Computational Humanities Research

www.cambridge.org/chr

Computational narratology with transformer embeddings‡

William Rathje

University of California, Berkeley, USA

Research Article

Cite this article: Rathje William. 2026. "Computational narratology with transformer embeddings." *Computational Humanities Research*, **2**:e15, https://doi.org/10.1017/chr.2026.10035

Received: 30 September 2025 UTC
Revised: 3 April 2026 UTC
Accepted: 24 April 2026 UTC

Keywords:
computational narratology; natural language processing; sentence transformers

Corresponding author:
William Rathje;
E-mail: wrathje@berkeley.edu

‡This article has been updated since original publication. A notice detailing the change has been published.

Abstract

This article presents a new method for analyzing narrative structure using sentence embeddings. We embed chunked text sequences to create narrative plot trajectory curves. This enables us to apply commonly used unsupervised word embedding techniques, such as semantic axes projection, to narrative sequence analysis at the scale of entire books. Our aim is to prototype several simple examples of the kinds of narratology techniques this method might enable. We apply a pretrained sentence transformer model to books, represented as embeddings of chunked sequences, from nearly 5 percent of the Project Gutenberg English Books Corpus filtered to works in the fiction category (totaling 2,965 texts). Texts date from the 1500s through the early 1900s, with the majority of texts concentrated in the nineteenth century. This period and dataset offer an opportunity to investigate the rise and development of the early novel form. Our analysis measures narrative trajectories along seven binary oppositions – class, gender, morality, collectivity/individual, nature/artifice, order/disorder, emotional/rational, and segments texts by decade, genre, and dialogue gender. We find trends broadly relevant to hypotheses from postclassical and feminist narratological theory and also present novel aggregate findings on narrative presentation and story shape.

Plain language summary

This article presents a method for recovering and comparing story trajectory curves. It draws on sentence embeddings, an extension of word embeddings, which represent words as points in a geometric space such that similarities in word meaning are denoted by similarities in geometric position. Contemporary embedding models, called "contextual embeddings" or "transformer embeddings", can represent full sequences according to proximity in meaning and differentiate between distinct meanings and uses of the same words (polysemy). We use contextual sentence embeddings to represent 2,965 texts from Project Gutenberg fiction, concentrated primarily on the nineteenth century but ranging from the 1500s to the early 1900s. To do this, we divide or "chunk" each book into 128-token units, represent these chunks using contextual sentence embeddings, and structure them into sequences. We then apply a technique regularly used in the literature to construct "axes" representing binary oppositions as continua (in our case gender, class, morality, collectivity/individual, nature/artifice, order/disorder, and emotional/rational) and measure the relative position of each text chunk along these "axes." This allows us to trace representational shifts along binary oppositions over the course of a book's narrative sequence. We compute average sequences across all books and segment them by decade, genre, and dialogue speaker gender. Our analysis finds clear trends for each segment, many of which confirm ideas from narratological theory. While word embeddings have been used in historical language analysis, for example, to trace how words change over time periods, their use in representing books as chunked sequences is less common and builds upon work in book plot arc modeling with techniques like sentiment analysis. Our method is statistically robust and provides a proof of concept for how to analyze texts individually and in aggregate as sequences of embeddings.

Introduction

Narratology developed as the application of structuralist theory to narrative analysis, with a specific and somewhat unusual focus relative to structuralism in general on temporality. Specifically, narratology treats time as conditional, approaching a physical, given temporal sequence of events as independent and separable from its contingent, descriptive rendering in narrative order.

In natural language processing, word embeddings, perhaps more than any other recent technique, are conceptually affiliated with structural linguistics (Lenci and Sahlgren 2023). Operationalizing embedding models for the analysis of text *sequences* has proven challenging, although recent transformer embedding models such as sentence-BERT make this tractable.

This article was awarded the Open Data badge for transparent practices. See the Data availability statement for details.

© The Author(s), 2026. Published by Cambridge University Press This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (https://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use and/or adaptation of the article.

In this article, we present a proof-of-concept framework for the analysis of embedded literary sequences. Using a dataset of nearly 3,000 fiction books from Project Gutenberg, we represent books as sequence trajectories by embedding fixed text chunks with a sentence transformer model and then projecting chunks onto seven predefined semantic axes. We then aggregate individual text sequence trajectories into composite, mean trajectories to segment by variables such as time period, genre, or author/speaker gender.

Although the use of embeddings for time-series analysis is becoming increasingly common in multimodal embedding applications, analyzing temporal sequences with embeddings at the level of the text rather than the corpus is relatively less common, although it has been done before with success (Toubia, Berger, and Eliashberg 2021). Sequence analysis using embeddings has multiple advantages, however, for studying narratology: (1) it enables comparisons between the structural corpus level and the level of the individual text and (2) it enables the recovery of semantic patterns of temporal ordering and movement rather than static semantic associations. At the same time, ensuring that trajectories of embeddings are statistically real effects distinct from random noise, especially in aggregate, is a key challenge in any sequential embedding analysis.

In this article, we present our approach to the analysis of sequence semantics with transformer embeddings. We outline the method and evaluate its robustness, and then we apply it to the study of several corpus-level variables. We find that the approach accurately captures known genre and period-specific effects and that statistically it is highly robust compared to null comparison models. We conclude by suggesting that this method shows promise for the analysis of narrative sequences in general. (This includes narrative as broadly construed in the social sciences and humanities, and one could imagine extending this pipeline to corpora of documents, scientific articles, speeches, etc.)

Adapting word embeddings to narratology

Word embedding models such as word2vec (Mikolov et al. 2013) have been used widely in computational linguistics and social science to quantify and measure semantics. Word2vec originally emerged as an interpretive adaptation of historically "black box" neural network language models trained to predict sequences. Word2vec learned a high-dimensional geometric representation of the word (token) distribution such that semantically similar words mapped onto geometrically similar vector positions (as measured using dot product). Embedding models were trained by iteratively sampling the token co-occurrence distribution (i.e., the words that most frequently co-occur when sampling from large corpora) and solving for an optimal vector parameterization grounded in these corpus-wide token co-occurrence relations. In practice, this allows for measuring word similarity using geometric measures like cosine similarity and was famously shown to solve analogies with simple arithmetic, for example, king + (woman − man) = queen.

Word embeddings are usually theorized in terms of distributional semantics (Lenci and Sahlgren 2023), and this literature generally highlights overlaps between contemporary distributional models and structural linguistics, focusing on early linguistics analysis in American structuralism, by, for example, Zellig Harris and Leonard Bloomfield, who defined word similarity as a statistical measure of word co-occurrence frequency (Lenci and Sahlgren 2023). Follow-up work using word embeddings and inspired by structuralism demonstrated the ability to recover subspaces or "semantic axes" from embedding representations by subtracting the average of vectors corresponding to binary oppositions (e.g., {good, ethical, moral} – {bad, unethical, immoral}). This was used to construct a direction in vector space between contrasting concepts and then measure arbitrary word positions along it (An, Kwak, and Ahn 2018; Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017; Ethayarajh, Duvenaud, and Hirst 2019; Kozlowski, Taddy, and Evans 2019). This approach, usually called "semantic axis projection," has been used frequently both to measure intrinsic model bias on properties like gender as well as to measure sociolinguistic variation for properties like gender, class, or stereotypes in language data. Embeddings have also been applied (Garg et al. 2018; Kozlowski, Taddy, and Evans 2019) and extended (Charlesworth, Caliskan, and Banaji 2022; Hamilton, Leskovec, and Jurafsky 2016) to the study of diachronic language change. These applications of word2vec to linguistic analysis focus on broad structural associations and diachronic shifts in large corpora. In many cases, they explicitly assume a structuralist orientation, tracing binary oppositions with semantic axis analysis in order to measure, for example, gender or class coding within corpora like Google Books (Charlesworth, Caliskan, and Banaji 2022; Kozlowski, Taddy, and Evans 2019).

