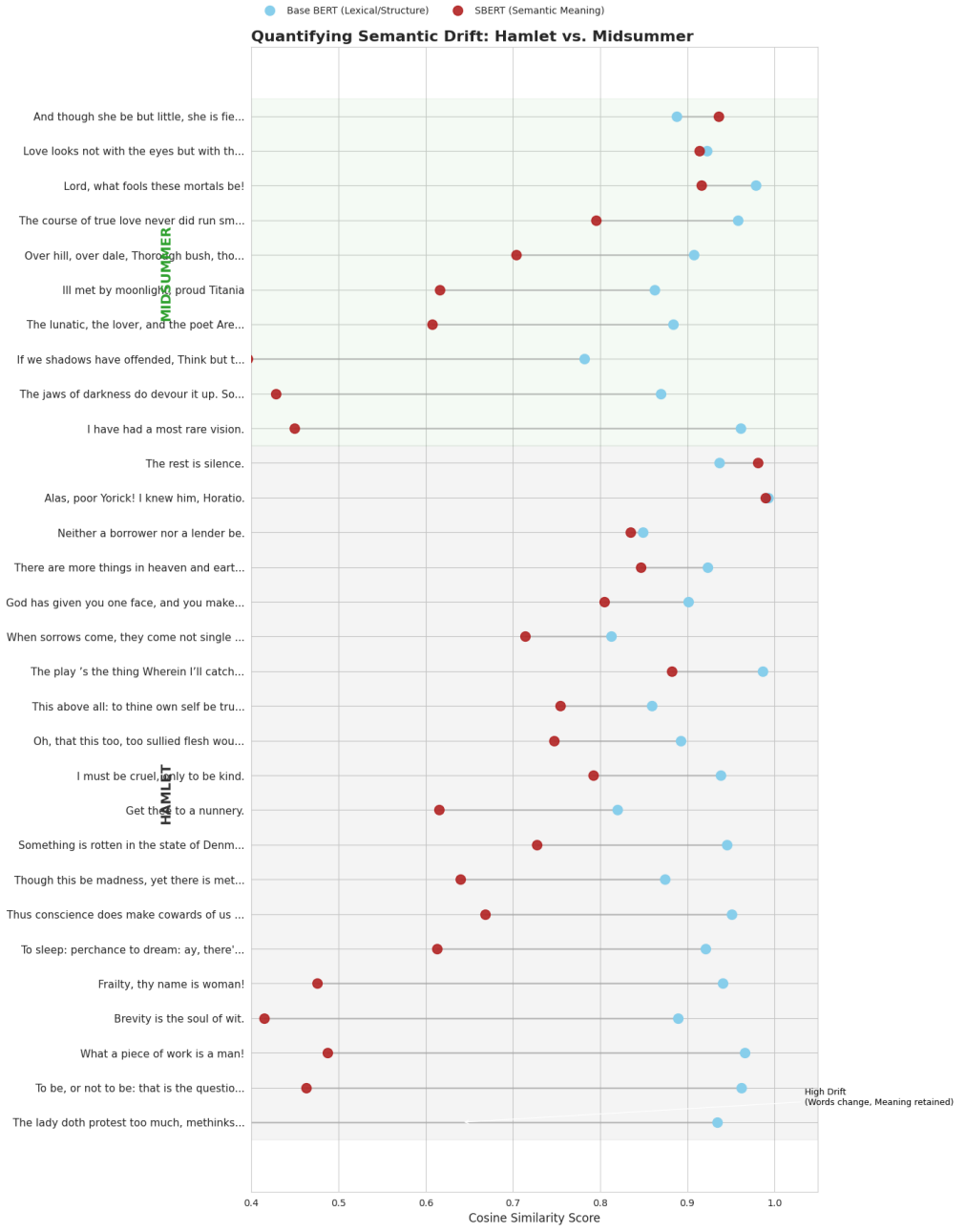


Figure 1: Visualizing semantic drift between original and modernized lines.

conference version

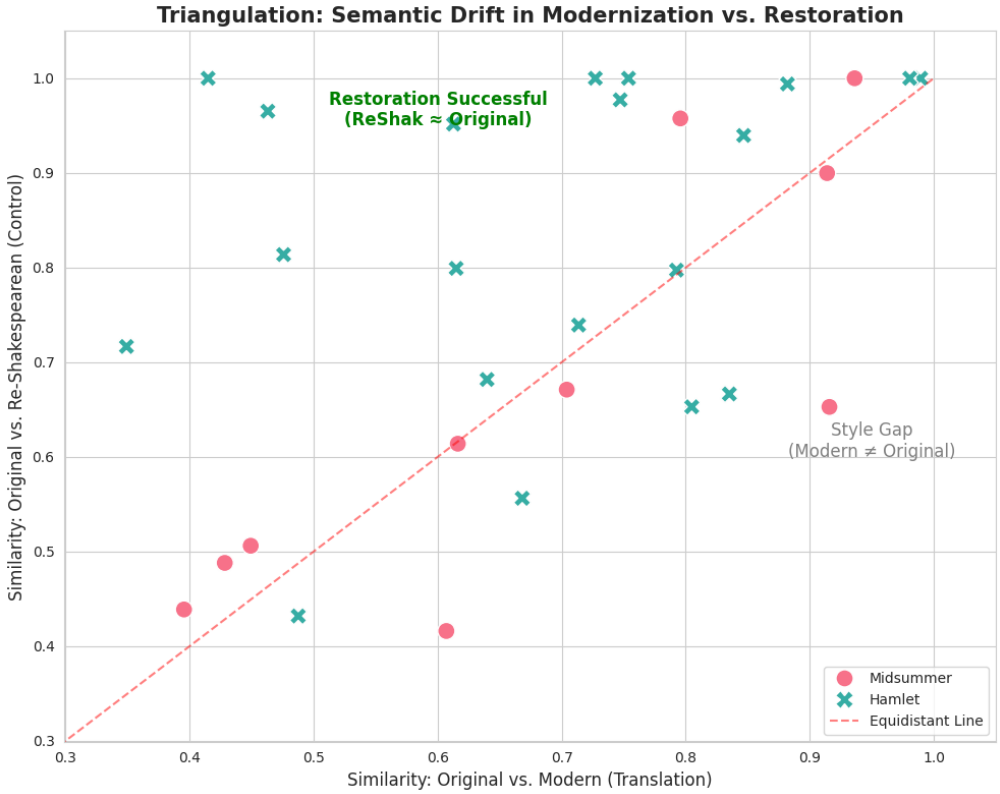


Figure 2: Measuring translation fidelity and restoration gain.

was comparable across genres ($p = 0.86$), the Restoration Gain revealed a statistically significant divergence ($p = 0.018$).

Table 1: Summary of Triangulation Analysis. Comparison of semantic preservation and reversibility across genres. Statistical significance was determined using Welch's t-test for unequal variances.

Metric	Hamlet (*Tragedy*)	Midsummer (*Comedy*)	*p*-value	Sig.
Sample Size (N)	20	10	-	-
1. Semantic Drift ($O \leftrightarrow M$)				
Mean SBERT Similarity (μ)	0.76	0.68	0.862	N.S.
Standard Deviation (σ)	0.18	0.21		
2. Restoration Gain ($O \leftrightarrow C$)				
Mean Gain (μ)	**+0.14**	**-0.01**	**0.018**	*
Standard Deviation (σ)	0.20	0.12		

While the triangulation method offers a novel quantitative framework for evaluating semantic drift, a limitation regarding the computational tools and corpus scope must be acknowledged. There is a potential for data contamination within the style-transfer model used to generate the "Re-Shakespeareanized" control. The Large Language Model (LLM) underlying this process was likely pre-trained on a vast corpus that included Shakespeare's original texts. Consequently, the "perfect restoration" of iconic lines like "To be, or not to be" (+0.50 Restoration Gain) may not reflect a successful stylistic reversal of the modern text, but rather an instance of memorization. We propose that a fully reproducible Third Angle would, in future research, be best served by an

LLM specifically fine-tuned on the Folger Shakespeare Library's digital editions.

5. Discussion

This section interprets the computational results in light of the study's theoretical framing, while remaining attentive to the limited scale and exploratory nature of the dataset. The following discussion should therefore be understood as provisional and hypothesis-generating rather than confirmatory, given the constrained corpus size and the pilot implementation of the triangulation framework. Across both Shakespeare corpora, the observed patterns of semantic drift and restoration suggest that pedagogical modernization operates not merely as a linguistic transformation, but as a restructuring of interpretive conditions. However, these computational measures (including SBERT-based similarity scores) should be treated as approximate proxies for semantic alignment rather than direct representations of meaning. Sentence embedding models are known to reflect patterns of lexical overlap and training distribution effects, and therefore provide only a partial account of interpretive or rhetorical change. Within this interpretive constraint, the results indicate that modernization tends to reduce semantic variance in passages characterized by metaphorical density, ambiguity, and rhetorical compression, particularly in the comedic corpus. Rather than treating this reduction as evidence of diminished meaning, it is more cautiously interpreted here as a compression of the range of permissible interpretive readings encoded in the text. In other words, modernization may not eliminate meaning, but it may narrow the space of interpretive possibility available to the reader. This observation can be situated at the intersection of hermeneutic and cognitive accounts of interpretation. In the tradition of Friedrich Schleiermacher, understanding is not a passive reception of semantic content but an active process of reconstructive engagement with linguistic and historical distance. From this perspective, interpretive difficulty is not incidental to understanding but constitutive of it. A complementary account emerges in cognitive psychology, particularly in the work of Bjork and Bjork, who describe learning as enhanced by "desirable difficulties" or conditions in which effortful processing, inference generation, and partial uncertainty support deeper encoding and transfer of knowledge. Interpreting Early Modern English may instantiate such conditions insofar as it requires readers to resolve syntactic ambiguity, infer pragmatic meaning, and negotiate historical linguistic distance. From this integrated perspective, pedagogical modernization can be understood as a form of instructional simplification that may increase immediate comprehension fluency while simultaneously reducing exposure to the kinds of interpretive challenges that support active meaning construction. Importantly, this does not imply that modernization inhibits learning in a deterministic sense, nor that interpretive difficulty is universally beneficial. Rather, it suggests a trade-off in the distribution of cognitive and interpretive labor: modernization shifts interpretive work away from the reader and into the editorial layer, potentially altering the conditions under which interpretive skill is practiced

The results reveal systematic patterns of localization. For example, the iconic line "To be, or not to be: that is the question" scored 0.96 in BERT but only 0.46 in SBERT when compared to LitCharts' rendering "To live, or to die? That is the question." Structurally, the sentences are nearly identical, but semantically the abstraction of "being" is reduced to the biological binary of living or dying. This shift illustrates how modernization

conference version

simplifies metaphysical ambiguity into concrete terms, aligning with a pedagogical goal of clarity but at the cost of interpretive depth.

Similarly, "The lady doth protest too much, methinks" was modernized as "The lady's promising a bit much, I think." Here, BERT scored 0.93 while SBERT dropped to 0.34. The semantic drift reflects the historical shift in the meaning of "protest": in Shakespeare's time, it meant "to vow or promise," whereas modern readers interpret it as "to object or complain." The modernization thus reverts to the older sense, but in doing so it alters the cultural resonance of the line, which has entered common parlance in its modern sense. Computational metrics capture this tension between historical fidelity and contemporary usage.

Other lines showed high semantic preservation. "Alas, poor Yorick! I knew him, Horatio" scored 0.99 in both models, indicating that LitCharts retained both structure and meaning. Likewise, "The rest is silence" scored 0.98 in SBERT, reflecting minimal drift. These cases suggest that modernization is selective: iconic lines with strong cultural currency are often preserved verbatim, while less familiar passages are more heavily localized.

These findings shed light on the cultural phenomenon of Shakespearean localization. Modernizations balance two competing imperatives: preserving iconicity for lines already embedded in cultural memory, and simplifying or reframing passages that risk alienating contemporary readers. The systematic reduction of SBERT variance observed in the *Midsummer* corpus raises a critical pedagogical question: does the removal of linguistic friction necessarily result in better comprehension, or does it result in cognitive flattening? The prevailing logic of Shakesfear assumes that difficulty functions primarily as a barrier to entry and that comprehension is best achieved by minimizing friction through paraphrase and syntactic simplification. From this perspective, editorial success is measured by immediacy of understanding and ease of informational retrieval.

From a cognitive learning standpoint this assumption is not self evident. A substantial body of research and educational psychology suggests that learning is often enhanced, not impeded by what Bjork and Bjork term desirable difficulties; obstacles that slow initial performance but promote deep processing, durable learning, and transfer of knowledge E. L. Bjork and R. A. Bjork (2011). Encountering, struggling with, and resolving difficulty requires active engagement with material, recruiting inferential reasoning, working memory, and metacognitive monitoring. In contrast, when interpretive challenges are preemptively resolved for the reader, opportunities for this form of cognitive labor are reduced.

Within this framework, the linguistic friction posed by Early Modern English may function as a desirable difficulty rather than an obstacle or barrier to learning. Parsing unfamiliar syntax, resolving semantic ambiguity, and negotiating historical/rhetorical density require readers to actively construct meaning rather than passively receive it. The observed reduction in semantic variants in the localized *Midsummer* corpus suggests that this process is systematically curtailed. By converging disparate lines towards a narrower band of semantic similarity, pedagogical localization may facilitate surface level comprehension while limiting the depth of interpretive engagement. This does not imply that accessibility and learning through desirable difficulty are in opposition, but

conference version

it does suggest that there is a qualitative difference between scaffolding that supports engagement with difficulty and localization that removes difficulty altogether. Localization risks transforming literary reading into a task of recognition rather than one of interpretation. In this sense, the computational signature of reduced SBERT variance indicates not only semantic simplification, but a pedagogical shift away from practices that cultivate sustained attention, interpretive resilience, and tolerance for ambiguity (skills central to both literary literacy and critical thinking).

The cognitive account of difficulty as productive also resonates strongly with Schleiermacher's Hermeneutic Theory of translation. For Schleiermacher, understanding a foreign text is never going to be a completed act but is instead an ongoing process of approximation sustained through effort, historical attentiveness and imaginative engagement Hermans (2019). Translation is not a shortcut to meaning but the articulation of a provisional understanding achieved through prolonged struggle. The impulse to eliminate difficulty (which Scheliermacher explicitly rejects as a conceptual fiction) misconstrues understanding as a transmissible object rather than a labor intensive process Hermans (2019). Through this lens, pedagogical localization does not only simplify Shakespeare's language but also alters the conditions under which understanding is expected to occur. By presenting meaning as immediately accessible and fully resolved, localized additions invert Schleiermacher's model, treating struggle as a defect rather than as the very medium through which interpretation and understanding unfolds. The computational reduction in semantic variants observed in the modernized corpus can be read as an empirical trace of this shift. A movement away from understanding as a hermeneutic labor and toward comprehension is informational extraction. In this sense the findings suggest that pedagogical localization does not simply mediate access to Shakespeare, but redefines what it means to understand a literary text. Computational metrics reveal how this balance is struck, quantifying the degree of semantic drift across different types of lines. The results matter because they demonstrate how editorial choices encode assumptions about audience interpretive capacity. By privileging readability, modernizations may inadvertently narrow the interpretive lens, reducing opportunities for students and readers to grapple with ambiguity and historical nuance.

This raises several interrelated questions about the broader implications of pedagogical localization. First, is the compromise between accessibility and interpretive richness ultimately beneficial? While simplification may make the text immediately approachable, it is unclear whether the intended accessibility gains are realized in practice. Second, does the process of simplifying complex passages limit opportunities for readers to develop or employ critical thinking skills? Engaging with syntactic complexity, rhetorical ambiguity, and historical linguistic variation is inherently a cognitive exercise. When the aforementioned challenges are preemptively resolved by the prose produced through localization, readers may receive answers rather than practice reasoning. Third, are some forms of literary meaning (such as humor, wordplay, or culturally contingent imagery) structurally resistant to localization? As the SBERT Triangulation analysis suggests, comedic lines often lose specificity that AI-based reconstruction can not recover, which points to an interpretive space that is very human and difficult to simplify or treat algorithmically. All three questions suggest that localization reshapes not only textual accessibility but the cognitive and interpretive practices that underlie meaningful engagement with Shakespeare.

conference version

Broadly, the Re-Shakespeareanized control text (C) generated by the style-transfer model presents a distinct linguistic profile: it functions as a pseudo-archaic hybrid that adopts the vocabulary of the Early Modern period while retaining the syntax of the modernization. The resulting text possesses a veneer of historical distance—making it sound arcane through the mechanical elevation of diction (e.g., swapping "know" for "perceive" or "you" for "thou"), but it lacks the structural density of the original text. While individual words are rendered more complicated, the sentence structure remains simple. The model abjures the complexity of Shakespearean syntax in favor of the direct Subject-Verb-Object (SVO) word order found in the modern translation. Crucially, the control text systematically fails to recover the formal constraints of the genre. The obligations to rhyme, meter, and lyricism are abandoned in favor of direct prose. In *A Midsummer Night's Dream*, where the original relies on the specific cadence of rhyming couplets to convey the magical or comedic tone, the Re-Shakespeareanized version produces utilitarian sentences with archaic vocabulary. As a control generated from the modernizations, this shows that the text becomes more direct, prioritizing the transmission of information over the restoration of aesthetic form.

From Fricker's lens, this suggests a genre-specific form of hermeneutical injustice: individuals are more likely to be structurally deprived of the resources needed to interpret comedic meaning than the rhetorical density of tragedy. Pedagogical localization inadvertently reshapes the interpretive experience differently depending on literary form, reinforcing systemic gaps in how readers encounter culturally and linguistically complex texts. The observation that AI-assisted reversion struggles to reconstruct the semantic specificity of comedy further illuminates the cognitive dimension of hermeneutical injustice. Where algorithmic style-transfer cannot recover lost meaning, the pedagogical and editorial choice to preemptively simplify may remove the interpretive scaffolding that audiences would otherwise exercise. In this sense, computational metrics make visible the structural constraints Fricker describes: the interpretive labor required to understand Early Modern English is systematically offloaded, creating conditions in which readers' epistemic agency is constrained. Over time, these structural constraints may produce a subtle but cumulative effect on literary literacy: consumers of localized texts may become adept at recognizing familiar phrases and plot points, yet have fewer opportunities to engage with syntactic ambiguity, rhetorical subtlety, and metaphorical density. Hermeneutical injustice operates not just in isolated moments, but as a patterned feature of the localized corpus, shaping how interpretive competence is cultivated or restricted across different contexts. Future implementations of the triangulation framework would replace this provisional component with a controlled sequence-to-sequence or neural style-transfer model trained on Early Modern English corpora, enabling parameter transparency, reproducibility, and systematic evaluation of reverse semantic recoverability independent of black-box constraints

The instance of the Re-Shakespeareanized version of *A Midsummer Night's Dream* drifting further from the original than the localized text itself suggests that comedy presents unique challenges for both human and algorithmic reconstruction. As Schleiermacher argues, understanding works in a distant tongue is inherently partial, requiring sustained hermeneutic effort. The Re-Shakespeareanized control illustrates that algorithmic reconstruction struggles with passages whose semantic and cultural density is particularly high, highlighting the limits of any intervention that aims to fully re-

conference version

store Early Modern English meaning. One possible explanation is that the stylistic and semantic choices in comedic passages are heavily tied to culturally specific imagery, wordplay, and pragmatic nuance, which are features that are often smoothed over or simplified in localization. Unlike the rhetorical density of tragedy (which may map more systematically to formal patterns that AI can exploit), the humor and fantastical imagery in comedy may lack readily accessible corollaries in contemporary language. This makes algorithmic style transfer less able to recover the original phrasing. In this sense, the AI's difficulty may reflect both the effects of localization (where cultural and poetic richness is intentionally simplified) and the context dependent, human nature of comedic expression. This suggests that localization does not operate uniformly across genres and suggests that semantic preservation in comedy is particularly vulnerable to loss during pedagogical simplification.

A significant anomaly occurs within the *Hamlet* corpus, where specific lines are restored to their exact original equivalent. While the iconic "To be, or not to be" demonstrated a near-perfect lexical and semantic restoration, a more complex anomaly emerges with the line: "The lady doth protest too much, methinks."

Node	Text Content	SBERT Score
Original (O)	The lady doth protest too much, methinks.	-
Modern (M)	The lady's promising a bit much, I think.	0.349
Control (C)	The lady doth promise o'er much, methinks.	0.716

Table 2: Semantic vs. Syntactic Restoration: The triangulation detects that while the definition changed, the syntax was restored.

The quantitative metrics here tell a legible story of philological correction at odds with cultural resonance. The Translation Fidelity is low at 0.349, this severe drop occurring because the modern editor has correctly identified that the Early Modern definition of "protest" meant "to vow or promise," whereas the contemporary definition implies objection. By translating "protest" to "promising," the editor achieves semantic accuracy but destroys the lexical link to the famous quotation. However, the Restoration Gain rebounds significantly to 0.716. The Re-Shakespeareanizes control did not revert to the iconic word "protest" as it did with "to be"). Instead, it took the modern concept ("promising") and successfully elevated the diction to a pseudo-archaic register ("promise o'er much") while restoring the syntactic tag ("methinks"). This indicates that the model is functioning generatively. The semantic proposition of the text–that the lady is making too many vows–is translated into Early Modern style without accessing the specific lexical token "protest." The result is a translation of the meaning of the original line that sounds wrong to the cultural ear.

The question of the exact re-creations in the triangulation process remains. This high fidelity may be an indicator of style transfer, though another possible explanation is that this re-creation represents an artifact of the model's training data. In these instances, the process seems to be more recollection than generation. Highly iconic lines may undergo some form of canonization. The selective preservation of culturally salient/canonized/iconic lines raises a deeper question about the co-construction of

conference version

culture and pedagogy. Are iconic Shakespearean quotations retained because they have already entered the fabric of everyday cultural understanding, or do curricula and study aids shape cultural valuation to reflect the perceived importance of these lines? This question is central to the hermeneutical injustice framework: the structural conditions that make some passages easily interpretable are not neutral; they are historically and socially situated, privileging certain semantic and cultural forms while marginalizing others. From this perspective, the consistent preservation of culturally prominent lines may amplify epistemic inequities. Readers repeatedly encounter passages whose meaning is largely pre-encoded by shared cultural knowledge, while less canonical lines are simplified or discarded, limiting opportunities to practice interpretive reasoning. In effect, cultural capital functions both as a guide and as a constraint, determining which lines retain their semantic richness and which are pre-digested for comprehension.

6. Conclusion

This study set out to operationalize the modernization of Shakespeare as a measurable act of localization. By constructing a parallel corpus of *Hamlet* and *A Midsummer Night's Dream* and applying a novel triangulation method, this research has quantified the semantic drift that occurs when Early Modern English is filtered for contemporary consumption. The results suggest that modernization is a selective, genre-dependent process: while the logical propositions of tragedy often survive the round-trip of translation, the specific aesthetic DNA of comedy suffers.

Ultimately, these computational metrics offer empirical support for the qualitative concerns regarding the pedagogy of simplification. By resolving ambiguities and flattening metaphors, modernizations enforce a single, pre-digested interpretation upon the reader. However, these findings are preliminary. Future research must situate itself more robustly within the NLP literature to refine the triangulation metrics, specifically addressing the effect observed in the restoration of iconic lines. A larger-scale analysis expanding the LitCharts corpus to the entire available Shakespearean theatrical canon is necessary to determine if the genre gap observed here is a universal feature of comedic localization or specific to the high fantasy of *Midsummer*. The addition of competitor aids such as No Fear Shakespeare and No Sweat Shakespeare would allow for the identification of distinct house codes of localization. By applying clustering algorithms to the drift metrics, research would be able to identify unique stylometric fingerprints, and by differentiating these house styles engender mapping the landscape of ed-tech and quantify how different pedagogical philosophies materialize in text.

Further, the emergent phenomenon of recursive relevance should be addressed. As students increasingly use LLMs to translate or summarize Shakespeare, they are engaging with models trained on these modernized and localized texts. This creates a feedback loop in which the simplified, pedagogically optimized versions of Shakespeare are propagated and amplified, which reinforces the interpretive scaffolding provided by localization. The consequences are twofold, not only does the original text risk further abstraction from its historical and rhetorical density, but the interpretive experience of learners may become increasingly constrained by the assumptions embedded in the localized/modernized corpora. Understanding and quantifying this recursive effect will

conference version

be critical for future research seeking to reconcile accessibility, pedagogy, and literary fidelity.

7. Data Availability

Data can be found here: https://doi.org/10.5281/zenodo.19700318.

8. Software Availability

Software can be found here: https://doi.org/10.5281/zenodo.19700318.

9. Acknowledgements

Thanks to Kent Chang for his primer in Cultural Analytics and willingness to be waylaid before and after office hours, and our parents, for everything.

10. Author Contributions

Ella Montgomery: Conceptualization, Data Curation, Formal Analysis, Software, Visualization, Writing – original draft

Alexandra Montgomery: Investigation, Writing - original draft

References

Albright, Michael Andrew (2019). "Modern, yet "full of forms, figures, shapes, objects": The Trouble with Translating Shakespeare's English into English". In: *Early Modern Culture Online* 7, 1–18. 10.15845/emco.v7i1.2833.

Bjork, Elizabeth Ligon and Robert A. Bjork (2011). "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning". In: *Psychology and the real world: Essays illustrating fundamental contributions to society*. New York, NY, US: Worth Publishers, 56–64. ISBN: 9781429230438. https://bjorklab.psych.ucla.edu/wp-content/uploads/sites/13/2016/04/EBjork_RBjork_2011.pdf.

Cross, Lezlie C. (2017). "Historicizing Shakesfear and Translating Shakespeare Anew". In: *Theatre History Studies* 36.1, 211–230. https://muse.jhu.edu/pub/181/article/679351 (visited on 01/08/2026).

Florman, Ben (May 11, 2014a). *A Midsummer Night's Dream: A Shakescleare Translation*. LitCharts LLC. https://www.litcharts.com/shakescleare/shakespeare-translations/a-midsummer-nights-dream.

— (May 11, 2014b). *Hamlet: A Shakescleare Translation*. LitCharts LLC. https://www.litcharts.com/shakescleare/shakespeare-translations/hamlet.

Florman, Ben and Justin Kestler (n.d.). *Our Story from Sparknotes to LitCharts*. LitCharts. https://www.litcharts.com/our-story-from-sparknotes-to-litcharts.

conference version

Fricker, Miranda (June 1, 2007). "Hermeneutical Injustice". In: *Epistemic Injustice: Power and the Ethics of Knowing*. Ed. by Miranda Fricker. Oxford University Press. ISBN: 9780198237907. 10.1093/acprof:oso/9780198237907.003.0008. (Visited on 01/08/2026).

Hermans, T. (2019). "Schleiermacher". In: *The Routledge Handbook of Translation and Philosophy*. Ed. by J. P. Rawling and P. Wilson. London, UK: Routledge, 17–33. https://doi.org/10.4324/9781315678481-2.

Hill-Madsen, Aage (2019). "The Heterogeneity of Intralingual Translation". In: *Meta: Translators' Journal* 64.2, 537–560. 10.7202/1068206ar. (Visited on 01/08/2026).

Mascarenhas, Natasha (June 10, 2021). *Course Hero acquires LitCharts, founded by the creators of SparkNotes*. TechCrunch. https://techcrunch.com/2021/06/10/course-hero-acquires-litcharts-founded-by-the-creators-of-sparknotes/.

McGann, Jerome J. (1983). *A Critique of Modern Textual Criticism*. University of Chicago Press.

Mityagina, Vera and Irina Volkova (2019). "Localization in translation theory and practice: historical and cultural view (the case of fiction adaptation)". In: *SHS Web of Conferences* 69, 00129. 10.1051/shsconf/20196900129.

Reimers, Nils and Iryna Gurevych (Aug. 27, 2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 10.48550/arXiv.1908.10084. http://arxiv.org/abs/1908.10084.

Tanselle, G. Thomas (1989). *A Rationale of Textual Criticism*. University of Pennsylvania Press.

Woodstein, B.J. (2024). *Translation Theory for Literary Translators*. Anthem Press. ISBN: 9781839992070. 10.2307/jj.15729456.

Zhang, Yuan, Jason Baldridge, and Luheng He (Apr. 1, 2019). *PAWS: Paraphrase Adversaries from Word Scrambling*. 10.48550/arXiv.1904.01130. http://arxiv.org/abs/1904.01130.

conference version

jcls. Journal of Computational Literary Studies

conference version

200 Years of Children in the Novel
On their Visibility, Value, and Agency

Andrew Piper[1]

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

OPEN ACCESS

Citation
Andrew Piper (2026). "200 Years of Children in the Novel. On their Visibility, Value, and Agency". In: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7987

Date published 2026-05-05 (preprint)
Date accepted tbc
Date received 2026-01-03

Keywords
Childhood Studies, Children in Literature, Narrative Agency, Large Language Models, Computational Humanities

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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. Across modern literary history, children have served as powerful symbolic figures through which societies negotiate core social values. Yet despite their prominence in canonical criticism, we lack a systematic account of how often children appear in texts, how they are valued, and what kinds of agency they are afforded over time. To address this gap, I conduct a large-scale computational analysis of 9,544 English-language novels published between 1800 and 2000. This study provides the first large-scale identification of child characters within the modern novel, combining classical natural language processing with large language models to measure their visibility, emotional valuation, narrative agency, and independence from kinship structures. Contrary to dominant histories of childhood, children and adolescents become progressively less visible in the novel over time, a decline driven by the systematic reduction of adolescent girls in particular. Children maintain a more positive emotional valence than adults even as their agency lags behind adult peers, suggesting a broader portrait of children as sentimentally valued, yet structurally constrained social actors.

1. Introduction

Children have long functioned as screens onto which cultures project some of their most deeply held social values. Innocence, vulnerability, discipline, transgression–these are just some of the ways children absorb and reflect back social norms. As a privileged site of cultural expression, the novel has long reflected these symbolic meanings of children and childhood. Yet while children populate some of the novel's most memorable scenes, from Charles Dickens to Toni Morrison, their broader presence remains largely undocumented. We know little about how often children appear, how they are valued, and what kinds of agency they are afforded in novels over the past two centuries.

This article begins to address this gap through a large-scale computational analysis of two centuries of English-language novels spanning the years 1800-2000 (N=9,544). Utilizing the affordances of classical methods in NLP combined with large language models, I track the visibility of child characters, measure the sentiments surrounding them, and examine their perceived agency over time. In doing so, I offer the first systematic account of how the modern novel has imagined children as social actors over the past two centuries.

conference version

Representations of Children

The cultural status of children has long been understood as a barometer of broader social values (Ariès 1962; Jenks 2020). Historians of childhood have traced the shifting notions of children's value within society, especially around issues of independence, labor and play (Jenkins 1998; Nelson et al. 2024; Zelizer et al. 1985). The nineteenth-century in particular is often seen as a watershed moment in the pivot from children understood as young workers to a protected class in need of time and space to develop. Across literary, visual, and media studies, scholars have consistently highlighted the motif of the "innocent child" as one of the more predominant ways that children come to be understood by the nineteenth century (Bernstein 2011; Gubar 2005; Jenkins 1998; Olson and Rampaul 2013; Wood 2012). As Jenkins 1998 writes, "The innocent child wants nothing, deserves nothing, and demands nothing – except, perhaps, its own innocence" (4). The innocent child is emptied of agency, but one who, precisely through its fixed nature, can serve as a powerful moral compass for adults. In response, scholars have drawn attention instead to the ways in which children are often depicted with desires and wants, such as the "erotic child" (Kincaid 1992), as they occupy a more ambiguous moral space. Such "degenerate innocents" (Thiel 2012), "threatening imps" (Olson 2013), "demons" (Pifer 2000; Renner 2016), or "revolting children" (Scahill 2015) have a long history in literature and the arts, often used as foils to hold up a proper view of the moral order.

Research has also highlighted how such cultural constructions of childhood are deeply gendered. Girls in particular are more often cast as embodiments of innocence, purity, and passivity, while boys are aligned with mischief, willfulness, or early labor potential (Bernstein 2011; Nelson et al. 2024; Pifer 2000). These gendered scripts not only shape the symbolic uses of children in literature but also structure the kinds of agency or lack thereof that can be attributed to them, reinforcing broader ideologies of masculinity and femininity.

Research on the representation of children and childhood thus converges around a general consensus of the modern significance of the innocent or angelic child. At the same time, there is also a consensus surrounding the emergence in the twentieth century of the so-called "developmental child," with a greater focus on life stages, rites of passage and caregiving (James et al. 1998; Jenks 2020; Steedman 1995). At the core of this work lies a central assumption surrounding the value and agency of children. Innocents and angels are emotionally and morally idealized, yet passive, while imps and demons are negatively coded, yet agentially empowered. In this view, agency and moral ambiguity are positively coupled, such that the more children gain the former the more the latter increases as well. The developmental child is in some sense a hybrid version: we should expect to see new forms of agency emerge that are still bounded by strong kinship ties and the distinctiveness of childhood.

Narrative Agency

As a theoretical construct, the concept of agency offers a means of analyzing how subjects are framed in relation to the symbolic and social structures that shape their capacity to act. One of the core ways human agency has been studied is through the psychological

conference version

construct of the *agentic self*. Building on Self-Determination Theory (Deci and Ryan 2012; Ryan and Deci 2000), scholars such as Little et al. (2002, 2006) define agency as the capacity for self-caused action—behavior that originates volitionally from the individual rather than being externally imposed. An agentic person is thus understood to be the origin of their own actions, acting in ways that are self-regulated and goal-directed rather than merely reactive. From this perspective, self-determination is the process by which individuals utilize their will to plot and navigate a chosen course through the uncertainties of their social environments. Agency, in this view, contributes to a sense of causal efficacy: people come to see themselves as authors of their actions by consistently engaging in behaviors that exert causal force on the world.

Central to such human agentic theories is an emphasis on the dynamic relation between person and environment. Agency is not simply an inner trait but a process of ongoing negotiation with contextual constraints and affordances (Little et al. 2002, 2006; Shogren et al. 2017). In this way, such psychological models are well-aligned with narratological constructs that focus on questions of *narrative agency*. In narratology, questions of agency have often been approached through models that formalize the roles characters play within stories. Greimas's actantial theory provides one of the most influential frameworks for analyzing narrative agency (Greimas 1984; Greimas and Porter 1977). Rather than focusing on characters as fully individuated psychological beings, the actantial model reduces narrative to a set of relational roles—subject, object, sender, receiver, helper, and opponent—that structure the dynamics of plot. Agency in this framework is not simply a matter of internal volition but of occupying a *functional position* in the unfolding of the story. In this sense, agency is relational and distributed, emerging from the configuration of roles that a narrative assigns rather than from individual psychology alone.

We can start to see how these theoretical constructs complement one another. Self-determination theory gives us a framework to understand and model the inner-drives of fictional characters, while narrative actantial theory provides a framework to study the semantic positioning of characters with respect to other characters. Building on these complementary theoretical perspectives, I conceptualize the representation of children's agency in the novel through four interlocking dimensions: their visibility, their value, the power and agential quality of their actions, and their independence from kinship ties. No single measure by itself captures the full complexity of narrative agency, but taken together these concepts can bring us towards a more holistic understanding of children's agency in fiction. I define these four concepts in the following way:

> **Visibility** refers to the frequency and narrative presence of child characters across the history of the novel. As narrative actantial theory suggests, the "child role" must be occupied and substantial for children to exercise narrative agency. Shifts in their narrative visibility offer an initial, but by no means sufficient index of their cultural relevance.
>
> **Value** captures how positively or negatively children are framed by their narrative environments. While visibility marks a character's presence, value reflects the narrative's orientation toward that presence. It offers a further contextual view of a character's narrative role. It can also help identify the potential moral valence of child characters so central to theories surrounding

their agency.

Agency captures how grammatically and semantically agential children are in a narrative. In psychological terms, agency entails self-determined, goal-directed action. In narrative terms, agency refers to the subjective positioning of characters (are they the initiators of action?). To capture this multidimensional construct I will be utilizing a series of measures that include grammatical analysis (agent-patient ratios), predicate-argument structures (i.e. verb frames), and emotional intensity (e.g. arousal and dominance) that help signal subjective agency in characters.

Independence refers to the extent to which child characters are embedded within explicit parentally structured social frameworks. From an actantial perspective, the presence of parents often signals that a child character is operating within a relational schema in which adult figures constrain the child's narrative trajectory, referred to in the psychological literature as "containment theory," where parents provide a structural frame to facilitate developmental success on the part of the child (Bion 1985; Douglas 2007; Schneider et al. 2003). Measuring parental presence offers an initial way to assess how frequently children are positioned within clear familial plots, offering another structural indicator of children's agential freedom.

By translating these theoretical constructs into measurable textual features, this study brings longstanding debates about representations of childhood into the domain of large-scale, empirical cultural analysis rather than focusing on a handful of emblematic texts. In doing so, I advance five hypotheses that structure my analysis:

H1. *The Visible Child.* As children become more socially important, they should appear more frequently in novels over time. Their "pricelessness" should correlate with increased presence, initially favoring boys with a trajectory towards greater gender equality.

H2. *The Good Child.* In keeping with prior work on the innocent or angelic child, children should begin by being more positively valued than adults in the nineteenth century, with girls exhibiting significantly more positive sentiment. This initial distinction should then decline over time as children become more agential and thus more morally ambiguous. Given historical norms that favor "good girls" we should expect girls to consistently maintain more positive valence than boys.

H3. *The Agentic Child.* We should observe all measures of agency increasing with respect to child characters with numbers rising faster for boys than girls given general historical gender stereotypes.

H4. *The Independent Child.* As children grow in prominence, we should see parents playing a less significant role in their lives, i.e. parents of children should be less frequent in novels where children become more frequent, though this is expected to have a weaker effect for girls, who are likely to be more strongly embedded within family structures.

H5. *The Finite Child.* Adulthood should gradually be decoupled from

conference version

parental presence, such that fewer adults are depicted with parents among the main characters of a novel, marking childhood as a bounded rather than lifelong condition.

Together, these hypotheses capture a broader expectation: that the novelistic imagination of childhood evolves from visibility without agency toward a more complex, independent, and finite conception of the child.

2. Materials and Methods

2.1 Data

Two collections of novels are used in this study. The first is a set of 8,916 English-language novels published between 1880 and 2000 collected by the Chicago Textual Optics Lab (Textual Optics Lab 2023). This collection represent a range of genres, styles, authors, and historical periods. The second is a collection of 628 English-language novels published between 1800 and 1900 collected as part of the Chadwyck-Healey Nineteenth-Century Fiction corpus collected by the Stanford Literary Lab (Algee-Hewitt 2024). This represents largely canonical novels published in the British Isles and North America.

2.2 Independent Variables

The following independent variables are used in this study:

- `decade` Decade of publication drawn from book metadata.
- `bookID` Unique identifier for each book in the corpus.
- `characterID` Unique identifier for each character of a book.
- `age_category` Three-level age category: child (0-12), adolescent (13-19), adult (20+).
- `predicted_gender` Two-level gender assignment: he/him and she/her.
- `is_child_of` Binary label of whether a character's parent appears in the top twenty characters.

To derive my independent variables, I use the following methods:

2.2.1 Character Detection and Gender Prediction

To identify characters in each novel and predict their gender, I use David Bamman's BookNLP "large model" (Bamman 2025), a natural language processing pipeline that includes: part-of-speech tagging, dependency parsing, entity recognition, character name clustering (e.g., "Tom," "Tom Sawyer," "Mr. Sawyer," "Thomas Sawyer" → TOM_SAWYER), co-reference resolution, and referential gender inference (TOM_SAWYER → he/him/his). Thus for every detected character I identify a predominant name, co-reference ID, predicted gender, and total number of mentions per book. For every character *occurrence* I also capture its part-of-speech (subject, predicate, prepositional object, etc.) and the full sentence in which it occurs.

conference version

2.2.2 Age Estimation and Parental Ties Detection

For the purposes of age estimation, I integrate the BookNLP output with a candidate large language model (Gemini 2.0 Flash) able to handle very long contexts appropriate to the novel data. For each novel, I supply the model with the full text of the book and a list of the top twenty named characters extracted by BookNLP. Character mentions generally follow a power law, with the top twenty characters accounting on average for over 80% of all character mentions in a book. This allows me to condition on significant characters to the narrative without introducing noise from rarely occurring entities.

Using the structured prompt shown below, the model was asked to assign each character in a book's list of characters to one of three age categories—child (0–12), adolescent (13–19), or adult (20+)—and provide a confidence level and brief textual justification for the classification. Characters were subsequently removed in cases where the model tagged the character as animals using a limited regex expression (`\b(?:pet|dog|cat)s?\b`). Horses were not included because there were numerous instances of horses being mentioned that were not related to the character.

In addition to estimating ages, the model was also instructed to identify explicit parent–child relationships among major characters (those in the top twenty). Each character was labeled as either is_child_of (true/false) with an associated parent name when true. The resulting dataset provides, for each character, a predicted age category, confidence score, and relationship status.

2.2.3 Gender and Age Validation

In order to evaluate the accuracy of this workflow, a team of three student annotators reviewed a random sample of 150 characters drawn evenly from our three age categories. Table 1 provides an overview of accuracy statistics across categories. The main source of error was adult characters predicted to be children by the model (out of 61 adults 9 were predicted to be children and 3 adolescents). We found that the majority of those errors (7) were cases of children/adolescents who became adults over the course of the novel, but whose predominant state was adult. One error was a general family name falsely attributed to a child, two were ambiguously defined young people at the boundary between adolescent and adult, and two were horses. Future improvements should include filtering for animals at the prompt level and providing instructions to assess age as a character's predominant state. As I will show, we can incorporate these levels of error into the study using sensitivity analysis.