It has been more challenging to extend word embedding models to the study of narrative. This is partly because narrative is highly focused on issues of structure and temporality that work at the level of the individual text rather than the corpus. Narratological theory, classical and postclassical, offers many and plural conceptualizations of narrative, but generally acknowledges that narratology involves the analysis of variability in how texts modulate formal properties, such as time, space, speaker position, or causation in order to represent events. In other words, narratives treat as conditional the kinds of mediating properties that would render the ordering of events into sequences an aesthetic construction rather than a natural and inevitable act.

This places the level of analysis in narratology somewhere between corpus structure and individual text, since it emphasizes how the formal sequencing of narrated content exhibits variation in the individual text that is also expected to adhere to broad intertextual patterns that emerge structurally in aggregate. If we conceptualize narrative this way, computational and statistical analysis would seem to have a place in narratology – provided that it can be operationalized to structure and analyze sequence over two temporal scales: the individual narrative sequence and the aggregate corpus sequence.

Most time-series analysis with word embeddings has focused on the aggregate corpus, as in the study of diachronic change in large corpora (e.g., to trace shifts in word usage over 100–200 years at the scale of Google Books). Static embeddings like word2vec generally struggle to embed time series at the scale of individual books, since the model is tuned for analyzing words in isolation rather than identifying how words come to assume distinct meanings in sequence context. Attempts to study individual text sequences by contrast have often drawn on methods like sentiment analysis on chunked sequence text (Jockers and Thalken 2020; Öhman et al. 2024; Reagan et al. 2016), LDA topic modeling (Elsner 2015), or tf-idf keyword analysis on trajectories (Boyd, Blackburn, and Pennebaker 2020). Newer multimodal models like CLIP embed individual videos as time series and allow prompt-based retrieval with cosine similarity (Radford et al. 2021). By contrast, there have been few comparable demonstrations to date of sequence analysis with word embeddings, although prior work that has tried this has been highly effective (Toubia, Berger, and Eliashberg 2021). Part

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

of the challenge with representing texts as sequences stems from the historical difficulty of modeling linguistic usage in sequence with static embeddings like word2vec – however this limitation has been ameliorated some in recent years with "contextual embeddings" (also called "transformer embeddings" to denote their use of the transformer architecture). Modern contextual embeddings using transformer language models like BERT (Devlin et al. 2019) and sentence-BERT (Reimers and Gurevych 2019) are specifically designed to disambiguate word meaning in sequence context.[1] This allows them to differentiate between the meaning of polysemous usages like 'river bank' versus 'financial bank'. It also allows them to infer word reference, such as the antecedent to whom the word "they" refers in a given sequence context. Recent work has demonstrated the extensibility of existing word embedding techniques to contextual models. Lucy, Tadimeti, and Bamman (2022) and Lucy and Bamman (2021) for instance recently demonstrated the applicability of familiar techniques used in static word embedding models such as word2vec, including semantic axis analysis and clustering, to the contextual embedding model BERT. There is also precedent for narratological analysis with contextual embedding models. Al-Laith et al. (2024), for instance, used BERT as a classifier to identify binary distinctions related to noise in textual sequences, and Rockmore et al. (2025) compared transformer contextual embeddings to tf-idf and other benchmarks in the context of literary text analysis. We explore here whether it is possible to adapt techniques for contextual embeddings to narrative analysis by modeling individual and aggregate text sequences in tandem. Our analyses are generally presented as proof-of-concept prototypes rather than fully developed evaluations, with the goal that they help establish the value and efficacy of sequence embedding analysis for future work.

Modeling sequences

Contextual embeddings such as sentence-BERT can effectively represent language at the two scales relevant to narrative: that of the individual text and that of the broader corpus. We suggest that this makes them specifically tailored to computational narratology. However, whereas measuring corpus-wide semantic associations and historical shifts has been well validated with static embedding models, representing text sequences with embeddings remains an emerging task in natural language processing.

To analyze how narratives are structured as temporal objects, we construct book sequence trajectory representations with embeddings derived from a pretrained sentence transformer model, which produces a single curve for each book in the corpus that we can then aggregate into pooled curve shapes and analyze along segments such as genre or decade. To do this, we start by chunking nearly 3,000 Project Gutenberg fictional texts into fixed 128-token segments and embedding each chunk with sentence embeddings. These yield book-level sequence representations. This allows us to apply embedding techniques like semantic axis analysis to textual sequences at the scale of individual books, exploring how various constructs vary throughout a text's narrative structure. Throughout, we measure seven semantic axes applied to each individual chunk: *gender, class, morality, collectivity/individual, nature/artifice, emotional/rational*, and *order/disorder.*

[1]Models such as Longformer (Beltagy, Peters, and Cohan 2020) are also designed for embedding long sequences. While we considered them, they are not tuned for semantic measurement or the analysis of short regions as used in this kind of task in the same way as sentence-BERT.

We then compile these projections into curves by resampling book sequences to the same normalized length and applying Gaussian smoothing to recover macro-level semantic patterns. We then pool these representations into aggregated curve shapes to further study how various subsets of the corpus such as genre, decade, or author gender tend to represent sequence in patterned forms. Since our approach relies on a somewhat new unit of analysis – embedded sequence segments – we outline our approach in more detail in this section and report several steps to validate its robustness prior to presenting the results.

Corpus acquisition and preprocessing

We use the gutenberg _english repository on HuggingFace's datasets repository, which covers over 80 percent of all English books on Project Gutenberg. We load the first 4,000 texts, representing slightly under 10 percent of the dataset (48,284 books total). We then filter within the metadata to members of the fiction category, yielding 2,965 texts. To do this, we extract "bookshelves" or "subjects" containing any of the keywords "fiction," "novel," "story," "drama," or "poet." This represents about 6 percent of the overall dataset, or around > 4.8 percent of all of Gutenberg's English fiction texts. A small number of nonfiction false positives appear to fall through but are rare and likely Gutenberg metadata labeling issues; this likely introduces some noise, but it also makes the fiction estimates conservative. Project Gutenberg also includes "complete works" and multivolume editions of texts, which could introduce some additional sequence noise. We estimate book year by extracting from metadata, whenever available, the author's birth year, which roughly designates relative authorship time periods for books when comparing between rather than within authors.

We next embed each book in full using the HuggingFace sentence transformer model *all MiniLM L6 v2* and use the final layer for analysis, as is conventional. We embed chunks of 128 tokens to encode individual books as sequences of embeddings. We leave unit score normalization off initially but normalize later when projecting axes using cosine similarity.