Category	Class	Precision	Recall	Sensitivity	Specificity	F1
Age	child	0.800	0.930	0.930	0.906	0.860
	adolescent	0.878	0.956	0.956	0.942	0.915
	adult	0.980	0.803	0.803	0.989	0.883
Gender	she/her	0.951	0.906	0.906	0.963	0.928
	he/him	0.929	0.963	0.963	0.906	0.946

Table 1: Validation metrics for age and gender classification

conference version

conference version

Prompt Structure for Character Age Classification

```
You are a literary analysis expert. Please analyze the following book text
and estimate the age category for each named character.

CHARACTERS TO ANALYZE:
[character_list]

AGE CATEGORIES:
- child: 0-12 years old
- adolescent: 13-19 years old
- adult: 20+ years old

PARENT-CHILD RELATIONSHIPS:
- is_child_of: true if the character is clearly identified as the child of
another important character in the novel (regardless of their current age)
- is_child_of: false if no clear parent-child relationship is established with
other important characters

INSTRUCTIONS:
1. Read the book text carefully
2. For each character, determine their age category AND whether they are the
child of another important character based on textual evidence
3. If age is unclear, make your best estimate based on context clues
(relationships, roles, behavior, etc.)
4. Return results in JSON format only

REQUIRED JSON FORMAT:
{
  "book_id": "[book_id]",
  "characters": [
    {
      "name": "character_name",
      "age_category": "adult|adolescent|child",
      "confidence": "high|medium|low",
      "evidence": "brief explanation",
      "is_child_of": true|false,
      "parent_name": "name or null"
    }
  ]
}

BOOK TEXT:
[book_text]
```

2.2.4 Sentence Sampling

As support for my measurements, I extract sample sentences for every detected character using the BookNLP tokenization results along with the character's grammatical role identified by the dependency parsing (nsubj, dobj, pobj, etc.). Given the imbalanced nature of the collections as well as adult versus child characters, I sample according to the following criteria:

Collection	Age	Characters (per book)	Sentences (per character)	Total Sentences
Stanford	Child	Top 20	100	105,285
Stanford	Adult	Top 20	10	107,224
Chicago	Child	Top 20	10	192,219
Chicago	Adult	Top 3	10	236,776

Table 2: Sentence sampling scheme by collection and age group.

2.3 Dependent Variables

A. Visibility

2.3.1 Child Character Mention Rate

The proportion of character mentions attributable to each age category within books. For a given book b, let $M_{a,b}$ denote the number of mentions of characters in age group a (restricted to the top twenty most-mentioned characters), and let M_b denote the total number of mentions for all such characters in the book. Then

$$\text{CharacterMentionRate}_{a,b} = \frac{M_{a,b}}{M_b}.$$

This continuous measure captures the relative narrative attention devoted to each age group and allows for the assessment of changes in the prominence of children across time and gender.

B. Value

2.3.2 Valence

The psychological tradition of Valence–Arousal–Dominance (VAD) modeling (Bradley and Lang 1994; Mohammad 2018) conceptualizes emotion along three continuous dimensions: valence (pleasant–unpleasant), arousal (calm–excited), and dominance (submissive–in control). This tripartite framework is widely used in computational linguistics and psychology as a parsimonious way to capture affective meaning, and it aligns closely with theories of agency: high dominance and arousal often correspond to characters portrayed as active, powerful, and agentic, while low scores indicate passivity or subordination.

To calculate scores, I apply the NRC VAD Lexicon (Mohammad 2018), which provides human-annotated valence, arousal, and dominance scores for over 20,000 English words. This resource has been validated across multiple corpora and has become a standard dictionary for large-scale affective text analysis. I then aggregate all sentences in which each character appears from the above-mentioned sentence tables, tokenize the resulting text into words, and match those tokens against entries in the NRC VAD Lexicon. Each matched word contributes to its **valence score**, which are then averaged across all words associated with a given character. The result is a continuous measure from 0 to 1 that captures positive and negative associations of a given character.

conference version

C. Agency

2.3.3 Subject-Object Ratio

Subject-object ratio is designed to capture the grammatical agency afforded to child characters. Linguistic research shows that, across languages, the syntactic subject of a transitive clause is typically the agent or initiator of an event, whereas the object often encodes patienthood or passivity (Langacker 1991), as can be seen in these two examples from the data with respect to the character "Lonnie":

1. *Lonnie* gazes at me. NSUBJ
2. I'll be up here all day with *Lonnie* and the children. POBJ

Moreover, functional and typological theories of syntax–semantics interface have long modeled this pattern as a mapping of thematic roles (agent, theme) onto syntactic functions (subject, object). A high subject/object ratio should therefore serve as a reliable proxy for narrative agency.

To measure grammatical agency, I calculate how often a character appears as the *subject* of a clause (the initiator of an action) versus as the *object* (the one being acted upon). Dependency parses from BookNLP are used to identify these roles: tokens marked with the dependency label `nsubj` are counted as subjects, while tokens with labels containing `obj` (e.g., `dobj` and `pobj`) are counted as objects.

Formally, for each character i in book b, I count the number of times they occur as subject (subj_{ib}) and as object (obj_{ib}), and compute the proportion of subject occurrences as:

$$\text{Subject–Object Ratio}_{ib} = \frac{\text{subj}_{ib}}{\text{subj}_{ib} + \text{obj}_{ib}}$$

This value ranges from 0 to 1, where 0 indicates the character only appears as an object, 1 indicates the character only appears as a subject, and intermediate values reflect the balance between being an initiator and a recipient of action.

2.3.4 Arousal and Dominance

Here I use same method as for valence for the other two dimensions of the VAD lexicon (Mohammad 2018) to capture emotional dimensions of agency.

2.3.5 Power-Agency Frames

The concept of power and agency connotation frames originates in work by Sap et al. (2017) on film scripts. Their central insight is that verbs implicitly project connotative meanings about the power relations and agency levels of the participants in an action. For example, an agent who *implores* a tribunal is portrayed as less powerful than the tribunal, whereas an agent who *demands* is portrayed as more powerful. Likewise, a character who merely *waits* is represented with low agency, while one who *decides* or *commands* is attributed high agency. In this framework, power is relational (whether the agent has more or less authority relative to the theme (object) of the verb), while agency captures the degree to which the agent is portrayed as active, decisive, and self-determining.

conference version

Using the power/agency connotation frames lexicon from Sap et al. (2017), I evaluate the verbs associated with each character's actions in order to measure their relative agency and power. The lexicon assigns each verb two kinds of connotative information: whether the agent of the verb is portrayed with high, neutral, or low agency, and whether the agent holds more, less, or equal power than the theme (object) of the verb. For every verb where a character is the subject/agent, I record the corresponding agency score and assign power accordingly (e.g., if the subject "commands" someone then they are assigned positive agency and power respectively). Conversely, when a character is the object/theme, the polarity is reversed (e.g., being the recipient of a command implies lower power and agency). Table 3 illustrates the directionality of assignments. Summing across all actions, this method yields two distributions of agency (positive/negative) and two for power (positive/negative) for each character.

Verb Type	Character Role	Assigned power
High power/agency verb	Agent/Subject	+ power/Agency
	Theme/Object	– power/Agency
Low power/agency verb	Agent/Subject	– power/Agency
	Theme/Object	+ power/Agency
Neutral power/agency verb	Agent/Subject or Theme/Object	Neutral

Table 3: Assignment of agency and power values based on verb type and character role.

conference version

D. Independence

2.3.6 Parent Rate

The proportion of characters who are explicitly identified as the child of another character in the same book. This binary outcome reflects whether a given character is embedded within a parent–child relationship in the narrative world. It serves as an indicator of social dependence and is used to analyze how such narrative framing varies across time, gender, and age.

2.4 Regression Modeling Framework

To analyze temporal and demographic variation in dependent variables, I employ a suite of regression models tailored to the structure and scale of each outcome. All models include the predictors `decade` (centered and scaled in 10-year units), `predicted_gender` (male, female), and `age_category` (adult, adolescent, child), along with their interactions. When appropriate, I include a random intercept for book (`bookID`) to account for repeated measures within texts.

Binary Logistic Regression. For the binary outcome of `Parent_Rate` I use a mixed-effects logistic regression model:

$$\text{logit}(\Pr(Y_i = 1)) = \beta_0 + \beta_1 \cdot \text{Decade}_i + \beta_2 \cdot \text{Gender}_i + \beta_3 \cdot \text{Age}_i + \ldots + u_{\text{BookID}[i]}$$

where Y_i is a binary indicator (e.g., `is_child_of`), fixed effects model historical and demographic predictors (with age classifications corrected through sensitivity analysis), and $u_{\text{BookID}[i]}$ represents a random intercept for each book. The ellipsis (...) denotes interaction terms among predictors.

Linear Mixed-Effects Models. For continuous outcomes at the character level (such as emotional valence, agency, and subjecthood), I use linear mixed-effects models:

$$Y_i = \beta_0 + \beta_1 \cdot \text{Decade}_i + \beta_2 \cdot \text{Gender}_i + \beta_3 \cdot \text{Age}_i + \ldots + u_{\text{BookID}[i]} + \epsilon_i$$

where Y_i is a standardized score (e.g., valence, agency), fixed effects model historical and demographic predictors (with age classifications corrected through sensitivity analysis), the ellipsis (...) denotes interaction terms among predictors, $u_{\text{BookID}[i]}$ is a book-level random intercept, and ϵ_i is residual error.

To address measurement error in age classification, I implement a sensitivity analysis framework that corrects for imperfect classification accuracy. Using error rates estimated from the validation data above—sensitivity (true positive rate) and specificity (true negative rate) for each age category—I draw corrected age labels across 100 iterations for each character. This approach propagates uncertainty from the classification process through to the final coefficient estimates, producing adjusted standard errors and significance tests that account for measurement error. All reported coefficients represent the mean estimates across iterations, with corrected standard errors reflecting both sampling variability and classification uncertainty. All coefficient tables are included in Appendix A and model diagnostics are reported in Appendix B.

3. Results

3.1 Base Rates

Overall, there are an estimated 167,220 characters across the corpus with approximately ten-percent labeled as either children or adolescents (Table 4). There are also meaningful differences in the fraction of child characters by collection (9.9% Chicago v. 13.3% Stanford) and child mentions (13.4 v. 21.7 respectively). In terms of gender, there is a highly skewed base rate of male (65.7%) and female (34.3%) characters as well as a within-gender rate of children (7.9% of male characters are boys v. 14.3% of female characters are girls). There are also differential levels of parental representation by age, with 60% of children having a parent in the top twenty characters of a book to 40% for adolescents and only 6% for adults. Finally, there is a stark contrast in the age distribution of children with another 60-40 split favoring adolescents versus young children, suggesting an overall narrative bias towards older and thus more agential child characters.

3.2 H1: Visibility (Figure 1, Row 1)

The Declining Child. Across the two centuries examined, overall character mentions remained largely stable. Yet contrary to expectations, both child and adolescent characters experienced gradual declines in their share of mentions relative to adults. Adolescents exhibit the sharper decrease, losing roughly 0.23–0.25 percentage points per decade, amounting to a roughly five-point decline over the 200-year span. Children experience a more modest but still consistent reduction of around 0.17 percentage points per decade, totaling just over three percentage points across the same period.

conference version

Measure	Category	Percentage
Child Mention Rate	All Characters	13.9
	Chicago	13.4
	Stanford	21.7
Overall Character Gender Rate	Male	65.7
	Female	34.3
Child Rate within Gender	Male	7.9
	Female	14.3
Age within Children	Child	36.8
	Adolescent	63.2
Parent Rate	Child	60.5
	Adolescent	41.0
	Adult	6.5

Table 4: Reference table showing overall and group-specific rates of children by gender, age, and collection. Chicago refers to 20C books and Stanford to 19C books.

Women Replacing Girls. Gender compounded these age patterns in sharply asymmetric ways. While adult women continued to gain representation across the two centuries, younger female characters experienced marked long-term losses. Female adolescents, who initially held a visibility advantage over their male counterparts, saw this advantage erode rapidly: their relative share declined by more than two percentage points per decade, amounting to a dramatic forty-plus–point drop across the full historical span. Female children experienced a milder version of this pattern, but the direction was the same. Their mentions rose less quickly than those of adults, and once differences in age and gender composition are accounted for, their relative share fell by roughly one percentage point per decade—amounting to a decline of about a quarter of their initial advantage over the two centuries examined.

Taken together, these patterns show that the historical rise in women's narrative prominence was overwhelmingly an adult-centered phenomenon. The gains accrued primarily to adult women, whose share of character mentions expanded steadily over the two centuries. By contrast, younger characters moved in the opposite direction: both children and adolescents became progressively less central to fictional worlds, and this retreat was most pronounced among adolescent girls, who lost visibility at substantially faster rates than any other group. Rather than a uniform broadening of representation, the long-run trajectory reflects a redistribution of narrative attention toward adulthood—especially adult women—at the expense of youth.

3.3 H2: Value (Figure 1, Row 4)

Adults and adolescents become more negatively coded, while children stay the same. The emotional valence associated with fictional characters has become progressively less positive over the past two centuries. Across all characters, valence scores show a clear and steady downward trend, indicating that the affective tone of fiction has grown increasingly negative over time. Female characters consistently exhibit slightly more positive valence than males across all life stages, and because their rate of decline is slower, the gender gap widens modestly across the period.

Age moderates these declines in notable ways. Adolescents begin at slightly lower

conference version

conference version

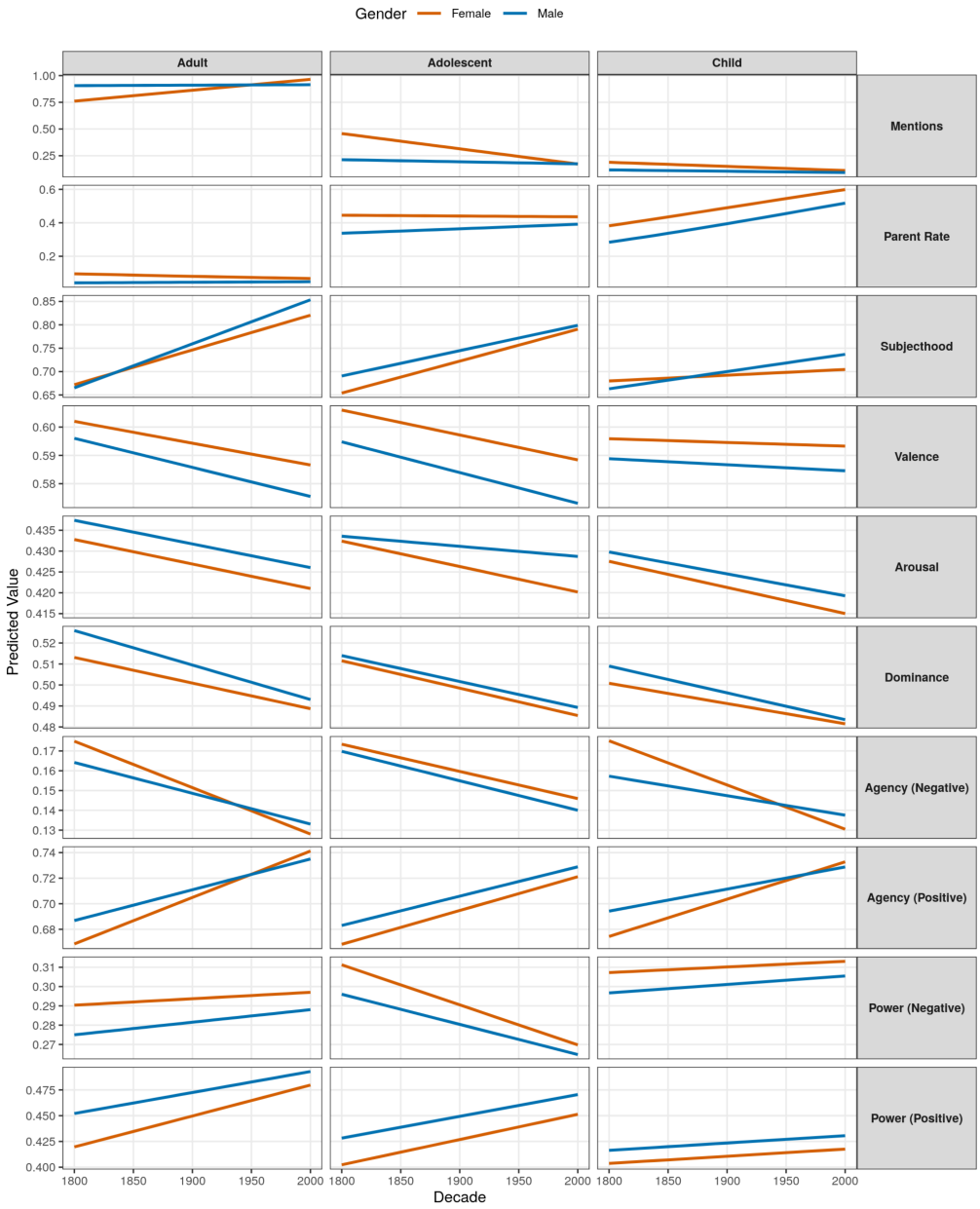


Figure 1: Predicted values by gender and age group across narrative measures. Each panel shows model-estimated trends over time for a specific narrative feature, faceted by character age group (Adult, Adolescent, Child). Gender lines represent male (blue) and female (orange) predicted values.

valence levels than adults and experience a similarly gradual decrease over time, with boys and girls following nearly parallel trajectories. In contrast, young children deviate from the general pattern: rather than declining, children show a small but significant increase in positive valence over the two centuries examined. This upward shift holds for both boys and girls and suggests that the emotional framing of childhood has grown more positive even as the broader narrative climate of fiction has darkened.

H3: Agency (Figure 1, Rows 3,5-10)

Overall Trends

Grammatical Agency Rises, Emotional Dominance Declines. Across all characters, two broad and somewhat opposing historical trends emerge. First, grammatical and semantic agency increases: characters are more often subjects, positive agency frames increase (on the order of +0.002 per decade), and negative agency frames decline (about -0.0015 per decade). On the other hand, indicators of affective control weaken: dominance shows a steady decrease (roughly -0.0017 per decade), and arousal also edges downward (around -0.0006 per decade). Power semantics show mixed results: positive power increases modestly over time (about +0.002 per decade), while negative power also ticks up slightly overall (approximately +0.0007 per decade), a rise driven largely by adults, as adolescents in particular become less likely to be framed in negative-power terms. In short, narratives grow more agentive in form (who acts and how they are framed) even as the emotional force and felt control attached to those actions temper over the two centuries.

Age-Related Trends

Children are less agential than adolescents who are less agential than adults. Age remains the strongest determinant of narrative agency and aligns with general social expectations. Children and adolescents are consistently less agentive than adults across core dimensions: they appear less often as grammatical subjects, exhibit lower dominance and positive power, and are more likely to be framed in constrained or dependent roles. Patterns of negative portrayal are more differentiated. Adolescents are especially prone to negative-agency framings, while children are more often linked to negative-power semantics than adults, even as both groups remain below adults in overall positive agency and control.

Over time, differences in agency between children and adolescents shift but do not disappear. Adolescents typically occupy an intermediate position between adults and children—making gains in subjecthood and positive power that bring them somewhat closer to adults, yet continuing to lag behind on most measures of control and initiative. Children remain distinctly less agentive: their growth in activity and power is slower, and the decline in their negative portrayals is more muted than for adults. In short, as expected adolescence marks a partial entry into adult forms of agency, whereas childhood continues to occupy a comparatively passive and protected position within the narrative hierarchy.

Gender-Related Trends

Female adolescents exhibit the lowest levels of agency and gain the least over time. Gender remains a consistent axis of differentiation across agency measures. Female characters begin the period with lower subjecthood, dominance, positive power, and positive agency than males, but they also experience faster gains on several of these dimensions. As a result, the gender gaps in both positive and negative agency narrow steadily across the two centuries, even if they do not fully disappear.

conference version

The strongest point of constraint is the adolescent girl whose gendered disadvantage is greatest relative to male age-peers. Female adolescents show the largest gender gaps in grammatical, semantic, and affective agency: compared with adolescent boys, they are notably less likely to appear as subjects, to display dominance, or to receive positive-agentive framing. These gaps are larger than those observed among children, even though adolescent girls may have higher absolute agency than younger characters. Moreover, the negative-agency disparity between female and male adolescents narrows only modestly over time, in contrast to the more substantial convergence seen among other groups.

Among children, gender differences are more mixed. Girls show faster improvement in positive-agentive framing and a slightly slower decline in dominance, suggesting a modest advantage in the affective quality of agency. At the same time, they lag behind boys in gains in subjecthood, becoming active grammatical agents more slowly even as their portrayals grow more positive. Differences in arousal and power for children remain small and comparatively stable. Overall, childhood offers emerging strengths for girls in how their actions are valued, paired with continued lags in who is permitted to act.

Summary

Taken together, the results reveal a double asymmetry in the literary construction of agency. First, a vertical asymmetry: children and adolescents occupy structurally less agentive positions than adults across most dimensions, with adolescents partially approaching adult patterns over time while children remain the furthest removed in subjecthood, dominance, and power. Second, a horizontal asymmetry: female characters continue to display lower grammatical and semantic agency than males, though they also experience faster long-run gains on several measures.

Importantly, these gendered disadvantages vary strongly by age. The largest gender gaps appear in adolescence, where female characters fall further behind their male peers than do female children relative to boys, even when adolescent girls display higher absolute levels of agency than younger characters. By contrast, female children and adult women show the clearest improvements: across the century, both groups gain in positive-agentive framing and reduce their exposure to negative-agency contexts. Overall, the long-term trajectory points toward partial convergence in some forms of agency while maintaining durable age- and gender-based stratifications in who is permitted to act, with what power, and under which narrative conditions.

3.4 H4: The Independent Child (Figure 1, Row 2)

Children are increasingly accompanied by visible parents. Contrary to the initial hypothesis, parental presence has increased markedly over time. This rise is strongest among children, who show a substantial upward shift in the likelihood that a parent is mentioned in their narrative vicinity. Girls begin with higher levels of parental presence than boys, but both genders display similarly steep increases across the two centuries.

Adolescent characters remain noticeably less parent-embedded than children at baseline, and their temporal trajectories grow only modestly. Both male and female adolescents

conference version

exhibit small but positive increases in parental presence, far short of the growth observed for younger characters. Taken together, these patterns point to a broad historical movement toward more family-embedded portrayals of young children, while adolescence remains relatively independent with only limited shifts toward increased parental framing.

3.5 H5: The Finite Child (Figure 1, Row 2, Left)

The gendered results of adults as children. The finite child is designed to ask whether adults, rather than children, are less likely to be depicted narratively with parents, i.e. whether adults are decreasingly represented as children. The results reveal a strongly gendered pattern in the long-term decline of parental presence among adults. Female adult characters begin the period far more likely than males to be associated with a parent, reflecting enduring conventions that link women more tightly to familial roles even in adulthood. Over time, however, this difference contracts sharply. Female adults show a pronounced decline in parental associations—falling at roughly three to four times the rate that male adults change—while male adults exhibit a small positive increase. Taken together, these trends suggest that the finitude of childhood is expressed most clearly through the narrative trajectory of female adults: the strong parental embeddedness that once characterized their roles diminishes steadily across the two centuries, bringing adult women progressively closer to the relative independence long associated with adult men.

4. Discussion

The results displayed here suggest a complex picture of children's narrative agency, one modulated by age, gender, and specific features.

4.1 The declining prevalence of children is led by adolescent girls

Across the two centuries examined, fictional youth become steadily less central to the narrative landscape. Both children and adolescents show long-term declines in their share of character mentions relative to adults, with adolescents experiencing the sharper contraction—nearly a five-point decrease overall—while children decline by a more modest but still notable three points. The sharpest losses occur among adolescent girls, who began the nineteenth century as one of the most visible youth groups but whose relative prominence erodes rapidly as adult women rise. The decline of the fictional child is thus best understood not as a generalized cultural shift away from youth, but as a redistribution of representational space toward adult women at the expense of adolescent girls.

On the one hand we can understand this as a shift in cultural values away from the angelic young heroine on the verge of marriage celebrated in mid-nineteenth century novels, such as Amelia Sedley in *Vanity Fair* or Amy Edmynsone in *The Heir of Redclyffe*. But we might also hypothesize that this shift is associated with the rise of a new genre, the young adult novel, in which adolescent girls play a central role. Which direction the causal arrow points is the subject of future work, but as women replace young women at the center of the modern novel the lives of young women become ensconced at the

conference version

heart of this burgeoning new form beginning in the 1960s and 70s.

4.2 The Angelic Young Woman Becomes the Good Child

Although overall sentiment becomes increasingly negative across the period, the magnitude and direction of this trend differ sharply by age. Adolescents—regardless of gender—show declines similar to adults, whereas children's emotional valence remains stable. This produces a notable reordering of evaluative positions. Adolescent girls, who begin the nineteenth century with the highest valence levels of any group (i.e. "angelic"), lose this advantage as their sentiment gradually converges toward adult female levels, though it remains notably higher than male adolescents. By contrast, young children—both boys and girls—maintain the most positive emotional positioning throughout the period. Their insulation from the long-term increased negativity that characterizes the modern novel suggests an intentional cultural attachment to the sentimental child, what might be termed a "constancy of goodness."

4.3 The Agency Lag

Although children show increases in grammatical and semantic agency over time—appearing more often as subjects, acquiring slightly more positive power, and experiencing modest reductions in negative-agentive framings—these shifts consistently lag behind those observed for adults. Across most dimensions, adults show the steepest gains, adolescents occupy an intermediate position, and children change the least. The familiar age hierarchy therefore remains intact: adults maintain the highest levels of dominance and subjecthood, adolescents follow, and children remain the least agentive group.

These upward movements in child agency thus appear to reflect broader stylistic shifts in narrative representation rather than a targeted historical investment in empowering child characters. Gender differences are notable. Girls gain positive and affective forms of agency somewhat faster than boys, especially in semantic framing, yet they continue to lag in grammatical subjecthood, where boys make more rapid gains. Overall, children do become more agentive over the two centuries—but at a slower pace than their gender-equivalent adult counterparts, and without substantial disruption to the longstanding age-based stratification of narrative power.

4.4 Children are increasingly situated within family ties

Contrary to theories of the increasingly autonomous child, our analysis shows that children have become progressively more embedded within kinship structures. Across the two centuries examined, narratives featuring young children now incorporate parents at significantly higher rates than they did in the nineteenth century, a shift observable for both boys and girls, though girls still lead in this category. Adolescents, by contrast, display only modest increases in parental presence, remaining substantially less parent-embedded than younger characters.

This narrative trend may be an indicator of a larger historical transition toward a more developmentally oriented depiction of childhood: where the Victorian child was defined largely by sentimental goodness and adult-threshholds, modern fiction places greater emphasis on the young child situated within family systems. Children remain morally

conference version

good, but they are increasingly framed through the relational context of parenting, consistent with the broader cultural rise of the "developmental child." Conversely, we might see this as a potential decentralization of children as they become narrative backgrounds to "parental" adults, there to shore-up the ideology of the nuclear family.

4.5 The decline of the three-generation women's novel

Representations of adults with living or narratively present parents show a markedly gendered trajectory. Female adult characters—who begin the nineteenth century far more closely tied to parents than male adults—experience a sharp decline in parental associations over time. This contraction diminishes the presence of three-generation narrative structures that were once common for women, drawing adult female characters into a more individuated representational space. Male adults, by contrast, show a slight increase in parental presence, producing a gradual narrowing of the gender gap rather than a uniform retreat from parent-linked adulthood. Women's parents thus become another casualty alongside adolescent girls tied with the rising narrative centrality of adult women characters.

4.6 "Women are to men as children are to adults"

Across nearly every domain of characterization, gender continues to structure fictional agency in ways that echo long-standing age hierarchies. Female characters exhibit less grammatical and semantic agency than males, while also carrying a greater share of relational and affective framing. They are evaluated more positively and are more closely tied to family relations, especially early in the period. In these respects, women's narrative positioning mirrors that of children relative to adults: less powerful, more relationally embedded, and framed through a more positive moral and emotional lens.

At the same time, historical trends also nuance this portrait. Women's agentive roles improve more rapidly than men's across the two centuries, narrowing gaps in positive and negative agency even as they continue to lag in dominance and subjecthood. Their overall narrative prominence also grows steadily, driven primarily by gains among adult women rather than youth. By the end of the twentieth century, women have reached parity with men in several narrative domains, even as key asymmetries persist.

Taken together, the resemblance between gender and age hierarchies remains striking but not static. Women and children share a historically relational, morally elevated, and structurally constrained narrative position, yet the trajectories diverge: women move toward greater independence and agency, while children retain their distinctive combination of moral positivity and limited power. The modern novel thus preserves a familiar structural analogy—women to men as children to adults—even as it moderates the terms of that relationship over time.

5. Conclusion: Limitations and Future Work

A host of questions are raised by these findings. The first and most important is that of **genre**: how would children in books expressly targeting youth-oriented audiences behave compared to this broader random sample? Is the failure of children to assume

conference version

increased levels of narrative agency or the declining prevalence of agential adolescents a potential explanation for the rising popularity of young adult fiction? Have children become any less agential in "youth" books over time? Future work will want to study specific sub-samples of writing targeting different audiences to complement these broader historical observations.

Another limitation is **measurement error**. While our sensitivity analysis suggests our results are robust against prediction errors by our LLMs, there is still a great deal of work that can be done to capture the measurement of agency. While this paper has proposed several different measures that capture a range of dimensions–prevalence, independence, semantic framing, valence, and emotional control–one could also imagine more narratively attuned measures of causal change: for example, what large-scale plot points or viewpoints change over narrative time due to the actions of child characters? While it would be surprising if such a measure moved directionally differently from semantic and grammatical framing, it would give us further insights into where and how agency is or is not being assumed on the part of children.

Finally, in addition to expanding the **cultural breadth** of the analysis, future work will want to condition on **narrative outliers**, i.e. cases where children play a stronger than expected role in the plot. While there is value in capturing broad behavioral measures across fictional storytelling, many questions remain surrounding narrative outliers. When children are as present as adults, do we see them assuming agential traits similar to adults or do they remain distinctive, continuing to be narrative anchors conveying positivity, goodness, and dependence? As with all averaging effects, we lose important granularity regarding a smaller subset of narratives built around child characters or extreme child characters. Future work will want to focus in on narratives where we see children playing a prominent role to better understand what these "extraordinary" children may teach us.

6. Data Availability

Data and code can be found here: https://figshare.com/s/ddc0bef13d69d6148b65.

7. Acknowledgements

This research was generously supported by the Social Sciences and Humanities Research Council of Canada (435-2022-0089).

8. Author Contributions

Andrew Piper: Conceptualization, Formal Analysis, Writing – original draft, Writing – review & editing

conference version

References

Algee-Hewitt, Mark (2024). "The Canon". In: *The Cambridge Companion to Literature in a Digital Age*. Ed. by Adam Hammond. Cambridge Companions to Literature. Cambridge University Press, 47–65.

Ariès, Philippe (1962). *Centuries of Childhood: A Social History of Family Life*. Trans. by Robert Baldick. Translated from the French *L'Enfant et la vie familiale sous l'ancien régime*. New York: Alfred A. Knopf.

Bamman, David (2025). *BookNLP: A natural language processing pipeline for books*. https://github.com/booknlp/booknlp. Accessed: 2025-04-01.

Bernstein, Robin (2011). *Racial innocence: Performing American childhood from slavery to civil rights*. New York University Press.

Bion, Wilfred R (1985). "Container and contained". In: *Group relations reader* 2.8, 127–133.

Bradley, Margaret M and Peter J Lang (1994). "Measuring emotion: the self-assessment manikin and the semantic differential". In: *Journal of behavior therapy and experimental psychiatry* 25.1, 49–59.

Deci, Edward L and Richard M Ryan (2012). "Self-determination theory". In: *Handbook of theories of social psychology* 1.20, 416–436.

Douglas, Hazel (2007). *Containment and reciprocity: Integrating psychoanalytic theory and child development research for work with children*. Routledge.

Greimas, Algirdas Julien (1984). *Structural Semantics: An Attempt at a Method*. Lincoln: University of Nebraska Press.

Greimas, Algirdas Julien and Catherine Porter (1977). "Elements of a narrative grammar". In: *Diacritics* 7.1, 23–40.

Gubar, Marah (2005). "The Victorian Child, c. 1837-1901". In: *Representing Childhood*.

James, Allison, Chris Jenks, and Alan Prout (1998). *Theorizing childhood*. ERIC.

Jenkins, Henry (1998). *The children's culture reader*. NYU Press.

Jenks, Chris (2020). *Childhood*. Routledge.

Kincaid, James R (1992). *Child-loving: The erotic child and Victorian culture*. Routledge.

Langacker, Ronald W (1991). *Foundations of cognitive grammar: Volume II: Descriptive application*. Stanford university press.

Little, Todd D, Patricia H Hawley, Christopher C Heinrich, and Katherine W Marsland (2002). "Three views of the agentic self: A developmental synthesis." In.

Little, Todd D, CR Snyder, and Michael Wehmeyer (2006). "The agentic self: On the nature and origins of personal agency across the lifespan". In: *Handbook of personality development*, 61–79.

Mohammad, Saif M. (2018). "Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words". In: *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*. Melbourne, Australia.

Nelson, Claudia, Elisabeth Wesseling, and Andrea Mei-Ying Wu (2024). *The Routledge Companion to Children's Literature and Culture*. Routledge.

Olson, Debbie (2013). "The Hitchcock Imp". In: *Lost and Othered Children in Contemporary Cinema*. Ed. by Debbie Olson and Andrew Scahill. Bloomsbury Publishing USA, 295–314.

Olson, Debbie and Giselle Rampaul (2013). "Representations of childhood in the media". In: *The Routledge International Handbook of children, adolescents and media*. Routledge, 49–56.

conference version

Pifer, Ellen (2000). *Demon or Doll: Images of the Child in Contemporary Writing and Culture*. Charlottesville: University of Virginia Press, 272.

Renner, Karen J (2016). *Evil children in the popular imagination*. Springer.

Ryan, Richard M and Edward L Deci (2000). "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." In: *American psychologist* 55.1, 68.

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*, 2329–2334.

Scahill, Andrew (2015). *The revolting child in horror cinema: Youth rebellion and queer spectatorship*. Springer.

Schneider, William Joel, Timothy A Cavell, and Jan N Hughes (2003). "A sense of containment: Potential moderator of the relation between parenting practices and children's externalizing behaviors". In: *Development and Psychopathology* 15.1, 95–117.

Shogren, Karrie A, Todd D Little, and Michael L Wehmeyer (2017). "Human agentic theories and the development of self-determination". In: *Development of self-determination through the life-course*, 17–26.

Steedman, Carolyn (1995). *Strange dislocations: Childhood and the idea of human interiority, 1780-1930*. Harvard University Press.

Textual Optics Lab (2023). *US Novel Corpus*. https://textual-optics-lab.uchicago.edu/us_novel_corpus.

Thiel, Liz (2012). "Degenerate 'Innocents': Childhood, Deviance, and Criminality in Nineteenth-Century Texts". In: *The Child in British Literature*. Ed. by Adrienne E. Gavin. London: Palgrave Macmillan, ---.

Wood, Naomi (2012). "Angelic, Atavistic, Human: The Child of the Victorian Period". In: *The Child in British Literature*. Ed. by Adrienne E. Gavin. London: Palgrave Macmillan, ---.

Zelizer, Viviana A Rotman et al. (1985). *Pricing the priceless child: The changing social value of children*. Vol. 24. Basic Books New York.

conference version

A. Supplementary Regression Tables

Table A1: H4-H5 Parental Presence. This table reports adjusted logistic regression coefficients predicting whether a character is associated with a parent figure among the top 20 characters in the narrative. Estimates are corrected for measurement error through sensitivity analysis. Standard errors are shown in parentheses. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Children/Adolescents	Adults
Intercept	0.1570**	−2.9987***
	(0.0582)	(0.0046)
Decade	0.0581***	0.0094***
	(0.0142)	(0.0009)
Gender:Female	0.2253**	0.6182***
	(0.0772)	(0.0051)
Age:Adolescent	−0.7119***	
	(0.0588)	
Decade × Gender:Female	−0.0048	−0.0263***
	(0.0184)	(0.0011)
Decade × Age:Adolescent	−0.0463**	
	(0.0143)	
Gender:Female × Age:Adolescent	0.0979	
	(0.0776)	
Decade × Gender:Female × Age:Adolescent	−0.0084	
	(0.0184)	

conference version

Table A2: Linear Mixed-Effects Models. This table reports adjusted fixed effects estimates from linear mixed-effects models predicting sentence-level character variables, corrected for measurement error through sensitivity analysis. Random intercepts for BookID account for repeated measures. Standard errors are shown in parentheses. Significance levels: *$p < 0.05$, **$p < 0.01$, *** $p < 0.001$.

	Mentions	Subjecthood	Valence	Arousal	Dominance	Agency (Negated)	Agency (Positive)	power (Negative)	power (Positive)
Intercept	0.9103***	0.7594***	0.5858***	0.4317***	0.5095***	0.1486***	0.7109***	0.2815***	0.4725***
	(0.0006)	(0.0003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
Decade	0.0004**	0.0094***	−0.0010***	−0.0006***	−0.0016***	−0.0016***	0.0024***	0.0007***	0.0020***
	(0.0002)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Gender:Female	−0.0476***	−0.0131***	0.0085***	−0.0048***	−0.0086***	0.0029***	−0.0060***	0.0122***	−0.0228***
	(0.0010)	(0.0005)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0003)	(0.0003)	(0.0003)
Age:Adolescent	−0.7168***	−0.0146***	−0.0018***	−0.0006***	−0.0079***	0.0063***	−0.0050***	−0.0011***	−0.0231***
	(0.0006)	(0.0003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
Age:Child	−0.8055***	−0.0593***	0.0009	−0.0072***	−0.0133***	−0.0012	0.0006	0.0196***	−0.0490***
	(0.0030)	(0.0035)	(0.0009)	(0.0006)	(0.0008)	(0.0018)	(0.0025)	(0.0025)	(0.0027)
Decade × Gender:Female	0.0098***	−0.0020***	0.0003***	0.0000	0.0004***	−0.0008***	0.0012***	−0.0003***	0.0010***
	(0.0002)	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0001)
Decade × Age:Adolescent	−0.0024***	−0.0040***	−0.0001***	0.0003***	0.0004***	0.0001***	−0.0001***	−0.0022***	0.0001**
	(0.0002)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Decade × Age:Child	−0.0017*	−0.0057***	0.0008***	0.0000	0.0004*	0.0006	−0.0007	−0.0002	−0.0013*
	(0.0007)	(0.0009)	(0.0002)	(0.0001)	(0.0001)	(0.0004)	(0.0005)	(0.0005)	(0.0006)
Gender × Age:Adolescent	0.1688***	−0.0093***	0.0047***	0.0000	0.0054***	0.0019***	−0.0052***	−0.0021***	0.0003
	(0.0010)	(0.0005)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0003)	(0.0003)	(0.0003)
Gender × Age:Child	0.0929***	0.0053	−0.0007	0.0016	0.0035***	0.0025	−0.0019	−0.0031	0.0099**
	(0.0052)	(0.0053)	(0.0012)	(0.0009)	(0.0009)	(0.0026)	(0.0032)	(0.0037)	(0.0036)
Decade × Gender:Female	−0.0221***	0.0034***	−0.0001***	−0.0003***	−0.0005***	0.0009***	−0.0009***	−0.0002***	−0.0006***
× Age:Adolescent	(0.0002)	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0001)
Decade × Gender:Female	−0.0124***	−0.0005	−0.0002	−0.0001	−0.0001	−0.0005	0.0000	0.0002	−0.0010
× Age:Child	(0.0013)	(0.0012)	(0.0002)	(0.0002)	(0.0002)	(0.0006)	(0.0008)	(0.0007)	(0.0008)

B. Model Diagnostics

H1-H3. Model assumptions for linear mixed-effects models were assessed through examination of standardized residuals and formal diagnostic tests. Diagnostics were conducted on a single iteration of the sensitivity-adjusted data (iteration 1) to evaluate model fit with error-corrected age classifications. Dependent variables comprise three distinct data types: continuous psychological ratings (`valence, arousal, dominance`), character-level linguistic proportion variables derived from feature counts (`subj, agency_pos, agency_neg, power_pos, power_neg`), and book-level proportion data (`mention_rate`).

Shapiro-Wilk tests of residual normality revealed differential patterns by variable type. The continuous psychological variables demonstrated excellent adherence to normality assumptions (valence: $W = 0.9968$, arousal: $W = 0.9907$, dominance: $W = 0.9964$). While all tests reached statistical significance due to large sample size, the W statistics approaching unity indicate practically normal distributions suitable for linear modeling. The book-level mention rate variable also showed strong adherence to normality (mentions: $W = 0.9792$), indicating that when aggregated to the book level as proportions of total character mentions, this measure meets normality assumptions well.

The character-level linguistic proportion variables showed greater deviations from normality (subj: $W = 0.9624$, agency_pos: $W = 0.8491$, agency_neg: $W = 0.7825$, power_pos: $W = 0.8816$, power_neg: $W = 0.8736$). Visual inspection of distributions revealed expected patterns for bounded proportion data: `agency_pos` and `subj` exhibited ceiling effects with most characters showing high proportions of positive agency and subject positions, while `agency_neg` and `power_neg` demonstrated floor effects consistent with characters predominantly exhibiting positive traits. These distributional characteristics reflect theoretically meaningful patterns in narrative character representation rather than model inadequacy.

Histograms (Figure B1) display the raw distributions of dependent variables prior to modeling. The mention rate variable exhibits a pronounced U-shaped distribution, with substantial frequencies at both extremes (near 0 and 1), reflecting that age groups either dominate narrative attention or receive minimal mention within books. The continuous psychological ratings (valence, arousal, dominance) demonstrate approximately normal distributions centered around moderate values, consistent with standardized psychological scales. Character-level proportion variables show theoretically expected patterns: `subj` and `agency_pos` display strong ceiling effects with most values concentrated near 1, while `agency_neg` exhibits a pronounced floor effect with the majority of characters showing minimal negative agency. The power variables demonstrate more dispersed distributions with moderate central tendencies.

Residual plots (Figure B2) assess model fit assumptions through standardized residuals plotted against fitted values. The continuous psychological variables (valence, arousal, dominance) exhibit excellent model fit characteristics: residuals are evenly distributed around zero, variance remains approximately constant across fitted values (homoscedasticity), and LOESS smoothers remain flat and centered near zero, indicating no systematic patterns in residual structure. The mention rate residuals display banding patterns at extreme fitted values (near 0.1 and 0.9), reflecting the U-shaped raw distribution, yet the LOESS smoother remains relatively flat near zero, suggesting

conference version

adequate model specification despite the non-normal raw distribution. This confirms that the model's predictors successfully account for the systematic variation in mention rates across demographic groups.