Constructing semantic axes

Semantic axis analysis is performed trivially with word2vec by measuring word cosine similarity to a direction vector usually defined by subtracting the mean of semantically opposite token vectors, for example, {*good, ethical, moral*} – {*bad, unethical, immoral*}. With contextual embeddings, single token subtractions tend to yield unstable axes, since contextual embeddings vary embedding position by surrounding sequence context, meaning that a given token like "good" will not inherently map onto its mean corpus usage. There is limited albeit emerging consensus on how best to operationalize semantic axes for contextual embeddings. Some studies use single words, or simple phrase templates, with the recognition of the potential for axis instability. Lucy et al. (2022) propose subtracting the means of numerous sampled contextual examples of word opposition pairs, drawing from, for example, the Wikipedia corpus (which maintains independence from their focal study corpora). Recent work applying sentence transformers such as Zeng et al. (2024) uses wordnet derived antonym pairs to construct initial semantic axes that are pruned with LLM assistance, then later projects masked contextual samples for numerous example texts about various groups onto them to measure how the axis position varies for the *contexts* surrounding the masked words denoting the group. In our case, fiction books, using Wikipedia

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

Table 1. Axis seed words

Axis	Positive	Negative
Gender	Man, male, he, him, his, guy	Woman, female, she, her, hers, gal
Morality	Good, virtue, noble, right, moral	Evil, bad, wicked, wrong, immoral
Class	Rich, wealthy, noble, manager, superior	Servant, worker, peasant, subordinate, poor, impoverished
Collectivity	Us, we, together, community, social	Separate, alone, apart, I, me, independent
Nature	Forest, beach, river, stream, mountain, sky, sun, earth, nature	Artificial, machine, technology, make, manufacture, art, writing, city, building, street
Emotional	Feel, feeling, sense, heart, passion, love, hate, desire, joy, pain, happy, sad, anger, love	Calculate, reason, rational, judge, measure, argue, plan, use, think, organize
Order	Organize, order, rules, laws, predict, stable, control, plan, hierarchy	Chaos, disorder, change, anarchy, crazy, disrupt

sampling such as in Lucy et al. (2022) is less directly applicable since word usage varies considerably between nineteenth century fiction and contemporary encyclopedia form.

One alternative to single token representations or Wikipedia sampling is to sample tokens over many contexts within the Project Gutenberg corpus itself. Although this risks some possible circularity since we are measuring the same texts used to define the axes, it is also more likely to capture historically specific semantics than if we were to draw on contemporary corpora. As a robustness evaluation, we construct axes from (a) individual word token oppositions and (b) random context samples from our Project Gutenberg corpus. We then compare axis similarities to test how closely they align. For the seven contextual semantic axes, we randomly sample the Gutenberg fiction corpus for up to 10,000 contextual examples[2] for each of the following individual seed words, then construct axis poles from the unit score normalized embedding means of those contextual samples. We then subtract these poles to compute the semantic axes. We report seed terms in Table 1.[3]

As noted, we construct axes from seed terms only versus from 10,000 per-term contextual examples and report the comparison in Table 2.

Contextual sampling from our corpus is modestly similar (0.737 cosine similarity) to single token subtractions, suggesting that the two capture relatively similar semantics, though there is some variation across the seven axes. Given that the two approaches appear generally comparable and since contextual axes are theoretically motivated, we elect to use Gutenberg context sampling for axis construction in the main analysis.

Table 2. Single token versus Gutenberg fiction contextual axis comparison

	Cosine similarity
Gender	0.816
Class	0.740
Morality	0.611
Collectivity	0.748
Nature	0.806
Emotion	0.674
Order	0.761
Overall	0.737

Table 3. Axis validation for 420 out-of-sample manually crafted antonym sentence pairs

Axis	Accuracy	Random axes
Gender	0.860	0.502
Class	0.734	0.495
Morality	0.817	0.499
Collectivity	0.790	0.502
Nature	0.800	0.496
Emotional	0.940	0.504
Order	0.700	0.497
Mean	0.800	0.499

We next validate the axis out of sample on a set of 420 curated antonym pair sentences, scaled with assistance from GPT-5. We present these results in Table 3.

These results are out-of-sample validations tested on contemporary examples, not historical Gutenberg fiction. Accuracy is based on whether the axis projection polarity (measured by its sign value) matches the labeled polarity (e.g., individualist vs. collective) for each sentence. Accuracies in all cases exceed 0.700 and are 0.800 on average. Despite using fixed linear projections rather than trained classification, accuracies are comparable to simple supervised baselines. We additionally compare to random null axes, which we expect to predict validation pairs at random chance. Across 1,000 random unit length axes per dimension, none approach the performance of the constructed axes. We compared these axes to the single-token derived seed axes as well and found comparable performance on this validation task (mean = 0.820). We report some of the top scoring sample chunks for face validity in Table 4.

Following Antoniak and Mimno (2021) who show that semantic axes are fragile to seed word selection, we also evaluate robustness to seed choice. We perform single token leave-one-out ablations for our contextual axis to evaluate robustness to seed term selection. Cosine similarities to the original axis are consistently high (cosine similarity 0.90) across all terms and axes. We report full results in the Supplementary Material. We also compare cosine

[2]Some words do not appear in at least 10,000 contexts; in these cases, we sample as many contexts as possible.

[3]The gender seed tokens adhere to a binary conceptualization of gender, since our focus is predominantly on representations from the nineteenth century and prior, although we recognize that this construct is historically specific and incomplete as a representation of gender identity. Contextual representations produce a continuous score, which avoids direct binary classifications – ambiguous cases, such as when female speakers adopt male-coded idioms, tend to score ambiguously on our inferred semantic axis, so while our initial seed tokens are built on binary samples, the operationalized score outputted ultimately captures variations in use that adhere to a more continuous and fluid gender conceptualization.

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

Table 4. Examples of top scoring chunks by axis for each pole

	Positive	Negative
Class	To do him justice, she had to admit that he did not even pretend much. He wanted her because she was a step up in the social ladder, and, in his opinion,. . .	The majority of us forget in a year or two all that we learnt at school because the conditions of our lives are such as to destroy all inclination for culture or refinement. . . .
Gender	Are you prepared to shut your eyes and ears henceforth to all that you may hear of her new life, when she is happy, rich, respectable, a courted heiress—perhaps. . .	This personage is the Register of Deeds. He is an unemotional character, living in his business almost as exclusively as the Scarabee, but without any of that eagerness and enthusiasm. . .
Collectivity	. . .elective body, changing from time to time according to the needs and ideas of those whom it represents, should refuse obedience to a bad law in the name of the people, well and good. . .	I understand that something has happened, Chérie, and that you want to be alone. I won't be a hindrance to you. . .
Morality	In his section of a few dozens of soldiers there were many superior to him in intelligence, in studiousness, in character; but they were all courageously undergoing the test. . .	all that he speak me fair—be twisting the words in his mind into evils I have not dreamed of. Sebert, I do not reproach you with it! I think it all the fault of my own blunders,—and therein I find a new terror. . .
Nature	The mountain was partly obscured by rapidly shifting masses of mist; the air was pungent and seemed to hum with a thousand tiny, electric voices.	It was thus that the magic qualities of the machine were made known to me, and out of it the following papers have grown.
Emotional	But if I lose you, do not speak to me of hope, for God knows I will not survive the loss." She turned on me like a flash. "Speak to me of your love," she said, "not of your grief." "Very well, I love you more than life. Beside my love, my grief is but a dream. Come with me to the end of the world, ...	But, as I shall try to show, a legal duty so called is nothing but a prediction that if a man does or omits certain things he will be made to suffer in this or that way by judgment of the court; and so of a legal right.
Order	Meanwhile, as the moon was about to rise, and as there was still floating in the zenith a remnant of the brightness of twilight, these clouds formed at the summit of the sky a sort of whitish arch, whence a gleam of light fell upon the earth.	He was growing sharper, finer, in a way younger, a man without grip on himself, a shapeless creature in process of transformation.

Table 5. Comparison to placebo axes

Axis	Placebo shuffle mean (cosine similarity)	Placebo shuffle max (cosine similarity)
Gender	0.040	0.988
Morality	0.084	0.816
Class	−0.165	0.289
Collectivity	−0.035	0.616
Nature	0.147	0.790
Emotional	−0.117	0.560
Order	0.122	0.526

similarities for each axis to placebo axes, which randomly reorder tokens, in Table 5.

Constructing semantic sequence curves

We next analyze chunked semantic axis projection sequences as sequence curve plots. To do this, we normalize books of varying lengths to the same length through resampling, then apply smoothing to recover macroscopic trends. We then pool curves into aggregated trajectory plots. We describe the procedure below, then report a set of robustness tests. Composite embedded sequences are a somewhat new artifact – while embedded sentences and time-series embeddings are sometimes analyzed, whether broad axis projection patterns at the scale of the book are statistically meaningful is an open question. We statistically test against a time-shuffled null model below, and we also statistically evaluate the robustness of various smoothing and resampling parameter selections; in all cases, we find strong statistical robustness. We additionally briefly compare our model to sentiment analysis and keyword-based trajectories, which have also been used in the literature to construct narrative curves. These results are presented in the remainder of this section.

First, however, we briefly discuss the theoretical validity of our approach. Our aim is to present a preliminary proof-of-concept method for analyzing embeddings at the level of book-length sequences. Our sequence curves are meant as minimal, parsimonious representations of semantic trajectories rather than definitive narrative arc shapes, which is why we limit processing to resampling and smoothing. This method is intended to recover statistical tendencies rather than definite forms, like the "six archetypal story narratives." Nevertheless, we do make a somewhat strong assumption that semantic tendencies at the scale of individual texts can aggregate into pooled sequences that reveal meaningful intertextual trends in semantic sequencing. Although we test this assumption statistically below, conceptually, it adheres to a genre-theoretic interpretation that intertextuality derives not only from borrowing a shared repertoire of topics, but also from a) sequencing them along shared temporal orderings and b) constructing them using shared, corpus-level structural binary oppositions. In other words, our statistical model is intentionally lightweight, which is intended to allow us to empirically test long-articulated but comparably stronger theoretical assumptions from narrative and structuralist theory.

Plotting procedure

For plotting embedded sequences as normalized and aggregated curves, we resample and smooth the raw embedding projection score sequences then plot pooled composite curves. We first resample chunked sequences to 300 time points by linearly interpolating between points and then sampling equally along the interpolated curve. Resampling normalizes books of distinct lengths, allowing

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

us to compare books on the same time scale. Conceptually, resampling normalizes books of differing lengths onto the same time scale by interpolating lines between neighboring points and sampling equally along the interpolated curve. In many cases, when resampling to 300 time units with 128-token chunks, this will downsample longer books onto a shortened normalized time scale with lower resolution. We considered alternative forms of interpolation, as well as non-interpolated alternatives, such as taking the means of standard buckets, but found linear interpolation to preserve the most signal while maintaining the simplest set of assumptions.

We then apply Gaussian smoothing to focus on the most salient trends and reduce noise. Smoothing functions like a low-pass filter that dampens the highest frequencies – often noisy local fluctuations – to recover slower moving trends comprising the underlying global curve shape. Bootstrapping with 95th confidence intervals was applied to plots when noted. Plots show axis similarities using cosine similarity except where they also use z-scoring when noted. Z-scoring removes the absolute scale and plots axis motion per chunk along a standard scale. We do not use it when trying to show variations in semantic axis value means or amplitudes. To construct aggregate mean trajectory plots, we average the projection scores for each chunk to compute a composite curve.

Time-shuffled null comparison

It is possible that averaging semantic axis projection trajectories across many books might result in curves that appear qualitatively meaningful but are indistinct from patterns that emerge by random chance. To test whether our aggregate curve shapes are different from random distributions, we compare to a within-book time shuffle null, which randomly permutes chunk order within books while maintaining the semantic axis projection values of the permuted chunks (Figure 1). This evaluates the degree to which the temporal sequence order of projection values makes a meaningful difference to the recovery of curve shapes. The observed aggregate curves are dramatically more structured than within-book shuffle null curves. For instance, the observed range exceeds the median null range by more than an order of magnitude on every axis, all of which are statistically significant at $p < 0.001$ (e.g., morality: 2.54 vs. 0.013; class: 1.60 vs. 0.013; and emotional: 1.52 vs. 0.013).

Chunking and smoothing stability

Chunking stability: We embed chunked text; chunking the sequences splits it into equal text intervals. It does not attempt to isolate meaningful narrative "beats," but simply samples on a consistent time unit. It is possible that shortening or lengthening chunk window sizes could influence semantic values. We evaluate the robustness of our results to chunk lengths of half (64 tokens) and twice (256 tokens) the size of our main text chunk window size on a reduced random subsample of 10 percent of the corpus and find strong overall median correlations to each (64 tokens = 0.920, 256 tokens = 0.934) and median correlation consistently greater than 0.90 across our seven semantic axes. We report full results in the Supplementary Material.

Smoothing stability: One significant potential threat to the validity of our analysis is the smoothing procedure. Smoothing is widely used in time-series analysis and necessary to produce interpretable curve forms, but it risks introducing structure not present in the data. To test for this, we check sensitivity to smoothing parameters by measuring Pearson's correlation between various smoothing coefficients and chunk sizes against a reference sequence. We run these for the diachronic, genre, and gender analyses reported in the main results, testing smoothing factors of 0.00, 0.05, 0.10, and 0.15 and resampling sizes of 100, 240, and 500. We also evaluate null control results, by shuffling chunk order randomly within books, to tests whether sequence structure randomizes. While correlations with the raw sequences are moderate, as expected, correlations to even lightly smoothed curves are consistently above 0.90–0.95. Null results clearly randomize sequence structure, although they maintain absolute semantic offset, as expected. Smoothing qualitatively tracks raw sequence trends, and the analysis is robust to smoothing parameter variation overall. We find the same effects when testing by various variables, including decade, genre, and gender. It is unlikely that smoothing plays a substantial role in synthesizing or disrupting sequence trends. Full smoothing validation results are presented in the Supplementary Material.

Even given strong statistical results, we briefly justify our choice of smoothing and raise possible limitations. Smoothing chunked sequences acts as a low-pass filter dampening the signal's highest frequency bands to recover the underlying curve shape. If we assume that rapid frequency changes are largely local effects, applying a low-pass filter to the sequence should extract genuine low-frequency curve structure if it exists rather than synthesize spurious patterns (unless the smoothing factor is extremely large). Conceptually, smoothing gives priority to long-term semantic trends over short term, reversible oscillations. Narratologically, curve smoothing assumes that thematic language, such as individual versus collective language, temporally repeats in a patterned way across several chunks even if local chunks show brief deviations – in other words, smoothing isolates moments where a semantic pattern is sustained for more than short-term intervals. However, narratives communicate not only through consistent repetitions but also through dialectical constructs (e.g., "I am far from the crowd"). Smoothing does not eliminate dialectical constructs if they persist over several chunks, but it means that we must interpret apparently neutral trends cautiously; a neutral trend does not necessarily indicate the absence of a binary opposition, but may also indicate the integration of opposites.