Character-level proportion variables exhibit characteristic banding patterns in residual plots, reflecting the discrete nature of underlying count data with limited observations per character. These horizontal bands correspond to the finite set of possible proportion values when numerators and denominators are small integers. Additionally, these variables show heteroscedasticity (non-constant variance) and slight trends in LOESS smoothers, indicating systematic residual patterns. However, these deviations from ideal linear model assumptions are well-documented in the statistical literature on proportion data analysis and do not invalidate inference when the research focus centers on effect directions and interaction patterns rather than precise point predictions.

The residual patterns observed are theoretically meaningful rather than indicative of model misspecification: the ceiling effects in positive agency and subjecthood reflect genuine narrative tendencies toward protagonistic character representation, while floor effects in negative traits align with documented positivity biases in fiction. The robust significance of effects across 100 sensitivity analysis iterations, combined with substantively interpretable interaction patterns, provide confidence in the reliability of findings despite these distributional characteristics.

H4,H5. Logistic mixed-effects models assessing parental embeddedness (H4-H5) were evaluated using DHARMa simulation-based diagnostics applied to the sensitivity-adjusted data (iteration 1). These models examine the likelihood that a character is situated in an explicit parent–child relationship. Diagnostics were conducted separately for children and adolescents (H4) and for adults (H5), following the model specifications used in the primary analyses.

DHARMa residual diagnostics indicate strong overall model fit. In both models, the simulated quantile residuals closely followed the expected uniform distribution (see Q-Q plots in Figure B3), with Kolmogorov–Smirnov tests showing no significant deviation from uniformity (H4: $p = 0.068$; H5: $p = 0.957$). These results indicate that the logistic link and predictor structure adequately capture the data-generating process.

Dispersion tests likewise showed no evidence of overdispersion or underdispersion in either model. For H4, the dispersion statistic was near unity (dispersion = 0.994, $p = 0.120$); for H5, dispersion was similarly well behaved (dispersion = 1.008, $p = 0.450$). Outlier tests detected no excess of extreme residuals beyond what would be expected under the fitted model (H4: $p = 0.203$; H5: $p = 0.909$). Taken together, these results confirm that both logistic mixed-effects models satisfy core assumptions and that the estimated effects for parental embeddedness in H4 and H5 are not compromised by violations of distributional assumptions or unmodeled structure in the data.

conference version

conference version

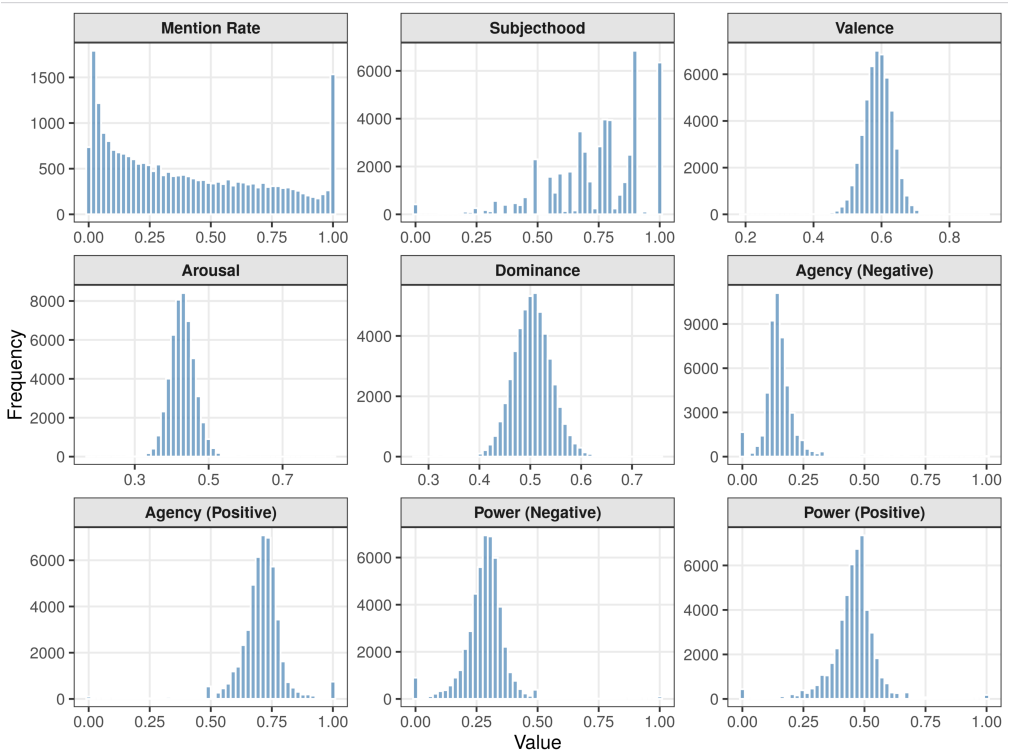


Figure B1: Distributions of dependent variables from sensitivity-adjusted data (iteration 1). Histograms display frequency distributions for count data (total mentions), continuous psychological ratings (valence, arousal, dominance), and linguistic proportion variables (subj, agency_pos, agency_neg, power_pos, power_neg) using error-corrected age classifications. Proportion variables show expected floor and ceiling effects consistent with characters predominantly exhibiting positive traits in narrative texts.

conference version

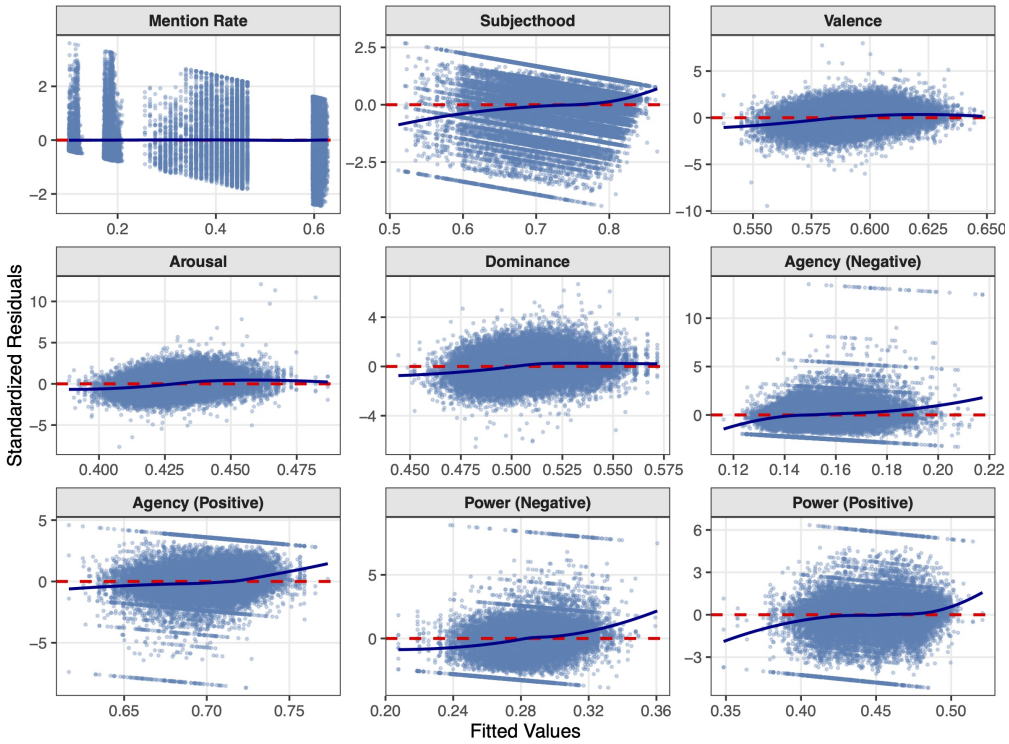


Figure B2: Residual diagnostics for linear mixed-effects models. Standardized residuals plotted against fitted values for nine linear models estimated using the sensitivity-adjusted data (iteration 1). Blue LOESS smoothers illustrate residual structure, and red dashed lines denote the zero baseline. The mention-rate model (top left), which is based on book-level proportions, shows pronounced banding at extreme values, reflecting the U-shaped distribution of raw mention shares across age–gender groups. Character-level proportion variables (`subj`, `agency_pos`, `agency_neg`, `power_pos`, `power_neg`) exhibit horizontal banding due to the discrete nature of the underlying count data and the small integer denominators that generate limited possible proportion values. In contrast, the continuous psychological measures (`valence`, `arousal`, `dominance`) show approximately homoscedastic scatter with LOESS curves remaining close to zero across the fitted-value range, indicating good model fit and no major violations of linearity or variance assumptions.

conference version

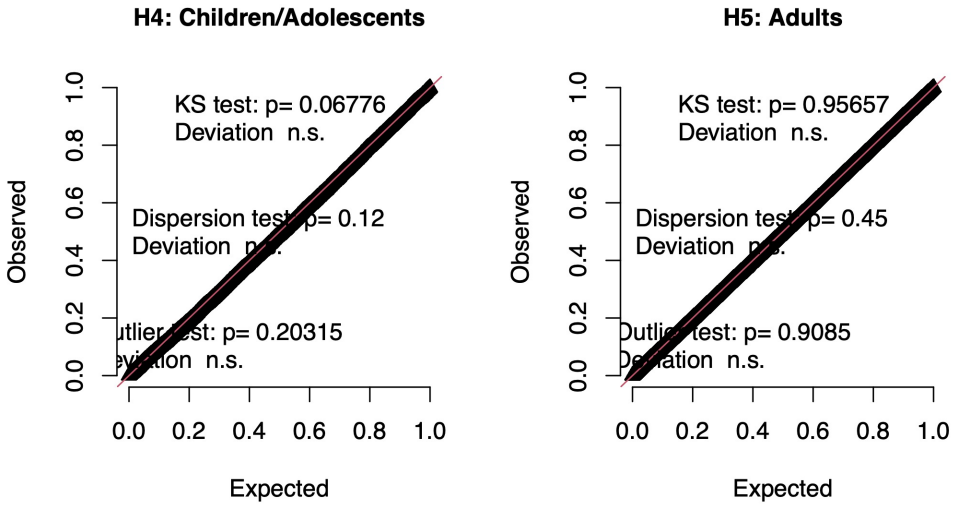


Figure B3: DHARMa Q-Q plots for logistic regression models. Quantile–quantile plots compare observed DHARMa residuals (y-axis) against the expected uniform distribution (x-axis) for models predicting parent presence. The left panel shows the H4 model for children and adolescents; the right panel shows the H5 model for adults. Points closely following the diagonal indicate good model fit. Diagnostic test results are displayed within each panel: KS (Kolmogorov–Smirnov) tests assess overall distributional conformity, dispersion tests evaluate variance structure, and outlier tests detect extreme values. All diagnostics indicate no significant deviations from expected behavior (p-values > 0.05), and visual inspection confirms that residuals closely follow the uniform reference line.

jcls. Journal of Computational Literary Studies

conference version

Where Empires End

Geography of the Poetic Formula "from A to B"

Antonina Martynenko[1]
Artjoms Šeļa[1]
Petr Plecháč[1]

1. Institute of Czech Literature, Czech Academy of Sciences, Prague, Czech Republic.

OPEN ACCESS

Citation
Antonina Martynenko, Artjoms Šeļa, and Petr Plecháč (2026). "Where Empires End. Geography of the poetic formula "from A to B"". in: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7986

Date published 2026-05-05 (preprint)
Date accepted tbc
Date received 2026-01-09

Keywords
poetry, geospatial analysis, literary geography, axis of orientation hypothesis

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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. The paper examines the imaginary geography of the poetic formula "from place_a to place_b" across six European languages. Using georeferenced PoeTree corpora, we analyze the types, distances, and directions of these spans, which reflect a "soaring view" rooted in the ode tradition and sustained in 16th-19th century poetry. We show that the formula functions as a tool of political and cultural boundary-making: Romantic-era national literatures favour more local spaces, while imperial traditions combine local and global spans. Long-distance formulas tend to align along the East-West axis, being an important framework for global geographical imagery. We use statistical modelling to distinguish geographical symbolic "centers" and "borders" and show that centers are typically political entities and borders are natural features, pointing to a shared European geographical imagination.

1. Introduction

A bird's-eye view of a battle in *The Lord of the Rings*. A screen hovering over lands and armies in a computer strategy game. An eighteenth-century panegyric ode, in which a poet's gaze "ascends" into the sky to deliver praise to an empress. What these three things have in common is a view through an aerial, "soaring" camera – a technique originally codified not in filmmaking, but in literature.

In European cultures, the soaring perspective most likely originates from poetry, arriving in modern literatures through the genre of ode. To briefly summarize its complex history: the ode genre in modern European literatures has its roots in ancient Greek Pindaric odes. Originally, Pindar's odes honored athletes and explored mythological themes, settings, and locations. Adapted into Romance literatures via Greek and Latin, early Italian odes incorporated panoramic view and Pindaric picturesqueness to praise private individuals, which in turn gave impetus to the Pléiade poets to establish the ode in France (Maddison 1960).

French poets – most notably Ronsard and Malherbe, who were crucial in establishing the genre – dedicated their odes to high officials. This gave the genre its political turn and transferred the soaring view from mythological locations to real geographical ones: cities and castles, rivers and mountains that outlined the boundaries of one's territory.

By the 16th-17th centuries, the French absolutist ode obtained its definitive form, with the soaring view already fixed as a key element. For example, in Nicolas Boileau's

conference version

famous work *The Art of Poetry* (*L'Art Poétique*, 1674), the ode is described as "raising its ambitious flight to the heavens" ("L'Ode [...] Élevant jusqu'au Ciel son vol ambitieux" – Boileau 1872, 215). It was also Boileau who translated (Pseudo-)Longinus' treatise *On the Sublime* and introduced the notion of the elevation of the mind through the use of language. For poets seeking the sublime, the mind's elevation grants, among other things, the ability to observe infinitely and speak of everything from above (Marin 1993; Ram 2003). The sublime notion was particularly influential in the 18th and early 19th centuries, and in some traditions the connection between the ode and the sublime resulted in panoramic views – where placenames were used as the markers on the map, charting and glorifying the imperial territories (Ram 2003).

During the 18th and 19th centuries, the soaring view was still embedded in the elevated tone of the ode. This poetic tradition gave rise to a common line formula "FROM place_a TO place_b", which uses real space to define, contest, or will into existence political abstractions; to observe vast territories from a vantage point so high that the poet could see the symbolic truth of historical events, unfold past and future. The names of the particular locations with the exact "from A to B" formula appeared already in the early writings of Malherbe, as for example, in the *Stances Récit d'un Berger au Ballet du Triomphe de Pallas...* (1615)[1], and can be found in the 18th-century odes in German and Russian (Pumpyanskiy 1983; Ram 2003).

The basic grammar of the "from A to B" formula, codified within ode, has diffused across genres and poetic practice, echoed in charged political verse, satire and lyrical subversion. Its core use – mobilizing imperial ambition and mapping dominion – necessarily follows Napoleon ("That eagle whose flight for twelve years grew weary, / From Cairo to the Capitol, and from the Tagus to the Volga!"[2]), Russian territorial threats , like in Pushkin's *To the Slanderers of Russia* ("Are we so few? Or from Perm to Taurida <...> Shall not Russia rise, / All bristling with steel?"[3]), Prussian state ("From the Neman to the Rhein"[4]) and, infamously, whole Germany ("From the Meuse to the Neman / From the Adige to the Belt"[5]). The formula maps and re-inscribes national borders of a land that lacks them – Czech national movement of the 19th century obsessively charts its territories in verse, moving between Bohemian Forest and Krkonoše. The poetic gaze extends to imaginary, wishful unities: Ján Kollár spends full two quatrains of a sonnet from *Slávy dcera* charting the geography of the pan-Slavic world ("From Mount Athos to Mount Triglav, to Pomerania") to "kiss each other / and behold our homeland: All-Slavia"[6]. Not all geography is exclusively national or ethnic: Karel Destovnik, a

1. Cf.: "Voyes des bordes de Loire et des bords de Garonne / Jusques à ce rivage où Thétis se courrone" ["From the banks of the Loire and the banks of the Garonne / Unto that coast where Thetis is crowned"] (Malherbe 1842, 211).
2. "Cet aigle dont le vol douze ans se fatigua / Du Caire au Capitole et du Tage au Volga!", Victor Hugo, *Au Colonel G.-A. Gustaffson*, PoeTree ID: fr-7430. This and the following examples are drawn from the PoeTree corpus version 1.0.0 (Plecháč et al. 2025a), which is also accesible online at `https://versologie.cz/poetree/browser/search`). The data repository accompanying this paper provides metadata for all cited examples, linked to the full dataset via IDs (see subsection 2.1). Throughout the paper, examples are identified by the author's name, poem title, and their PoeTree IDs in the footnotes. Full bibliographic details and compete texts are accessible through the PoeTree corpus.
3. "Иль мало нас? Или от Перми до Тавриды <...> Стальной щетиною сверкая, / Не встанет русская земля?...", Alexander Pushkin, Клеветникам России, ru-38723.
4. "Vom Njemen bis an den Rhein", Theodor Fontane, *Kaiser Wilhelms Rückkehr*, de-11850.
5. "Von der Maas bis an die Memel / Von der Etsch bis an den Belt", August Heinrich Hoffmann, *Das Lied der Deutschen*, de-18548.
6. "Od Athosa k Trigle, k Pomořanům <...> Líbejme se při tom vespolek, To hle vlast je naše: Všeslávia!", Ján Kollár, *Od Athosa k Trigle, k Pomořanům...*, cs-21209.

conference version

Slovene Partisan poet who died at 22 fighting Nazi Germany, lists his comrades "from Jesenice to Trbovlje"[7] – major historical mining and ironwork cities – and overlays class on top of space.

This intensely rhetorical and political dimension of the formula opens it to subversion and inversion of authority: Ambrose Bierce promises "golden reign of Reason" that fills the world "from Walnut Creek to San Jose" – a span of around 80 kilometers in California (en-1373). Soviet poet Pavel Antokolsky replaces Russian imperial geography with the span between places that are also tragic events: from Khodynka (the infamous crowd crush) to Tsushima (the devastating naval defeat). His post-monarchy rewrite of the *Erlkönig*, filled with political ghosts, ends with "Father, have we arrived? Where are we? — In Russia. / We are buried in the earth, Alyosha" [8].

However, perhaps the most significant transformation happens in lyrical verse, where the space between A and B loses empires, armies, ghosts, and corpses. A quiet poem *Čas deževja* ("Rain season") by Črtomir Šinkovec charts Slovenia ("Od Ljubljane do Kočevja", sl-5217) to track the passage of endless rain clouds over those who walk to and back from school. The new point of view, which matches that of a weather channel, becomes, perhaps, more haunting if we keep in mind the ideology of poetic soaring and the historic usage of this formula.

Despite the long presence of the formula across European tradition and its ideological use, to our knowledge, there is no comprehensive large-scale study focusing on its role, meaning, and the way it structures imaginary geography.

Large-scale digitization of texts, alongside advances in GIS, automatic named-entity recognition, and geocoding, has transformed the spatial humanities. From the pioneering *Atlas of the European Novel* by Franco Moretti (Moretti 1998) to its critics (Döring 2013) and more recent theoretical (Juvan 2015) and practical work on literary space (e.g., Evans and Wilkens 2018; Wilkens 2021; Skorinkin and Orekhov 2025), geography has become embedded in the modern methodological toolkit. These studies have demonstrated, on a large scale, how literary geography reflects nation-building, as well as perceptions of and attitudes toward a state's inner and outer borders.

For narrative prose, not only can the full set of mentioned locations be mapped as static points, but spatial spans – such as the paths traced by characters moving from one location to another – can also be modeled (Wilkens et al. 2024). Because narrative space in poetry is far more limited for such tracing, we propose that the "soaring view" originating in the ode constitutes a valuable source for modeling poetic geography not as isolated points, as in previous studies (Kuzmenko and Orekhov 2016; Gavin and Gidal 2017), but as geographical spans.

This study investigates the imagined geography of the "from A to B" formula – its span and direction – in poetry across six European languages. By comparing several modern poetic traditions, we examine how the direction, the covered distance, and the geographic imagination are structured by the formula. We show how natural environments and political entities act differently in the collective map of Europe, charted by the soaring

7. "Od Jesenic do Trbovelj", Karel Destovnik, *Kralj Matjaž*, sl-2225.
8. "Отец, мы доехали? Где мы? – В России / Мы в землю зарыты, Алеша.", Pavel Antokolsky, Последний, ru-23766.

poetic view. We use formal models to differentiate between "centers" and "borders" of the literary imaginary that, taken together, reflect back the "implied" map of Europe.

2. Data and Methods

2.1 Data

We look for formulas in the georeferenced PoeTree corpora[9] (Plecháč et al. 2024). Named entities were recognized using best-performing models at the time of tagging for specific languages after their evaluation against the gold standard, namely: NameTag 2 for Czech (Straková et al. 2019), spaCy (Honnibal et al. 2020) for English and Slovene (`en_core_web_trf` and `sl_core_trf` pipelines), and few shot prompting of GPT-4 Turbo for German, French, and Russian (Plecháč et al. 2025b). Recognized unique toponyms were manually linked to Wikidata, which was used to extract their geographic coordinates (latitude, longitude); we tested multiple coordinates sets for rivers – including source, mouth, and intermediate points – to assess the robustness of the approach for entities that are not confined to a single point (see subsection 3.3). For each toponym we also added available Wikidata classification ("country", "continent", "river", etc.) that we later cleaned, refined and adjusted manually (see subsection 3.2).

We then extracted line-based formulas that fit variations of "from place_a to place_b" using a simple set of rules:

1. A set of prepositions of origin (PO) was defined for each language as: {z, od} in Czech, {von, vom} in German, {from} in English, {de} in French, {с, от, из} in Russian, and {iz, od} in Slovene.

2. A set of prepositions of destination (PD) was defined for each language as: {do, po, k} in Czech, {nach, bis, zu} in German, {to} in English, {à} in French, {до, на, в} in Russian, and {do} in Slovene.

3. Each sequence of lemmata fitting the pattern `PO x G x PD x G` where "G" denotes geographical entity and "x" stands for zero to four other lemmata was considered a formula.

The resulting dataset, after manual checks and cleaning, included 1,061 formulas found in 881 poems by 380 authors; language-specific data is summarized in Table 1. The number of formulas is much smaller than the number of mentioned geographical locations overall: only part of the most frequently mentioned geographical entities are also frequent in the "from–to" formulas (average Kendall rank correlation between the top-20 locations between formulas and all entities in each corpus is ~0.27).

2.2 Overall approach

Since the data presents a "soaring view" between places, often spanning thousands of kilometers, we calculated the Haversine distance between coordinates, approximating a

9. Although PoeTree currently includes corpora in eleven languages, not all are supplied with the geographical annotation required for this study. Our selection of six corpora reflects data availability at the time of the research. While we acknowledge the need to extend this work to other poetic traditions, the six traditions represented here already span a broad typological and cultural range, and the results are unlikely to be an artifact of this particular selection.

conference version

Corpus	Time span	Tokens total	N geo entities	NER model	N formulas
Czech (cs)	1700–1900	18,540 k	31,515	NameTag	308
German (de)	1500–1900	12,482 k	23,737	GPT-4T	58
English (en)	1700–1900	17,506 k	33,971	Spacy	294
French (fr)	1600–1900	7,877 k	19,283	GPT-4T	217
Russian (ru)	1700–1920	9,235 k	21,097	GPT-4T	162
Slovenian (sl)	1800–2000	875 k	2,709	Spacy	22

Table 1: Corpus overview.

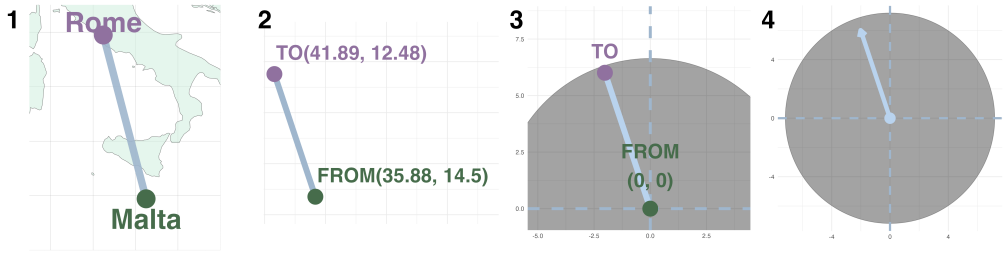


Figure 1: Transformation of a single "from–to" formula to the "clock" plot.

bird or plane flight distance in kilometers.

To summarize directions, we transposed each formula's "from" place to (0,0) coordinates and moved the "to" place accordingly, namely subtracting the original "from" values from the "to" coordinates. The example on Figure 1 presents how the line "from Malta's temples to the gates of Rome"[10] resulted in the coordinates `from_Malta(35.88, 14.5)` and `to_Rome(41.89, 12.48)`, and after the transposition these coordinates became `from(0, 0)` and `to(6.01, -2.02)`. We then visualize each poetic tradition on a "clock" plot (step 4 on Figure 1), where each clock hand represents one transposed "from–to" formula, its direction and distance.

We then aggregate the directionality of the formulas and the covered distance of the gaze for various summary and descriptive statistics. The exploratory analysis focuses on three things: 1) distances and directions in the formula; 2) types of places the formula connects; 3) structure that the formula imposes on locations and the European map. We use formal modeling to test the differences between natural and administrative locations; the model will be described in the corresponding section.

3. Results

3.1 Long vs short distances and their directions

The "from–to" formulas encompass both local and global panoramas. The distances observed between paired locations range from very short spans (less than 1 km) to extremely long ones (maximum distance: from Andes to Tibet[11], 18,209 km). The mean and median distances vary between the corpora (Table 2), yet the relative proportions of distances shorter and longer than the mean remain similar, with approximately 70% of formulas in all corpora describing shorter spans. At the same time, both mean and median distance values are particularly low in Czech and Slovenian corpora. If we are

10. John Greenleaf Whittier, *The Christian Tourists*, en-37986.
11. "Des Andes au Thibet", Victor Hugo, *XX (Ce que vous appelez dans votre obscur jargon...)*, fr-8160.

conference version

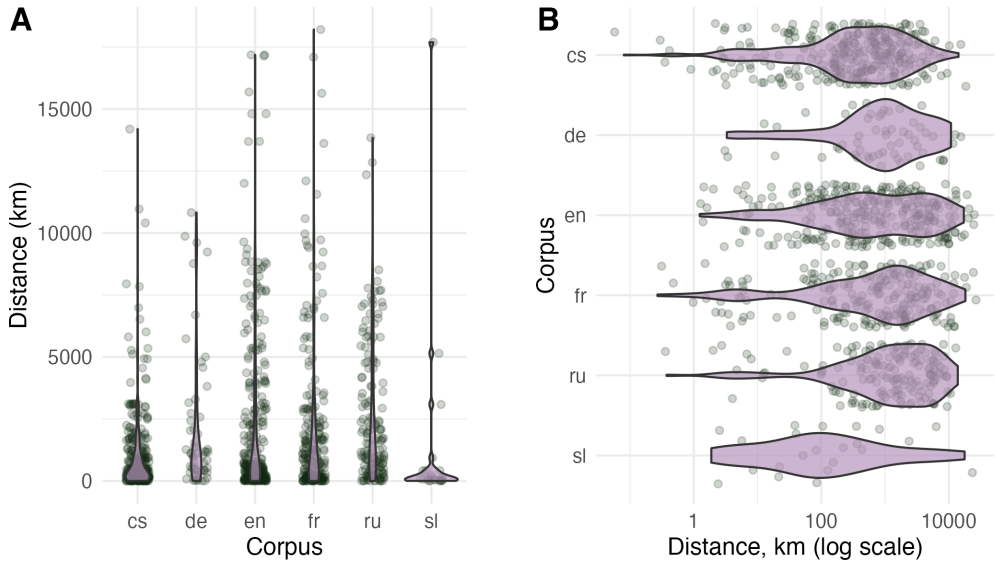


Figure 2: Distribution of distances: a) raw values; b) log scaled.

conference version

to apply an external threshold uniformly across corpora – say, arbitrarily considering distances over 1000 km as "long" – the proportion of long-distance spans in Czech and Slovenian becomes even smaller relative to the other corpora.

Corpus	Mean distance, km	Median distance, km	% dist > mean	% dist > 1000 km
cs	1054	436	27.9 %	29.2 %
de	2143	1100	27.6 %	55.2 %
en	2335	648	31.3 %	42.2 %
fr	2213	974	28.6 %	49.8 %
ru	2551	1478	33.3 %	57.4 %
sl	1321	90.9	13.6 %	13.6 %

Table 2: Distance statistics for each language.

There is no single correct way to split spans into shorter and longer ones: an approach based on a single threshold for all languages enforces a single global perspective onto national traditions, while an approach based on language-specific averages allows literatures to function in their own scales of national spaces, not overshadowed by imperial distances of England or Russia. When referring to "short" and "long" distances below, we adopt the second approach ("national averages"), prioritizing variability and sacrificing some uniformity in the interpretation of the results.

Although the number of formulas found in the Slovenian corpus is too small to be reliable, the substantially larger Czech dataset allows us to suggest a distinction between "younger" national poetic traditions established during 19th-century national revivals and "imperial" literatures, such as Russian, English, and French. The former tend to explore shorter distances and local spaces, while the latter frequently feature very long-distance spans, extending across imperial and continental borders.

Across all examined corpora, shorter distances extend in multiple directions without a dominant axis (Figure 4), whereas longer spans tend to organize along the horizontal

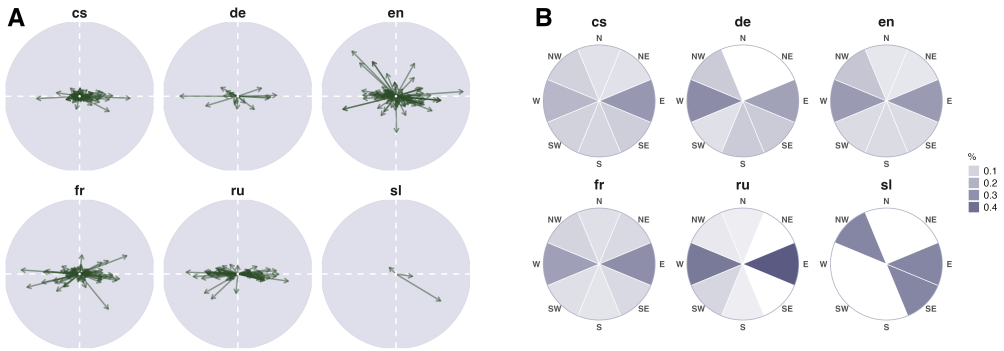


Figure 3: Long distances in "from–to" formulas: a) the "clock" plots, true distances, same scale; b) percentage of formulas looking into each of the eight directional segments.

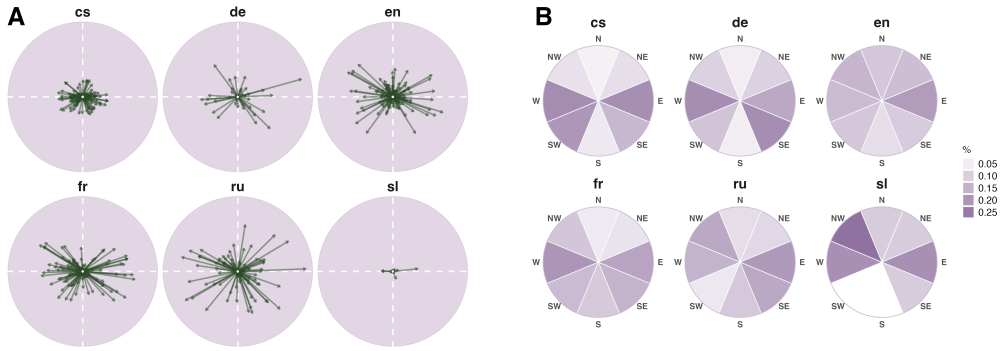


Figure 4: Short distances in "from–to" formulas: a) the "clock" plots, true distances, same scale; b) percentage of formulas looking into each of the eight directional segments.

axis (Figure 3): they are visibly "squished" from the top and bottom[12]. To summarize the directional distribution of these formulas, Figure 3 and Figure 4 show the percentage of formulas falling into each of the eight "clock" segments, defined by cardinal and intercardinal directions such as North, Northeast, and East.

Russian poetry provides the most pronounced example: 46% of spans longer than the mean are oriented eastward and 37% westward, meaning that 82% of long spans lie along the horizontal axis. A strong East-West orientation among long-distance formulas is also evident in the French and German corpora, where 55% of long spans align with this axis. The English corpus constitutes a partial exception, with a notable share of very long distances directed toward the Northwest (14% of long spans). Nevertheless, the majority of long spans (54%) are still oriented horizontally, from East to West or vice versa. It is worth noting that even in short distances, the tendency to orient the gaze by East–West is still visible: southern and northern spaces remain less occupied.

3.2 Types of locations

Each geographical location mentioned in the formulas was manually assigned to one of 22 categories, which were further grouped into six major classes and two overarching

12. These results remain stable when using different coordinate sets (sources, mouths, or intermediate points) for rivers.

conference version

groups: human-established political or administrative entities and natural features[13]. There is no clear division whereby particular classes are used exclusively in short- or long-distance formulas: the same types of locations and formulaic patterns can describe both local and global spaces. The classes most commonly occurring in all corpora are cities and water-related entities: the most frequent spans are "from {city} to {city}" and "from {river} to {river}", both used to express shorter as well as longer distances.

Table 3 shows that political entities appear in the formula slightly more frequently than natural features in most corpora, with the exception of Czech. The latter strongly favors formulas that include mountains (accounting for 22% of all geographical entities), making "from {mountain} to {mountain}" the second most frequent formula type after the common "from {city} to {city}".

Counterintuitively, water-related entities are least frequent in the English corpus, where political entities – particularly countries – are more prevalent. At the same time, countries are relatively rare in the formulas overall, and very long spans such as "from Australia to Norway" are often recast as spans between two cities (e.g., "From Hobart Town to Hammerfest"[14]). Another common way to convey a very broad "soaring view" is to use water boundaries, most frequently rivers and straits (e.g., "from the River Ganges to the River Amazon"[15], "from the Bering Strait to the Strait of Magellan"[16]).

	cs	de	en	fr	ru	sl
Political entities (total)	44.8	59.5	59.9	63.4	52.5	65.9
City	32.3	45.7	39.5	52.3	44.1	56.8
Country	7.0	6.9	**12.6**	4.4	8.0	–
Settlement	5.2	6.9	7.3	6.2	–	9.1
Natural entities (total)	55.2	40.5	40.1	36.6	47.5	34.1
Land	10.9	7.8	16.5	11.5	11.1	–
Mountain	**22.6**	2.6	4.6	4.1	7.1	11.4
Water	20.6	25.0	17.5	20.5	27.8	20.5
Other	1.5	5.2	2.0	0.7	1.2	–

Table 3: Distribution of location types (all values are %).

Local spaces are often described with the same formulaic patterns, such as "from {city} to {city}" and "from {river} to {river}". However, smaller settlements such as villages are used exclusively to convey short-distance perspectives (the median distance described by "from {village} to {village}" pattern is 9 km). The shortest distances described by the "from–to" formula are those between parts of a city. In most cases, these are paths running through the capital cities, for example, from the Prague district Hradčany to Karlín[17] or through the Paris city center – "from the Boulevard du Temple to La

13. Political entities: I. City: cities (733), ancient cities (59), city parts (87); II. Country: countries (170); III. Settlement: castles/forts (14), villages (93), palaces (6), ancient sites (5). Natural entities: IV. Land: regions (197), islands (40), ancient regions (7), continents (16); V. Mountains: mountains (208), hills (2), volcanos (4); VI. Water: rivers (358), seas (48), bays (12), lakes (11), capes (8), straits (10); VII. Other (34). The number in brackets indicates the total number of occurrences out of 2,122 geographical entities, with two entities appearing in each of the 1,061 formulas. The taxonomy is based on Wikidata classes and was not parallel annotated, ambiguous cases were resolved jointly by the authors. Although the number and composition of the classes therefore involve a degree of subjectivity, we believe that these groupings still capture major differences between the most used geographical entities, such as cities and water-related features.
14. William Mackay MacKeracher, *Canadian-born*, en-21959.
15. "Du Gange à l'Amazone", Victor Hugo, *XLI (Qui que tu sois qui tiens un peuple dans ta main...)*, fr-7986.
16. "De Behring à Magellan", Victor Hugo, *Le Nid*, fr-7374.
17. "Z Hradčan do Karlína", Josef Svatopluk Machar, *Bože Požehnaní*, cs-29419.

conference version

Madeleine"[18]. Besides denoting the space inside own capitals, references to city parts may be used as metonymy and evoke specific historical episodes, such as wars: for example, among the French formulas, two explicitly refer to the Moscow Kremlin in connection to Napoleonic wars ("from Naples to the Kremlin"[19], "from the Kremlin to El Escorial"[20]).

3.3 Symbolic common locations: centers and borders

The geographic locations that appear in "from–to" formulas cannot be inferred from the lists of the most frequent mentioned places in poetry in general. Rather, the use of this formula produces a specific set of locations that serve as points of departure or arrival in a poet's imagined flight. These locations are shared across the corpora, being both frequent and stable across languages (Table 4).

Locations appearing in 5 corpora	Political entities: Athens (7), Cairo (9), Paris (27), Rome (27) Natural entities: Baltic Sea (22), Carpathian Mountains (9), Danube (24), Euphrates (8), Ganges (13), Rhine (21), Tiber (14)
Locations appearing in 4 corpora	Political entities: Babylon (7), Bethlehem (8), Egypt (12), France (11), India (10), Japan (5), Moscow (16), Syria (5), Vienna (12) Natural entities: Asia (5), Don (8), Nile (12), Tagus River (12)
Top-20 "from" locations (all data)	Bohemian Forest (18), Paris (15), Baltic Sea (14), Giant Mountains (12), Tagus River (11), Volga (11), Prague (10), Danube (10), Rome (10), Moravia (9), Ganges (9), Moscow (9), Vienna (8), Tatra Mountains (8), Nile (8), Vistula (8), Neva (8), River Thames (7), Alps (6), Tiber (6)
Top-20 "to" locations (all data)	Bohemian Forest (29), Rome (17), Prague (16), Rhine (15), Danube (14), Tatra Mountains (12), Paris (12), Adriatic Sea (11), Tiber (8), Baltic Sea (8), Carpathian Mountains (7), Moscow (7), Vltava (6), Czech Republic (6), Italy (6), Peru (6), India (6), Egypt (6), Cairo (6), Beijing (6)

Table 4: The most frequent locations across corpora.

In most cases, the entities appearing in "from–to" formulas across the corpora are cities (both modern and ancient), countries, or rivers. Although cross-corpus frequency alone cannot provide evidence for the same meanings or functions of these locations, we observe that different traditions also agree in the direction of these spans. Figure 5 presents "clock" plots for the locations selected from the list above (only "from" directions shown). Certain locations consistently point in the same direction, thereby functioning as symbolic borders – for instance, formulas beginning at the Tagus River point eastward in all corpora (e.g. "from the Tagus to the Rhine"[21], "from the Tagus to the Volga"[22], while the Baltic Sea serves as a northern boundary, with its directions oriented towards the south ("from the Baltic Sea to the Black Sea"[23], "from Baltic to the Nile"[24]). By contrast, some locations – most notably Rome – aim in multiple directions across all traditions, functioning as symbolic centers ("from Rome to Paris"[25], "from Rome to Jamaica"[26], "from Rome to China"[27]).

18. "Du Temple à la Madeleine", Alphonse Daudet, *La Double Conversion*, fr-3754.
19. "De Naples au Kremlin", Albert Angot, *Sedan*, fr-112.
20. "Du Kremlin à l'Escurial", Victor Hugo, *Les Funérailles de Louis XVIII*, fr-7428.
21. John Lawson Stoddard, *The Death of Antoninus Pius*, en-33309.
22. "Du Tage au Volga", Victor Hugo, *Au Colonel G.-A. Gustaffson*, fr-7430.
23. "От Балтики до Черноморья", Pavel Antokolsky, Германия, ru-23588.
24. Angus Mackay, *The Sultan at the Kaiser's Kourt*, en-21727.
25. "De Rome à Paris", Alfred de Musset, *Conseils à une Parisienne*, fr-12336
26. "De Rome à la Jamaïque", Paul Verlaine, *Air et Duo*, fr-17403.
27. "С Риму до Китая", Antiochus Kantemir, Сатира *V.* На человека, ru-25667.

conference version

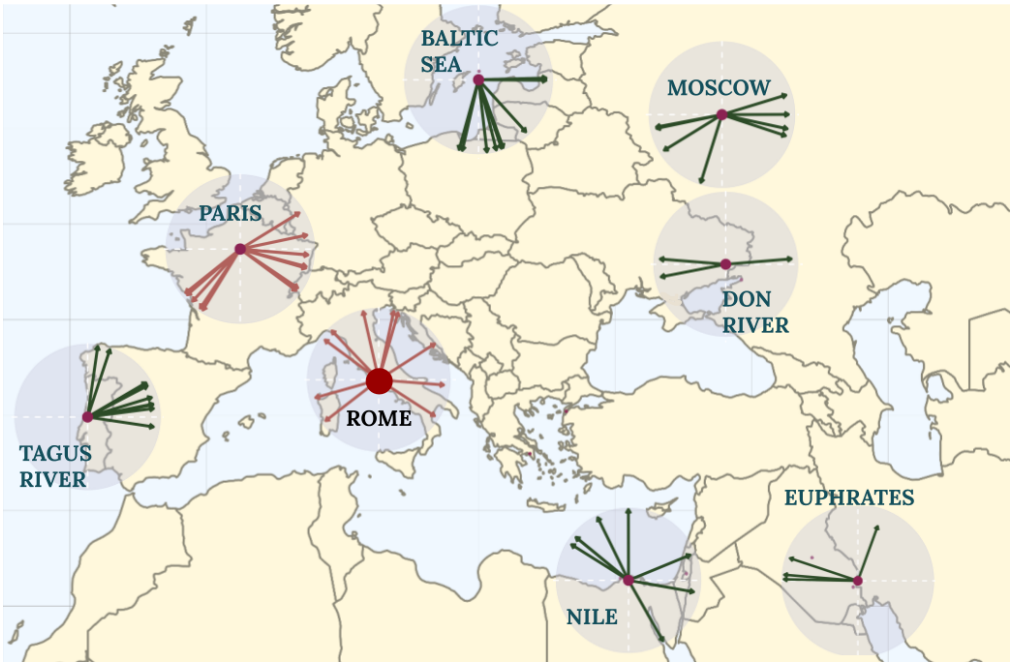


Figure 5: "Clock" plots for selected most common locations in the "from–to" formulas across languages; standardised distances showing directions when the location serves as "from" point.

Model. To operationalize the difference between, for example, Rome and Tagus river seen on Figure 5, we model how concentrated the "clock hands" are in certain directions. High directional concentration can be interpreted as "borderness" and low concentration as "centrality". To quantify these properties, we implement an entropy-based measure[28] as follows:

1. Take a location that appears multiple times either as a starting point ("from"), or as an endpoint ("to") of a formula.
2. Calculate pairwise circular absolute distances between angles of the "clock hands", which gives us a distribution of angle differences in the $(0°, 180°)$ range.
3. We bin the resulting continuous distribution into 36 bins and calculate the proportion of distances falling in each bin. This results in a probability distribution over the "difference" bins.
4. Finally, we measure how "unequal" the resulting distribution is using both Shannon entropy and the Gini coefficient[29], a standard measures of distributional inequality.

The underlying intuition of this approach is simple: the more concentrated are "A to B" vectors in a certain direction, the more concentrated there will be the differences between their angles, the more skewed, or inflated around certain values will be the distribution leading to lower entropy measures. Conversely, in the cases such as Rome, where the gaze travels in many directions, the differences between angles will cover a

28. Statistical modelling was conducted in R; the full implementation, including formulas and scripts, is available in the accompanying repository.
29. For binned distributions, Gini and entropy are highly correlated (-0.93); we prefer using entropy, as it is not bounded to the $(0,1)$ interval and is slightly easier to model.

conference version

larger range, resulting in a flatter distribution which will be reflected in higher entropy values.

From			To		
Place	Type	Entropy	Place	Type	Entropy
Ganges	natural	2.07	Tatra Mountains	natural	1.34
Bohemian forest	natural	2.23	Adriatic sea	natural	2.25
Baltic sea	natural	2.37	Baltic sea	natural	2.70
Tagus River	natural	2.48	Danube	natural	2.75
Prague	political	2.78	Bohemian forest	natural	2.77
Volga	natural	2.80	Paris	political	2.86
Moscow	political	2.84	Tiber	natural	2.87
Danube	natural	2.90	Prague	political	3.26
Paris	political	3.07	Rhine	natural	3.27
Rome	political	3.21	Rome	political	3.30

Table 5: Ten most frequent places in "from" and "to" positions, arranged by their borderness.

The approach seems to generally capture the structural difference between centers and borders (Table 5): Rome and Paris score the highest as centers (entropy values 3.2 and 3.1, respectively), while the Tagus (2.5), the Thames (2.2), the Ganges (2.0), and the Ural Mountains (1.3) tend to act as borders of the imagined space.

Does natural geography, then, generally tend to mark the borders of the poetic space, as opposed to political and administrative entities, like cities, that coalesce into centers? To test this, we fit a Bayesian regression to our "borderness" measure: the model predicts the amount of structure in the direction of the "gaze" vectors depending on the general type of a location ("natural" or "political").

The obvious complication here arises from data scarcity: there are many locations that participate in the formula only a few times, so that the measured entropy for them would be naturally small, and no definitive conclusion about their structure can be made. Instead of filtering them out, we include the count of each location (either as "from" or "to") in the model explicitly and allow it to interact with the location type.

The resulting model can be expressed as: `entropy ~ location type * log(number of occurrences) + number of languages`. The model predicts the "borderness" of a location (entropy) based on the location type, the number of occurrences (log-transformed), and their interaction, allowing the effect of frequency to differ by location type. The number of languages in which a location appears is included as an additional predictor. Entropy is modeled using a Gaussian likelihood with regularizing priors. We restrict the analysis to locations with at least three occurrences in the "from A to B" formulas. Two separate models were fitted: for locations that participate in formulas as "A" ("from", n=78) and locations that are "B" (endpoints, "to", n=79).

For both models we find that the difference between natural and political entities can be seen, but only at a higher number of occurrences: sometimes there are just not enough observations for drawing any conclusions about the structure using our entropy-based measure. However, for the data we have, frequent locations do differ: administrative and political entities are predicted, on average, to have higher entropy, that is, more "centrality", than natural entities. In more than 95% of the draws, posterior means for natural geographic locations are lower, and they lose on average ~0.3 points of entropy

conference version

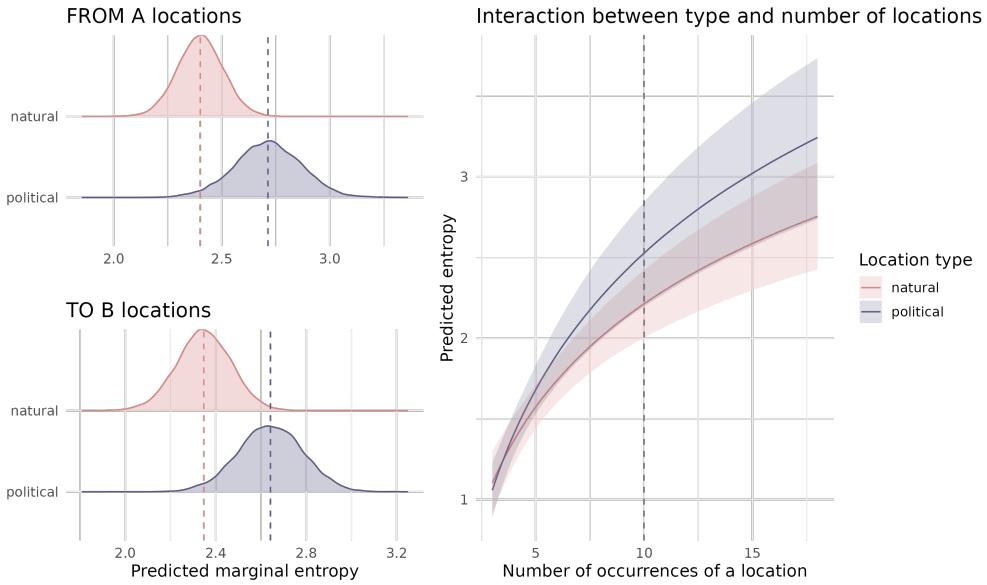


Figure 6: Model predictions (on the left) and interaction (on the right). The grey vertical line on the right pane marks the value of n that was used for predictions on the left. Colored lines on the left are the means of both distributions.

(predictions are made for `n_occurrences = 10`, `n_languages = 3`).

It is worth noting that the center/border distinction holds if we fit a series of models on randomly determined coordinates for rivers (either as mouth, source, or midpoint). Average decrease in entropy remains at 0.29 and the average difference in draws lands at 94%. Single-coordinate coding of rivers is, of course, inadequate, but, as we see, it remains a useful proxy for charting the poetic gaze that clearly travels not in a straight line between two coordinates, but rather as an area on the map: an area with a cardinal direction (not unlike a weather front). All the coordinate-picking and line-drawing within this area remains a hopeful approximation, that can, nevertheless, be validated, when we start to sample multiple *possible* lines. Even if particular origin points are unhelpful, the general direction still can be discerned.

These results should be treated as a weak tendency that we are able to cautiously infer given our very limited (even scarce) data. Natural borders can easily be centers: the river Rhine (entropy 3.27), being a border from many perspectives and drawing poetic gaze from many directions, turns into an European center; the Danube (2.9), a traditional border of Western Europe (and, before that, a Roman frontier) becomes a center of (Central) Europe. On the other hand, Beijing (1.3), a "northern capital", one of the oldest cities on Earth and an indisputable center of the region, for European poets soaring over the globe, serves as an eastern stopping point, the easternmost border.

It is significant, then, that *no single formula starts with Beijing*: it always ends there (n=6). The symbolic and semantic origin point of the soaring perspective lies in the West. This is reflected in a curious asymmetry between starting and endpoint locations of a formula, already partly seen in Table 4. Among top-20 "from" locations, only the Nile and the Ganges lie outside of European space (yet still mark the Classic world, the *ecumena*). Among top-20 "to" locations there are Peru, India, Egypt, Cairo, Beijing. Almost a plane departure schedule from Frankfurt airport. Perhaps the most resistant

conference version

city to this asymmetry is Rome (and other transnational centers, like Paris): not only do all roads lead there, but also all roads start there. With some preference, perhaps, for being an endpoint (n "from Rome" = 10; n "to Rome" = 17).

4. Conclusion and Discussion

The poetic formula "from A to B" reveals a number of commonalities in geographical imagery across six examined European traditions. Originating in the widely used poetic genre of the ode, the formula is employed to map both local spaces and global boundaries. Unlike the "small worlds" and domestic spaces characteristic for fictional character movements in prose (Wilkens et al. 2024), poets' ability to observe infinitely from above allows scaling from the corners of a capital city's to distant, cross-continental spans.

Shorter "from–to" spans are slightly more characteristic of national literatures established during Romanticism, which demonstrate a greater interest in exploring their own spaces, and, in the case of the Czech corpus, even introduce their own geographical features (i.e., mountains) into the formulaic pattern. However, this observation requires further confirmation using poetic data from other "younger" national literatures. Similarly, the nature of the data – a commonly used but rather rare formula – makes it impossible to trace chronological changes in the directions and distances of the spans: current number of formulas do not allow for separation into finer time periods, so we were unable to test whether the imperial literatures have also begun with slightly smaller-scale "from–to" spans similar to those in the Czech corpus.

The long-distance spans, which are present in all examined corpora, demonstrate the shared global imagination existing in different – historically and geographically speaking – poetic traditions across the European continent. We observed that long spans are typically organized along the East–West axis, and even short-distance formulas tend to be drawn to this organizing direction. In cultural evolution, this axis is sometimes regarded as a factor that allegedly enabled rapid cultural transmission and the spread of innovations across Eurasia, contributing to the flourishing of its regions (Diamond 1997). However, recent research shows that there is no quantitative evidence that similar latitudes along the East–West axis alone constitute a decisive factor in faster or easier cultural transmission (Chira et al. 2024). Our findings, which reveal the persistent presence of East–West orientation in modern European poetry, suggest that this axis holds significant *cultural* value, particularly as a framework for global imagination and generalized representations of Europe. A larger dataset with non-European poetic traditions may clarify whether this pattern of imagining the global extends beyond the West.

The six examined European traditions unanimously pointed to a number of geographical locations as "centers" or "borders" of a shared poetic map of Europe. Our statistical models suggest that natural entities, particularly water features such as rivers and seas, are more likely to function as borders, i.e., to have a more pronounced common orientation. In contrast, political entities, such as cities, tend to serve rather as "centers" or, in other words, the points of departure and arrival from multiple directions. Our current data, however, indicate that this pattern is mostly true for the locations in

conference version

Europe, while cities such as Beijing are perceived as exotic outer spaces. Nevertheless, the dispersion of geographical locations common to most of our corpora suggests that the 16th–19th centuries poetic imagination of space was somewhat broader and more interconnected than might be assumed from national borders alone.

Fernando Pessoa, writing as Caeiro, said that people looking at Tagus see "everything which isn't there" (Pessoa 1998, n.p.), referring to the semantization of space through history and literature. This is what the "soaring spirit" of a poet sees as well, even if it refers to locations that, technically, have coordinates. We see this textualization of space not only from a global perspective, but also from the national ones: an abundant data from Czech corpus brings regional markers to the top of frequency lists: Bohemian Forest, Prague, Moravia. National space emerge and orients itself against the shared European and global spatial imagination. After all, "from A to B" formula and the close alignment with the historical ode tradition, survives in texts everyone is constantly exposed even today: national anthems.

5. Data Availability

Data and code are available at: https://doi.org/10.5281/zenodo.19570538.

6. Acknowledgements

This article was supported by the Czech Science Foundation (project ga23-07727S).

7. Author Contributions

Antonina Martynenko: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing - review & editing

Artjoms Šeļa: Conceptualization, Formal analysis, Methodology, Visualization, Writing – original draft, Writing - review & editing

Petr Plecháč: Conceptualization, Data curation, Funding acquisition, Methodology, Writing – original draft

References

Boileau, Nicolas (1872). *Œuvres poétiques*. Vol. 1. Paris: Imprimerie générale.

Chira, Angela M., Russell D. Gray, and Carlos A. Botero (2024). "Geography is not destiny: A quantitative test of Diamond's axis of orientation hypothesis". In: *Evolutionary Human Sciences* 6, e5. 10.1017/ehs.2023.34.

Diamond, Jared M. (1997). *Guns, germs and steel: the fates of human societies*. New York: Norton.

Döring, Jörg (2013). "How Useful Is Thematic Cartography of Literature?" In: *Primerjalna književnost* 36.2, 139–149.

Evans, Elizabeth and Matthew Wilkens (2018). "Nation, Ethnicity, and the Geography of British Fiction, 1880-1940". In: *Journal of Cultural Analytics* 3.2. 10.22148/16.024.

conference version

Gavin, Michael and Eric Gidal (2017). "Scotland's Poetics of Space: An Experiment in Geospatial Semantics". In: *Journal of Cultural Analytics* 2.1. 10.22148/16.017.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd (2020). "spaCy: Industrial-strength Natural Language Processing in Python". In: *Zenodo*. 10.5281/ZENODO.1212303.

Juvan, Marko (2015). "From Spatial Turn to GIS-Mapping of Literary Cultures". In: *European Review* 23.1, 81–96. 10.1017/S1062798714000568.

Kuzmenko, Elizaveta and Boris Orekhov (2016). "Geography Of Russian Poetry: Countries And Cities Inside The Poetic World". In: *Digital Humanities Conference*. https://dh2016.adho.org/abstracts/3.

Maddison, Carol (1960). *Apollo and the Nine: A History of the Ode*. London: Routledge and Kegan Paul.

Malherbe, François (1842). *Poésies de François Malherbe*. Paris: Charpentier.

Marin, Louis (1993). "1674 – Le sublime, l'infini et le "je ne sais quoi"". In: *De la littérature française*. Paris: Bordas, 327–332.

Moretti, Franco (1998). *Atlas of the European novel: 1800-1900*. London; New York: Verso.

Pessoa, Frenando (1998). *Fernando Pessoa & Co.: Selected Poems*. Ed. by Richard Zenith. New York: Grove Press.

Plecháč, Petr, Silvie Cinková, Robert Kolár, Artjoms Šeļa, Mirella De Sisto, Lara Nugues, Thomas Haider, and Neža Kočnik (2024). "PoeTree: Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian and Spanish". In: *Research Data Journal for the Humanities and Social Sciences* 9.1, 1–17. 10.1163/24523666-bja10044.

Plecháč, Petr, Artjoms Šeļa, Helena Bermúdez Sabel, Klemens Bobenhausen, Silvie Cinková, Ingerid Løyning Dale, Éliane Delente, Mirella De Sisto, Thomas Haider, Benjamin Hammerich, Péter Horváth, Ranveig Kvinnsland, Neža Kočnik, Robert Kolár, Kirill Korchagin, Antonina Martynenko, Adiel Mittmann, Benjamin Nagy, Borja Navarro Colorado, Lara Nugues, Gábor Palkó, Vladimir Plungian, Richard Renault, Pablo Ruiz Fabo, Levente Seláf, and Dmitri Sitchinava (2025a). *PoeTree. Poetry Corpora in Czech, English, French, German, Hungarian, Italian, Norwegian, Portuguese, Russian, Slovenian, and Spanish*. 10.5281/ZENODO.17414036.

Plecháč, Petr, Artjoms Šeļa, Silvie Cinková, Mirella De Sisto, Lara Nugues, Neža Kočnik, Robert Kolár, and Thomas Haider (2025b). "Named Entity Recognition and Linking in PoeTree Corpora". In: *Studia Metrica et Poetica* 12.2, 7–18.

Pumpyanskiy, Lev V. (1983). "Lomonosov i nemetskaya shkola razuma". In: *XVIII vek. Sbornik 14. Russkaya literatura XVIII – nachala XIX veka v obshchestvenno-kulturnom kontekste*. Leningrad: Nauka, 3–44.

Ram, Harsha (2003). *The imperial sublime: a Russian poetics of empire*. Madison: University of Wisconsin press.

Skorinkin, Daniil and Boris Orekhov (2025). "The Outward Turn. Geocoding the Expansion of Fictional Space in Russian 19th-Century Literature". In: *Journal of Computational Literary Studies* 4.1. 10.48694/JCLS.4228.

Straková, Jana, Milan Straka, and Jan Hajic (2019). "Neural Architectures for Nested NER through Linearization". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 5326–5331. 10.18653/v1/P19-1527.

conference version

Wilkens, Matthew (2021). "Too isolated, too insular: American Literature and the World". In: *Journal of Cultural Analytics* 6.3. 10.22148/001c.25273.

Wilkens, Matthew, Elizabeth Evans, Sandeep Soni, David Bamman, and Andrew Piper (2024). "Small Worlds: Measuring the Mobility of Characters in English-Language Fiction". In: *Journal of Computational Literary Studies* 3.1. 10.48694/JCLS.3917.

conference version

jcls. Journal of Computational Literary Studies

conference version

Stylometry or Embeddings? Authorship Attribution for Russian and Italian Poetry

Maria Levchenko[1]

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

OPEN ACCESS

Citation
Maria Levchenko (2026). "Stylometry or Embeddings? Authorship Attribution for Russian and Italian Poetry". In: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7985

Date published 2026-05-05 (preprint)
Date accepted tbc
Date received 2026-01-09

Keywords
authorship attribution, stylometry, large language models, embeddings, poetry

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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. Large Language Model (LLM) embeddings achieve strong performance in authorship attribution, yet it remains unclear which aspects of literary style they encode. We address this question through a residualization analysis of two poetry corpora: 5,800 Russian poems (29 authors) and 10,400 Italian poems spanning seven centuries (52 authors). Using a progressive residualization waterfall, we subtract interpretable stylometric features and high-dimensional lexical controls from embedding representations to quantify their contribution to attribution accuracy.
For Russian poetry, residual signal collapses to near chance (1.1×) after accounting for character n-grams and word bigrams, indicating that embeddings largely compress orthographic and lexical distributions already exploited in classical stylometry. For Italian poetry, a reduced but significant residual persists (4.6× chance), consistent with diachronic or dialectal variation not fully captured by standard features. We conclude that embeddings and stylometry rely on overlapping signals but differ in how they weight lexical, semantic, and historical variation.

1. Introduction

Large language models have become objects of investigation n computational literary studies, including probing experiments on whether their representations capture formal properties of verse such as meter and rhythm (Glaser 2025; Jannidis et al. 2025). This raises a natural question: can LLM representations serve as instruments for literary analysis?

The most direct access to how LLMs represent text is through embeddings — dense, high-dimensional vectors that encode distributional properties of input texts. In a literary context, embeddings provide a compact representation of textual variation that can be evaluated empirically. Preliminary experiments on poetry corpora show that simple classifiers trained on embeddings achieve high accuracy in pairwise authorship attribution, indicating that embedding spaces encode author-discriminative information.

But a 3,072-dimensional vector offers no obvious interpretation. We cannot inspect an embedding and determine which textual properties — lexical preferences, syntactic habits, prosodic structure, or orthographic conventions — contribute to attribution performance. High classification accuracy alone therefore does not establish that embeddings capture stylistically meaningful patterns; it may instead reflect sensitivity to surface regularities that are not easily interpretable in literary terms.

Classical stylometry approaches authorship attribution through explicitly defined feature spaces — character n-grams, lexical distributions, and morphosyntactic patterns — whose contributions can be inspected and quantified. This design enables diagnostic analysis: attribution decisions can be decomposed into concrete stylistic signals and evaluated for stability across texts, genres, and historical conditions. At the same time, predefined feature inventories limit coverage: attribution performance has repeatedly improved as new feature classes are incorporated, suggesting that no taxonomy is ever fully exhaustive.

High-dimensional embeddings, trained on large corpora to model distributional structure, motivate an empirical test of whether they capture stylistic patterns beyond those accessible to existing feature inventories: how do embedding-based representations relate to established stylometric methods, and what does each contribute?

This paper investigates this question through authorship attribution across two typologically distinct corpora: Russian poetry (1850–1930) and Italian poetry spanning seven centuries. This multilingual design follows recent evidence that LLM representations vary substantially across languages, motivating evaluation beyond English-only benchmarks (Ahia et al. 2023; Kim et al. 2025; Lundin et al. 2025).

Our analysis proceeds in four stages. We first compare attribution performance across embeddings, stylometric features, and their combination (§5). We then apply a residualization waterfall, progressively removing interpretable feature tiers from embedding representations to quantify their contribution to attribution accuracy (§6). We probe embedding vectors directly to test whether specific textual properties are linearly decodable (§7). Finally, we analyze the complementarity between the two approaches through bidirectional residualization and error analysis (§8).

2. Stylometric Features, Neural Attribution, and Explanatory Gaps

Contemporary stylometry is characterized by an explicit concern for explanatory control: identifying which textual features carry authorial signal, under what conditions, and why. Foundational work established frequency-based feature sets as robust baselines for authorship attribution (Burrows 2002), while subsequent comparative studies refined these methods and assessed their stability across corpora and distance measures (Evert et al. 2017; Neal et al. 2017). Toolkits such as stylo (Eder et al. 2016) further standardized experimental practice, enabling systematic evaluation and reproducibility.

At the same time, stylometric research has repeatedly shown that no single feature inventory exhausts authorial style. Character n-grams proved especially effective for morphologically rich languages, capturing orthographic and morphological regularities missed by word-based features (Stamatatos 2013). Syntactic distributions, POS patterns, and rhythmic features each contribute incremental signal, but with distinct failure modes and text-length requirements. Crucially for poetry, attribution reliability degrades sharply for short texts: Eder (Eder 2013) estimates that several thousand words are typically required for stable stylometric attribution, far exceeding the length of most poems. Moreover, Rybicki and Eder (Rybicki and Eder 2011) demonstrate that

conference version

different frequency bands encode different kinds of information: for prose, function words dominate, while content vocabulary often introduces topical noise. Whether this hierarchy holds for poetry has remained largely untested.

Neural authorship attribution shifts this landscape by replacing explicit feature modeling with implicit representation learning. Early shared task evaluations showed that simple n-gram baselines remained competitive with or superior to neural approaches in cross-domain conditions (Kestemont et al. 2018), but transformer-based models achieve strong attribution performance without predefined stylistic features, raising the possibility that they capture patterns beyond established stylometric taxonomies. However, this performance comes at the cost of interpretability. Brad et al. (Brad et al. 2022) show that BERT-based attribution often relies on topical correlations rather than stylistic generalization, with accuracy degrading when topic and authorship are decorrelated. Contrastive approaches such as LUAR (Rivera-Soto et al. 2021) mitigate some topical bias by training author-invariant representations, but remain opaque with respect to which stylistic properties are encoded.

Parallel work in NLP interpretability has introduced probing classifiers as diagnostic tools for neural representations (Belinkov and Glass 2019). Probing studies have identified syntactic and morphological information in contextual embeddings (Hewitt and Manning 2019), and more recently examined authorship-relevant signals in prose (Wegmann et al. 2022). Yet these studies typically operate independently of stylometric feature theory and rarely address poetry or large-scale embedding models. Recent large-scale embedding models differ substantially from earlier Transformer encoders in both dimensionality and training regime; whether this increased scale yields qualitatively different stylistic representations, or merely amplifies topical and lexical correlations, remains an open empirical question.

Across these strands, a central gap remains. Stylometry provides interpretable feature inventories but no framework for explaining modern embedding-based attribution; neural methods deliver high accuracy but limited insight into stylistic encoding; probing reveals accessible information but lacks connection to established stylometric taxonomies. What is missing is a feature-grounded decomposition of embedding representations that can determine how much neural attribution performance is explained by known stylometric signals — and what, if anything, remains beyond them. This study addresses that gap through residualization, probing, and comparative analysis across poetic traditions.

3. Data

We analyze two poetry corpora from the PoeTree project (Plecháč et al. 2024, 2023), selected to test generalizability across languages, literary traditions, and historical spans.

The Russian corpus consists of 5,800 poems by 29 poets (1850–1930). To ensure balanced representation, we restrict the dataset to authors with at least 200 poems and sample 200 poems per author. The resulting corpus spans a relatively narrow historical window in which poets shared cultural context, social networks, and often thematic concerns, making authorship attribution primarily a test of stylistic discrimination rather than

conference version

Table 1: Corpus statistics for Russian and Italian poetry.

Metric	Russian	Italian
Authors	29	52
Poems	5,800	10,400
Poems per author	200	200
Historical span	1850–1930	1200–1900
Words per poem		
Mean	100.4	295.7
Median	81.0	100.0
IQR (Q1–Q3)	56–116	91–108
IQR width	60	17
Range	4–1,525	17–27,726
Lines per poem		
Mean	21.2	44.3
Median	16.0	14.0

historical or topical separation.

The Italian corpus comprises 10,400 poems by 52 poets spanning seven centuries (1200–1900), again with 200 poems sampled per author. This corpus serves as a contrastive test case, evaluating whether findings generalize across a much longer historical span and a different linguistic tradition. Its temporal breadth introduces additional sources of variation — orthographic conventions, vocabulary, and poetic forms evolve substantially over time — but also provides a stringent robustness check. Table 1 summarizes corpus statistics.

The Italian corpus has substantially higher variance in poem length, driven by the presence of long-form poems such as epics and narrative compositions. Median values therefore provide a more representative measure of typical poem length. In both corpora, median poem length lies well below the 2,500-word threshold identified by Eder (Eder 2013) as necessary for stable stylometric attribution, making our datasets a demanding test case for both stylometric and embedding-based methods.

Morphosyntactic annotation was obtained from PoeTree via its API. For the Russian corpus, we additionally incorporate expert-annotated prosodic metadata from the Russian National Corpus (Grishina et al. 2009, including meter (e.g., iambic, trochaic, ternary, dolnik), metrical foot count, clausula type, rhyme scheme, and stanzaic structure). These annotations yield 15 prosodic features, which constitute Tier 4 (Prosody) in our residualization framework and allow direct assessment of how explicitly prosodic information is reflected in embedding representations.

No comparable prosodic annotation is available for Italian poetry. This asymmetry limits direct cross-linguistic comparison of prosodic encoding; accordingly, we treat the Italian corpus primarily as a test of generalizability for analyses that do not rely on explicit prosodic features.

Each poem is embedded as a single text unit, without chunking or truncation. All poems fall well within the context limits of the evaluated models, ensuring that embeddings reflect complete texts rather than aggregated fragments. Embeddings and associated metadata will be released to support replication. We evaluate several contem-

conference version

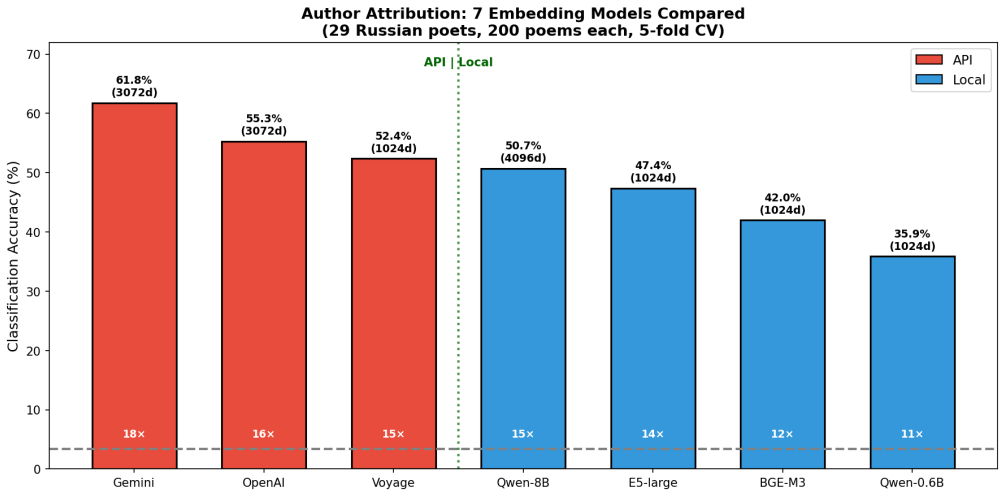


Figure 1: Multiclass authorship attribution accuracy across embedding models (Russian corpus; 29 authors, 200 poems per author; 5-fold cross-validation)

conference version

porary embedding models representative of current large-scale representation learning. These include OpenAI text-embedding-3-large, Google Gemini text-embedding-001, and voyage-3-large, all of which produce 3,072-dimensional embedding vectors. In addition, we tested several open-source embedding models (Qwen-3-8B, Qwen-3-06B, E5-large, and BGE-M3) to assess whether observed patterns generalize beyond proprietary systems.

Across models, attribution performance was broadly consistent. Gemini text-embedding-001 achieved the highest multiclass attribution accuracy on both the Russian and Italian corpora (see Figure 1), while the open-source Qwen-3 embeddings yielded the strongest performance among non-proprietary models. We therefore report results for both representations in the main analysis.

4. Methodology

We employ three complementary analytical strategies: authorship attribution experiments to assess performance, residualization to quantify the contribution of interpretable features to embedding variance, and probing analyses to test which textual properties are linearly accessible in embedding space.

Authorship attribution experiments. We formulate attribution as both a multiclass and a binary classification task. In the multiclass setting, each poem is assigned to one of K authors (K=29 for Russian; K=52 for Italian), reflecting realistic attribution scenarios and yielding a single accuracy metric. Corresponding chance performance is 3.4% for Russian and 1.9% for Italian.

In the binary setting, we train classifiers for all author pairs (406 pairs for Russian; 1,326 for Italian). Pairwise attribution enables finer-grained analysis, allowing us to distinguish failures driven by specific confusable author pairs from more general limitations of the representation. All attribution experiments use logistic regression. We compare three feature configurations: (i) embeddings alone; (ii) classical stylometric features alone (character n-grams of length 2–4, word n-grams of length 1–2, and function-

word frequencies); and (iii) concatenated representations combining embeddings with stylometric features.

To assess which interpretable features account for embedding-based authorship signal, we apply a residualization waterfall that progressively removes feature contributions from embedding representations. At each step, we evaluate how much authorship information remains after subtracting the variance explained by a given feature tier.

Concretely, for each tier of features, we fit a Ridge regression model that predicts embedding vectors from feature values. Residual embeddings are obtained by subtracting the predicted embeddings from the original embeddings. Classification accuracy on these residuals quantifies how much authorship signal survives after removal of that feature tier, while the coefficient of determination (R^2) measures the proportion of embedding variance explained.

Ridge regression provides a principled linear mapping from feature space to embedding space, balancing goodness of fit with regularization of coefficient magnitudes. Given a feature matrix X and an embedding matrix E, we estimate

$$\hat{\beta} = \arg\min_{\beta} \left(\|E - X\beta\|^2 + \lambda \|\beta\|^2 \right) \tag{1}$$

and compute residual embeddings as

$$E_{\text{resid}} = E - X\hat{\beta} \tag{2}$$

While residualization quantifies how much embedding variance is explained by known feature tiers, probing evaluates the inverse relationship: whether specific textual properties are linearly decodable from embedding representations. High probing performance indicates that a given property is directly accessible in embedding space.

We probe embeddings for lexical, punctuation, prosodic, and topical properties using linear models. For each property, we train a Ridge regression model for continuous targets and a logistic regression model for categorical targets, using 5-fold cross-validation. We report R^2 for continuous properties and classification accuracy for categorical properties.

Our residualization framework assumes that stylometric information is encoded in embedding space in a manner that is linearly accessible. To assess whether this assumption omits substantial non-linear structure, we compared linear Ridge regression against Kernel Ridge Regression with an RBF kernel. If stylistic information were primarily encoded in a complex non-linear manifold, the kernel method would be expected to remove substantially more authorship signal during residualization.

In practice, the kernel model exhibited severe overfitting in this high-dimensional setting: while achieving near-perfect fit on the training data, it failed to generalize to held-out data, leaving residual attribution accuracy high (47.76%). By contrast, linear Ridge regression generalized reliably, reducing residual attribution accuracy to near chance levels (4.34%). This outcome is consistent with prior probing studies showing that many linguistic properties in large language models are encoded in approximately linear subspaces (Hewitt and Liang 2019).

conference version

Accordingly, residualization in this study measures the extent to which stylistic information is linearly accessible in embedding space. It does not imply that embeddings lack non-linear stylistic structure, only that such structure does not support generalizable authorship attribution under the present experimental conditions.

To assess whether attribution performance reflects topical rather than stylistic cues, we evaluate models under topic-controlled conditions following Brad et al. (Brad et al. 2022). Topics are induced via LDA clustering, and attribution is tested under three regimes: (i) within-topic, where training and testing are restricted to poems from the same topic cluster; (ii) cross-topic, where models are trained on one topic cluster and evaluated on another; and (iii) leave-one-topic-out, where each topic cluster is held out in turn during training.

Performance degradation from the baseline to cross-topic conditions is interpreted as evidence of topic dependence, whereas residual accuracy under topic shift indicates stylistic signal that generalizes beyond topical content.

All experiments use 5-fold stratified cross-validation with a fixed random seed (42). Statistical significance is assessed via bootstrap confidence intervals.

5. Results

We report attribution results organized around three questions: (i) how embeddings compare to classical stylometry, (ii) whether the two approaches are complementary, and (iii) how text length mediates performance.

5.1 Multiclass attribution

Multiclass authorship attribution is challenging for both corpora due to short text length (median 81 words for Russian, 100 words for Italian). Table 2 reports multiclass attribution accuracy using classical stylometric methods. Despite having nearly twice as many authors, the Italian corpus yields higher attribution accuracy across methods, reflecting the greater stylistic diversity introduced by its longer historical span.

Character-level features dominate performance in both languages. Character n-grams (with or without function words) substantially outperform word-based and function-word-only baselines, while Burrows' Delta variants perform poorly under these short-text conditions. These results align with prior findings that character-level representations are more robust for short and morphologically rich texts.

5.2 Embeddings and complementarity

Table 3 summarizes multiclass attribution accuracy for embeddings, stylometry, and their combination, while Figures 2 and 3 visualize these relationships across embedding models. For Russian, embeddings and stylometry achieve comparable standalone performance (61.3% vs. 59.2%), but their combination yields a substantial improvement (+9.2 percentage points). In contrast, for Italian, classical stylometry clearly outperforms embeddings (+12.6 pp), and combining the two provides only modest additional gains (+2.1 pp).

conference version

conference version

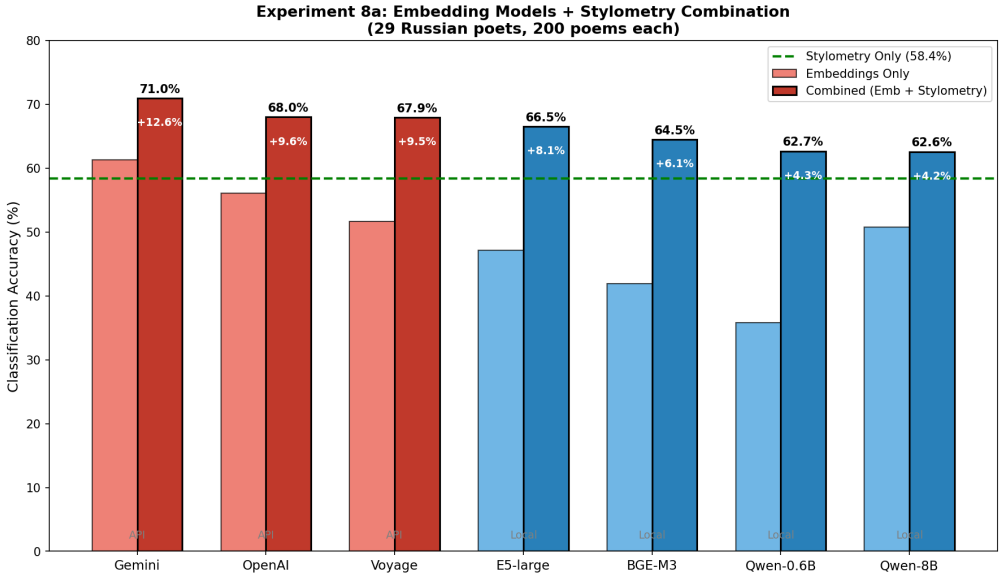


Figure 2: Multiclass authorship attribution on the Russian poetry corpus

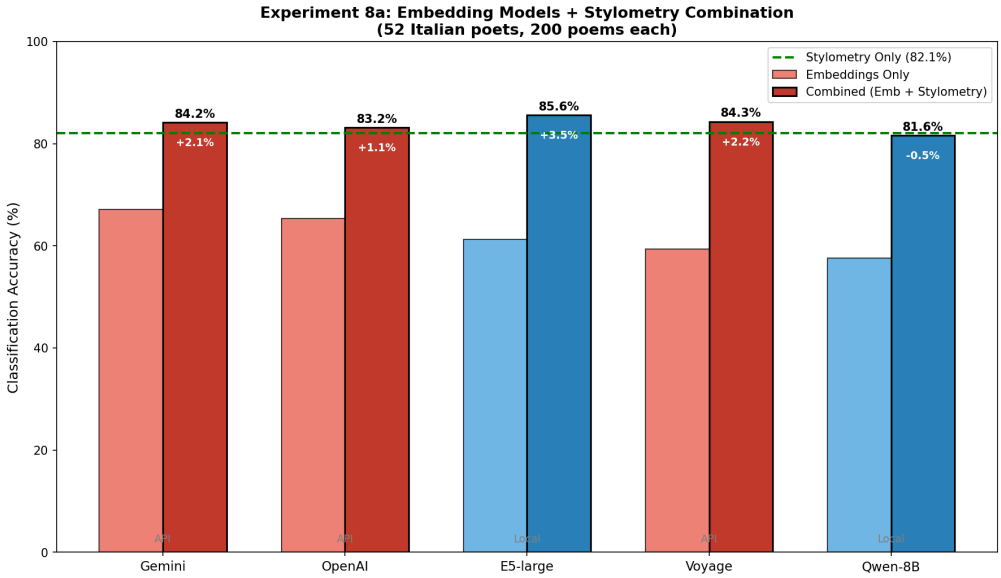


Figure 3: Multiclass authorship attribution on the Italian poetry corpus

Table 2: Multiclass authorship attribution accuracy for classical stylometric methods.

Method	Russian (29 authors)	Lift vs. chance	Italian (52 authors)	Lift vs. chance
Full combined	61.7%	18×	80.9%	42×
Char 3-grams + function words	59.2%	17×	79.8%	42×
Char n-grams (2–4)	58.3%	17×	78.6%	41×
Char 3-grams	57.9%	17×	65.8%	35×
Word n-grams (1–2)	33.8%	10×	55.3%	29×
MFW + Logistic Regression	19.2%	6×	45.5%	24×
Cosine Delta	19.0%	6×	39.4%	21×
Function words	16.2%	5×	39.4%	21×
Burrows' Delta (100 MFW)	13.0%	4×	33.3%	18×
Burrows' Delta (500 MFW)	7.1%	2×	24.1%	13×
Chance	3.4%	1×	1.9%	1×

Table 3: Multiclass authorship attribution accuracy for embeddings, stylometry, and their combination.

Method	Russian (29 authors)	Italian (52 authors)
Embeddings (Gemini)	61.3% ± 1.3%	67.1% ± 0.5%
Stylometry (char 3-grams + function words)	59.2% ± 1.1%	79.7% ± 1.3%
Combined	**70.5% ± 0.9%**	**81.8% ± 1.1%**
Chance	3.4%	1.9%

5.3 Binary attribution

Pairwise attribution reveals a different performance profile. Across all 406 Russian author pairs, stylometry achieves higher mean accuracy than embeddings (93.6% vs. 89.6%). The gap widens for Italian, where stylometry reaches 98.0% accuracy compared to 91.4% for embeddings across 1,326 pairs.

The contrast between near-parity in multiclass attribution and stylometry's advantage in binary settings suggests that embeddings are relatively better at handling many-way confusion, while stylometric features excel at fine-grained pairwise discrimination.

5.4 Effect of text length

Text length strongly constrains attribution accuracy for both stylometry and embedding-based methods. Single-poem attribution performance (approximately 61% for Russian and 67–79% for Italian) reflects a lower bound imposed by limited textual evidence rather than a ceiling of methodological capability. Table 4 illustrates the effect of text length on multiclass attribution accuracy in the Russian corpus. Concatenating as few as five poems (≈500 words) raises accuracy above 95% for both stylometry and embeddings, with near-perfect attribution achieved beyond 2,000 words.

As shown in Figure 4, error rates decrease monotonically with increasing word count in binary classification. With approximately 500 words, attribution accuracy exceeds 95% for both stylometry and embeddings; at 2,000 words — approaching the threshold identified by Eder (Eder 2013) — both methods achieve near-perfect performance. These

conference version

Table 4: Effect of text length on multiclass authorship attribution accuracy (Russian corpus).