Comparison to sentiment analysis and tf-idf keywords

There are two opposing analytic risks with embedding sequences for narrative analysis: (1) that they fail to add additional information beyond existing methods, such as sentiment analysis or keyword counts and (2) that they are uninterpretable and fail to track existing sequence analysis techniques, such as sentiment analysis and keyword counts. To check robustness to both of these, we compare our embedded sequence projections to VADER sentiment scores on chunked sequences and to tf-idf keyword measures, per chunk, of the seed words used to construct our semantic axes. When comparing to sentiment analysis, we find moderate to strong correlations between VADER sentiment trajectories and all of our axes at the median. The highest correlations are to the collectivity (0.854), gender (0.706), nature (−0.728), morality (0.686), and order (0.691) axes, with moderate correlations to the class (−0.459) and emotional (−0.545) axes. These correlations reflect the fact that our axes are constructed through valenced oppositions that often embed evaluation. However, they are also potentially consistent with the possibility that many binary contrasts in nineteenth century fiction are structured through evaluative distinctions. The moderate to high correlations generally suggest that a substantial fraction of variance along most semantic axes aligns with valence, while the semantic axes further capture more specific semantic

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

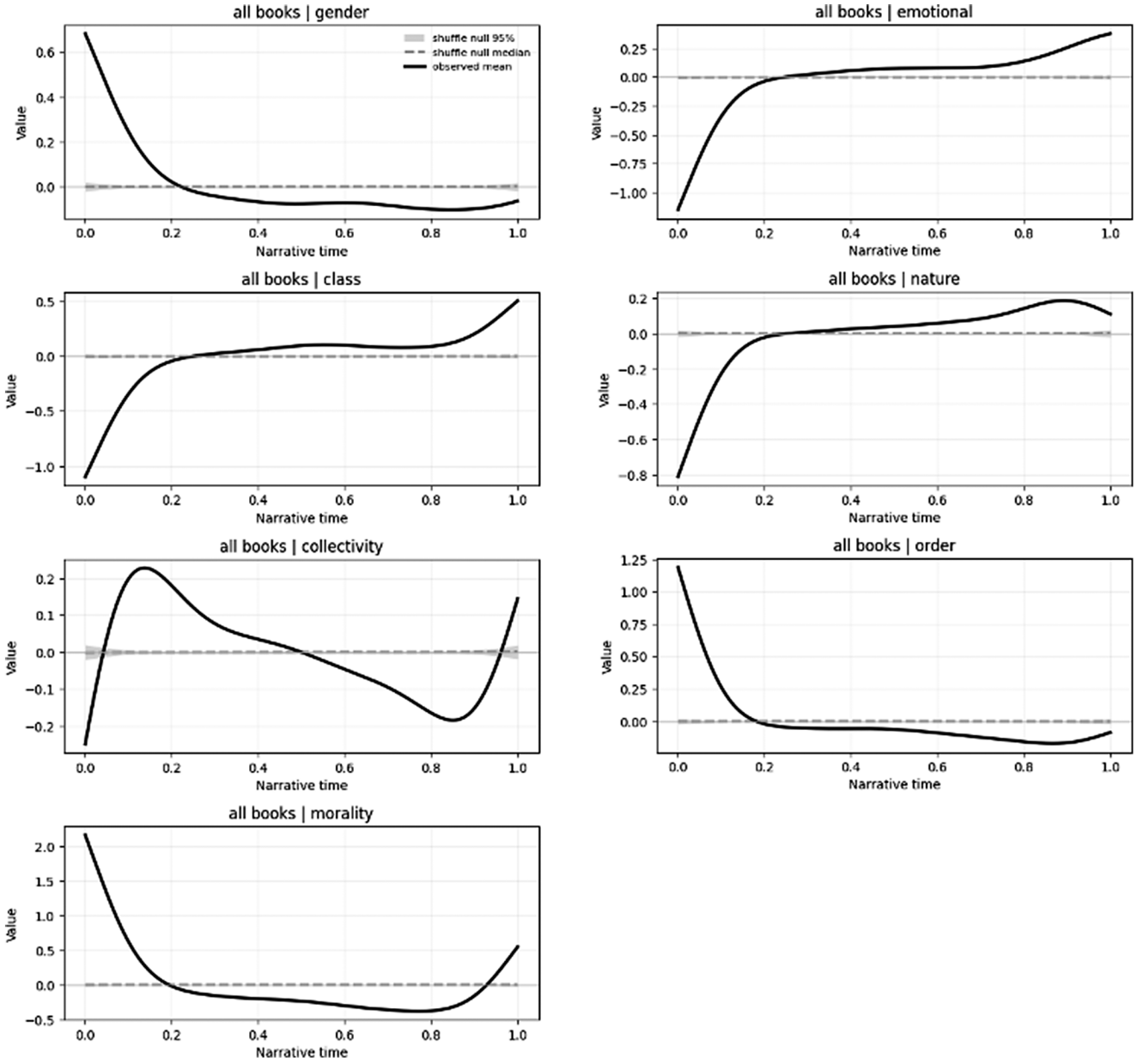


Figure 1. Observed z-scored corpus-level mean trajectories (black) compared with a within-book time shuffle null (gray), obtained by randomly permuting chunk order within each book while preserving the set of chunk-level semantic axis projection values. Shaded bands show the central 95 percent of shuffled aggregate curves. Across all axes, observed trajectories have substantially larger temporal range than the shuffled null trajectories ($p < 0.001$), indicating that the observed trajectories depend on narrative ordering rather than semantic composition only.

directions that are irreducible to overall sentiment. Conceptually, this potentially suggests, consistent with structural linguistics, that semantic axes show how narratives move through different "categories" of concepts that are each constructed in structurally valenced terms as relatively good or bad.

Our tf-idf keyword analysis shows small albeit modest correlation with all but the "order" axis (median class = −0.296, collectivity = 0.279, emotional = 0.607, gender = 0.206, morality = −0.340, nature = 0.345, and order = 0.041), with the highest correlation with the "emotional" axis (0.607). This is expected since contextual embeddings measuring text projections onto semantic axes should be adding much more information than raw keyword counts, though it also suggests that contextual axes remain semantically grounded relative to a simpler baseline.

Analytic strategy

We briefly outline the set of analyses we perform in the following sections.[4] First, similar to Charlesworth, Caliskan, and Banaji (2022), Hamilton, Leskovec, and Jurafsky (2016), Kozlowski, Taddy, and Evans (2019), or Al-Laith et al. (2024), we report aggregate diachronic trends for the seven semantic axes by decade (although again here we differ in analyzing down-sampled mean narrative *sequences* per decade). We next visualize mean

[4]In all plots, we apply Gaussian smoothing, and in some instances, we bootstrap results; in those cases, we smooth prior to bootstrapping to avoid modifying bootstrapped confidence intervals after the fact.

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

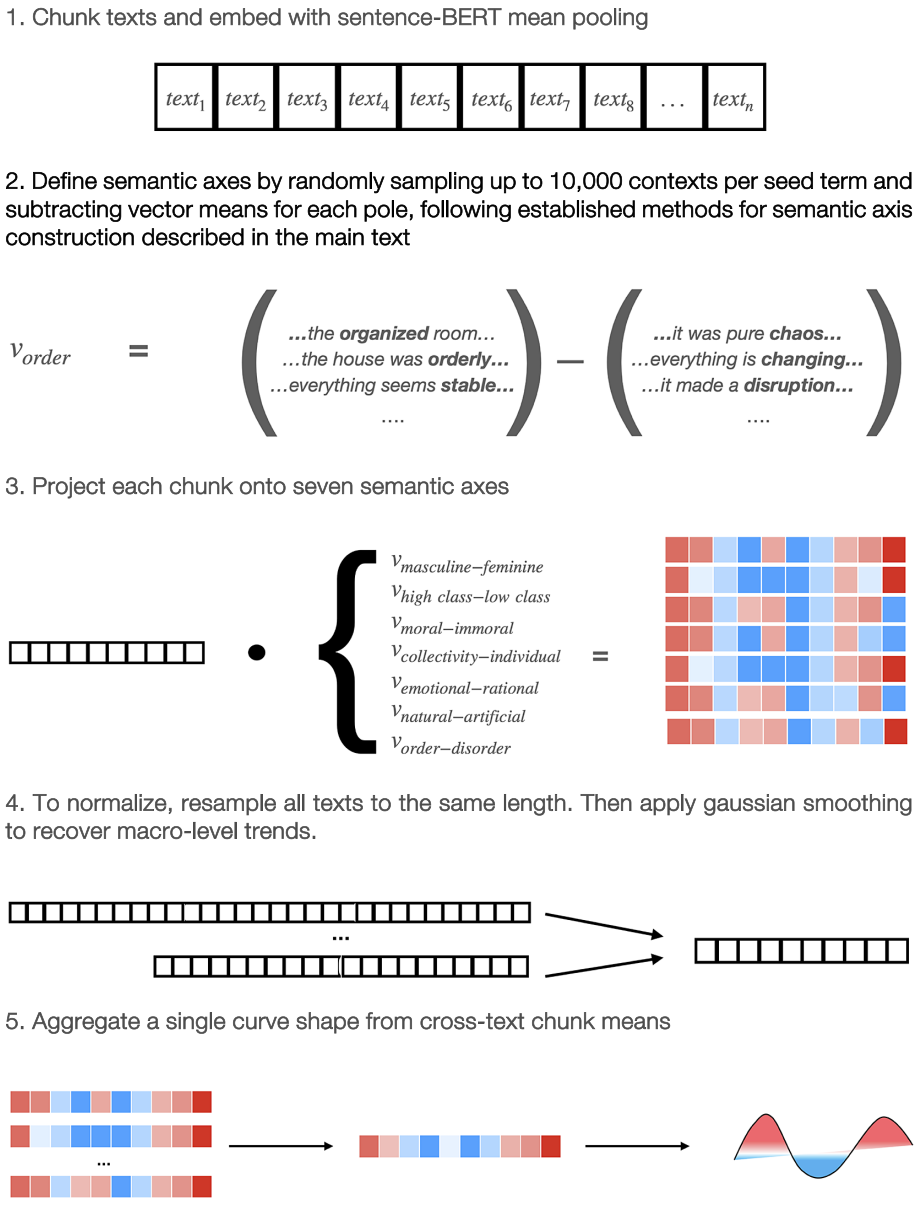


Figure 2. Schematic diagram of our approach to semantic trajectory embedding.

sequence axis projections by genre, to determine how genre modulates the same content categories through varying narrative form.

Following computational work from Jockers and Kirilloff (2016) and Jockers and Mimno (2013) on nineteenth century gendered novel representations, we next analyze the influence of author and character gender on our semantic narratological arcs. We label authors by gender using a simple names list from the Natural Language Toolkit (NLTK).[5] We then extract all dialogue quotations per book and check the speaker's gender by scanning for name matches within a ±20-word window[6] and labeling them by gender. This allows us to segment sequence semantic axis projections by author gender versus speaker gender, enabling us to trace patterns in the representations of gendered dialogue. (The purpose of this, described further in the next section, is to trace how character voice is constructed by standpoint and how this bears on sociolinguistic gender constructions in the emerging novel form, following arguments proposed by feminist postclassical narratological theory.)

Lastly, prior to presenting results, we briefly operationalize a provisional definition of narrative as we approach the construct here, as distinct from pure literary temporality. We earlier informally suggested that narrative involves the conditional ways in which a set of events can be mediated through, for example, temporal order, speaker position, awareness of the reader, or causal sequence. Prince (2019, 2012) gives a parsimonious and formal definition: "A narrative is a verbal or non-verbal entity taken to constitute the representation of at least two related asynchronous events (or one state of affairs and one event) that do not presuppose or imply each other." Prince is here defining narratives as a set of events (or event–state pairs) that are a) related through an ordering

[5]This is intended as a general proof of concept rather than a definitive classifier.

[6]The +/- 20-word window includes only words outside the quote body but pivots around the opening of the quote.

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

in time b) not simultaneous c) and not logically implied. By this Prince separates descriptions like "when I poured water to the top of the glass it was full" (not a narrative but a logical implication) or uncorrelated simultaneous incidents like "the sun set as the glass broke" from "related" sequences like "after I poured water in the glass, it broke" (Prince 2012). So narrative by this definition is again the study of how meaning is encoded in ordered sequences. In other words, analyzing sequence is analyzing narrative – with the caveat that not all sequences are narrative sequences. Our method does not strongly endeavor to separate out narrative and non-narrative representations of time at the micro-level – for example, our sequence models are not filtering out chunked sequences using Prince's criteria. At statistically large scales, however, when comparing across narrative forms (which all of our texts are), we suggest that many non-narrative utterances should be controlled out as randomly distributed, leaving intentional narrative construction explaining a non-trivial amount of the patterned variance. We also anticipate that non-narrative sequences (like logical entailment) should be more rare than narrative sequence. However, we briefly evaluate this assumption empirically in the Supplementary Material with a small classifier trained on positive and negative examples following Prince's definition applied to a sample of the Gutenberg fiction corpus.

Finally, we provide a visual schematic diagram of our approach in Figure 2.

Results

General diachronic trends

Generalizing, the rise of the novel has often been understood as concentrating around eighteenth century works focused on moral instruction that develop in the nineteenth century into early realist and increasingly individualist portrayals (Watt 2001). Similarly, the broadly increasing entry of women writers into novel publication in the nineteenth century played a significant role in constituting the form. The period also coincides with a shift from Enlightenment era literature to Romantic era literature, highlighting nature, emotions, and the individual. Our modeling strategy allows us to test for the presence of these trends in the Gutenberg data, by observing changes in mean plot trajectory shape along our seven semantic axes by decade. These sequence trends when measured in aggregate are broad and general, but they align with widely observed shifts in narrative form occurring during this time period.

We observe in Figure 3 a matching, gradual shift over decades toward more individualist than collectivist, less moralizing, more natural than artificial, and more disordered rather than ordered language, along with a slightly less pronounced trend toward more emotional than rational language. This shift seems to match the general trend toward Romantic era themes (concerning emotions, nature, the individual, and a critique of order, reason, and conventional morality) and away from earlier Enlightenment era themes (more closely focused in this data on artifice, the collective, morality, order, and reason).

We also see a shift in sociodemographic variables. On the gender axis, we observe books adopting, on average, more female gender-coded language throughout the century. The class axis shows a shift, generally, from lower to higher classed language, with a trend toward increasing class trajectories (suggestive, possibly, of class aspirational novels), although there is some leveling off in this trend over time.

Genre trends

We next evaluate our analysis on genre exemplars, by selecting texts widely recognized as comedic or tragic and measuring mean sequence differences. We start with all Shakespearean tragedies and comedies (Figure 4), since these are known to adhere to a consistent and reasonably strict stylistic form. We then evaluate genre differences on widely recognized nineteenth century comedies and tragedies (Figure 5). We finally measure the similarity between all books and our genre exemplars in a single plot (Figure 6). We confirm that sequence measurements on our seven axes differentiate genre in broadly recognizable ways.

Shakespearean genre categories, while complicated in many texts, follow a consistent typology over a single corpus so are a helpful narrative benchmark for our method. Comedies are often observed as breaking away from collective moral norms into individual freedom (typically to another geographic location, e.g., the pastoral "green world") and then returning to collective norms again, whereas tragedies often involve a shift from collective moral integration toward exclusion and individual isolation alongside a shift from order to disorder. This is recognized broadly, including in, for example, Frye's well-known genre taxonomy (Frye 1957).