Concatenation	Words	Gemini Embeddings	Stylometry
1 poem	100	61.3%	59.2%
5 poems	502	95.3%	96.2%
10 poems	1,004	99.1%	99.5%
20 poems	2,008	99.7%	100.0%

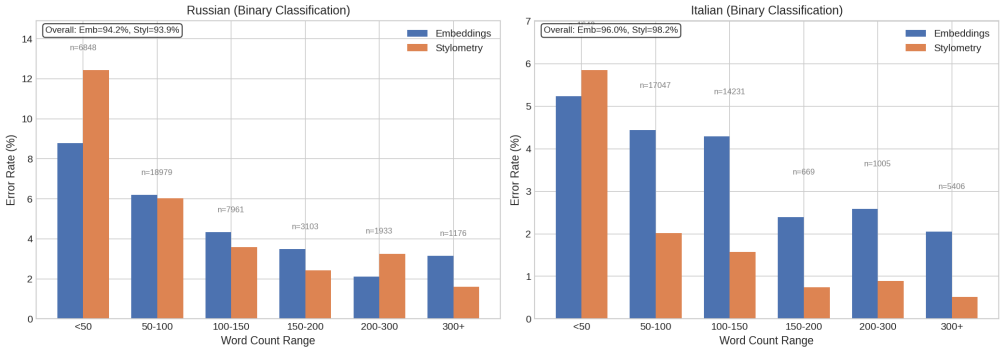


Figure 4: Binary attribution error rates as a function of text length

results demonstrate that text length, rather than representational capacity, is the primary limiting factor in single-poem authorship attribution.

6. Embeddings Deconstruction: Residualization Waterfall

We organize features into six tiers of increasing complexity and interpretability:

- Tier 1: Surface and Rhythm (20 features). Simple orthographic markers including punctuation density, line length variance, and Type-Token Ratio (TTR).
- Tier 2: Coarse-Grained Content (40 features). To distinguish thematic focus from stylistic habit, we utilized the top 20 TF-IDF terms and a 20-topic LDA model. This tier captures broad thematic clusters (e.g., "Religious," "Nature").
- Tier 3: Morphosyntax (100 features) – Captures grammatical fingerprints using distributions of Universal Dependency tags (17), morphological features (Case, Gender, Tense), and syntactic dependency types, extracted using PoeTree.
- Tier 4: Prosody (15 features, Russian Only) – Expert-annotated prosodic features from the Russian National Corpus (e.g., meter type, foot count, clausula structure).
- Tier 5: Character n-grams (top 2,000 features) – TF-IDF weighted character n-grams (2-4 characters).
- Tier 6: Word bigrams (top 2,000 features) – TF-IDF weighted word pairs.

After observing substantial residual signal in the four-tier waterfall, we added two additional "stylometry" tiers to capture higher-order features. These tiers are less interpretable than linguistic and prosodic features but allow us to test whether the residual signal represents genuinely novel encoding or merely features our taxonomy missed.

conference version

conference version

Table 5: Residualization waterfall summary for multiclass authorship attribution accuracy.

Metric	RU Gemini	RU Qwen3-8B	IT Gemini	IT Qwen3-8B
Number of authors (K)	29	29	52	52
Chance accuracy	3.4%	3.4%	1.9%	1.9%
Baseline embeddings	61.3%	50.7%	67.0%	57.6%
After interpretable features (T1–T4)	40.0%	32.2%	46.0%	41.1%
After character n-grams (T5)	10.2%	8.3%	16.1%	16.8%
After word bigrams (T6)	3.9%	4.0%	8.7%	9.5%
Final residual (kernel control)	3.7%	3.9%	8.9%	9.4%
Final lift vs. chance	1.1×	1.1×	4.6×	4.9×
Total R^2 explained	80.1%	79.9%	85.1%	74.4%

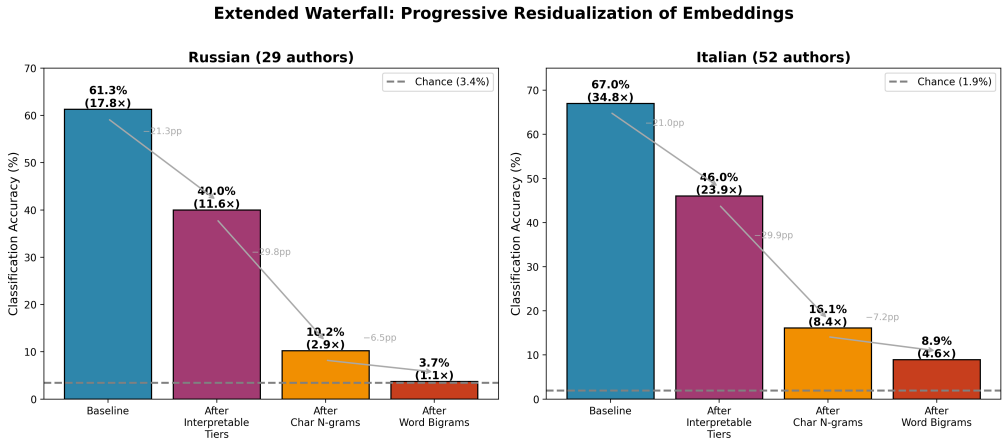


Figure 5: Extended Waterfall: Progressive Residualization of Embeddings

To prevent data leakage, all components — TF-IDF vectorizer, LDA model, Ridge residualizer, and classifier — are trained exclusively on the training data within each fold of cross-validation. Test-fold features are transformed using training models. All residualization results below in Table 5 use linear Ridge regression, which we found to generalize better than kernel methods (see Methodology).

6.1 Progressive feature removal

Progressive removal of interpretable features reveals how much variance in embedding representations is explained by known features. After removing all four tiers (surface, content, grammar, prosody), accuracy drops from 61.3% to 40.0% for Russian and from 67.0% to 46.0% for Italian, still substantially above chance (11.6× and 24.2× chance, respectively). However, these 175 features explain only 18.5% of embedding variance.

Adding higher-dimensional n-gram features (character n-grams and word bigrams) brings the residual accuracy closer to chance for Russian (≈1.1×), but not for Italian, where the residual remains significantly above chance (≈4.6–4.9×, see Figure 5), despite explaining more than 80% of the variance.

For Russian, the final residual accuracy (3.7%) is statistically indistinguishable from chance. A permutation test with shuffled author labels (n = 10) yields a null distribution with a mean of 3.5% and 95% CI [3.2%, 3.9%]; the observed residual falls within

this interval ($p = 1.0$), confirming that no discriminative authorial signal remains after removing all six tiers. The apparent residual observed after removing only the interpretable features is thus an artifact of excluding character-level distributions.

For Italian, the final residual accuracy (8.9%) remains significantly above chance. Under the same permutation protocol, the null distribution has a mean of 1.9% and 95% CI [1.8%, 2.1%], and the observed residual exceeds this range ($p < 0.001$). Therefore, while extended stylometric features account for most of the embedding variance, a non-trivial author-discriminative signal persists. This asymmetry suggests that the explanatory sufficiency of extended stylometric features is corpus-dependent: for synchronous poetry corpora, stylometric features fully explain embedding-based attribution, while for diachronic corpora like Italian, embeddings retain additional signal beyond the feature space considered here. Identifying historical or dialectal features that could close this gap remains future work.

The waterfall analysis initially suggested that embeddings capture deeper authorial patterns beyond interpretable linguistics — e.g., the 11.6× chance residual after removing surface, content, grammar, and prosody features. However, adding character n-grams and word bigrams — features previously excluded as "uninterpretable" — collapsed the residual for Russian to 1.1× chance. The "mysterious" residual signal was due to character-level orthographic features that our original feature taxonomy failed to capture.

This finding carries a methodological lesson: claims about unexplained neural encoding should be tested against a full stylometric feature inventory, not just linguistically motivated subsets. What appears novel may reflect gaps in feature engineering rather than genuinely emergent representation.

The asymmetry in residual accuracy between Russian and Italian complicates this conclusion. For Italian, the residual, while reduced by 80% through extended features, remains significant at 4.6× chance. This suggests that embeddings capture stylistic aspects of Italian poetry that neither traditional features nor character n-grams fully recover. Three potential factors contribute to this residual:

- Orthographic variation: The 700-year temporal span of the Italian corpus introduces spelling conventions that character n-grams may not fully capture.
- Dialectal variation: Italian poets like Belli and Meli use dialect markers, which may encode identity through phonological patterns beyond our features.
- Prosody: Without prosodic annotation, we cannot assess whether meter and rhyme schemes explain additional variance.

This residual reflects a genuine limitation of our analysis: embeddings encode authorial variation that traditional features do not fully capture.

To understand why character n-grams are so discriminative, we analyze logistic regression coefficients across 406 binary author pairs. For each pair, we identify the n-grams with the highest absolute coefficients and categorize them by punctuation, morphological endings, or lexical fragments. Stable n-grams, which appear in the majority of comparisons for a given author, constitute a "fingerprint." For example, Tsvetaeva's

conference version

Table 6: Continuous property probes using Ridge regression (5-fold cross-validation).

Property	R^2	Encoding strength
Vocabulary diversity (500–2000 MFW)	0.77	Strong
Vocabulary diversity (100–500 MFW)	0.68	Strong
Vocabulary diversity (1–100 MFW)	0.63	Strong
Exclamation rate	0.53	Moderate
Ellipsis rate	0.44	Moderate
Em-dash rate	0.30	Moderate
Hyphen rate	−0.37	Not encoded
Colon rate	< 0	Not encoded

most discriminative feature is the em-dash, appearing in 28 of 28 comparisons (mean coefficient = 2.64). These orthographic choices encode deep stylistic signatures, such as Tsvetaeva's fragmented intensity, Blok's flowing traditionalism, and Baltrušaitis's use of ellipses.

The sufficiency of linear residualization has important theoretical implications. By stripping embeddings of their predictive power using a simple linear transformation, we effectively collapse Russian authorship attribution to chance (1.1×). This implies that LLMs perform a linear projection of traditional stylometric features. While the Transformer architecture's complexity is required for text generation, the representation of style it constructs remains surprisingly Euclidean.

7. What Do Embeddings Encode? (Probing Analysis)

Probing classifiers test which textual properties are linearly accessible in embedding space. We train simple models — Ridge regression for continuous properties and logistic regression for categorical properties — to predict linguistic attributes directly from embedding vectors. High performance indicates that a property occupies a linear subspace of the embedding representation; low performance suggests that it is weakly encoded or only accessible non-linearly.

Lexical and surface properties. We first probe continuous surface-level properties using Ridge regression (Table 6). Lexical diversity measures are strongly encoded: vocabulary diversity across frequency bands yields R^2 values between 0.63 and 0.77. These results indicate that embeddings function primarily as vocabulary encoders, with strong sensitivity to lexical distribution.

Punctuation encoding is selective. Em-dashes ($R^2 = 0.30$) and ellipses ($R^2 = 0.44$) are linearly accessible, while hyphens and colons show no recoverable signal ($R^2 < 0$). This selectivity has practical consequences: Tsvetaeva's distinctive em-dash usage is encoded in embedding space, whereas Blok's preference for standard hyphenation is not. The asymmetry may reflect differences in frequency or semantic markedness in training data, but we lack direct evidence to confirm this interpretation.

Semantic vs Prosodic Properties. Binary classification probes reveal a clear hierarchy between semantic and prosodic information (Table 7). Topic classification achieves 81.3% accuracy, substantially outperforming all prosodic probes. Rhyme, meter, and

conference version

Table 7: Classification probes for semantic and prosodic properties using logistic regression (5-fold cross-validation).

Property	Type	Accuracy	Chance	Lift
Topic (Inner vs. Nature)	Semantic	81.3%	50%	1.6×
Rhyme (ABAB vs. AABB)	Prosodic	70.1%	50%	1.4×
Meter (Iamb vs. Trochee)	Prosodic	67.5%	50%	1.4×
Feet (4 vs. 5)	Prosodic	66.0%	50%	1.3×
Clausula (regular vs. free)	Prosodic	58.6%	50%	1.2×
Meter (5-class)	Prosodic	32.8%	20%	1.6×

Table 8: Authorship attribution accuracy by vocabulary frequency band.

Frequency band	Word type	Accuracy
1–100	Function words	76.4%
100–500	Style markers	76.6%
500–2000	Topic vocabulary	**77.6%**
2000–5000	Rare vocabulary	74.0%

conference version

foot count yield only moderate lift over chance, and fine-grained meter classification (five classes) reaches just 32.8% accuracy (1.6× chance).

These results indicate that embeddings encode what poems are about more readily than how they structure verse. This hierarchy aligns with our ablation analysis: masking content words alters embeddings nearly five times more than shuffling word order (cosine distance 0.134 vs. 0.027), confirming that semantic content dominates embedding geometry.

Poetry Differs from Prose. Rybicki and Eder (Rybicki and Eder 2011) found that for prose, function words (top 100 most frequent) provide the most stable authorship signal, while topic vocabulary (ranks 500–2000) introduces noise. We tested whether this hierarchy holds for poetry by restricting stylometric features to specific frequency bands (Table 8).

The pattern inverts. Topic-associated vocabulary (500–2000 MFW) yields the highest attribution accuracy, outperforming function words and rare vocabulary. This inversion is consistent with our probing results: if embeddings encode topic strongly and poets exhibit stable thematic preferences, topic vocabulary naturally becomes author-discriminative. In poetry, authors are distinguished more by what they write about than by sentence-level grammatical habits — the opposite of prose attribution.

While two poetic corpora cannot establish a universal poetry–prose distinction, the consistent inversion across languages and literary traditions suggests that this effect is not corpus-specific and warrants broader investigation.

Topic Confound Analysis. To assess whether embedding-based attribution relies on topical shortcuts, we evaluate performance under topic-controlled conditions following Brad et al. 2022 (Table 9). Cross-topic attribution accuracy drops by 25.6 percentage points relative to baseline, confirming substantial topic dependence. However, accuracy remains 10.3× above chance, indicating that a significant portion of authorial signal

Table 9: Cross-topic validation results for authorship attribution (Russian corpus).

Condition	Accuracy	Δ from baseline
Baseline (random CV)	61.3%	—
Leave-one-topic-out	55.8%	−5.5 pp
Within-topic	55.5%	−5.8 pp
Cross-topic (train A, test B)	35.7%	−25.6 pp

generalizes across thematic boundaries.

Approximately 40% of discriminative capacity is topic-dependent, while 60% reflects topic-independent stylistic patterns. This decomposition clarifies the relationship between embeddings and stylometry: embeddings encode topic strongly, and topic correlates with authorship, but they also capture persistent author-specific patterns — vocabulary preferences, punctuation habits, and other stylistic markers — that survive topic shifts.

These results do not challenge the known role of topic in authorship attribution; rather, they quantify its contribution and show that embedding representations retain non-topical authorial signal that persists across thematic shifts.

8. How Do Methods Differ? (Complementarity Analysis)

We explore how embeddings and stylometry contribute uniquely to authorship attribution. The combined performance gains (+9 pp for Russian, +2 pp for Italian) suggest that the two methods capture different aspects of authorial style. We investigate this complementarity through bidirectional residualization, error analysis, and case studies.

8.1 Bidirectional Residualization

To quantify the unique contributions of embeddings and stylometry, we regress each representation from the other and evaluate authorship attribution on the resulting residuals (Table 10). Residual accuracy directly measures how much author-discriminative signal remains after removing the information predictable from the other representation.

In the Russian corpus, embeddings consistently retain more unique signal after removing stylometric features (10.5–13.1% accuracy; 3.0–3.8× chance) than stylometry retains after removing embedding-predictable information (4.4–7.6%; 1.3–2.2× chance). This asymmetry indicates that embeddings capture additional authorial signal beyond character n-grams and function-word distributions.

In contrast, the Italian corpus shows a more balanced pattern: residual accuracy is similar in both directions (13.8–19.2%), suggesting that embeddings and stylometry capture comparably sized but partially distinct components of authorial signal. This balance is consistent with the greater stylistic, orthographic, and semantic diversity introduced by seven centuries of Italian poetic production.

conference version

Table 10: Bidirectional residualization: residual authorship attribution accuracy after removing the other representation.

Language	Embedding model	After removing stylometry	After removing embeddings
RU	OpenAI	13.1% (3.8×)	6.0% (1.7×)
RU	Gemini	12.5% (3.6×)	7.6% (2.2×)
RU	Qwen3-8B	10.5% (3.0×)	4.4% (1.3×)
IT	Gemini	19.1% (9.9×)	19.2% (10.0×)
IT	Qwen3-8B	18.6% (9.7×)	13.8% (7.2×)

Table 11: Error rates by text length bin for embedding-based and stylometric authorship attribution.

Length bin	N poems	Embedding error	Stylometry error	Δ
Very short (≤30 words)	252	41.7%	43.0%	−1.3 pp
Short (31–50 words)	822	38.5%	41.2%	−2.7 pp
Medium (51–80 words)	1804	37.7%	39.6%	−1.9 pp
Long (81–120 words)	1,590	35.2%	38.2%	−3.0 pp
Very long (>200 words)	410	36.4%	35.3%	+1.1 pp

conference version

8.2 Error analysis

Embeddings and stylometry fail on different texts, with only 31% of errors overlapping (Jaccard similarity). Stylometry performs poorly on short, vocabulary-rich poems (correlation with TTR: $r = -0.14$), while embeddings show more uniform error rates across text lengths (Table 11).

For short poems, embeddings have a slight advantage; for long poems, stylometry performs slightly better. This crossover indicates that embeddings efficiently capture semantic content, even in short texts, while stylometry requires more material to establish stable frequency patterns.

To make the complementarity between embeddings and stylometry concrete, we briefly discuss two representative examples in which one method succeeds while the other fails.

Embedding-distinctive example (Akhmatova; 100% embeddings / 7% stylometry).

> Не знаю, где ты и где я.
> Те ж песни и те же заботы.
> Такие с тобою друзья!
> Такие с тобою сироты!
>
> И так хорошо нам вдвоем:
> Бездомным, бессонным и сирым...
> Две птицы: чуть встали — поём.
> Две странницы: кормимся миром.

Embeddings correctly attribute this poem to Akhmatova by capturing its semantic signature: motifs of homelessness, wandering, intimacy, and the recurrent figure of "two". These thematic patterns align closely with Akhmatova's broader poetic persona

and are strongly represented in embedding space. Stylometric features, by contrast, offer little discriminative power here: word length, punctuation, and function-word usage are typical of Silver Age poetry and do not distinguish the author reliably.

Stylometry-distinctive example (Nadson, 0% emb / 96% stylo):

> Ночевал ушкуйник в заозерье,
> Натыкал на остры стрелы перья.
>
> Полонянка на песке лежала,
> Под жгутами язвы ныли ало...

Stylometric attribution succeeds here due to highly distinctive surface features. Archaic lexical items (ушкуйник, полонянка) generate characteristic character n-gram distributions, and the rhyme patterns (-нье / -рья, -ала / -ало) further reinforce author-specific orthographic and phonological regularities. These features are rare across the corpus and therefore strongly discriminative.

Embedding-based attribution, by contrast, fails in this case. The poem's semantic content — a medieval / historical narrative centered on brigands and captivity — overlaps with themes frequently employed by other poets. As a result, semantic similarity dominates the embedding representation, obscuring the author-specific surface patterns that stylometry exploits effectively.

The complementarity between embeddings and stylometry is genuine rather than redundant. Bidirectional residualization shows that each method retains unique authorial signal after accounting for the other: embeddings preserve 12.5% accuracy (3.6× chance) beyond stylometry, while stylometric features retain 7.6% (2.2× chance) beyond embeddings. Error analysis corroborates this asymmetry, with only 31% overlap in misclassifications and distinct failure modes — stylometry struggles with short, vocabulary-rich poems, whereas embeddings exhibit more uniform error rates across text lengths.

The practical implications are clear. Combining both approaches yields consistent gains in attribution accuracy (+9 percentage points for Russian, +2 for Italian). For applications prioritizing predictive performance, ensemble methods are therefore advisable. For interpretive contexts — such as forensic attribution or literary analysis — stylometric features remain essential due to their transparency, while embedding-based methods help identify semantic forms of distinctiveness that stylometry alone may miss.

9. Limitations

Three limitations warrant acknowledgment. First, the Italian corpus lacks prosodic annotation, preventing a fully symmetric cross-linguistic evaluation of our feature framework. Evidence from Russian suggests that prosodic features contribute relatively little to attribution accuracy, making this an unlikely explanation for the residual signal observed in Italian; however, this cannot be directly tested within the present data.

Second, our corpora represent specific literary traditions: Russian poetry (approximately 80 years, with a shared cultural and historical milieu) and Italian poetry spanning seven centuries. While this contrast is methodologically informative, our findings may not

conference version

generalize to prose, to other languages, or to contemporary literary production.

Third, and most fundamentally, character n-grams identify discriminative patterns without explaining them. Knowing that Tsvetaeva favors em-dashes reveals what distinguishes her computationally, not why she adopted this stylistic habit or what expressive work it performs. The gap between statistical identification and literary interpretation therefore remains open.

For widely canonical authors (e.g., Dante, Mandelstam), prior exposure during model pretraining is effectively guaranteed. As a result, authorship attribution in such cases should not be interpreted as a test of generalization to unseen authors. This limitation does not undermine our central objective, however, which is not to assess whether models can recognize authors, but to analyze how authorial signal is represented in embedding space.

If attribution performance were driven primarily by direct memorization or implicit author identifiers, residualization against surface, lexical, and character-level features would not be expected to systematically reduce performance. The observed collapse of attribution accuracy after feature removal — particularly for the Russian corpus — therefore supports the conclusion that embeddings encode distributed textual regularities rather than relying on direct recall of author labels.

10. Conclusion

This study examines how LLM embeddings capture authorial style compared to classical stylometry and identifies the interpretable features that drive their discriminative power. The analysis leads to three primary conclusions.

First, embeddings and stylometry encode overlapping but distinct signals. For Russian poetry, both methods achieve comparable accuracy on single poems (61% vs. 59%), but their combination provides substantial gains (+9 percentage points), particularly for shorter texts. For Italian, stylometry clearly outperforms embeddings (+12.6 pp), though combination still improves performance (+2.1 pp). The methods fail on different texts, with only 31% overlap in misclassifications — confirming complementarity, not redundancy.

Second, what initially seemed to be a 'deep residual' signal was largely an artifact. After removing interpretable linguistic features, embeddings retained 11.6× chance accuracy for Russian, indicating that what appeared as novel encoding was largely explained by orthographic and lexical patterns captured by classical stylometry. Character-level patterns — punctuation preferences, morphological endings, orthographic habits — explain 44% of embedding variance, more than all linguistic features combined. The finding partially replicates for Italian, where the residual drops 80%, but remains significant (4.6× chance), likely reflecting dialect markers and orthographic variation our features incompletely represent.

Third, embeddings primarily encode vocabulary and semantic content. Probing reveals strong linear encoding of lexical diversity ($R^2 = 0.63$–0.77) and topic (81% accuracy), while prosodic features, such as meter and rhyme, remain less accessible (58–70%).

conference version

Punctuation encoding is selective: em-dashes and ellipses are captured; hyphens and colons are not. For literary applications, this implies that embeddings will capture thematic signatures more reliably than formal structure.

These findings suggest a division of labor. For applications prioritizing accuracy — large-scale attribution, corpus organization — combined approaches are advisable. For applications requiring interpretability — forensic analysis, pedagogical feedback, close reading — stylometric features remain essential due to their transparency, while embedding-based methods help identify semantic forms of distinctiveness that stylometry alone may miss. The patterns of misclassification confirm this division: confusions align with literary-historical distinctions (school, era, influence), suggesting the methods are working with meaningful signals, rather than arbitrary statistical noise.

What remains open is connecting statistical patterns with literary meaning. Identifying Tsvetaeva's em-dash signature or Baltrušaitis's ellipses highlights computational distinctions, but understanding why these poets adopted these styles and their expressive significance requires interpretive scholarship that computational methods can inform, but not replace. Future work should focus on bridging this gap, exploring how these computational signals can be mapped to deeper literary interpretation.

conference version

11. Data Availability

Data can be found here: https://doi.org/10.5281/zenodo.18260458.

12. Software Availability

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

13. Author Contributions

Maria Levchenko: Conceptualization, Data curation, Methodology, Software, Writing – original draft, Writing – review & editing

References

Ahia, Orevaoghene, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov (Dec. 2023). "Do All Languages Cost the Same? Tokenization in the Era of Commercial 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, 9904–9923. 10.18653/v1/2023.emnlp-main.614. https://aclanthology.org/2023.emnlp-main.614/.

Belinkov, Yonatan and James Glass (2019). "Analysis Methods in Neural Language Processing: A Survey". In: *Transactions of the Association for Computational Linguistics* 7. Ed. by Lillian Lee, Mark Johnson, Brian Roark, and Ani Nenkova, 49–72. 10.1162/tacl_a_00254. https://aclanthology.org/Q19-1004/.

Brad, Florin, Andrei Manolache, Elena Burceanu, Antonio Barbalau, Radu Ionescu, and Marius Popescu (2022). *Rethinking the Authorship Verification Experimental Setups*. arXiv: 2112.05125 [cs.CL]. https://arxiv.org/abs/2112.05125.

Burrows, John (Sept. 2002). "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship". In: *Literary and Linguistic Computing* 17.3, 267–287. ISSN: 0268-1145. 10.1093/llc/17.3.267. eprint: https://academic.oup.com/dsh/article-pdf/17/3/267/2743069/170267.pdf. https://doi.org/10.1093/llc/17.3.267.

Eder, Maciej (Nov. 2013). "Does size matter? Authorship attribution, small samples, big problem". In: *Digital Scholarship in the Humanities* 30.2, 167–182. ISSN: 2055-7671. 10.1093/llc/fqt066. eprint: https://academic.oup.com/dsh/article-pdf/30/2/167/21517531/fqt066.pdf. https://doi.org/10.1093/llc/fqt066.

Eder, Maciej, Jan Rybicki, and Mike Kestemont (Aug. 2016). "Stylometry with R: A Package for Computational Text Analysis". In: *The R Journal* 8.1. Open access. License: CC BY 3.0 Unported, 107–121. ISSN: 2073-4859.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt (June 2017). "Understanding and explaining Delta measures for authorship attribution". In: *Digital Scholarship in the Humanities* 32.suppl$_2$, ii4–ii16. ISSN: 2055-7671. 10.1093/llc/fqx023. eprint: https://academic.oup.com/dsh/article-pdf/32/suppl_2/ii4/21298943/fqx023.pdf. https://doi.org/10.1093/llc/fqx023.

Glaser, Ben (2025). "TrochAIc: Metrical Tools for AI Interpretability". In: *Anthology of Computers and the Humanities* 3. Ed. by Margherita Fantoli Taylor Arnold and Ruben Ros, 1438–1453. 10.63744/K9Mwiiszu7QL.

Grishina, Elena, Kirill Korchagin, Vladimir Plungian, and Dmitry Sitchinava (2009). "Poeticheskii korpus v ramkakh NKRJa: obshchaya struktura i perspektivy ispol'zovaniya". In: *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy*. Saint Petersburg: Nestor-Istoriya, 71–113.

Hewitt, John and Percy Liang (Nov. 2019). "Designing and Interpreting Probes with Control Tasks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, 2733–2743. 10.18653/v1/D19-1275. https://aclanthology.org/D19-1275/.

Hewitt, John and Christopher D. Manning (June 2019). "A Structural Probe for Finding Syntax in Word Representations". 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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, 4129–4138. 10.18653/v1/N19-1419. https://aclanthology.org/N19-1419/.

Jannidis, Fotis, Rüdiger Kleymann, Jens Schröter, and Heike Zinsmeister (2025). "Do Large Language Models Understand Literature? Case Studies and Probing Experiments on German Poetry". In: *Journal of Computational Literary Studies* 4.1. 10.48694/jcls.4225.

Kestemont, Mike, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast (2018). "Overview of the Author Identification Task at PAN-2018: Cross-Domain Authorship Attribution and Style Change Detection". In: *Working Notes Papers of the CLEF 2018 Evaluation Labs*. Ed. by

conference version

Linda Cappellato et al. Vol. 2125. CEUR Workshop Proceedings. Avignon, France, 1–25. https://ceur-ws.org/Vol-2125/invited_paper_2.pdf.

Kim, Yekyung, Jenna Russell, Marzena Karpinska, and Mohit Iyyer (2025). *One ruler to measure them all: Benchmarking multilingual long-context language models*. arXiv: 2503.01996 [cs.CL]. https://arxiv.org/abs/2503.01996.

Lundin, Jessica M., Ada Zhang, Nihal Karim, Hamza Louzan, Victor Wei, David Adelani, and Cody Carroll (2025). *The Token Tax: Systematic Bias in Multilingual Tokenization*. arXiv: 2509.05486 [cs.CL]. https://arxiv.org/abs/2509.05486.

Neal, Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard (Nov. 2017). "Surveying Stylometry Techniques and Applications". In: *ACM Comput. Surv.* 50.6. ISSN: 0360-0300. 10.1145/3132039. https://doi.org/10.1145/3132039.

Plecháč, Petr, Silvie Cinková, Radek Kolář, Artūrs Šeļa, Michele De Sisto, Laurence Nugues, Thomas Haider, and Nejc Kočnik (2024). "PoeTree: Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian and Spanish". In: *Research Data Journal for the Humanities and Social Sciences*. Online ahead of print. 10.1163/24523666-bja10044. https://doi.org/10.1163/24523666-bja10044.

Plecháč, Petr, Radek Kolář, Silvie Cinková, Artūrs Šeļa, Michele De Sisto, Laurence Nugues, Thomas Haider, Nejc Kočnik, Bálint Nagy, Étienne Delente, Romain Renault, Klemens Bobenhausen, Benjamin Hammerich, Andreas Mittmann, Gábor Palkó, Péter Horváth, Borja Navarro Colorado, Pablo Ruiz Fabo, Helena Bermúdez Sabel, Kirill Korchagin, Vladimir Plungian, and Dmitry Sitchinava (2023). *PoeTree: Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian, and Spanish*. Version 0.0.2. Zenodo. 10.5281/zenodo.10008458. https://doi.org/10.5281/zenodo.10008458.

Rivera-Soto, Rafael A., Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews (Nov. 2021). "Learning Universal Authorship Representations". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 913–919. 10.18653/v1/2021.emnlp-main.70. https://aclanthology.org/2021.emnlp-main.70/.

Rybicki, Jan and Maciej Eder (July 2011). "Deeper Delta across genres and languages: do we really need the most frequent words?" In: *Literary and Linguistic Computing* 26.3, 315–321. ISSN: 0268-1145. 10.1093/llc/fqr031. eprint: https://academic.oup.com/dsh/article-pdf/26/3/315/3955977/fqr031.pdf. https://doi.org/10.1093/llc/fqr031.

Stamatatos, Efstathios (Jan. 2013). "On the robustness of authorship attribution based on character n-gram features". In: *Journal of Law and Policy* 21, 421–439.

Wegmann, Anna, Marijn Schraagen, and Dong Nguyen (May 2022). "Same Author or Just Same Topic? Towards Content-Independent Style Representations". In: *Proceedings of the 7th Workshop on Representation Learning for NLP*. Ed. by Spandana Gella, He He, Bodhisattwa Prasad Majumder, Burcu Can, Eleonora Giunchiglia, Samuel Cahyawijaya, Sewon Min, Maximilian Mozes, Xiang Lorraine Li, Isabelle Augenstein, Anna Rogers, Kyunghyun Cho, Edward Grefenstette, Laura Rimell, and Chris Dyer.

conference version

Dublin, Ireland: Association for Computational Linguistics, 249–268. 10.18653/v1/2022.repl4nlp-1.26. https://aclanthology.org/2022.repl4nlp-1.26/.

conference version

jcls. Journal of Computational Literary Studies

conference version

Modeling and Reasoning over Observations:
An Ontology for Literary Criticism

Emilio M. Sanfilippo[1]
Claudio Masolo[1]
Alessandro Mosca[1]
Gaia Tomazzoli[2]

1. Institute of Cognitive Sciences and Technologies, Laboratory for Applied Ontology, National Research Council of Italy, Trento, Italy.
2. Dipartimento di Studi Europei, Americani e Interculturali, Sapienza University of Rome, Rome, Italy.

OPEN ACCESS

Citation
Emilio M. Sanfilippo, Claudio Masolo, Alessandro Mosca, and Gaia Tomazzoli (2026). "Modeling and Reasoning over Observations. An Ontology for Literary Criticism". In: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7993

Date published 2026-05-05 (preprint)

Date accepted tbc.

Date received 2026-01-07

Keywords
Semantic Web, ontology, literature, literary criticism

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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. In the scope of the Digital Humanities, it is useful to represent and reason about data that make explicit *observations* (*claims*) on the subjects being studied. This is particularly interesting in *criticism*, where scholars debate different topics, sometimes grounding their claims on what others have said, some other times rejecting and even contradicting them. In this paper we present an *ontology*, in the form of a Semantic Web model, to support the documentation, analysis, and management of *observational data*. Our ontology focuses on literary criticism, although its design follows a modular architecture, enabling its use in other application contexts. We discuss throughout the paper the motivations for our research, as well as its potential and limitations.

1. Introduction

In the Digital Humanities (DH) research landscape various parties call for the formal treatment of what one might call *observations* (*claims*) on domain entities (Di Pasquale et al. 2024; Pianzola et al. 2025; Sartini et al. 2023). For example, in the scope of archaeological research, experts may need to model observations on the production date of the artifacts found in an excavation site; or, in musicology, scholars may express observations relative to the analysis of scores for recurring patterns and other features of their interest (Sanfilippo et al. 2025a). In this paper, we consider the observations that scholars propose with the purpose of *interpreting* literary texts. Specifically, our interest lies in *ontologies* (in the computer science sense, see Guarino et al. 2009) to represent *observational data* and (logically) reason over them (Sanfilippo and Ferrario 2024).

From a DH standpoint, the objective of this research is twofold. On the one hand, the development of a computational ontology for observations can contribute to better understand how scholars formulate their claims (e.g., which arguments, data, methodologies, etc. they use). This aligns with similar efforts in other fields. For example, in archaeology, according to Gardin, the adoption of formal methods was not meant to introduce information technologies in the discipline, but "to gain a better control of archaeological reasoning per se [...]" (Gardin, quoted in Dallas 2016). On the other hand, our research can lead to the creation of datasets for literary studies and applications for their perusal. In particular, by modeling observations in formal terms, the presence of

alternative and potentially conflicting perspectives can be documented and analyzed, as is the case with interpretations of the same text by scholars from different traditions and cultures. This is advocated as an important strategy to develop computational systems that can make sense of plural views on the same phenomena, reflecting debates and controversies in research (Sandri et al. 2023; Tomasi et al. 2021).

Our work is based on previous research by Masolo et al. 2025, where the authors present an observational ontology in first-order logic. In this paper, we provide, instead, a formal representation in Semantic Web languages, which are the W3C standards for computational ontologies. In particular, the ontology is specified in the Web Ontology Language (OWL)[1] and includes SWRL rules[2] to enhance reasoning.[3] By adopting Semantic Web languages, despite their formal limits in comparison with first-order logic, we aim – in the long term – to exploit technologies to organize and store data in open access repositories compliant with the FAIR principles for data management (Wilkinson et al. 2016). In addition, Semantic Web languages are useful to automatically reason over the data through formal axioms, which is required in our case to automatically detect relations between multiple sources or observations, as we will see. Last but not least, the use of these languages opens the way to the integration with subsymbolic computational approaches, including Large Language Models (LLMs), thus paving the way to potentially adapting multiple techniques to analyze and compare data and texts.

Before introducing the ontology, some words are necessary to clarify our approach. First, observations represent how domain entities are classified through processes such as formal analysis, empirical measurement, and scholarly argumentation, among others. Consequently, observations are not necessarily veridical.[4] Second, observations are the *results* of the processes rather than the processes themselves. For example, the observation ascribing the authorship of a manuscript to a certain person may be provided by different scholars, who may even adopt different input data for their studies. Third, for observations to be uniformly modeled and analyzed, they must be expressed using controlled vocabularies shared by the collaborators on specific tasks. We call these *observational vocabularies*. An observational vocabulary consists of a finite set of classes that are taxonomically organized at different levels of generality. Each class groups specific observations that correspond to the attribution of a property to one or more individuals.[5] For example, the class `ProductionDateObs` (the suffix 'Obs' stands for observation) can be used to represent the individual observations attributing to the observed item its likely production date. Depending on the axiomatic constraints between the classes in the taxonomy, one can infer how observations on the same entity relate. In particular, the ontology is able to represent and reason over overtly incompatible observations without leading to logical inconsistency.

Finally, a consideration with respect to the specific scope of literary studies is due. As we said, we attempt to formally treat the claims that scholars make in their investigations. This does not mean that our ontology aims at the same expressivity of scholarly discourse, or that we aim at "translating" criticism in formal symbolic terms; quite the contrary.

1. https://www.w3.org/TR/owl2-overview/
2. https://www.w3.org/submissions/SWRL/
3. A beta version of our ontology in OWL is presented by Sanfilippo et al. 2024b.
4. In domains like literary criticism one might even question the concept of veridicity (see Lamarque 1990).
5. An observational vocabulary can be understood as a specific type of vocabulary for annotation in NLP, with the key feature that its terms represent observations.

conference version

The ontology can express only certain portions of scholarly claims supporting their analysis, comparison, documentation, and interoperability. Also, we do not claim that criticism *must* be expressed in formal terms. We just claim that ontologies can be used in parallel to traditional, informal research methods to support interpretive work. Last but not least, in using ontologies as formal models we are not committing to *formalism* in the sense of the 20th-century school of literary criticism.

The paper is structured as follows. Section 2 introduces some insights on the case study that we will use throughout the paper to exemplify our approach. Section 3 presents some of the main features of the ontology. Section 4 digs into the case study, showing how the ontology can be used to document and analyze observational data in literary criticism. Finally, Section 5 concludes the paper.

2. Case Study: Part I

Our case study belongs to Medieval Italian Literature and concerns the interpretation of four female characters. Here, we discuss the interpretation of Dante's character of Beatrice proposed by Barolini 2012. The goal is to support the study of the history of criticism on Beatrice, enabling scholars to review, compare, and analyze what readers across time have claimed about this crucial figure.

From a literary standpoint, Beatrice is a *fil rouge* in Dante's literary texts. Scholars are often interested in understanding how she was interpreted in various epochs and cultures, as this can also shed light on the reception of Dante's whole oeuvre (Abreu de Lima 2025). In particular, since she appears in multiple works of Dante, from his early writings (his lyric poems and the *Vita nova*) to those of maturity (notably the *Divine Comedy*), it is also interesting to see how Beatrice is interpreted in different textual contexts.

In addition, tracing the history of criticism devoted to this character raises multiple questions that extend beyond the scope of literary scholarship and are worth exploring. For example, an assumption that seems to be shared in the history of Dante studies is that Beatrice is a *single* character appearing in various texts. In principle, this means that she exhibits some peculiar *traits* that, according to interpreters, prove that she is always the same character, despite some differences in her development from one text to the other. However, it is not simple to single out these traits, primarily because the sameness of Beatrice across Dante's production is, more often than not, only tacitly assumed. The presumed identity of Beatrice as depicted in Dante's works is of great interest to philosophers interrogating the nature of fictional characters (see, e.g., Paolini Paoletti et al. 2025), and is also central for the design of formal models (see Section 4).

It is important to stress that what we will present about the case study is only at a preliminary stage: the purpose is to exemplify the methodology and the approach we intend to adopt.[6] When the number of texts and related data increase, together with their conceptual and terminological diversity, a formal approach can help navigate through texts and interpretations. In the specific case of Beatrice, a large dataset from

6. The corpus we work with consists of around 600 critical texts on Italian medieval literature. These are mainly in Italian and English, but also include some texts in French and Spanish. The meta-data of the corpus is available through Zotero at https://www.zotero.org/groups/5305334/mite_-_make_it_explicit/library.

conference version

multiple critical texts can support scholars interested in the history of the character's reception in analyzing argumentation strategies, recurrent topics, commonalities and diversities in scholarly claims, etc., hence in looking at the critical tradition from a meta-level perspective.

3. The Ontology

The ontology has been designed with a modular architecture. The version presented in the paper consists of two modules: the core module, which contains the main elements, and an additional module that imports and extends the core module with elements demonstrating its application to a specific case in literary studies. The ontology is publicly available;[7] here we present only some key notions, but interested readers can refer to the OWL files for an in-depth view of the axiomatization. For readability, we will mostly use the syntax of Description Logic (Baader 2003).

Figure 1 shows the most general classes of the core module and their taxonomic organization. We will present two general types of observations, namely, *basic* (Section 3.1) and *source* (Section 3.3) observations, as well as the notions of *text* and *report* (Section 3.2). The ontology also includes so-called *argumentative* observations, which we leave out from the paper due to space limitations. Sibling classes, like Observation and Text, are disjoint.[8] As usual, Thing is the most general class in OWL.

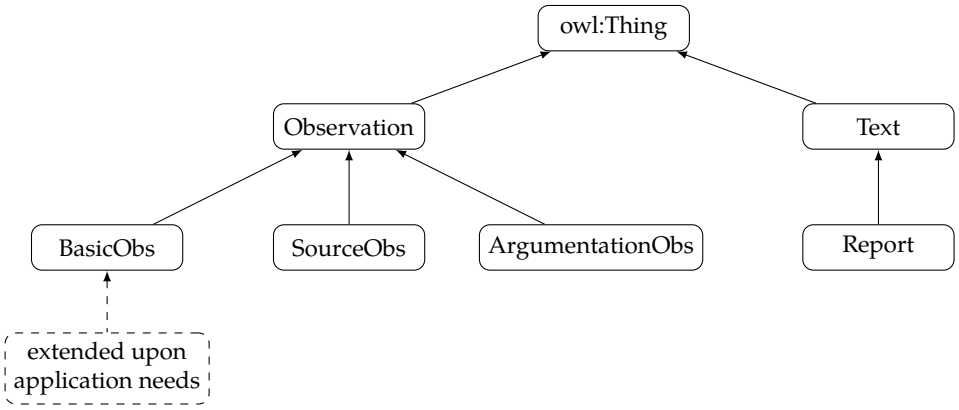


Figure 1: Solid boxes represent classes; solid arrows stand for subsumption between classes

3.1 Basic observations

Basic observations are the simplest types of observations. Unlike source (and argumentative) observations (see the next sections), they do not involve other observations, i.e., they are not *observations of observations*. Basic observations are used to represent claims by specific observers and, consequently, the core module does not include subclasses of the general BasicObs class, because they are always domain-dependent. Basic observations are expressed through (taxonomically-organized) observational vocabularies. Different communities can produce alternative vocabularies that can be later aligned to

7. See paragraph "Software Availability": the file with the core module is core-module.rdf; the other module is in literary-module.rdf; the exemplificative instantiation (i.e., the Abox level) is case-study.rdf
8. It should be therefore clear that the ontology includes classes for observations, as well as classes for things that are not observations like texts.

conference version

enable meaningful data interchange.

From a methodological perspective, for the introduction of basic observations, users of the ontology should follow the following steps:

Step 1 Identify the observational vocabulary and formally represent it as a taxonomy of (disjoint) classes whenever possible. This involves making choices about what to represent as observation and what to represent through a standard logical predication mechanism. In principle, observations are to be used for potentially debated information, whereas regular predicates are preferable for information that can be taken as stable.[9]

Step 2 For each observation class, determine the cardinality and type of arguments, i.e., the entities involved in instances of the observation class.

Step 3 Characterize relations and dependencies between the observation classes, e.g., their incoherence or correlation.

We go through these steps with illustrative examples from the selected case study.

Step 1. In this first step, the interaction with domain experts leads to the creation of a shared observational vocabulary to express claims on both literary and critical texts. Figure 2 offers an example of such a vocabulary.[10] Observations of symbolic agency (SymbolicAgencyObs) describe the extent to which a literary character functions symbolically, representing abstract ideals or theological concepts, among others. Narrative agency observations (NarrativeAgencyObs) describe narrative features like being talkative, authoritative, or a character with an active agency, namely, a character that performs actions and drives the narrative forward. Gender observations (GenderObs) describe gender-related traits. Observations of ethical agency (EthicalAgencyObs) are used to describe ethical traits such as being virtuous or sinful. Finally, observations of affective agency (AffectiveObs) express traits describing how a character is associated with emotions, sexual desire, etc.

For the sake of shortness, we show here only some axioms holding among the classes in the taxonomy.[11] By (a1)–(a2), EthicalAgencyObs and NarrativeAgencyObs are disjoint subclasses of BasicObs, i.e., they do not share common instances, hence an individual observation cannot instantiate both EthicalAgencyObs and NarrativeAgencyObs. Similar axioms are used to characterize the other classes in Figure 2.

a1 $\mathsf{EthicalObs} \sqcup \mathsf{NarrativeAgencyObs} \sqsubseteq \mathsf{BasicObs}$
a2 $\mathsf{EthicalAgencyObs} \sqcap \mathsf{NarrativeAgencyObs} \sqsubseteq \bot$

Step 2. In the second step we qualify the *arguments* of the observations. We use hasArgument as the relation[12] to represent the entities involved in an observation, i.e., the entities which the observation predicates about. By axioms (a3)–(a4), observations of ethical- and narrative agency have only one argument, which must be an instance

9. A knowledge representation system can evolve in time. Therefore, information that was initially represented as stable may be later reinterpreted in observational terms at later stages. Ontology evolution techniques can be adopted in these scenarios; we will not discuss this topic in the paper.
10. The taxonomy is not exhaustive; it is used only for exemplificative purposes.
11. Elements of the ontology are prefixed with *a*- for axioms and *r*- for rules. Formulas for examples are prefixed with *e*-.
12. If not otherwise specified, all relations are OWL object properties.

conference version

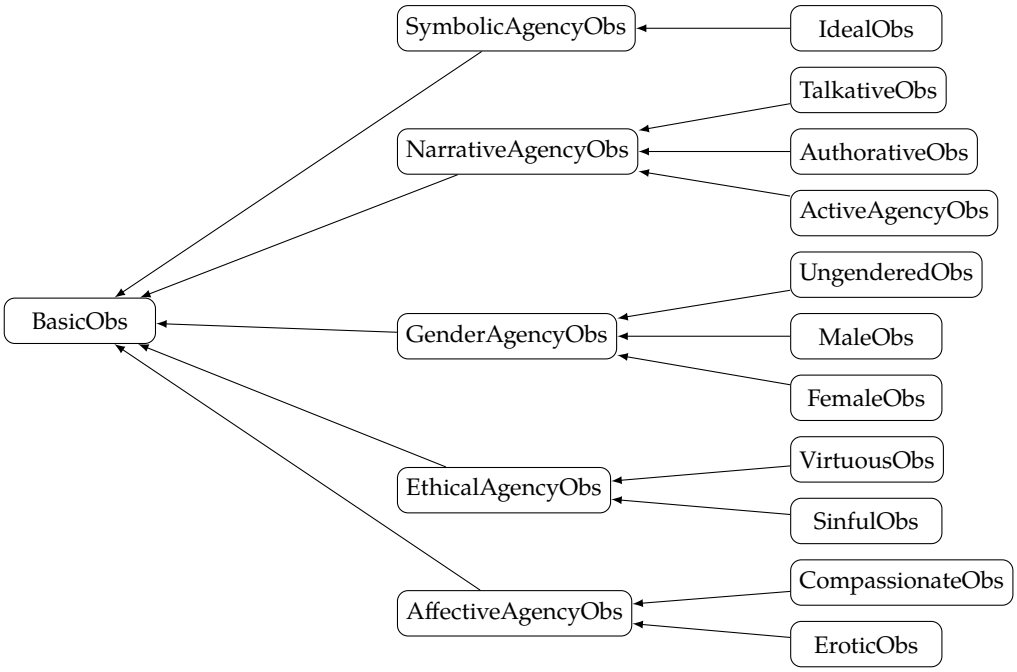


Figure 2: An illustrative fragment of a segment of our taxonomy of basic observations.

of the (non-observational) class Character.[13] Because of the subsumption relations on the taxonomy, subclasses of EthicalAgencyObs and NarrativeAgencyObs inherit the constraints in (a3)–(a4).

a3 $\mathsf{EthicalAgencyObs} \sqsubseteq \exists^{=1}\, \mathsf{hasArgument.Character}$

a4 $\mathsf{NarrativeAgencyObs} \sqsubseteq \exists^{=1}\, \mathsf{hasArgument.Character}$

To give an example of a basic observation, (e1) specifies the basic observation o of kind TalkativeObs with argument (the literary character of) Dante's Beatrice. In other words, the observation o represents, in the terms of the ontology, that Beatrice is a talkative character. To simplify the notation, we will write $\mathbf{p}(a_1, \dots, a_n)$ to denote the unique observation with minimal kind PObs and with arguments $a_1, \dots, a_n$.[14] In the case of the observation o specified in (e1), we can therefore simply write **talkative**(*beatrice*).

e1 $\mathsf{TalkativeObs}(o) \sqcap \mathsf{hasArgument}(o, \mathit{beatrice})$

Step 3. We now further enrich the observational vocabulary to also enhance the reasoning capabilities of the ontology.

First, we introduce SWRL rules to automatically infer relations of *incoherence* (INC) and *correlation* (COR) between observations. Incoherence means that two (or more) observations are conflictual when applied to the same entity. An example is provided by rule (r1), according to which VirtuousObs and SinfulObs are incoherent if they have the same argument. Correlation instead is a form of implication between (instances of) different observations.[15] For example, by (r2), observations of active agency correlate with observations of talkative. This is intended to capture the meaning of being an active agent as an agent that is talkative, among other traits.

13. For the sake of this paper, we do not dig into the ontological analysis of literary characters but simply assume them to be among the entities in the application domain.
14. The minimal kind of an observation is the most specific class it is an instance of.
15. Correlation does not have to be confused with probabilistic correlation.

conference version

r1 $\mathsf{VirtuousObs}(?x), \mathsf{hasArgument}(?x, ?z), \mathsf{SinfulObs}(?y), \mathsf{hasArgument}(?y, ?z) \rightarrow \mathsf{INC}(?x, ?y)$

r2 $\mathsf{ActiveAgencyObs}(?x), \mathsf{hasArgument}(?x, ?z), \mathsf{TalkativeObs}(?y), \mathsf{hasArgument}(?y, ?z) \rightarrow \mathsf{COR}(?x, ?y)$

Second, (r3) establishes that instances of VirtuousObs with the same arguments are the same observation.[16] This rule clarifies that observation classes are defined by their arguments (see discussion below).

r3 $\mathsf{VirtuousObs}(?x), \mathsf{hasArgument}(?x, ?z), \mathsf{VirtuousObs}(?y), \mathsf{hasArgument}(?y, ?z) \rightarrow \mathsf{SameAs}(?x, ?y)$

Discussion. In the taxonomy in Figure 2, characters' traits are represented as observations. An alternative consists in treating them as non-observational disjoint classes like (e2), saying that Virtuous and Sinful do not have common instances.

e2 $\mathsf{Virtuous} \sqcap \mathsf{Sinful} \sqsubseteq \bot$

Assume now that a user of the ontology represents a literary character as being virtuous, whereas another user represents it as sinful, e.g., because they rely on different literary sources. By (e2) this leads to a logical contradiction. Differently, an approach on observations like the one presented above preserves consistency while taking into account incompatible claims. In particular, when one represents two observations instantiating VirtuousObs and SinfulObs, respectively, to state that the same character is observed as virtuous and sinful, the two observations are classified as incompatible by the application of (r1). This brings us back to the discussion on how to decide whether to treat a concept as an observation or a non-observation class. As the example shows, it is possible to use observations to model and reason about conflicting data without compromising the consistency of the system. In a sense, observations are a way to operationalize the modeling of contradictory debates without sacrificing logical consistency.

In the presence of correlation (COR) rules like (r2), one could think of implementing an off-line algorithmic procedure based on the INSERT operation from SPARQL 1.1 UPDATE[17] that, given a report including one of the two observations, exhaustively saturates it by introducing a new instance of the correlated one with the appropriate arguments binding. Following our example, according to (r2), a report containing the observation $\mathsf{ActiveAgencyObs}(o) \sqcap \mathsf{hasArgument}(o, \mathit{beatrice})$ would then be automatically saturated with the insertion of the new $\mathsf{TalkativeObs}(o') \sqcap \mathsf{hasArgument}(o', \mathit{beatrice})$ observation, grasping the idea formally represented in (r2) by which an active agent is also talkative. In this manner, one would expand the observational data in the knowledge base according to the constraints in the ontology.[18]

Regarding rule (r3), notice that by having a taxonomy of observation classes at different levels of generality, users of the ontology can choose the granularity of the observations that they wish to express. For example, they can decide whether they wish to state that

16. A similar rule applies to all leaf-nodes of the taxonomy in Figure 2. Recall that SameAs is a built-in SWRL construct to model identity between individuals.
17. https://www.w3.org/TR/sparql11-update/#deleteInsert/
18. The use of a SPARQL INSERT is needed in these cases because SWRL rules do not allow the introduction of new individuals in a knowledge-base.

conference version

an entity is virtuous or simply has an ethical trait. From a practical stance, our suggestion is to use the most specific leaves of the taxonomy whenever possible, as they convey richer information and can be more precisely characterized. For example, given the taxonomy in Figure 2, a rule for the sameness of EthicalObs would be incorrect, because it cannot discriminate between the more specific observations that it subsumes. This consideration leads to a broader comment on the level of detail that an observational vocabulary should capture. While general observation classes can be useful for modeling information extracted from multiple texts, they may oversimplify the claims made in the texts. Conversely, specific observation classes may help to reflect fine details, but they may be difficult to generalize and may not be useful for comparing claims from multiple texts. Developing vocabularies for formally dealing with literary scholarship remains a challenge. A recommendation is to collaborate closely with domain experts to find a satisfactory level of granularity that neither trivializes information nor becomes useless for computational tasks.

In our case, the taxonomy shown in Figure 2 was obtained through a great deal of painstaking work on a selection of texts from the corpus. We first (manually) identified the set of sentences in the texts which we considered relevant for the analysis. We then extracted a vocabulary from the texts (e.g., the term *loquax* used by Barolini 2012), and finally reworked the vocabulary to make it potentially applicable to the various texts at hand. For example, we replaced *loquax* with *talkative*, as the former seemed to apply specifically to Barolini 2012.[19] Hence, the purpose was, firstly, to develop a vocabulary that can be used to represent information from multiple texts, and secondly, to develop a shared and intersubjective semantics for its terms in order to make the ontology transparent to multiple users. Further work is needed to both improve the methodology and integrate multiple computational techniques for ontology-based information extraction.[20]

Finally, it should be clear that the development of rules for inferring incompatibility, correlation, sameness, and other relations between observations must reflect debates within the relevant community and be based on the intended meaning of the types of observations included in the observational vocabulary. We suggest always checking with experts that the inferences reflect their requirements and knowledge in such a way to avoid undesirable information being present in the knowledge base.

In the next section we introduce the notions of text and report, which are relevant for other observation types.

3.2 Texts and Reports

Texts are abstract sequence of words in a (formal, natural) language. *Abstract* means that a text is not the physical configuration of words in (printed or digital) exemplars. The same text can be indeed "realized" in multiple copies or other editorial instances.[21] As sequences of words, texts that display even slightly different words differ from each other, although they can be similarly interpreted. Examples of texts that concern our

19. One could track the provenance of the concepts (relations) in the vocabulary to the text(s) they come from, paying however attention to which semantic characterization is given to this concept-to-text relation.
20. The reader can refer to the Sections 4–5 for more reflections on these aspects.
21. We do not discuss this aspect further in the paper. Readers can refer to (Mizoguchi and Borgo 2025; Sanfilippo 2021) for readings on this topic in the context of knowledge representation.

conference version

application focus are literary texts and critical essays. In the scope of the ontology, texts have a prominent role because they are either the sources of observations or the objects of interpretation. For example, Barolini 2012 is the source for Barolini's observations on Dante's texts. The focus on texts also clarifies that we are interested in *documented* observations that can be shared and understood by others, rather than in mental states held by their authors.

Reports are specific types of texts written in observational vocabularies, i.e., they record interpreters' observations. A report can be therefore understood as a collection of formally specified observations. As we will see, reports are particularly relevant because all observations must have reports as their ultimate sources.

We assume that texts written in a natural language can be ambiguous and therefore lead to different, perhaps even incompatible, interpretations (Sanfilippo et al. 2024a). Conversely, because reports are formal texts written in observational vocabularies, they are not ambiguous, in the sense that their contents reflect the axiomatic structure provided by the ontology. For example, if a report makes a particular observation, it cannot be claimed that the report does not make, or, more strongly, rejects that observation. However, it is possible for a report to be incoherent, for example if it contains incoherent observations (e.g., in the sense of (r1)).

3.3 Source observations

Source observations make explicit *who* (or better *what*) expresses an observation, namely, its source. They are therefore relevant to document the provenance of observations.

In the taxonomy of source observations, we distinguish between two (disjoint) high-level branches: *illocutionary* and *illocutionary interpretative* observations. The fundamental difference between the two is that only the latter model how observers interpret texts. The emphasis on the *illocutionary* dimension of source observations originates from philosophy (Marsili 2024); we find it useful to emphasize that they represent types of *propositional attitudes* through which observers commit to their claims, either positively (with assertions) or negatively (with rejections). Importantly, source observations are observations of other observations (see examples below).

Illocutionary observations. These observations are used to record what a text claims. Formally, illocutionary observations have two arguments: (*i*) a text (the source), and (*ii*) the claimed observation. In particular, the relation hasSource (with range Text) specializes hasArgument to relate an illocutionary observation to its textual source, whereas hasObservation (with range Observation) to relate a source observation to the asserted or rejected observation. Axiom (a5) says that illocutionary observations (IllocutionaryObs) have exactly two arguments, one is a text (a6), the other is an observation (a7).

a5 $\mathsf{IllocutionaryObs} \sqsubseteq \exists^{=2}\, \mathsf{hasArgument}$
a6 $\mathsf{IllocutionaryObs} \sqsubseteq \exists^{=1}\, \mathsf{hasSource.Text}$
a7 $\mathsf{IllocutionaryObs} \sqsubseteq \exists^{=1}\, \mathsf{hasObservation.Observation}$

Assert- (AssertObs) and reject (RejectObs) observations are disjoint subclasses of illocutionary observations, see (a8)–(a10). Notice that the class for illocutionary observations

conference version

is not equivalent to the (disjoint) union of assertion and rejection observations. This is because we leave it open the possibility of having other types of illocutionary observations standing for other sorts of propositional attitudes (the same considerations holds for interpretative observations, see the next paragraph). However, when populating the ontology, the class IllocutionaryObs is meant to be instantiated only through its subclasses.

a8 $\mathsf{AssertObs} \sqsubseteq \mathsf{IllocutionaryObs}$
a9 $\mathsf{RejectObs} \sqsubseteq \mathsf{IllocutionaryObs}$
a10 $\mathsf{AssertObs} \sqcap \mathsf{RejectObs} \sqsubseteq \bot$

To give an example of assertion, $\mathbf{asr}(brl, \mathbf{talkative}(beatrice))$, corresponding to the observation o in (e3),[22] states that the talkative feature of Beatrice is asserted by the text *brl*, standing for Barolini's essay (Barolini 2012); *brl* is therefore the source of the observation $\mathbf{talkative}(beatrice)$.

e3 $\mathsf{AssertObs}(o) \sqcap \mathsf{hasSource}(o, brl) \sqcap \mathsf{hasObservation}(o, o') \sqcap \mathsf{TalkativeObs}(o') \sqcap \mathsf{hasArgument}(o', beatrice)$

Because illocutionary observations are themselves observations, they can be nested to form *chains* of observations (see also Section 4 on this).[23] This can be useful in the humanities to refer to indirect sources (a source asserts that another source asserts that ...). Assertions (or rejections) having reports as their sources are granted a special meaning. The idea is that the authors of reports *infer* (or do *not* infer in the case of rejections) the claimed observation from their beliefs, i.e., the observations make (partially) explicit what the observers belief. An example is the chain of assertions $\mathbf{asr}(r, \mathbf{asr}(brl, \mathbf{talkative}(beatrice)))$ which says that the report *r* asserts that Barolini's critical text asserts that Beatrice is talkative. Hence, according to the intended meaning of reports' assertions, the author of *r* makes explicit their beliefs on Barolini's text.[24]

According to our methodology, illocutionary chains must have reports as their ultimate sources. This is because, as said in Section 3.2, reports are the formal texts collecting observations. In other words, illocutionary observations made by reports make it explicit that reports' authors (e.g., researchers using the ontology) are producing a set of observations on the texts at stake. To recall the example above, this formal mechanism makes it clear that it is not Barolini's essay *per se* claiming that Beatrice is talkative; rather, it is the author of the report ascribing to Barolini's text that specific claim, formally representing it through the observational vocabulary of the ontology.

Finally, (r4) is a SWRL rule for the sameness of assert observations (the same for reject observations). As for the example of (e3), a rule like (r4) is useful to check the compliance of the data against some quality constraints. In this case, to guarantee that assert (reject) observations with the same arguments (text and observation) are the same.

22. In $\mathbf{asr}(t, o)$ (we write **asr** rather than **assert**), *t* is the source text while *o* is the asserted observation.
23. This means that one can model assertions of assertions, rejections of rejections, assertion of rejections, etc. In line with previous work on argumentation theory (Dung 1995), the rejection of a rejection does not amount to an assertion.
24. Readers can refer to the research by Masolo et al. 2025 for an in-depth reading on the semantics of source observations.

conference version

r4 $\mathsf{AssertObs}(?x), \mathsf{hasSource}(?x, ?t), \mathsf{hasObservation}(?x, ?z),$
$\mathsf{AssertObs}(?y), \mathsf{hasSource}(?y, ?t), \mathsf{hasObservation}(?y, ?z) \rightarrow \mathsf{SameAs}(?x, ?y)$

Illocutionary interpretative observations. The intended meaning of this type of observations is to represent *interpretations* of texts (Masolo et al. 2025). This is the reason why illocutionary interpretative observations have three arguments: (*i*) the text that is the source of the observation; (*ii*) the text that is interpreted; and (*iii*) the either asserted or rejected observation. In particular, the second argument plays a fundamental role, because it stands for the information that must be taken into account for the observation in the third argument. When the source of an illocutionary observation is a report, its intended meaning is that the report's author *infers* the asserted observation (or does not infer it for rejection) from both its beliefs and the information from the interpreted text. This recalls research work in (computational) literary studies (e.g., Gius and Jacke 2017), where text's interpretation is a form of inference.

Axioms (a11)–(a14) fix the arguments for interpretative illocutionary observations (IllocutionaryIntObs). The relation hasInterpretedText (with range Text) specializes hasArgument to relate the observations to the interpreted text.

a11 $\mathsf{IllocutionaryIntObs} \sqsubseteq \exists^{=3}\, \mathsf{hasArgument}$
a12 $\mathsf{IllocutionaryIntObs} \sqsubseteq \exists^{=1}\, \mathsf{hasSource.Text}$
a13 $\mathsf{IllocutionaryIntObs} \sqsubseteq \exists^{=1}\, \mathsf{hasInterpretedText.Text}$
a14 $\mathsf{IllocutionaryIntObs} \sqsubseteq \exists^{=1}\, \mathsf{hasObservation.Observation}$

Axioms (a15)–(a17) represent illocutionary assertion (IntAssertObs) and rejection (IntRejectObs) as disjoint subclasses of interpretative observations.

a15 $\mathsf{IntAssertObs} \sqsubseteq \mathsf{IllocutionaryIntObs}$
a16 $\mathsf{IntRejectObs} \sqsubseteq \mathsf{IllocutionaryIntObs}$
a17 $\mathsf{IntAssertObs} \sqcap \mathsf{IntRejectObs} \sqsubseteq \bot$

For instance, $\mathbf{iasr}(r, brl, \mathbf{talkative}(beatrice))$[25], corresponding to (e4), says that (the author of) the report r asserts the observation $\mathbf{talkative}(beatrice)$ by interpreting the text brl by Barolini. Hence, according to the intended meaning of interpretative assertions with reports as sources, this means that the author of the report r infers $\mathbf{talkative}(beatrice)$ by their interpretation of the text brl.

e4 $\mathsf{IntAssertObs}(o) \sqcap \mathsf{hasSource}(o, r) \sqcap \mathsf{hasInterpretedText}(o, brl) \sqcap$
$\mathsf{hasObservation}(o, o') \sqcap \mathsf{TalkativeObs}(o') \sqcap \mathsf{hasArgument}(o', beatrice)$

It is important to notice that because interpretative observations consider how observers interpret texts, it is possible that a single text is interpreted in different and incompatible ways. For example, an observer may not infer or may even reject $\mathbf{talkative}(beatrice)$ with respect to brl. Also, similarly to illocutionary observations, interpretative observations can be nested into chains. What is fundamental, for the same reason as the one provided in the previous section, is that semantically meaningful interpretative observations have reports as their ultimate sources.

Rule (r5) constrains sameness for interpretative assert observations (the same for inter-

25. In the compacted notation $\mathbf{iasr}(t_s, t_i, o)$ (we write **iasr** rather than **intassert**), t_s is the source text, t_i is the interpreted texts, and o is the inferred observation.

conference version

pretative rejections).

r5 IntAssertObs(?x), hasSource(?x, ?t), hasInterpretedText(?x, ?tt), hasObservation(?x, ?z), IntAssertObs(?y), hasSource(?y, ?t), hasInterpretedText(?y, ?tt), hasObservation(?y, ?z) → SameAs(?x, ?y)

Discussion. We expand upon the distinction between illocutionary and interpretative illocutionary observations by comparing **asr**(r, **asr**(brl, **talkative**($beatrice$))) with **iasr**(r, brl, **talkative**($beatrice$)).

The first observation represents what the author of the report r infers about the text *brl* by Barolini, i.e., it makes explicit part of their beliefs on *brl*. The observation is to be read as: "(the author of) r infers that *brl* asserts **talkative**($beatrice$)." According to the intended meaning of illocutionary assertions, the author of r formally represents an information they are aware of, but they do so *without* interpreting or even accessing the text *brl*. For example, the observer may merely document some information that was relayed to them.

The second observation is read as: "(the author of) r infers **talkative**($beatrice$) by interpreting *brl*." In this case, the observer has access to the text *brl* and interprets it; in this sense, the information in *brl* is necessary for inferring **talkative**($beatrice$). In summary, illocutionary observations (AssertObs, RejectObs) can be used to report the observers' beliefs on texts, whereas interpretative observations (IntAssertObs, IntRejectObs) represent how readers interpret texts. From a modeling perspective, the difference in the use of source observations depends on what one wishes to say: the record of a belief *vs.* the explicit interpretation of a text. To operationalize this distinction in order to treat it systematically when populating the ontology, one may explore the use of natural language processing techniques to analyze texts, checking the way in which they present their claims, in particular, checking if textual evidences provide elements towards one or the other type of observation.[26]

4. Case Study: Part II

We now come back to the case study to demonstrate how the ontology can be used to represent observational data in literary studies and automatically reason about them. As said, the purpose is to (partially) exemplify the approach, not to provide an exhaustive view on the analysis of our corpus.

The ontology enables the reiteration of observations, forming *chains* of various length and complexity. To reason over these chains, we find it useful to focus on certain *patterns* that are among the most common for practical application. The presented approach can be extended to cover further cases. The pattern schemas are the following ones (where r is a report, t^c, $\bar{t}^c$ are critical texts, t^l is a literary text, and o is a basic observation):[27]

p1 **iasr**(r, t^c, o): the (author of) the report r infers the basic observation o from the critical text t^c.

26. Thanks to one of the anonymous reviewers for the suggestion. Further work in this direction is needed.
27. (p1)–(p4) stand for general observations with given forms. We consider only assertions, but similar considerations can be made for rejections.

conference version

p2 $\mathbf{iasr}(r, t^c, \mathbf{asr}(t^l, o))$: (the author of) r infers that t^c asserts that t^l asserts o.
p3 $\mathbf{iasr}(r, t^c, \mathbf{iasr}(t^c, t^l, o))$: (the author of) r infers that t^c asserts o *by interpreting* t^l.
p4 $\mathbf{iasr}(r, t^c, \mathbf{iasr}(t^c, \bar{t}^c, \mathbf{asr}(t^l, o)))$: (the author of) r infers that t^c, by interpreting $\bar{t}^c$, asserts that t^l asserts o.

First, the four patterns focus on interpretative observations because they are particularly interesting types of observations, making explicit how texts are interpreted. Also, because they have reports as sources, they cannot be interpreted in ways different from what they state. Second, patterns in (p2) and (p3) differ in that the nested observation is an illocutionary observation (**asr**) in (p2), and an interpretative observation (**iasr**) in (p3). The intended meaning is that a pattern like (p2) represents an information on t^l that (the author of) t^c reports without emphasis on their interpretation of t^l. Differently, in (p3) the emphasis is on the interpretation of t^l by t^c.

We now consider the following examples based on Barolini 2012. The basic observations are organized along the lines of what we presented in Section 3.1. The two groups of observations (*i*) (e5)–(e9) and (*ii*) (e10)–(e14) capture some aspects in the characterization of Beatrice provided by Barolini's essay (*brl*) in the context of Dante's lyric poetry (*lyr*) and *Divine Comedy* (*dv*). In particular, Barolini argues that Beatrice is a continuous presence in Dante's literary production. In his lyrics, she is portrayed as a *courtly lady*, a woman celebrated in poetry as an ideal of perfection the poet aspires to through his love, but which can never be truly attained. Beatrice is also depicted as having erotic traits, but in these texts she does not exhibit an active agency, i.e., a personality, nor she speaks. In the *Comedy*, she becomes a much more complex character, according to Barolini. She retains some of her original traits, such as the erotic traits. However, she also displays characteristics such as being talkative and authoritative. For instance, she explains important theological and philosophical doctrines to Dante. In addition, she qualifies as an agent with a proper personality, and a strong and complex one. According to Barolini, these are traits associated to different literary traditions, among which the Old French lyric form of pastoral poetry, axed around the character of the *pastourelle* (shepherdess).

Observations (e5) and (e10) instantiate the pattern (p2) seen above. Here, the report r_1 infers from Barolini's essay that Beatrice is observed as being a female character in both the lyrics (e5) and the *Comedy* (e10). The use of **asr** stresses that there is not emphasis on Barolini's interpretation of the texts; the scholar provides a piece of information that is useful to their argument. All other cases but (e7) and (e8) instantiate the pattern (p3).[28]

e5 $\mathbf{iasr}(r_1, brl, \mathbf{asr}(lyr, \mathbf{female}(beatrice)))$
e6 $\mathbf{iasr}(r_1, brl, \mathbf{iasr}(brl, lyr, \mathbf{erotic}(beatrice)))$
e7 $\mathbf{iasr}(r_1, brl, \mathbf{idny}(brl, lyr, \mathbf{talkative}(beatrice)))$
e8 $\mathbf{iasr}(r_1, brl, \mathbf{idny}(brl, lyr, \mathbf{activeAgency}(beatrice)))$
e9 $\mathbf{iasr}(r_1, brl, \mathbf{iasr}(brl, lyr, \mathbf{ideal}(beatrice)))$
e10 $\mathbf{iasr}(r_1, brl, \mathbf{asr}(dv, \mathbf{female}(beatrice)))$
e11 $\mathbf{iasr}(r_1, brl, \mathbf{iasr}(brl, dv, \mathbf{erotic}(beatrice)))$
e12 $\mathbf{iasr}(r_1, brl, \mathbf{iasr}(brl, dv, \mathbf{talkative}(beatrice)))$
e13 $\mathbf{iasr}(r_1, brl, \mathbf{iasr}(brl, dv, \mathbf{activeAgency}(beatrice)))$

28. The pattern instantiated in (e7) and (e8) nests an interpretative denial, and it is not discussed in the paper.

conference version

e14 $\mathbf{iasr}(r_1, brl, \mathbf{iasr}(brl, dv, \mathbf{authoritative}(beatrice)))$

As an analytic task, by reasoning over these observations, one may wish to infer that Beatrice is classified as a certain type of character, explicitly defined in the ontology. To operationalize this, we need to introduce formal mechanisms to reason over observation chains. For shortness, we show here only some SWRL rules; readers can refer to the available files for an in-depth reading.

The following rule (r6) applies to pattern chains like (p3). Accordingly, when an interpretative assert observation x nests a second interpretative assert observation o, such that o has basic observation bo and interpreted literary text t, then t assertively claims bo. A similar rule is used to infer that a text rejects a basic observation (rejClaims), as in the cases of (e7) and (e8).

r6 IntAssertObs$(?x)$, hasObservation$(?x, ?o)$, IntAssertObs$(?o)$, hasObservation$(?o, ?bo)$, BasicObs$(?bo)$, hasInterpretedText$(?o, ?t)$, LiteraryText$(?t) \rightarrow$ assClaims$(?t, ?bo)$

We can now apply reasoning to infer that a character is of a certain type with respect to the interpreted text. For instance, rule (r7), reasoning over both assertion and rejection claims allows to infer that a character is a *pastourelle* in the context of a text (pastourelleCharacterIn). Specifically, the rule represents the *sufficient* conditions for x to be classified as a *pastourelle*. In a similar vein, rule (r8) classifies an entity as a *courtly lady* within a text (courtlyLadyCharacterIn). Finally, (r9) represents the sufficient conditions to consistently merge *pastourelle* and *courtly lady*, hence to classify entities that are interpreted as being complex realizations of both types (hybridCourtlyPastourelleCharacterIn). Also, for the complexity intrinsic to this novel character type, the rule adds the classification under the observation ActiveAgencyObs, which is neither included in *pastourelle* nor in *courtly lady*.

r7 assClaims$(?t, ?o1)$, assClaims$(?t, ?o2)$, assClaims$(?t, ?o3)$, rejClaims$(?t, ?o4)$, TalkativeAgencyObs$(?o1)$, FemaleObs$(?o2)$, EroticObs$(?o3)$, AuthoritativeObs$(?o4)$, hasArgument$(?o1, ?x)$, hasArgument$(?o2, ?x)$, hasArgument$(?o3, ?x)$, hasArgument$(?o4, ?x) \rightarrow$ pastourelleCharacterIn$(?x, ?t)$

r8 assClaims$(?t, ?o1)$, assClaims$(?t, ?o2)$, assClaims$(?t, ?o3)$, FemaleObs$(?o1)$, EroticObs$(?o2)$, IdealObs$(?o3)$, hasArgument$(?o1, ?x)$, hasArgument$(?o2, ?x)$, hasArgument$(?o3, ?x)$, rejClaims$(?t, ?o4)$, rejClaims$(?t, ?o5)$, TalkativeObs$(?o4)$, ActiveAgencyObs$(?o5)$, hasArgument$(?o4, ?x)$, hasArgument$(?o5, ?x) \rightarrow$ courtlyLadyCharacterIn$(?x, ?t)$

r9 assClaims$(?t, ?o1)$, assClaims$(?t, ?o2)$, assClaims$(?t, ?o3)$, assClaims$(?t, ?o4)$, assClaims$(?t, ?o5)$, TalkativeAgencyObs$(?o1)$, FemaleObs$(?o2)$, ActiveAgencyObs$(?o3)$, AuthoritativeObs$(?o4)$, EroticObs$(?o5)$, hasArgument$(?o1, ?x)$, hasArgument$(?o2, ?x)$, hasArgument$(?o3, ?x)$, hasArgument$(?o4, x)$, hasArgument$(?o5, ?x) \rightarrow$ hybridCourtlyPastourelleCharacterIn$(?x, ?t)$

By reasoning over the rules and the data in (e5)–(e9), Beatrice is classified as a courtly lady character in the context of the lyrics, whereas she is classified as a hybrid character

conference version

in the *Comedy* based on the data in (e10)–(e14).[29] This is meant to formally grasp that, following Barolini's essay, Beatrice is a complex character whose traits are inherited from multiple literary traditions. As mentioned, the ontology is intended for use with a corpus of critical texts on Medieval Italian literature. However, to expand its scope, inferences could explicitly take literary genres, production times and other information into account in order to model the dependency between character types and literary contexts.

Discussion. Rules (r7)–(r9) give some hints on the logical reasoning that can be done on the data; in these cases, to state that a character, interpreted with respect to a literary text, is classified in a certain way, as it is explicitly characterized in the SWRL rules.

From the technical side of automated reasoning, although it is well known that *unrestricted* combinations of OWL 2 ontologies and rules are undecidable, in the last 20 years a few theoretical studies have proved that the combination of highly expressive ontology languages, such as OWL 2 DL, and DL-safe SWRL rules (i.e., rules that apply only to individuals which are explicitly given in the knowledge base) can be done while preserving decidability (Hitzler and Parsia 2009). However, complexity results characterizing the combination of languages from the DL-Lite family, as well as other restricted fragments of OWL 2 DL with rules are still in investigation, and we cannot add further details on that.

It is important to stress that (r7)–(r9) are just examples with the purpose of showing how specific observational data can trigger inferences. Hence, the rules are not general purpose reasoning mechanisms on literary texts. For a more robust analysis on these lines, one must introduce and properly characterize character types and their features by means of new classes and rules, relying on experts' knowledge and their *desiderata*. More interesting results can be obtained by expanding the knowledge base with additional elements to reason over multiple reports and texts. For instance, if different reports record information on different critical texts which interpret the same literary text, one can compare the data in the reports to analyze the convergences and divergences in criticism.

Also, (r7)–(r9) infer binary relations between a character and the interpreted text. These relations hold by virtue of the interpretation that a report records with respect to a critical text. Hence, it would be more appropriate to infer an n-ary relation, such as: x is a courtly lady character in the literary text t, given the interpretation of t' provided by report r. However, Semantic Web languages lack sufficient expressivity for directly modeling n-ary relations. The strategy of *reification* could be adopted (for a reading, see Porello et al. 2025b), but this would affect the kind of reasoning that can be performed and would require a more complex knowledge-base architecture.[30] We leave this issue to future work.

29. For this task we used the reasoner HermiT, version 1.4.3.456. The file with the instantiation of the ontology is available online. As a technical note, rule (r6), upon which (r7)–(r9) rely, reasons over interpretative observations, whereas–as said–(e5) and (e10) include both **iasr** and **asr**. Observations (e5) and (e10) require indeed to trigger other rules, which are not shown here for the sake of shortness (see the available OWL files).

30. A rule should infer the existence of a reified relation holding among the desired n relata. As said in Section 3.1, this form of inference cannot be expressed in SWRL. Therefore, the system architecture would need to integrate both SWRL rules and additional methods like the use of SPARQL INSERT functions, as previously discussed.

conference version

A final remark on Beatrice is in order, given what we said in Section 2. From a formal perspective, in the examples above we introduced a single logical constant to refer to her across both the lyrics and the *Comedy*. This choice entails that Beatrice is *a priori* treated as the same entity in both textual contexts. An alternative strategy would be to introduce two distinct constants, one for each text, and then decide *a posteriori* whether they co-refer. These two approaches are not as different as they may initially appear, because analyzing the data may always lead to revising the initial assumptions and concluding that what was modeled as a single (or as two distinct) character(s) should instead be treated as two different (or identical) entities (see Porello et al. 2025a). From a conceptual perspective, questions of identity remain challenging, especially when crucial information is left implicit. In Barolini's essay, it is claimed that Beatrice functions as the *fil rouge* of Dante's oeuvre, despite the variability of her characterizations. Accordingly, the theoretical task becomes that of identifying the (relational) traits that warrant treating these differences as belonging to one and the same character. Even more challenging is that, because we deal with critical texts, scholars may provide non-overlapping traits for identity. In this sense, questions of identity lose their absolute status and become subjective to interpreters. A strategy to address this latter issue can consist in comparing alternative views on what seems to be a single character, and finding relevant (dis-)agreements among interpretations to ground identity claims among groups of scholars (see, e.g., the work of Sanfilippo et al. 2025b). This is not only a matter for speculation, but is relevant for formal modeling, hence, to decide upon which elements a model quantifies.

5. Conclusions

We presented an ontology for representing and reasoning about observational data, with a focus on literary scholarship. Because of its modular architecture, the ontology can be applied to other domains, though it would likely demand adaptations depending on the domain's requirements. We discussed a preliminary case study to show case our approach, although further work is needed to support a more in-depth study, including an evaluation of the proposed vocabulary and the related reports.

The ontology exhibits some key features that distinguish it from related efforts (e.g., Bruno et al. 2024; Daquino et al. 2020; Doerr et al. 2011; Sartini et al. 2023). First, it explicitly addresses the plurality of perspectives that characterizes the humanities. Different communities can define their own observational vocabularies to express observations, and subsequently provide alignments to enable data interchange or even integration. Second, the ontology supports the modeling of incompatible observations on the same subjects, thereby accounting for controversies in debates. Moreover, incompatibility among observations can be *automatically classified* through reasoning mechanisms, whereas other approaches have so far paid only limited attention to inference capabilities. Third, since one can (logically) quantify over observations, they can be nested into chains of different length and structure. This is also a relevant point of departure with respect to existing approaches. Reasoning over chains can be useful for analyzing observations drawn from multiple sources and focused on different texts. In this paper, we considered only a limited set of chain patterns, although the analysis can be extended in several directions.

conference version

conference version

Given the subtleties and intricacies of literary inquiry, our proposal faces a number of challenges. A primary one concerns the production of a structured dataset, grounded in the ontology, that can effectively support scholarly inquiry on a selected corpus. This task is difficult because critical essays vary widely in their results, concepts, and methodologies (Tomazzoli and Sanfilippo 2025). Approaches of machine learning and natural language processing could be explored to develop a pipeline for the (semi-) automatic population of a knowledge base. Furthermore, combining these techniques with ontologies allows for a more in-depth examination of the data by leveraging the strengths of multiple techniques (Pan et al. 2024).

As mentioned in Section 3.1, a second challenge lies in the creation of observational vocabularies for representing the data. It is well known that literary scholars rarely make the intended meaning of their concepts explicit in their essays (Pichler and Reiter 2022). If observational vocabularies are not sufficiently well characterized, the results obtained through the ontology or other methods are correspondingly limited. Interpretative approaches grounded in narratology, like the ones presented by Flüh et al. 2021, as well as studies in (computational) linguistics on the development of controlled vocabularies (Ide and Pustejovsky 2017) and computational literary studies on corpora and their analyses (Schöch et al. 2023), may offer useful resources in this respect, as they provide vocabularies, concepts, and methodologies that have been studied in a systematic and schematic manner.

In conclusion, the proposed method is neither an attempt to formally represent literary criticism *tout court*, nor a novel way of practicing criticism through formal means. Rather, the ontology provides a framework to represent selected portions of the literary debate in a way that makes them *analyzable*, *comparable*, *replicable*, and *interoperable* through computational tools. Hence, its potential to support scholarly inquiry can be ultimately demonstrated only through close collaboration with domain experts who are willing to examine their corpora through the lens of formal models.

6. Software Availability

The ontology files can be found here: https://github.com/appliedontolab/MITE

7. Acknowledgements

This research is supported by the projects: *Make it explicit: Documenting interpretations of literary fictions with conceptual formal models* (MITE) funded by the European Union – Next Generation EU, Mission 4, Component 1, CUP B53D23028830001; *Future Artificial Intelligence Research* (FAIR) funded by the European Union – Next Generation EU, PNRR MUR PE0000013. We would like to thank the anonymous reviewers for their feedback on an earlier version of this paper.

8. Author Contributions

Emilio M. Sanfilippo: Investigation, Conceptualization, Data curation, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Funding acquisition,

Project administration

Claudio Masolo: Investigation, Conceptualization, Methodology, Formal analysis, Writing – review & editing

Alessandro Mosca: Investigation, Conceptualization, Methodology, Formal analysis, Writing – review & editing

Gaia Tomazzoli: Investigation, Methodology, Writing – review & editing, Resources, Validation, Funding acquisition

References

Abreu de Lima, Heloísa (2025). "Interpreting Beatrice: The Critical Reception of the Character in the Last Twenty-Five Years". In: *Humanities* 14.6, 1–18.

Baader, Franz (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge university press.

Barolini, Teodolinda (2012). "Per una storia della letteratura italiana in considerazione del genere sessuale, con una discussione sulla *Beatrix loquax* di Dante". In: *Il secolo di Dante*. Milano: Bompiani. Chap. 16, 553–583.

Bruno, Enrica, Valentina Pasqual, Francesca Tomasi, et al. (2024). "Italo Calvino's 'Destini incrociati'. An Experiment of Semantic Narrative Modelling and Visualisation". In: *Umanistica Digitale* 17.8, 47–69.

Dallas, Costis (2016). "Jean-Claude Gardin on Archaeological Data, Representation and Knowledge: Implications for Digital Archaeology". In: *Journal of Archaeological Method and Theory* 23.1, 305–330.