Our sequence models identify a similar pattern in the structure of narrative arcs segmented by genre. On the collectivity, order, and morality axes, comedies follow a roughly u-shaped arc, consistent with the idea that comedies represent a break from and return to collective moral order punctuated by a shift toward relative individualism and disorder. By contrast tragedies observe a gradual decline on the order, collectivity, and morality axes throughout. Comedies sit higher on the class and emotional/rational axis than do tragedies. The strongest distinction between genres is on the gender axis, where the u-shaped curve for comedies is potentially consistent with a tendency to cross-gender characters or invert the patriarchal order in the "green world" segment of most Shakespearean comedies. These trends do not of course "prove" genre-theoretic claims about formal narrative structure, but they are broadly consistent with such claims as aggregate patterns, suggesting the method is at least directionally compatible with existing structuralist narratological theory.

When we move to analyzing post-1700s genre, these sequence trends persist but compress – in other words, comedy and tragedy start to move toward similar formal structures. This is consistent with general understandings of genre and the rise of the novel, which see the novel form as collapsing what was once strict comedy and tragedy into increasingly hybrid narrative form (Bakhtin 2010; Lukács 1974). One of the larger changes compared to Shakespearean genre occurs on the emotional axis, where nineteenth century tragedy appears to shift further away from reason and toward emotion and the nature axis (i.e. away from artificial and social). Finally, the composite genre plot in Figure 6 suggests that extending genre exemplars to the whole corpus along our semantic axes seems to differentiate the forms roughly. We include this example as illustrative, but it is not meant as a formal classification.

Gender trends

We next segment trajectories by gender.[7] To do this, we label author by gender using the NLTK names list. We then scan for quotations

[7]We use the terms "women authors" and "male"/"female" in this section to reference those positioned in the nineteenth century gendered system, while recognizing that these categories are socially constructed and historically specific rather than fixed.

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

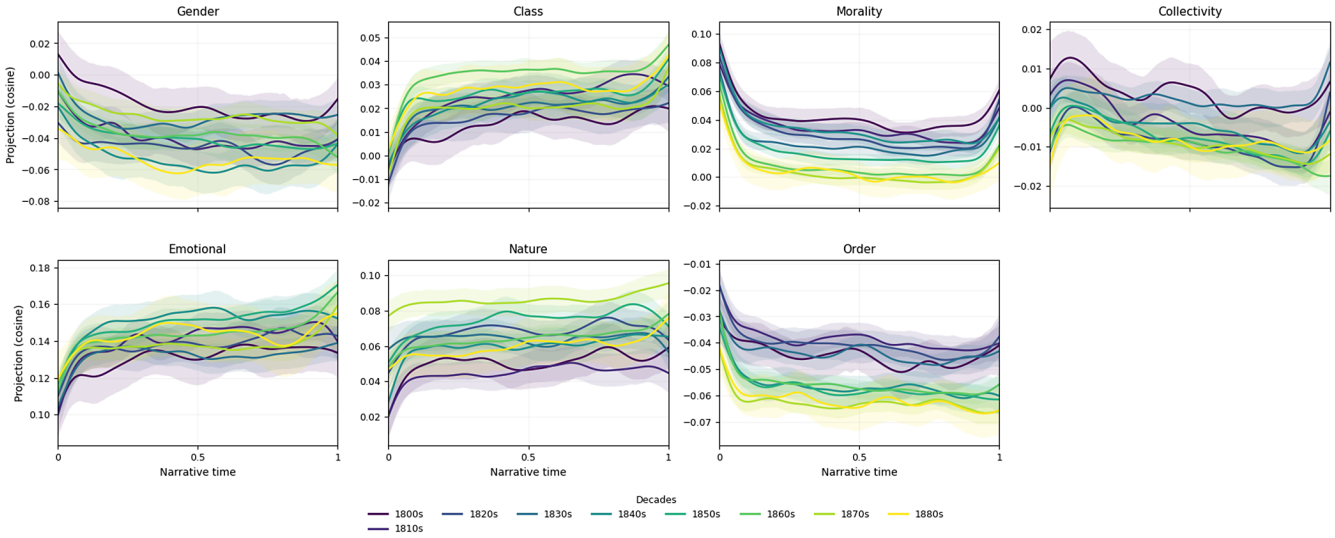


Figure 3. Diachronic change (1800–1899) in narrative arcs, based on mean plot trajectories by decade (bootstrapped 95% CIs, Gaussian smoothing fraction = 0.1).