Daquino, Marilena, Valentina Pasqual, and Francesca Tomasi (2020). "Knowledge Representation of Digital Hermeneutics of Archival and Literary Sources". In: *JLIS: Italian Journal of Library, Archives and Information Science= Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione: 11, 3, 2020*, 59–76.

Di Pasquale, Alessio, Valentina Pasqual, Francesca Tomasi, and Fabio Vitali (2024). "On Assessing Weaker Logical Status Claims in Wikidata Cultural Heritage Records". In: *Semantic Web* 15.6, 2395–2417.

Doerr, Martin, Athina Kritsotaki, and Katerina Boutsika (2011). "Factual argumentation—a core model for assertions making". In: *Journal on Computing and Cultural Heritage (JOCCH)* 3.3, 1–34.

Dung, Phan Minh (1995). "On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-person Games". In: *Artificial intelligence* 77.2, 321–357.

Flüh, Marie, Jan Horstmann, Janina Jacke, and Mareike Schumacher (2021). *Toward Undogmatic Reading: Narratology, Digital Humanities and Beyond*. Hamburg University Press.

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.

Guarino, Nicola, Daniel Oberle, and Steffen Staab (2009). "What is an Ontology?" In: *Handbook on ontologies*. Springer, 1–17.

conference version

Hitzler, Pascal and Bijan Parsia (2009). "Ontologies and rules". In: *Handbook on ontologies*. Springer, 111–132.

Ide, Nancy and James Pustejovsky, eds. (2017). *Handbook of Linguistic Annotation*. Springer, IX–1459.

Lamarque, Peter (1990). "Reasoning to What is True in Fiction". In: *Argumentation* 4.3, 333–346.

Marsili, Neri (2024). "The Definition of Assertion: Commitment and Truth". In: *Mind & Language* 39.4, 540–560.

Masolo, Claudio, Emilio M. Sanfilippo, Emanuele Bottazzi, Roberta Ferrario, Alessandro Mosca, and Marta M Vilardo (2025). "An Observational Approach to Representing Interpretation". In: *Applied Ontology*, 1–23.

Mizoguchi, Riichiro and Stefano Borgo (2025). "An Ontology of Representation". In: *Applied Ontology* 20.2, 137–152.

Pan, Shirui, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu (2024). "Unifying large language models and knowledge graphs: A roadmap". In: *IEEE Transactions on Knowledge and Data Engineering* 36.7, 3580–3599.

Paolini Paoletti, Michele, Jansan Favazzo, and Francesco Orilia (2025). *The Identity of Fictional Characters: A Philosophical Survey*. EUM Edizioni Università di Macerata.

Pianzola, Federico, Luotong Cheng, Franziska Pannach, Xiaoyan Yang, and Luca Scotti (2025). "The GOLEM Ontology for Narrative and Fiction". In: *Humanities* 14.10, 193.

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.

Porello, Daniele, Emilio M. Sanfilippo, and Alessandro Mosca (2025a). "Measuring Similarities of Literary Characters". In: *Formal Ontology in Information Systems*. IOS Press, 110–124.

Porello, Daniele, Walter Terkaj, Laure Vieu, Emilio M. Sanfilippo, and Francesco Compagno (2025b). "Approximating DOLCE in OWL: The DOLCEbasic and DOLCEnaryRel Core Modules". In: *Applied Ontology* 20.4, 298–329.

Sandri, Marta, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek (2023). "Why don't you do it right? analysing annotators' disagreement in subjective tasks". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2428–2441.

Sanfilippo, Emilio M. (2021). "Ontologies for Information Entities: State of the Art and Open Challenges". In: *Applied ontology* 16.2, 111–135.

Sanfilippo, Emilio M. and R. Ferrario (May 2024). *D3.1 - Observations Modeling: State of the Art*. Project Report. Funded by European Union – Next Generation EU. MITE – Make it explicit: Documenting interpretations of literary fictions with conceptual formal. https://www.loa.istc.cnr.it/mite/wp-content/uploads/2024/09/MITE_D3_1.pdf.

Sanfilippo, Emilio M., Richard Freedman, and Alessandro Mosca (2025a). "Ontological modeling of music and musicological claims. A case study in early music". In: *International Journal on Digital Libraries* 26.2, 10.

Sanfilippo, Emilio M., Claudio Masolo, Emanuele Bottazzi, and Roberta Ferrario (2024a). "Interpreting Texts and Their Characters". In: *Formal Ontology in Information Systems*. SAGE Publications, 119–133.

conference version

Sanfilippo, Emilio M., Claudio Masolo, Alessandro Mosca, and Gaia Tomazzoli (2024b). "Operationalizing Scholarly Observations in OWL". In: *Proceedings of the Fourth International Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH)*. Vol. 3809. CEUR Workshop Proceedings, 1–12.

Sanfilippo, Emilio M., Claudio Masolo, and Gaia Tomazzoli (2025b). "Interpreting Literary Characters Through Diagnostic Properties". In: *Humanities* 14.11, 213.

Sartini, Bruno, Sofia Baroncini, Marieke van Erp, Francesca Tomasi, and Aldo Gangemi (2023). "ICON: An Ontology for Comprehensive Artistic Interpretations". In: *ACM Journal on Computing and Cultural Heritage* 16.3, 1–38.

Schöch, Christof, Julia Dudar, and Evgeniia Fileva, eds. (2023). *Survey of Methods in Computational Literary Studies (= D 3.2: Series of Five Short Survey Papers on Methodological Issues)*. CLS INFRA. https://methods.clsinfra.io/.

Tomasi, Francesca, Gioele Barabucci, and Fabio Vitali (2021). "Supporting complexity and conjectures in cultural heritage descriptions". In: *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*. Vol. 2810. CEUR Workshop Proceedings, 104–115.

Tomazzoli, Gaia and Emilio M. Sanfilippo (2025). "Dal testo letterario al testo critico: sfide per un modello formale sull'interpretazione". In: *Letteratura e intelligenza artificiale: un dialogo interdisciplinare*. Ed. by Tiziana Catarci, Agnese Macori, and Daniel Raffini. Lithos.

Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1, 1–9.

conference version

jcls. Journal of Computational Literary Studies

conference version

From Literary Criticism to Literary Studies

Topic Modeling Argentine Academic Journals (1982–2024)

Federico Gabriel Cortés[1]
Matei Chihaia[1]
Juan Manuel Franca[2]

1. Interdisziplinäres Zentrum für Editions- und Dokumentwissenschaft, Bergische Universität Wuppertal, Wuppertal, Germany.
2. Facultad de Informática, Universidad Nacional de La Plata, La Plata, Argentina.

OPEN ACCESS

Citation
Federico Gabriel Cortés, Matei Chihaia, and Juan Manuel Franca (2026). "From Literary Criticism to Literary Studies. Topic Modeling Argentine Academic Journals (1982–2024)". In: *CCLS2026 Conference Preprints* 5 (1). 10.26083/tuda-7981

Date published 2026-05-05 (preprint)
Date accepted tbc.
Date received 2026-01-05

Keywords
Argentine Literary Studies, Computational Literary Studies, Topic Modeling, Academic Journals

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 5th Annual Conference of Computational Literary Studies in Potsdam, Germany, in May 2026. Please check jcls.io for the final journal version.

Abstract. Argentine literary criticism has been defined since the 1950s by an orientation toward literature's outside – the practice of addressing literature through its social, cultural, and political dimensions. This article examines how that orientation transformed across two transitions: from literary criticism to academic literary studies after 1983, and from print to digital open access between 2008 and 2015. We analyze 1,967 articles from seven Argentine academic journals – transitional publications that experienced both transformations – using BERTopic topic modeling. Following Katherine Bode's performative approach to computational literary studies, we interpret the resulting topic configurations as operative vocabulary: naming practices through which Argentine literary studies has approached its object. The analysis traces a shift in this vocabulary after 2015 – from culture and politics to gender, body, and violence.

1. Introduction

Since the 1950s, literary criticism in Argentina has been defined by an orientation toward literature's outside[1] – the practice of addressing literature through its social, cultural, and political dimensions. The literary journal *Contorno* (1953–1959) marks the foundational moment of this orientation: confronting the aesthetic formalism of *Sur* – where Jorge Luis Borges was a founding member – its critics reinterpreted the Argentine literary tradition by linking literature to historical and social processes, under the decisive influence of Sartre's theory of committed literature (Cella 1999). This critical position established a lasting pattern: from *Contorno* onward, the most influential traditions in Argentine literary criticism have consistently privileged literature's external dimensions over its internal procedures. Critical historiography has traced how this orientation evolved: from 'literature and society' in the 1950s to 'culture and politics' from the 1980s onward, a shift shaped by the reception of cultural studies in Argentina (Dalmaroni 2004).

This article traces how that orientation transformed across two institutional transitions. The first, from traditional literary criticism to academic literary studies, accompanied

1. We use "outside" following the standard English translation of Blanchot's *dehors* (see, e.g., Allen, *Blanchot and the Outside of Literature*, 2018). Our use of the concept draws specifically on Blanchot's account of the tension between literature and culture in "Qu'en est-il de la critique?" (Blanchot 1949), as developed in the Critical History section.

the professionalization of the discipline following the restoration of democracy in 1983, when universities regained prominence and specialized journals consolidated within disciplinary frameworks. The second, from print to digital open access publishing, occurred between 2008 and 2015, coinciding with broader transformations in Argentina's cultural and political landscape. The academic journals in our corpus experienced both transitions: founded as print publications in the 1980s and 1990s, they transitioned to digital open access while maintaining institutional continuity.

Using BERTopic, we analyze 1,967 articles from seven Argentine academic literary journals spanning 1982 to 2024. Following the performative approach of Bode (2023), we conceive computational analysis not as a transparent instrument for revealing pre-existing patterns, but as a practice that participates in constituting its object of inquiry. The clusters that emerge from the model can be read as configurations of operative vocabulary through which Argentine literary studies has approached and constructed its object.

Our hypothesis is that the orientation toward literature's outside persisted through both transitions, even as the vocabulary through which that outside is named was significantly reconfigured. Our findings suggest a subsequent transformation: the decline of explicitly cultural and political vocabulary coincides with the emergence of a new configuration organized around gender, violence, and body. What persists through these reconfigurations is the practice of addressing literature through its outside, even as the specific vocabulary shifts. This continuity remains largely unexamined from a computational perspective, and existing accounts of the field's transformation – focused either on print journals that ceased publication or on journals born digital – have not addressed the trajectory of transitional academic journals that experienced both institutional shifts.

2. Critical History: from Literary Criticism to Literary Studies

The origins of literary criticism in Argentina are fundamental to understanding the corpus analyzed in our research, since from its inception the field has been defined by a representational dilemma: should critics explain literature by referring to reality or by attending to its internal logic.

In the introduction to the influential volume 10 of *Historia crítica de la literatura argentina*[2], Cella studies the beginning of critical discourse on literature in Argentina and establishes that the 'irruption of criticism' took place in 1955[3]. The main reference for this inaugural moment is the literary and cultural journal *Contorno* (1953-1959), whose primary task was to reinterpret the Argentine literary tradition from a particular position: they confronted *Sur* magazine (Jorge Luis Borges was one of its most famous members) in order to leave aside the romantic idea of literature and review the relationship between art and politics, linking literature to historical events in a new way. Thus, the irruption of literary criticism in Argentina was characterized by relating literature to society (Cella 1999).

2. *Critical History of Argentina's Literature.*
3. 1955 marks the military dictatorship that ended the first Peronist government, a period in which the various social, political and economic changes that took place meant a significant improvement for the working class and the middle class (Cella 1999).

conference version

conference version

From the 1970s, there was an expansion of the frontiers of the literary (Panesi 1998), with the appearance of renowned critics like Beatriz Sarlo, who began to study not just literature but also the mass media and the cultural industry. However, the literary critic's role remained the same as in the previous period: to intervene in public debate and demystify ideologies through the reading of literary texts and other cultural objects – deeply influenced by the reception of Roland Barthes's work (Podlubne 2021). One of the most representative journals of the period is *Los libros* (1969-1978). This period was marked by a strong politicization of critical discourse during determining historical events for Argentine society. The end of *Los Libros* coincided with the beginning of the last military dictatorship, as the closing of many literary and journals was due to the political violence, human rights violations, and censorship the dictatorship imposed. This moment is also characterized by the circulation of French structuralism, alongside Marxism, to study language and literature. The essential role of Structuralism was to introduce a scientific approach into critical discourse, which was linked to the expansion of the object of study beyond the strictly literary. Nevertheless, literature remained the privileged model for understanding other objects and discourses, such as those of the mass media.

Research on *Los Libros* indicates that the initial emphasis on developing a scientific approach to literature began to shift towards an interest in politics. Studies speak of a shift from literary criticism to a "political criticism of culture" (Panesi 1998), which considered the subject who reads and writes as a social entity and literature as an institutionally produced code in correlation with other surrounding codes and discourses.

The broader interest in culture can be found in the creation of *Punto de vista*, the journal that succeeded *Los Libros* and shared many of its members, among whom Beatriz Sarlo stands out. *Punto de vista* (1978-2008) became one of the most influential journals in Argentina's literary and intellectual history. The central challenge for modern criticism since its emergence in the 1950s, as identified by Dalmaroni (2004), is determining the most reliable method for establishing a connection between 'literature' and 'society'; but from *Punto de vista* onwards, this dichotomy shifts to 'culture' and 'politics'. This transformation was strongly influenced by the reception and circulation of English cultural studies in Argentina, a process that occurred in the pages of *Punto de vista* but also extended to other Latin American countries (Richard 2001).

Although *Punto de vista* was published for three decades, it gradually lost the strong influence it had in its origins. This shift coincided with wider changes in Argentina's cultural and institutional landscape following the restoration of democracy in 1983. The reopening of public institutions and the revitalization of national universities in the 1980s marked a new phase in the country's intellectual life. Significantly, all the academic journals in our corpus were founded during this period, as universities regained prominence as centers of critical thought and cultural production.

While this institutional transformation led to greater professionalization and academicization of literary criticism – transforming it into what we now recognize as academic literary studies – our hypothesis is that the constitutive orientation toward literature's outside persisted through these transformations, even as the vocabulary through which that outside is named was significantly reconfigured. We understand the orientation toward 'literature's outside' following Blanchot's description of the tension between

literature and culture: where culture operates as a dialectical machine of assimilation and identification of values, literature not only constitutes a transgression force but also functions as impugnation – a *puissance* that contests established powers, including itself as power (Blanchot 1971).

Criticism inherits this tension between literature and culture. In "Qu'en est-il de la critique?" (Blanchot 1949), Blanchot identifies a constitutive duality: criticism operates as mediator within institutional spaces – the university, journalism – yet simultaneously bears the task of liberating thought from the regime of 'value' in the Nietzschean sense. Criticism thus participates in culture's machinery while potentially suspending it. This essay was translated and published in the Argentine literary journal *Sitio* (1981-1987), a publication whose historical significance for the period of democratic recovery has been studied by Giordano (2015) and Cortés (2022). The reception of Blanchot's theoretical perspective in this context shaped how Argentine literary criticism understood its own practice: not merely as the identification of literary works with cultural values, but as an exercise capable of interrupting that very identification.

Argentine literary journals have historically enacted this tension. While *Contorno*, *Los libros* and *Punto de vista* intervened directly in political and cultural debates, largely independent of university structures, the academic journals in our corpus institutionalized this critical orientation within disciplinary frameworks. Their transition to digital open access constitutes the second shift our research examines. Our computational analysis tracks how the vocabulary through which this orientation is articulated – from the 'social' of earlier criticism to the 'culture' and 'politics' consolidated during the *Punto de vista* era – evolved through both transitions.

Studies on the history of Argentine literary criticism have addressed this transition. The end of influential print journals such as *Punto de vista*, *El ojo mocho*, *Diario de Poesía* and *Confines* between 2008 and 2013 marked a change of era (Bernini 2016). Subsequent scholarship has traced how born digital journals, emerging in the aftermath of the 2001 crisis (Saítta 2013), renewed critical discourse while expanding toward new literary objects, even as the imperative to intervene in the public sphere persisted (Vigna 2021; Vilar 2014; Hernaiz 2012).

Thus, the new scenario of academic publications on literature since 2002 has been described not only by the socio-political-economic crisis that has affected the country, but also by the expansion of Internet networks and digitalization. These studies, however, focus either on journals that ceased publication or on journals born digital. The academic journals in our corpus represent a different trajectory: transitional publications that maintained institutional continuity across both formats.

The significance of this shift toward open, digital dissemination becomes clear within the context of the transformation of Argentina's scientific and academic landscape during the 21st century. As Panesi (2015) argues, literary criticism in Argentina has always been closely linked to the political events that frame it. The defining feature of the shift from literary criticism to literary studies is the profound institutionalization of discourses on literature and language in Argentina between 1990 and 2023. This period saw the consolidation of specialized academic chairs in national universities and the expansion of scientific research careers within the National Scientific and Technical

conference version

Research Council (CONICET), leading to an exponential rise in doctoral research on literary studies (Gerbaudo 2024). This institutionalization not only formalized the field but also diversified its scope, embedding literary inquiry within a broader academic and scientific framework.

Finally, while the transition of literary journals to digital open access publications opens new possibilities for study, enriched by digital tools, we must acknowledge that computational literary studies (CLS) in Argentina remain in an early stage. The reasons for the near absence of CLS in a field where literary studies have such a rich tradition are beyond the scope of this paper. However, we suggest this gap can be partially attributed to the broader challenges Del Río (2015) outlined in her analysis of digital humanities difficulties in the 'Global South'. These challenges include limited access to technological infrastructure, funding constraints, and the need for specialized training, which hinder the growth of computational approaches. In fact, until now there is only one article that can be placed in the CLS field in Argentina: "Estudios literarios y lectura distante: un primer acercamiento a la actualidad de la investigación en las revistas académicas argentinas"[4] published by Lacalle and Vilar in the journal *Anclajes* in 2019. Lacalle and Vilar (2019) were the first to apply distant reading to this field. Although limited by a small corpus (only digital articles published between 2014 and 2015) and the basic lexical features of Voyant Tools, their findings are significant. By tracking the prevalence of the 'novel' and words like 'life' and 'body,' they argue that Argentine criticism tends to privilege the literature-reality pair rather than internal literary procedures.

3. Corpus Description

This paper examines 1,967 scientific articles on language and literature drawn from a selection of Argentinean academic journals (see Table 1). We have considered all the articles published since the first issues of the journals until January 2025 (we excluded from the analysis interviews and book reviews). These journals represent a diverse array of institutional affiliations, geographic locations, and scholarly literary traditions within Argentina. While the journals formally cover language and literature, their orientation is predominantly literary; as we will see in the Results, the marginal presence of linguistic studies in the analysis reflects this disciplinary profile.

The journals were selected on the basis of the following criterion: only those that digitized their complete print archive were included, since the analysis requires continuous coverage across both publication formats. The resulting corpus spans six universities across five Argentine provinces. Three journals belong to institutions in Buenos Aires province — two at the Universidad Nacional de La Plata, located in the Río de la Plata region, and one at the Universidad Nacional del Sur, based in Bahía Blanca, near the border with the Patagonian region — while the remaining four are based in La Pampa, Mendoza, Córdoba, and Santa Fe.

Beyond its analytical function, the corpus constitutes an infrastructural contribution: a structured, freely available dataset of Argentine academic literary production across four decades, with temporal and institutional metadata, that enables further computational

4. "Literary Studies and Distant Reading: a First Approach to current research in Argentine Academic Journals".

conference version

research.

Name	Institution	Province	Total articles	First printed number	First digital number
Anclajes	Universidad Nacional de La Pampa	La Pampa	390	1997	2016
Auster	Universidad Nacional de La Plata	Buenos Aires	156	1996	2018
Cuadernos de Literatura	Universidad Nacional del Sur	Buenos Aires	250	1982	2020
Cuadernos del CILHA	Universidad Nacional de Cuyo	Mendoza	366	1999	2011
Orbis Tertius	Universidad Nacional de La Plata	Buenos Aires	473	1996	2015
Revista de Culturas y Literaturas Comparadas	Universidad Nacional de Córdoba	Córdoba	191	2007	2015
Saga. Revista de Letras	Universidad Nacional de Rosario	Santa Fe	141	1987	2014

Table 1: Journals

Despite their heterogeneity in content focus, regional representation, and editorial practices, these journals share at least three notable characteristics that provide a cohesive framework for analysis. First, the majority of the journals began publication in the late 1980s or, predominantly, during the 1990s. This period aligns with a significant phase of institutionalization in literary studies in Argentina, driven by the strengthening of humanities programs and research centers at public universities (Gerbaudo 2024). The proliferation of academic journals during this era reflects the broader cultural and intellectual context of Argentine post-dictatorship, where democratic recovery fostered renewed interest in scholarly and cultural production (Patiño 1997).

In second place, all journals transitioned to digital editions between 2010 and 2020, a shift significantly influenced by the enactment of Law 26,899 on the Creation of Open Access Institutional Repositories[5]. This was a major step for Argentine academic publishing as it introduced the requirement for the creation of institutional repositories to guarantee open access to publicly funded research. The transition to digital formats not only expanded the accessibility of academic journals but also enabled retrospective digitization of earlier print issues. The journals in the corpus digitized their complete archives, making decades of academic contributions readily available online. This effort highlights the journals' commitment to preserving their intellectual legacy simultaneously leveraging the capabilities of digital technology.

Third, a particularly significant feature of the corpus is that all the journals are Open Access, published via the Open Journal System (OJS)[6], a characteristic that aligns with a broader trend within the Argentine academic publishing landscape. In Argentina, the majority of scientific journals, irrespective of discipline, adhere to Open Access principles (Beigel and Salatino 2015). This open model ensures the dissemination of knowledge without financial barriers, promoting an inclusive academic environment where researchers, professors and students can engage with scientific discourse. The

5. Accessible on: https://www.argentina.gob.ar/normativa/nacional/ley-26899-223459
6. For a description of OJS implementation at FaHCE-UNLP – the institution that contributes the largest number of articles to our corpus – see Unzurrunzaga et al. (2015)

conference version

accessibility of these journals amplifies their reach, allowing them to serve as platforms for dialogue not only within Argentina but also across Latin America and beyond, enhancing their visibility and impact.

4. Limitations

Although the digitalization and the implementation of diamond Open Access in Argentinean academic journals represented a leading public policy in the Latin American academic context (Cabrera Peña 2015), it is necessary to note that the process had its limitations. Journals frequently lack consistent metadata, and articles are published in unstructured formats, which hinders computational processing and limits the application of other analytical methodologies, such as bibliometrics. Some journals provide HTML versions of articles, but this was often limited only to the most recent issues, excluding the complete historical archives.

Moreover, the extensive corpus of traditional literary and cultural journals from the twentieth century, which had a decisive impact in the following institutionalization of literary studies as explained in section 'Critical History', has only been made digitally available in their facsimile editions, through the project *Archivo Histórico de Revistas Argentinas*[7]. In this sense, despite the decisive role these journals played in the country's intellectual history, until today there are no editions encoded according to TEI standards.

In this context, it is necessary to point out that other implementations of BERTopic on interdisciplinary scientific article corpora, such as Wang et al. (2023), Samsir et al. (2023), and Kim et al. (2024), rely on scientific databases like Scopus, Web of Science, and Library and Information Science, which have robust data structures. Meanwhile, topic modeling implementations in literary studies with Romance-language texts, such as Schöch (2021) and Völkl et al. (2022), not only draw on TEI-encoded texts but explicitly affirm that structured encoding is useful for their analysis.

Taking this into account, the main limitation of our research is that article information is not properly structured (title, keywords, bibliography, content), meaning all this information is embedded along with metadata within each text file. To achieve an interpretable topic distribution, our methodological choices were specifically designed to overcome these challenges. As a starting point, our data processing involved converting the .pdf versions of the articles into .txt format. We then semi-automatically extracted the publication year and journal for every article.

A further methodological consideration concerns the relationship between computationally generated clusters and scholarly categories. Even when clustering operates in semantic embedding space rather than through lexical co-occurrence, an interpretive gap remains between document proximity in a high-dimensional model and the identification of coherent configurations of scholarly practice. As Dobson (2021) argues, computational models in the humanities require interpretability at every level of the workflow: features, weights, and parameters must be exposed so that readers can evaluate whether the model's outputs are semantically meaningful. In our case, the vocabulary configurations presented in Table 2 are computationally mediated entry

7. See: https://ahira.com.ar/

conference version

points into the corpus, not transparent representations of pre-existing thematic categories. They require — and receive — qualitative validation against the documents themselves and the critical historiography of the field. The risk of apophenia — projecting coherent meaning onto statistically distinctive but potentially arbitrary word groupings (Shadrova 2021) — is addressed through our performative framing, which treats topics not as discovered themes but as configurations of operative vocabulary co-constituted by the computational process and the interpretive decisions that structure it.

5. Methodology

This section describes both the process and output structure of our BERTopic implementation, following Bode's proposal on performative computational literary studies. Our research goal is to analyze the literary studies landscape in Argentina, attending to two transitions: from traditional literary criticism to academic literary studies, and from print to digital open access format in academic and scientific publishing. Following Bode (2023), we do not consider computational analysis as a transparent instrument for revealing pre-existing patterns in the corpus, but as a material practice that participates in constituting – and is constituted by – its object of inquiry.

Bode departs from the observation that, despite their apparent differences, arguments both for and against computational literary studies are united by their insistence that computation only relates to literary phenomena by representing them. Representationalism holds that what is represented is independent of the practices of representation. Furthermore, representations, in their mediating function, are more directly accessible to the knowing subject than what is allegedly represented. This epistemological logic resonates with the Argentine critical landscape discussed in the Critical History section: criticism has consistently approached the literary object through its outside – social, cultural, political dimensions that function as more accessible mediations than the literary object itself. For Bode, computational literary studies, rather than maintaining the separation between computation and literary phenomena – already dismantled theoretically through critiques of Cartesian dualism – should consider how these two heterogeneous instances can participate in constituting each other. She thus proposes a performative approach to subjectivity, materiality, and technology, grounded in the premise that computational methods and objects are co-constituted.

In this respect, our approach conceives disciplinary vocabulary as constituting how the field approaches and constructs its object: the decision to name and address literary phenomena from a particular theoretical position shapes what becomes visible and what remains unexamined in research and teaching in the Argentinean context (Cortés et al. 2024). We consider the topics that emerge from our analysis as configurations of operative vocabulary – instances of terminological deployment detected across the corpus through the intersection of transformer embeddings, clustering algorithms, and the critical historiography that structures our interpretive decisions.

Furthermore, in order to situate this approach within performative computational literary studies, we strictly detail the logic behind our parameter selection. Given BERTopic's modularity, these configurations are decisive, as they actively shape the

conference version

resulting output.

BERTopic operates through a two-stage architecture (Grootendorst 2022). The first stage groups documents by semantic similarity through three sequential steps: embedding, dimensionality reduction, and clustering. The second stage extracts characteristic vocabulary for each cluster.

This two-stage architecture has significant epistemological implications. Document clustering operates in transformer embedding space before any vocabulary is extracted; the c-TF-IDF procedure that generates characteristic terms operates on already-constituted clusters, providing interpretive entry points rather than defining the groupings themselves. As Shadrova (2021) has argued, statistical distinctiveness does not guarantee thematic coherence — a critique that applies directly to this vocabulary extraction stage: c-TF-IDF identifies terms that are statistically characteristic of each cluster, but the resulting vocabulary does not necessarily constitute a coherent scholarly category. The vocabulary configurations presented in our analysis are labels for semantic structures detected in embedding space — not the structures themselves. The computational pipeline configures its object through a sequence of interpretive decisions (Dobson 2022), from the choice of embedding model to the parameterization of clustering, that precede and condition the vocabulary output.

Central to this first stage is the use of embedding models, which distinguishes BERTopic from techniques such as LDA. These transformer-based models produce vector representations that encode meaning such that semantically similar texts are closer in vector space (Grootendorst 2022). Document embeddings were generated using the paraphrase-multilingual-mpnet-base-v2 model, which combines the MPNet architecture (Song et al. 2020) with the Sentence-BERT training methodology (Reimers and Gurevych 2019), extended to more than 50 languages via the technique Knowledge Distillation (Reimers and Gurevych 2020). Recent evaluations confirm that the paraphrase-multilingual-mpnet-base-v2 excels at preserving data relationships between texts, as measured by rank correlation coefficients (Pavlyshenko and Stasiuk 2025). In addition, Borčin and Jose (2024) concluded that sentence transformers like MPNet benefit from preprocessing the input data.

The preprocessing pipeline performs four operations. First, text normalization removes accents and converts the text to lowercase. Second, a function reconstructs words split by hyphens across line breaks – a common artifact of PDF extraction. Third, tokenization and lemmatization using spaCy's Spanish model (es_core_news_lg) reduces words to base forms. Fourth, stopword removal filters vocabulary at two levels: standard Spanish stopwords and a custom list developed for this corpus.

From a performative perspective in computational literary studies, these preprocessing decisions shape what the model can detect. Lemmatization collapses morphological variants into base forms — políticas, político, política become a single vocabulary item — so that the pipeline operates at the level of conceptual vocabulary rather than surface forms. Documents that share naming practices but deploy them in different grammatical contexts are recognized as proximate. This aligns the unit of analysis with our research question, which concerns operative vocabulary — which concepts the corpus deploys — rather than how those concepts are inflected. A pipeline that preserved morphological

conference version

variation would foreground a different analytical object.

The custom stopword list constitutes a significant intervention where critical knowledge shapes the computational process. It includes metadata artifacts – journal names, publication identifiers, and other traces of digitization that would otherwise contaminate clustering. It also includes general literary terms – 'novel','literature', 'writer', 'work', 'text', 'author', etc. – words so common in literary studies that they appear across all scholarly practices regardless of methodology or approach. By removing these terms, we shape which vocabulary can characterize topics, allowing the analysis to capture singular differences in scholarly practice. Terms such as 'cultura' and 'cultural', though frequent, were deliberately retained because their distribution across topics is uneven: as we will see in the results, they characterize some vocabulary configurations but not others. Since stopword removal operates only at the vocabulary extraction stage — document clustering is determined by transformer embeddings before any vocabulary is extracted — retaining or removing these terms does not alter which documents group together, only which terms emerge as characteristic labels. Removing them would have eliminated a key axis of differentiation that the critical historiography identifies as central to the corpus's development.

Following preprocessing, each document is transformed into a 768-dimensional numerical vector using the embedding model. Documents with similar semantic content receive similar vectors, positioning them near each other in this high-dimensional space. However, as clustering algorithms perform better with lower-dimensional data (Kim et al. 2024), we reduce the 768 dimensions in two stages.

First, Principal Component Analysis (PCA) reduces the embeddings from 768 to 50 dimensions. Second, Uniform Manifold Approximation and Projection (UMAP) further reduces dimensionality while preserving the comparative distance and density of the high-dimensional data (Kim et al. 2024). We configure UMAP with n_neighbors=15, which considers 15 nearest neighbors when constructing local approximations, and min_dist=0.0, which determines how closely points can be packed together in the low dimensional representation – low values produce densely packed regions adequate for density-based clustering (McInnes et al. 2020, p. 23).

Documents are then grouped using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), an algorithm that generates clusters based on the density of data points (Kim et al. 2024) and models noise as outliers, preventing unrelated documents from being assigned to any cluster (Grootendorst 2022). The min_cluster_size parameter, set to 100, establishes that a cluster must contain at least 100 documents to constitute a topic; smaller groupings are designated as outliers. With approximately 2,000 documents, following (Samsir et al. 2023) we seek to identify substantial scholarly tendencies rather than small variations.

Once documents are assigned to clusters — or to the outlier category — the second stage extracts the vocabulary that characterizes each grouping. Documents are converted to word-frequency matrices using CountVectorizer. We configure this with ngram_range=(1, 2), capturing both single words and two-word phrases. The min_df=2 parameter requires words to appear in at least two documents.

To identify which words best characterize each cluster, we employ a class-based TF-IDF

conference version

transformation with √TF-BM25(IDF) weighting, following methodological recommendations from Borčin and Jose (2024). Standard TF-IDF identifies words that are frequent within a cluster but rare across the corpus. The square root transformation of term frequency reduces the impact of extremely frequent words, improving topic quality by enhancing both coherence and diversity (Borčin and Jose 2024). This weighting scheme produces, for each cluster, a ranked list of characteristic terms from which we extract the top 10 words. Taking into account that the corpus is constituted by scientific articles on language and literature, each topic can be understood as a situated vocabulary: a configuration of terms that indexes the naming practices through which the corpus constitutes its objects and articulates its critical positions (Cortés et al. 2024).

As discussed in the section on 'Limitations', the attachment of temporal and institutional metadata to each document, via a JSON file, is what enables our subsequent analyses. Publication year, derived from article metadata, allows us to trace topic evolution across the 1982-2025 period. Journal attribution allows us to classify documents by publication format – print or digital – enabling analysis of the material transition to open access publishing. These metadata fields do not influence clustering; they are attached afterwards to enable interpretation of pre-formed clusters in relation to our research questions.

Reliance solely on unsupervised machine learning risks misclassification, which reinforces the critical need for qualitative validation (Kim et al. 2024). Our validation process systematically examined topic interpretability. This included reviewing the characteristic vocabulary of each topic to confirm a recognizable thematic configuration and analyzing documents to verify that semantic groupings align with discernible scholarly practices.

After training and validating, we proceed to examine how the prevalence of the identified topics has shifted over the corpus's temporal span, from 1982 to 2025. This temporal period is divided into nine quantile-based intervals to ensure a balanced number of documents in each segment. For each period, the analysis calculates the percentage of documents assigned to each topic, producing time series visualized as a stacked area chart showing how vocabulary configurations expand, contract, or remain stable over time.

Beyond tracking individual topics, the analysis examines the prevalence of specific recurrent vocabulary across the topic distribution. We first identify which topics contain target terms in their characteristic vocabulary. These target terms – 'cultura', 'cultural', 'social', 'político' – are selected on the basis of the critical historiography developed in the 'Critical History' section, which traces the emergence and transformation of literary criticism and literary studies in Argentina. Terms that appear across multiple topics simultaneously cannot be reduced to either a thematic or a methodological index. Rather, their cross-topic presence reveals that vocabulary is precisely the site where this distinction becomes unstable: when critics deploy these terms, they can function both thematically and methodologically. The shared presence of this vocabulary across otherwise distinct topics can be interpreted as the common ground on which different configurations of literary studies operate.

The second analysis tracks the proportion of documents assigned to topics containing each target term, calculating for each period what percentage of the corpus clusters

conference version

around vocabulary configurations that share these lexical markers. This measure captures not word frequency, but the extent to which documents group around topics characterized by specific naming practices. The third analysis traces individual trajectories for each topic containing target vocabulary, showing how these specific vocabulary configurations rise or decline across periods.

Finally, to examine the material transition from print to digital open access publishing, the analysis classifies each document according to its publication format. Classification relies on journal-specific transition dates: the year when each journal shifted from print to digital. Documents published before a journal's transition year are classified as 'Print'; those published in the transition year or later are classified as 'Digital'. For each format category, we calculate the percentage of documents assigned to each topic, enabling direct comparison of topic distributions across publication regimes. The analysis then computes two metrics for each topic: the shift (Digital percentage minus Print percentage) and the growth factor (Digital percentage divided by Print percentage).

6. Results and Discussion

The BERTopic model trained on the corpus of 1,967 articles produced nine topics, with 205 documents (10.4%) assigned to the outlier category, which is consistent with optimization tests performed by Borčin and Jose (2024). Before examining temporal dynamics and the print-to-digital transition, it is necessary to characterize the vocabulary configurations that structure the corpus's landscape.

In the context of our BERTopic implementation, we understand the resulting topics as computationally detected configurations of operative vocabulary within Argentine literary studies. We proceed from the assumption that articles cluster together because they deploy similar naming practices for addressing the literary object – in the sense that criticism constitutes its object through naming, from particular theoretical and methodological perspectives (Cortés et al. 2024). Literary studies in Argentina, deeply connected to the practice of literary criticism, can thus be understood as specific and relatively autonomous modes of naming, and thus interacting with, literary phenomena. Table 2[8] presents the nine vocabulary configurations that emerged from the analysis.

Under the performative framework outlined in the Methodology, claims about the development of Argentine literary studies emerge from the encounter between what the model detects in the corpus and the critical historiography that informs our reading. The performative approach grounds this encounter in the corpus, where computational patterns and historiographic knowledge converge. The interpretation that follows emerges from this convergence and belongs to neither source alone.

The characteristic terms extracted via c-TF-IDF serve as distinctive labels — vocabulary that differentiates each cluster from the rest of the corpus. The terms listed in Table 2 are ranked by their c-TF-IDF weight, which measures how characteristic a term is of a given cluster relative to the corpus as a whole. This means that the order reflects statistical distinctiveness, not absolute frequency or thematic centrality. A term ranked first is the most distinguishing marker of that cluster, not necessarily the most frequent word

8. Representations are translated to English.

conference version

within it.

Vocabulary configurations

Topic	Count	Name	Representation
-1	205	-1_cultural_cultura_nacional_lengua	['cultural', 'culture', 'national', 'language', 'work', 'write', 'latinoamerican', 'nostalgia', 'social', 'intelectual']
0	403	0_politico_nacional_social_cultura	['political', 'national', 'social', 'culture', 'cultural', 'city', 'image', 'body', 'relation', 'memory']
1	254	1_social_discurso_cultura_relacion	['social', 'discourse', 'culture', 'relation', 'political', 'point', 'idea', 'god', 'cultural', 'human']
2	217	2_violencia_sexual_cuerpo_genero	['violence', 'sexual', 'body', 'gender', 'social', 'femenin', 'mexico', 'masculinity', 'mother', 'discourse']
3	185	3_dios_romano_griego_seneca	['god', 'roman', 'greek', 'seneca', 'death', 'virgilio', 'enea', 'human', 'rome', 'image']
4	181	4_poesia_poema_poeta_poetico	['poetry', 'poem', 'poet', 'poetic', 'verse', 'ovidio', 'voice', 'horace', 'poetic', 'image']
5	156	5_america_colonial_cultural_latinoamericano	['america', 'colonial', 'cultural', 'latinoamerican', 'mexico', 'cultura', 'latino', 'national', 'argentina', 'social']
6	135	6_teatro_memoria_teatral_baudelaire	['theater', 'memory', 'theatric', 'baudelaire', 'aira', 'scene', 'story', 'art', 'version', 'museum']
7	117	7_ruben_cuento_voz_modernismo	['ruben', 'story', 'voice', 'modernism', 'narrator', 'love', 'culture', 'infant', 'body', 'reality']
8	114	8_lengua_espanol_linguistico_lenguaje	['language', 'spanish', 'linguistic', 'language', 'verb', 'linguistic', 'grammar', 'doi', 'structure', 'lexic']

Table 2: Topics Representation

Table 2 provides a static overview of the topic distribution. The most immediately evident characteristic is the dominance of Topics 0 and 1, which together account for 33.4% of the corpus. These two configurations share significant vocabulary overlap: both include 'cultura' and 'social' among their characteristic terms. They differ, however, in some distinctive markers: Topic 0 (403 documents, 20.5%) is defined by 'político' and 'nacional' while Topic 1 (254 documents, 12.9%) foregrounds 'discurso'. At first glance, this shared yet differentiated vocabulary suggests two related but distinct modes of addressing literature as a cultural and social phenomenon.

The remaining topics present more distinctive configurations. Topic 2 (217 documents, 11.0%) clusters around vocabulary of gender, violence, and the body. Topic 3 (185 documents, 9.4%) groups articles deploying terminology of classical antiquity, while Topic 4 (181 documents, 9.2%) constitutes a poetry-centered configuration with references to both theory and canonical authors. Topic 5 (156 documents, 7.9%) is oriented toward Latin American and colonial thematics. Topic 6 (135 documents, 6.9%) combines theatrical and memory vocabulary alongside proper names suggesting attention to both European and Argentine literary traditions. Topic 7 (117 documents, 5.9%) groups vocabulary associated with modernism and narrative. Finally, Topic 8 (114 documents, 5.8%) clusters linguistic terminology.

The hierarchical clustering visualization (Figure 1) offers insight into the semantic distances between topics as computed by the algorithm. Hierarchical clustering measures the similarity between topic representations in the embedding space and arranges them

conference version

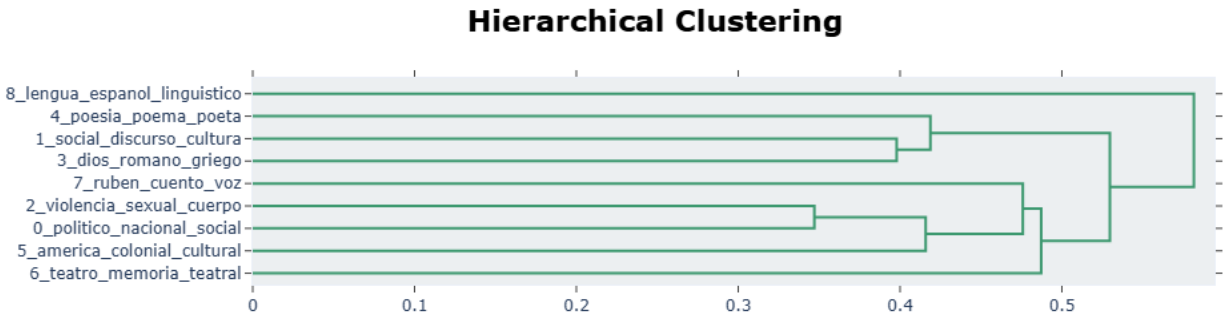


Figure 1: Hierarchical Clustering

in a dendrogram: topics that merge at lower distance values share more similar semantic architectures, while those that merge at higher values are more distinctive. What stands out is that Topics 0 and 2 constitute the closest pair in the entire model (distance 0.347). This suggests that articles addressing topics of gender, violence and body, despite their specialized vocabulary, share a semantic architecture with the broader cultural-political configuration that dominates the corpus. Also, this proximity may be explained, from the perspective of traditional criticism, by the fact that articles in Topic 2 tend to explore how literature addresses themes of gender-based violence, the objectification and control of bodies, and the social implications of these representations.

Equally significant is what the dendrogram separates: Topics 0 and 1, despite their shared surface vocabulary ('culture', 'social') and our first interpretation, do not cluster together. Topic 0 groups with the gender studies, Latin American, and theater configurations (Topics 2, 5, 6, 7), while Topic 1 clusters with classical philology and poetry studies (Topics 3, 4). This separation suggests that the term 'discurso' and 'social' which distinguishes Topic 1, indexes a different mode of addressing the literary object – one that, in semantic terms, aligns more closely with traditional literary-critical approaches than with the cultural studies orientation. At the opposite extreme, Topic 8 (linguistic studies) joins the tree last (distance 0.582), confirming its status as the most distinctive vocabulary configuration in the corpus.

Gender, violence, and body

The temporal distribution of topics (Figure 2) and the stacked area chart (Figure 3) present complementary views of the corpus's evolution across the four decades under study. Both visualizations confirm the same central transformation: the substantial expansion of Topic 2. Marginal until 2013 (1.77% of documents in the earliest period), this configuration accelerates after 2015, reaching 31.52% by 2023-2024 – becoming the single largest topic by the final period. This particular configuration can be connected to two related developments. First, the emergence of a strong feminist academic discourse in Argentina. The feminist movement gained significant visibility and influence during the last decade, a process that crystallized in the publication of the *Historia feminista de la literatura argentina* in 2020, edited by Arnés, De Leone, and Punte. The title deliberately echoes the *Historia crítica de la literatura argentina* mentioned above, signaling a rewriting

conference version

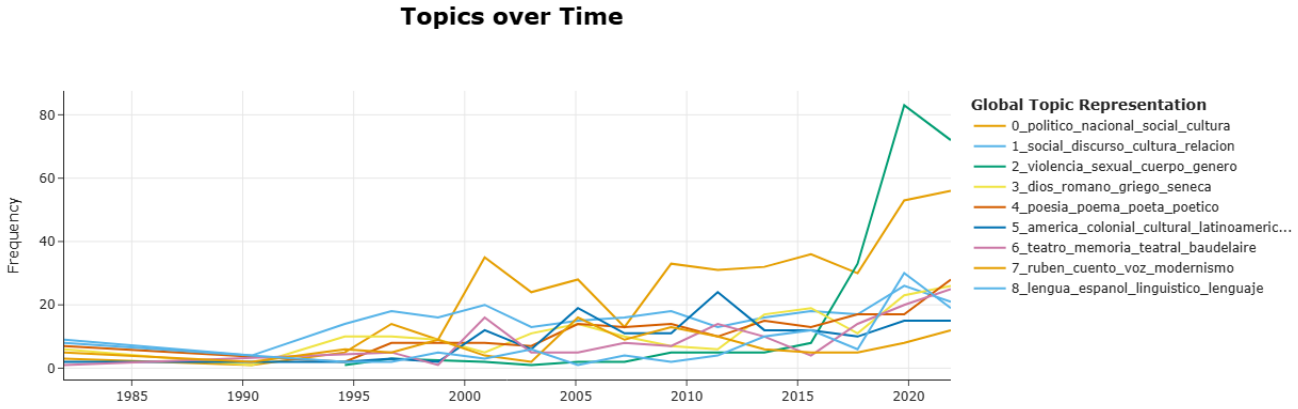


Figure 2: Topics over Time

conference version

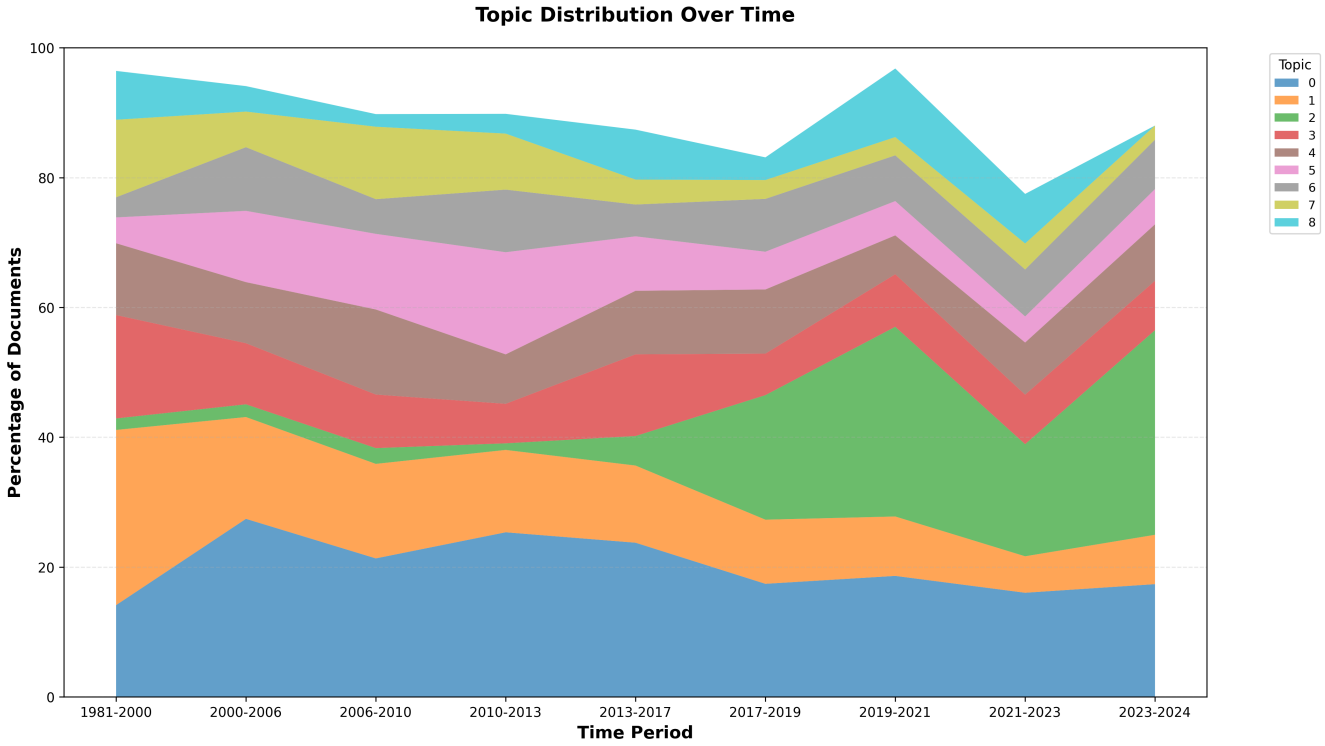


Figure 3: Topics Distribution over Time

of the critical tradition from a feminist position. Second, a broader return to violence as an interpretative framework for literary interpretation across Latin American studies. As Chihaia (2019) argues, the interest in violence marks a return to the committed criticism that characterized Latin American studies in the 1970s and 1980s. The particularity of this 'new commitment' however, is that it must be reevaluated in a context where social utopia crumbles under persistent inequality and postcolonial conditions.

This trajectory traces the emergence and rapid institutionalization of gender, violence, and body vocabulary within Argentine literary scholarship. The turning point coincides with the interpretation proposed in the beginning of *Historia Feminista de la literatura argentina*, "from 2015 onwards, as the feminist movement gained strength, the heterocis-patriarchal structure of Argentine society and culture became strikingly evident once again"[9] (Arnés et al. 2020: 15).

The stacked area chart renders visible what this expansion displaces. The growth of Topic 2 coincides with the contraction of Topic 1, which declines from 26.99% to 7.61%, and Topic 7, which falls from 11.95% to 2.17%. The area corresponding to Topic 2 occupies space previously held by discourse-analytical and modernist configurations. Yet Topic 0 – organized around 'political', 'national' and 'culture' – remains relatively stable throughout, fluctuating between 14% and 27%. Topics 3 and 4, oriented toward classical philology and poetry, also maintain consistent presence, indicating that traditional literary-critical approaches persist alongside the expanding configurations.

This distribution suggests that the cultural-political orientation identified in the critical history does not disappear but persists, and partially transforms into, the newer gender and body configuration. Within the corpus, the orientation toward literature's outside remains constant; what changes is the specific vocabulary through which that outside is addressed.

The distribution of topics across journals (Figure 4) reveals significant concentrations that correlate with editorial profiles. Topic 2 shows its strongest association with *Anclajes*: 176 of its 217 documents (81%) appear in this journal. This concentration does not indicate that gender and body vocabulary is absent elsewhere, but rather that *Anclajes* functioned as the primary editorial space for this configuration's institutionalization – a finding corroborated by the journal's publication of several special issues on gender and violence during the period under study. Similarly, Topic 8 concentrates in *Cuadernos de Literatura* and *Saga*, journals with explicit linguistic orientations, while Topic 3 finds its strongest presence in *Auster*, consistent with its classical philology profile.

Orbis Tertius presents a different case. As the journal with the largest number of articles in our corpus, it contributes substantially across multiple topic configurations. This distribution reflects the journal's scope and institutional weight within Argentine literary studies, which can be seen in the fact that it is the journal that published the highest number of articles from Topic 0 and 1, the most frequent ones.

9. All translations of *Historia Feminista de la literatura argentina* are our own

conference version

conference version

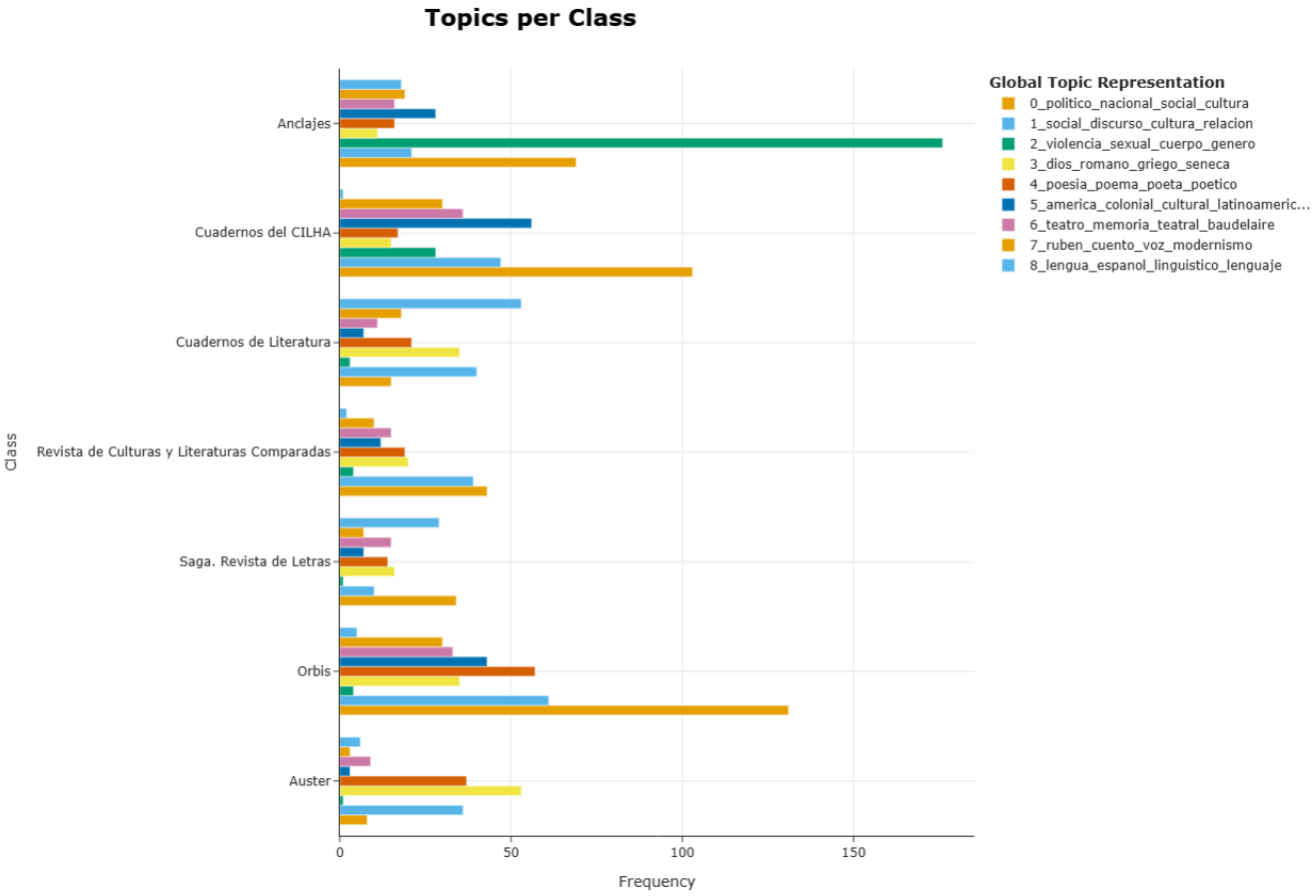


Figure 4: Topics per Class

Cultural-political vocabulary

The preceding analysis characterized vocabulary configurations and their temporal distribution. We now turn to tracking the prevalence of specific terms across these configurations: 'culture', 'cultural', 'político' and 'social'. These four terms correspond to what we identified in our hypothesis as the vocabulary through which Argentine literary criticism has historically articulated its orientation toward literature's outside (Blanchot 1971; Cortés 2021). This orientation – the practice of addressing the literary object through its social, cultural, and political dimensions – has crystallized throughout intellectual history under various methodological designations: sociological, ideological, culturalist criticism. The transition from literary criticism to literary studies institutionalized this orientation: as Dalmaroni (2004) observed, from *Punto de vista* the dichotomy shifted from 'literature and society' to 'culture and politics' shaped by the reception of cultural studies in Argentina.

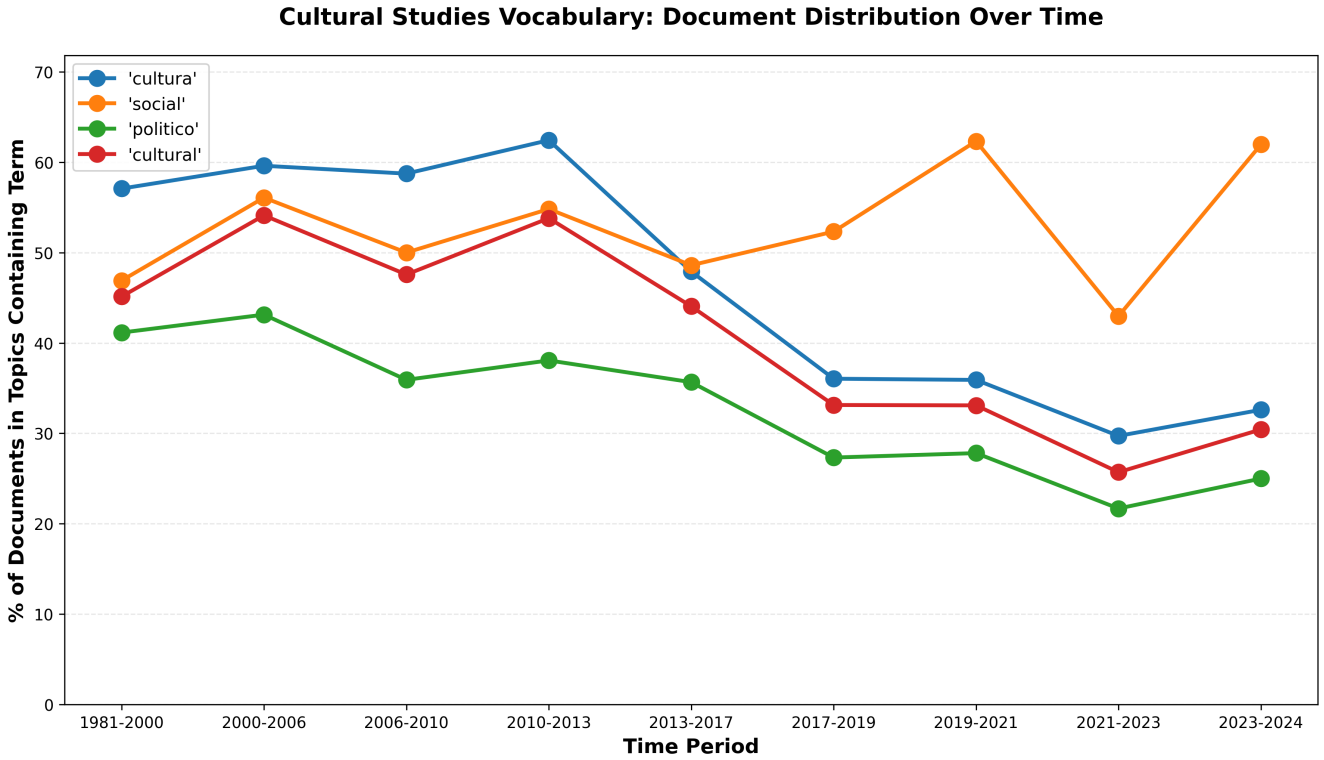


Figure 5: Cultural Studies Vocabulary: Document Distribution over Time

Figure 5 tracks the prevalence of the four target terms across the nine periods. The results partially support and partially complicate our hypothesis. Three of the four terms – 'culture', 'cultural' and 'political' – remain relatively constant until 2013, then follow a consistent downward trajectory. This decline might initially suggest a weakening of the orientation toward literature's outside within the corpus. However, the fourth term – 'social' – diverges from this trajectory, rising in recent periods rather than declining.

This divergence gains significance when read against the critical history described above. If the shift from 'literature and society' to 'culture and politics' characterized the *Punto de vista* era, our findings suggest a partial reversal of that trajectory: the decline of 'culture' and 'political' coincides with the expansion of Topic 2, which carries only 'social' among the target terms. In line with Chihaia (2019), we consider that the emergent gender

conference version

and body configuration does not represent an abandonment of the field's orientation toward literature's outside, but a return – in reconfigured form – to the social dimension that characterized the earliest moment of modern Argentine criticism. What returns, however, is not the Sartrean 'committed literature' of *Contorno*, but a 'social' mediated by contemporary concerns with violence, gender, and the body.

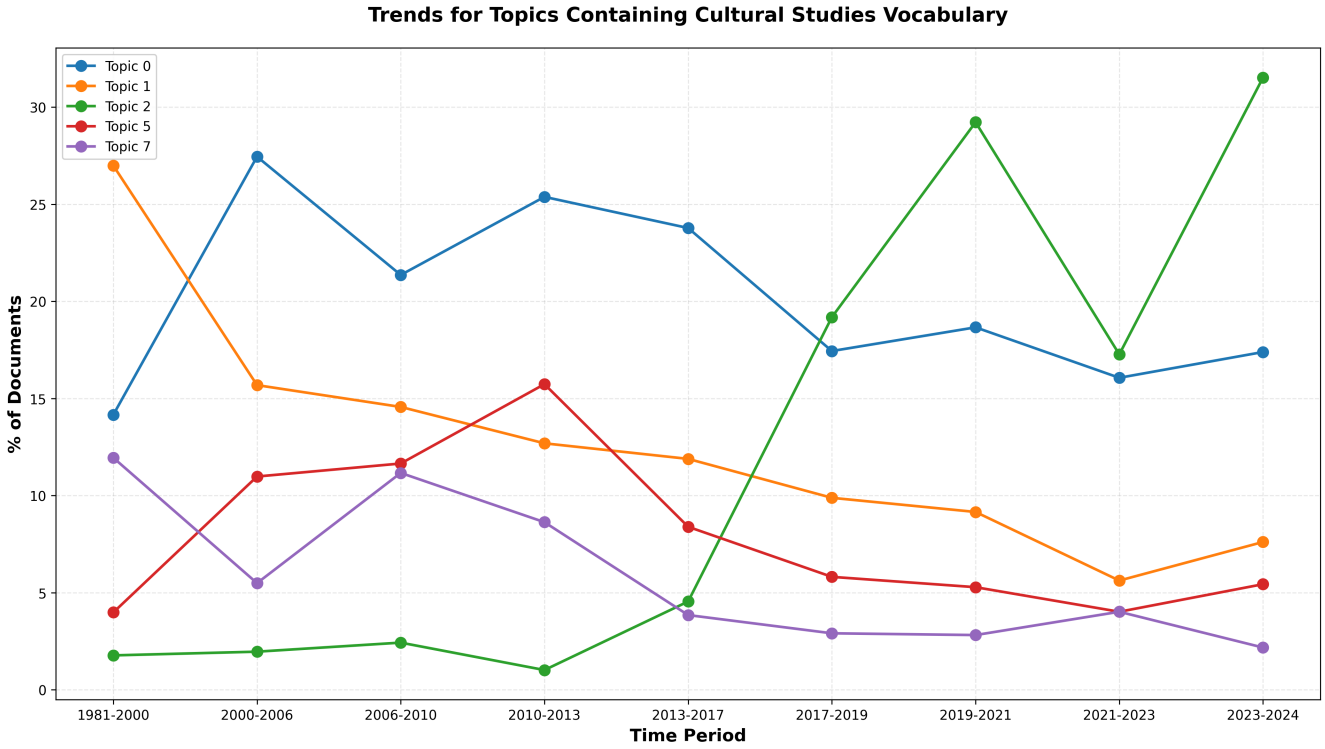


Figure 6: Trends for Topics Containing Cultural Studies Vocabulary

Figure 6 disaggregates this distribution by tracking the evolution of specific topics that contain target vocabulary. The data reveals a crucial asymmetry: while Topics 1, 5, and 7 include multiple target terms in their characteristic vocabulary, Topic 2 contains only 'social', and Topic 0 – despite being organized around 'political', 'national' and 'culture' – does not include 'social' among its characteristic terms. The decline of 'culture', 'cultural' and 'political' corresponds to the contraction of Topics 1 (from 27.0% to 7.6%), 5 (from 4% to 15.7%, then to 5.4%), and 7 (from 11.9% to 2.2%). The persistence of 'social', by contrast, is driven almost entirely by the expansion of Topic 2. Topic 0 remains stable but does not contribute to the 'social' trajectory. This separation suggests that the earlier vocabulary of culture and politics and the emergent vocabulary of the social now coexist as distinct modes of orienting critical discourse toward literature's outside.

The persistence of literature's outside

Finally, the comparison between print and digital eras (Figure 7 and Figure 8) provides a general view of how the transition in publication format intersects with the vocabulary transformations described above. Topic 2 exhibits the most pronounced shift: marginal during the print era, it becomes the dominant configuration in digital publications – confirming that the emergence of gender, violence and body vocabulary is concentrated in the period following the digital transition. Conversely, Topics 1 and 7 are approxi-

conference version

conference version

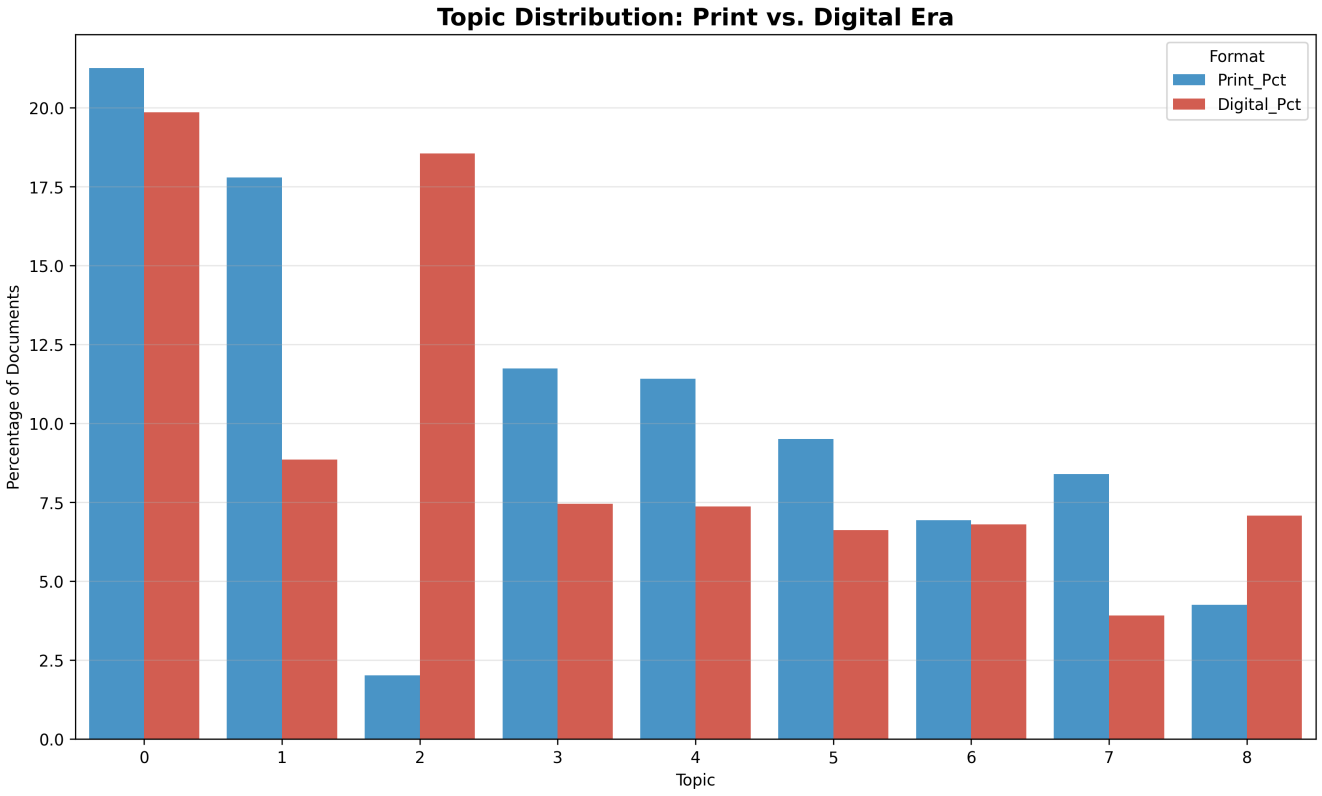


Figure 7: Topic Distribution: Print vs. Digital Era

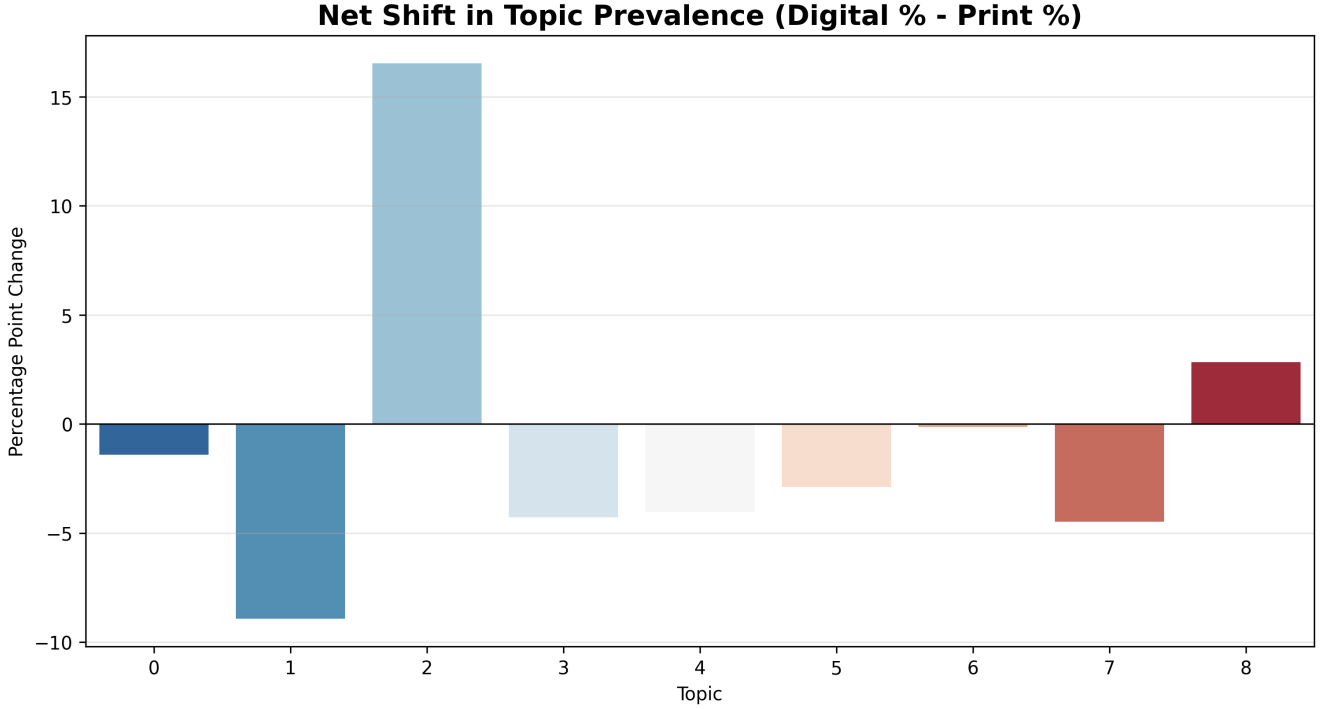


Figure 8: Net Shift in Topic Prevalence (Digital % - Print %)

mately halved, while Topic 0 remains stable across both formats. Besides providing a new platform for academic publishing, the print-to-digital transition coincided with a substantial reconfiguration of the vocabulary through which Argentine literary studies addresses its object.

These findings must be understood within the performative framework outlined in the Methodology. As established there, c-TF-IDF extracts what is distinctive to each configuration, not simply what is present. The decline of 'culture' and 'political', and the rise of 'social' in connection with Topic 2, are products of a computational process that foregrounds semantic distinctiveness. Read performatively, this does not indicate that gender and body studies lack cultural or political commitment – a claim that would contradict the deep implication of these approaches with different forms of political activism.

As Arnés et al. (2020) argue, contemporary feminist and dissident textualities constitute themselves as inseparable from politics and affects; they produce what the authors call an "affective reinvention of the political" (p. 21). Violence is now understood as "stratified and always gendered" (p. 19), and the central questions have shifted from ideological demystification to "what stories are narratable, what bodies are visible, and what narratives are legible" (p. 22). The computational analysis renders visible a reconfiguration in which the cultural and political dimensions of literary studies are increasingly articulated through vocabularies of gender, violence, and body – through social rather than explicitly political terminology.

The computational analysis thus traces a fundamental continuity beneath the transformation. Arnés et al. (2020) acknowledge that contemporary feminist textualities "renew a debate that is never fully settled: the relationship between literature and politics" (p. 29). Our findings confirm this renewal: the vocabulary shifts from culture and politics to gender, body, and violence.

The vocabulary configurations identified through topic modeling can be understood as historically specific modulations of literature's outside within Argentine critical practice. This does not imply that literature's outside is reducible to any set of dimensions – 'social', 'cultural', 'political' or 'gender' and 'violence' – but rather that criticism, operating within particular institutional and historical conditions, necessarily articulates that outside through specific naming practices. What our analysis tracks is not the outside itself, which for Blanchot remains radically heterogeneous to any designation, but the characteristic vocabulary through which Argentine literary studies has approached it.

7. Conclusion

This article analyzed the vocabulary of Argentine literary studies across two institutional transitions – from literary criticism to academic literary studies, and from print to digital open access – by applying BERTopic to 1,967 research articles from seven academic journals selected for their transitional nature: founded as print publications, they experienced both the institutionalization of literary studies and the shift to digital open access. This deliberate construction gives the corpus both an infrastructural value, as a structured and freely available dataset, and a conceptual one, since existing scholarship

conference version

had not addressed the particular trajectory of these transitional journals.

Conceived in performative terms, the findings presented here emerge from the encounter between what the model detects in the corpus – semantic proximities, vocabulary configurations – and the critical historiography that makes those configurations legible as modulations of the field's orientation toward literature's outside. The analysis demonstrates a shift after 2015 – from culture and politics to gender, body, and violence – but the underlying orientation persists: literature continues to be valued for its connection to what exceeds it. What the analysis captures is not abandonment but reconfiguration: a return, in transformed form, to the social dimension that characterized Argentine criticism from its origins in the 1950s. The semantic proximity between the gender-and-body configuration and the cultural-political vocabulary – the closest pair in the hierarchical clustering – suggests that the newer configuration inherits rather than displaces the earlier orientation. Meanwhile, the divergence of 'social' from 'culture' and 'political' complicates any account of a unidirectional shift: as the vocabulary of culture and politics contracts, the social persists through reconfigured naming practices organized around gender, violence, and the body.

The print-to-digital transition coincided with this reconfiguration: the vocabulary of gender, body, and violence, marginal during the print era, became dominant in digital publications. Its rapid emergence opens questions that the present analysis identifies but cannot answer. The embedding model detects that these documents converge, and c-TF-IDF tells us which vocabulary distinguishes them from other configurations, but neither stage reveals what produced the convergence – through which intellectual networks and institutional conditions this vocabulary consolidated. Future work will combine document embeddings with knowledge graphs and bibliometric analysis to move beyond vocabulary configurations toward these relational structures: how authors, concepts, and institutional affiliations connect across the corpus, how citation patterns track or diverge from the proximities the model detects, and how the corpus's intellectual networks have reconfigured alongside its vocabulary.

Computational literary studies in Argentina remain in an early stage. This analysis demonstrates that such approaches can contribute to the history of the discipline by identifying the historically specific vocabulary through which the discipline has approached its object.

8. Data Availability

Data can be found here: https://zenodo.org/records/19367447

9. Software Availability

Software can be found here: https://github.com/fedexx1/BERTopic_ArgentineLiteraryStudies

conference version

10. Acknowledgements

This research was funded by the National Council for Scientific and Technical Research of Argentina and by a Georg Foster grant from the Alexander Von Humboldt Foundation of Germany.

11. Author Contributions

Federico Gabriel Cortés: Project administration, Conceptualization, Writing – original draft, Writing – review & editing, Formal analysis

Matei Chihaia: Conceptualization, Writing – original draft, Formal analysis

Juan Manuel Franca: Conceptualization, Writing – review & editing, Formal analysis

References

Arnés, Laura A., Nora Domínguez, and María José Punte, director (2020). *Historia feminista de la literatura argentina*. Eduvim.

Beigel, Fernanda and Maximiliano Salatino (2015). "Circuitos segmentados de consagración académica: las revistas de Ciencias Sociales y Humanas en la Argentina". In: *Información, cultura y sociedad* 32. http://revistascientificas.filo.uba.ar/index.php/ICS/article/view/1342.

Bernini, Emilio (2016). "Presentación." In: *El matadero* 9. http://revistascientificas.filo.uba.ar/index.php/matadero/article/view/2968/2576.

Blanchot, Maurice (1949). *Lautréamont et Sade*. Les Éditions de Minuit.

— (1971). *L'amitié*. Gallimard.

Bode, Katherine (2023). "What's the Matter with Computational Literary Studies?." In: *Critical Inquiry* 29.4. https://www.journals.uchicago.edu/doi/10.1086/724943.

Borčin, Martin and Joemon Jose (Mar. 2024). "Optimizing BERTopic: Analysis and Reproducibility Study of Parameter Influences on Topic Modeling". In: 147–160. 10.1007/978-3-031-56066-8_14.

Cabrera Peña, Karen Isabel (2015). "Comparative analysis of public policies in open access models in Latin America. Brazil and Argentina cases." In: *RUSC. Universities and Knowledge Society Journal* 12(1). http://dx.doi.org/10.7238/rusc.v12i1.1947.

Cella, Susana (1999). *Historia crítica de la literatura argentina vol. 10*. Emecé Editores.

Chihaia, Matei, ed. (2019). *La violencia como marco interpretativo de la investigación literaria*. Narr Francke Attempto.

Cortés, Federico Gabriel (2021). "La palabra errante en Maurice Blanchot: Literatura como impugnación". In: *Landa* 10(1). https://revistalanda.ufsc.br/vol10-n1-20212/.

— (2022). "Maurice Blanchot en la crítica literaria argentina : recepción y resistencia". PhD thesis. Universidad Nacional de La Plata. Facultad de Humanidades y Ciencias de la Educación. https://www.memoria.fahce.unlp.edu.ar/library?a=d&c=tesis&d=Jte2448.

conference version

Cortés, Federico Gabriel, Miguel Dalmaroni, Verónica Delgado, Analía Gerbaudo, Verónica Stedile Luna, and Santiago Venturini, eds. (2024). *Un vocabulario de teoría: Literatura, enseñanza, investigación*. Ediciones UNL and EDULP. https://www.memoria.fahce.unlp.edu.ar/library?a=d&c=libros&d=Jpm6941.

Dalmaroni, Miguel (2004). *La palabra justa: Literatura, crítica y memoria en la Argentina, 1960- 2002*. Melusina. https://www.memoria.fahce.unlp.edu.ar/libros/pm.1/pm.1.pdf.

Del Río, Gimena (2015). "Humanidades Digitales. Mito, actualidad y condiciones de posibilidad en España y América Latina". In: *Revista Digital de Artes y Humanidades* 1. https://ri.conicet.gov.ar/handle/11336/13563.

Dobson, James E. (2021). "Interpretable Outputs: Criteria for Machine Learning in the Humanities". In: *Digital Humanities Quarterly* 15.2. https://dhq.digitalhumanities.org/vol/15/2/000555/000555.html.

— (2022). "Vector hermeneutics: On the interpretation of vector space models of text". In: *Digital Scholarship in the Humanities* 37.1, 81–93. 10.1093/llc/fqab079.

Gerbaudo, Analía (2024). *Tanto con tan poco: los estudios literarios en Argentina 1958-2015*. Ediciones UNL.

Giordano, Alberto, ed. (2015). *El discurso sobre el ensayo en la cultura argentina desde los 80*. Santiago Arcos Editor.

Grootendorst, Maarten (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv: 2203.05794 [cs.CL]. https://arxiv.org/abs/2203.05794.

Hernaiz, Sebastián (2012). *Rodolfo Walsh no escribió Operación masacre y otros ensayos*. 17 Grises Editora.

Kim, Keungoui, Dieter Kogler, and Sira Maliphol (May 2024). "Identifying interdisciplinary emergence in the science of science: combination of network analysis and BERTopic". In: *Humanities and Social Sciences Communications* 11. 10.1057/s41599-024-03044-y.

Lacalle, Juan Manuel and Mariano Vilar (2019). "Estudios literarios y lectura distante: un primer acercamiento a la actualidad de la investigación en las revistas académicas argentinas". In: *Anclajes* 23.1. https://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1851-%2046692019000100002.

McInnes, Leland, John Healy, and James Melville (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv: 1802.03426 [stat.ML]. https://arxiv.org/abs/1802.03426.

Panesi, Jorge (1998). *Críticas*. Grupo Editorial Norma.

— (2015). "La seducción de los relatos: diez años de crítica argentina (2004- 2014)". In: *Celehis - Revista del Centro de Letras Hispanoamericanas* 24.29. https://fh.mdp.edu.ar/revistas/index.php/celehis/article/view/1257.

Patiño, Roxana (1997). "Intelectuales en transición: las revistas culturales argentinas (1981- 1987)". In: *Cuadernos de Recienvenido* 4. https://dlm.fflch.usp.br/sites/dlm.fflch.usp.br/files/recienvenido04.pdf.

Pavlyshenko, Bohdan and Mykola Stasiuk (Aug. 2025). "SEMANTIC SIMILARITY ANALYSIS USING TRANSFORMER-BASED SENTENCE EMBEDDINGS". In: *Electronics and Information Technologies* 30. 10.30970/eli.30.4.

Podlubne, Judith (May 2021). "Barthes en Sarlo". In: *Cuadernos de Literatura* 24. 10.11144/Javeriana.cl23-47.prbs. https://revistas.javeriana.edu.co/index.php/cualit/article/view/33685.

conference version

Reimers, Nils and Iryna Gurevych (Jan. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: 3973–3983. 10.18653/v1/D19-1410.

— (2020). *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*. arXiv: 2004.09813 [cs.CL]. https://arxiv.org/abs/2004.09813.

Richard, Nelly (2001). "Globalización académica, estudios culturales y crítica latinoamericana". In: *CLACSO*. https://biblioteca.clacso.edu.ar/clacso/gt/20100912041222/11richard.pdf.

Saítta, Sylvia (2013). "En torno al 2001 en la narrativa argentina". In: *Literatura y Lingüística* 29. http://dx.doi.org/10.4067/S0716-58112014000100008.

Samsir, Samsir, Reagan Saragih, Selamat Subagio, Rahmad Aditiya, and Ronal Watrianthos (July 2023). "BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory". In: *JURNAL MEDIA INFORMATIKA BUDIDARMA* 7, 1514. 10.30865/mib.v7i3.6426.

Schöch, Christof (2021). *Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama*. arXiv: 2103.13019 [cs.CL]. https://arxiv.org/abs/2103.13019.

Shadrova, Anna (2021). "Topic models do not model topics: epistemological remarks and steps towards best practices". In: *Journal of Data Mining & Digital Humanities*. 10.46298/jdmdh.7595.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu (2020). *MPNet: Masked and Permuted Pre-training for Language Understanding*. arXiv: 2004.09297 [cs.CL]. https://arxiv.org/abs/2004.09297.

Unzurrunzaga, Carolina, Cecilia Rozemblum, Cristian Pucacco, Gustavo Parente, and Mauricio Esterellas (2015). "OJS Implementation and development of the Scientific Journals Site of the School of Humanities and Education Sciences at the Universidad Nacional de La Plata". In: *Scholarly and Research Communication* 6(1). http://sedici.unlp.edu.ar/handle/10915/89015.

Vigna, Diego (2021). "Breve historización de las revistas digitales de cultura y literatura en Argentina: El caso No Retornable". In: *El Taco en la Brea* 2(14). https://ri.conicet.gov.ar/handle/11336/165085.

Vilar, Mariano (2014). "Revistas, letras, Luthor". In: *Luthor* 22. https://revistaluthor.com.ar/ojs/index.php/luthor/article/view/130.

Völkl, Yvonne, Sanja Sarić, and Martina Scholger (Nov. 24, 2022). "Topic Modeling for the Identification of Gender-specific Discourse: Virtues and Vices in French and Spanish 18th Century Periodicals". In: *Journal of Computational Literary Studies* 1.1. 10.48694/jcls.108.

Wang, Zhongyi, Jing Chen, Jiangping Chen, and Haihua Chen (July 2023). "Identifying interdisciplinary topics and their evolution based on BERTopic". In: *Scientometrics* 129. 10.1007/s11192-023-04776-5.

conference version