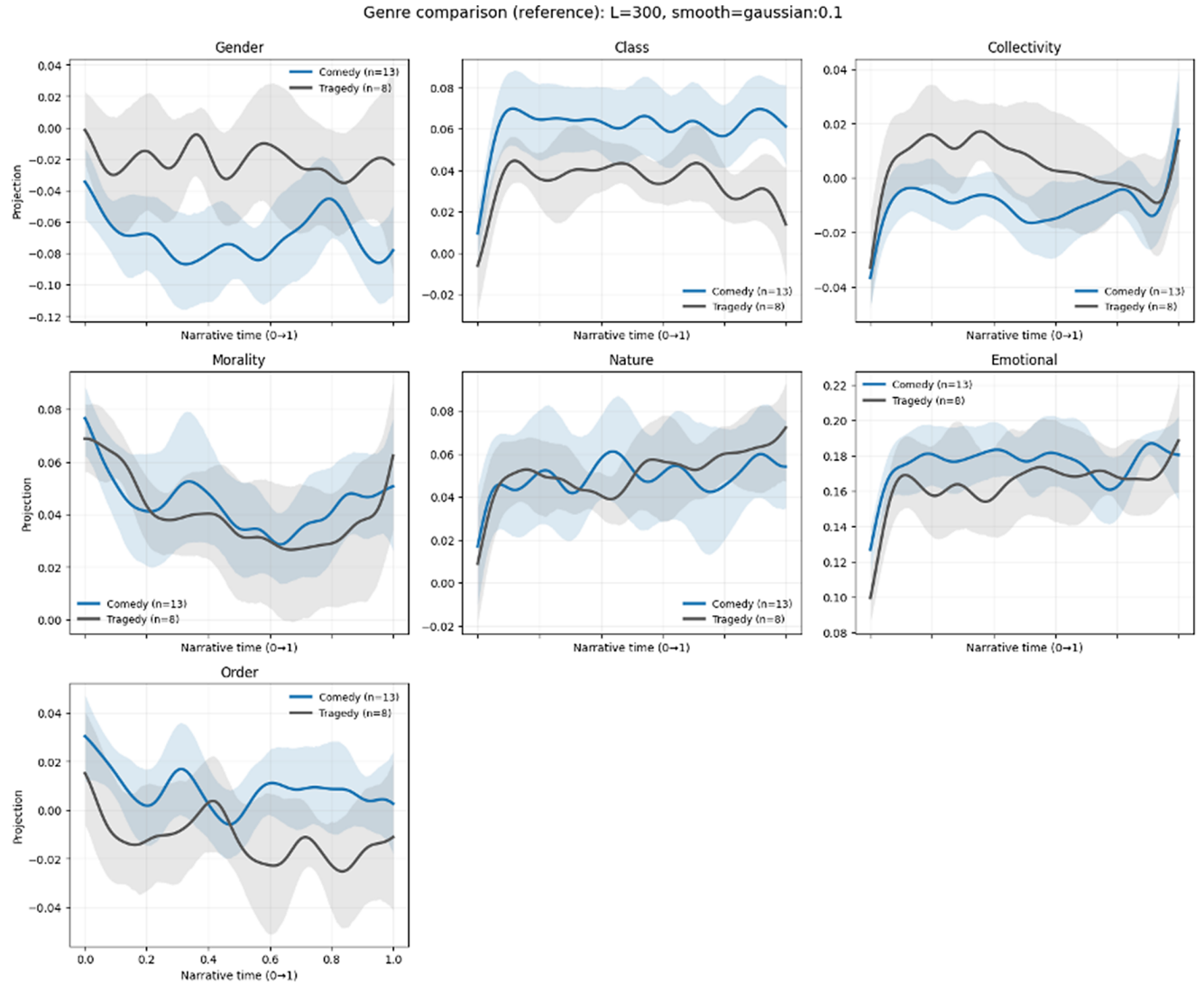


Figure 4. *Shakespearean genre.* Mean sequences (smoothed, bootstrapped 95% CIs) for all Shakespearean comedies versus tragedies. **Comedies:** "All's Well That Ends Well," "As You Like It," "The Comedy of Errors," "Love's Labour's Lost," "Measure for Measure," "The Merchant of Venice," "The Merry Wives of Windsor," "A Midsummer Night's Dream," "Much Ado about Nothing," "Pericles, Prince of Tyre," "The Taming of the Shrew," "The Tempest," "Twelfth Night; Or, What You Will," "The Two Gentlemen of Verona," "The Winter's Tale," "Cymbeline." **Tragedies:** "Antony and Cleopatra," "Coriolanus," "Hamlet, Prince of Denmark," "Julius Caesar," "King Lear," "Macbeth," "Othello, the Moor of Venice," "Romeo and Juliet," "Timon of Athens," "Titus Andronicus," "Troilus and Cressida."

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

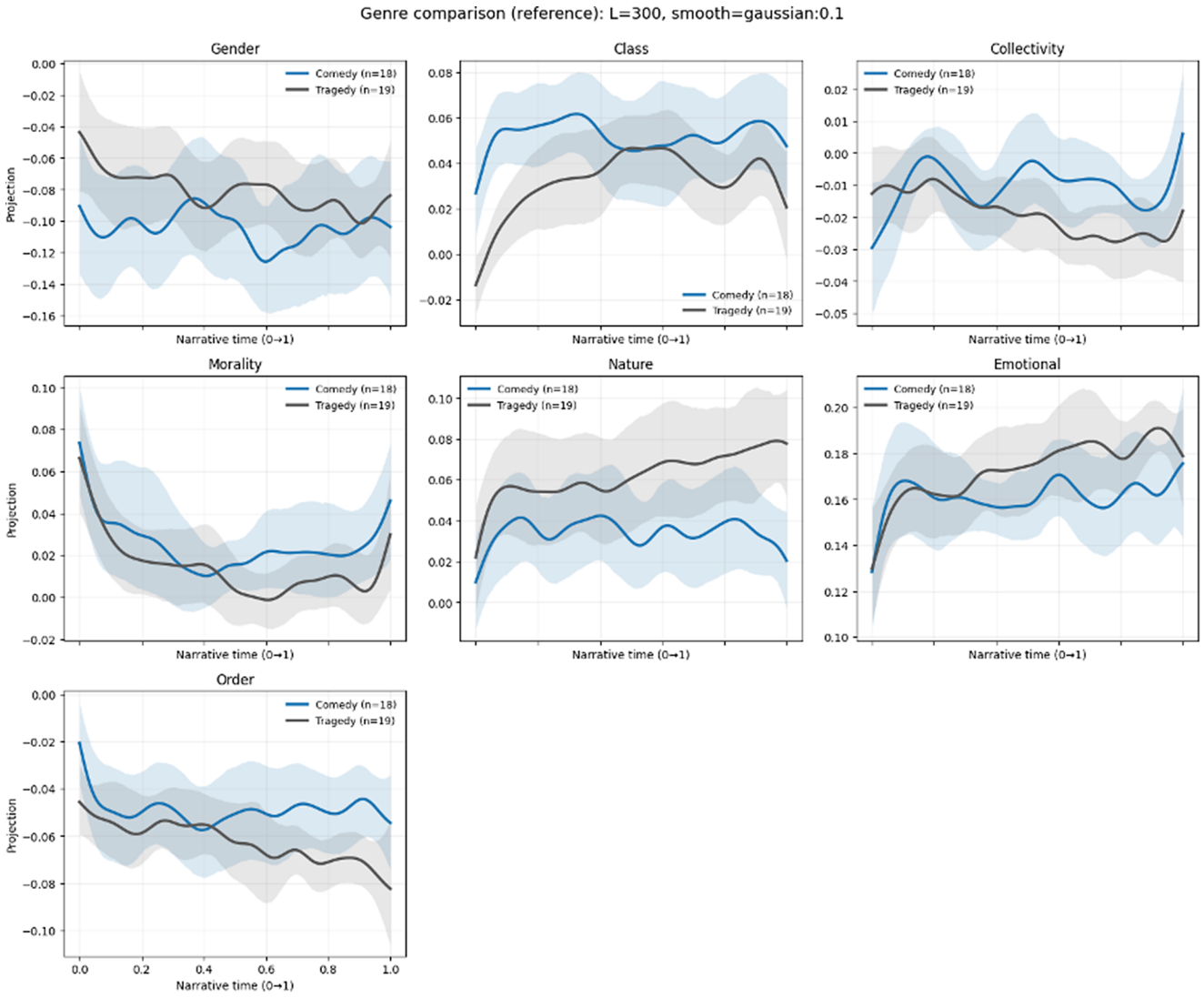


Figure 5. *Nineteenth century genre.* Mean sequences (smoothed, bootstrapped 95% CIs) for nineteenth century comedies and tragedies. **Comedies**: "Vanity Fair," "Pickwick," "Emma," "Three Men in a Boat," "The Old Curiosity Shop," "The Pickwick Papers," "Northanger Abbey," "Pride and Prejudice," "The Importance of Being Earnest," "The Adventures of Tom Sawyer," "Tom Jones," "Alice's Adventures in Wonderland," "Sense and Sensibility." **Tragedies:** "The Scarlet Letter," "Tess of the D'urbervilles," "Hedda Gabler," "The Mayor of Casterbridge," "The Sorrows of Young Werther," "Crime and Punishment," "Far From the Madding Crowd," "Agnes Grey," "Jude the Obscure," "Anna Karenina," "A Doll's House," "Madame Bovary," "Ghosts," "Moby Dick," "The Idiot," "Wuthering Heights."

in all texts to extract dialogue and label the speaker using a heuristic method. To detect speaker gender identity, we use a heuristic approach wherein we scan for quotations with length greater than two tokens, then check within a 20-token window outside the quotation for gendered names or for the pronouns "he" or "she." We scan from the start of the quotation, for example, 10 tokens back and 10 tokens forward. This heuristic is expected to introduce some mismatches, such as when multiple names interact. It will fail to tag cases in long dialogue sections without name designations, which are common in fiction. Our results should thus be thought of as a conservative baseline that systematically overlooks dialogue turn-taking examples and complex coreference. However, we find our approach is generally robust as a conservative sampling heuristic; we manually evaluate a 200-label sample and find the accuracy is high at 0.94. We then plot mean trajectories, for our seven semantic axes, for female/male writers and female/male speakers (Figure 7). This allows us to measure variation in the representation of female versus male speakers.

The aim of this analysis is to trace historical representational constraints on how gender is voiced and constructed, how standpoint influenced constructed representational positioning, and how gendered positions were articulated by women writers developing the early novel form's construction. We contextualize trajectory results: women authors in this sample (at pooled mean) voice male and female characters closer to the female-coded axis than male authors, suggesting that while (as structuralist and narrative gender theory predicts) a gender-coded voicing repertoire existed, women writers may have shifted its representational center of expression across genders in this sample of the corpus. We further observe this when, for instance, women writers appear to construct male character dialogue that ranges further on the emotional versus rational axis than male writers do in this sample generally, suggesting a possible semantic re-centering with respect to how masculinity was constructed and engaged inter-textually (male characters start out around parity but follow arcs toward heightened affectivity, congruent with some recent arguments

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

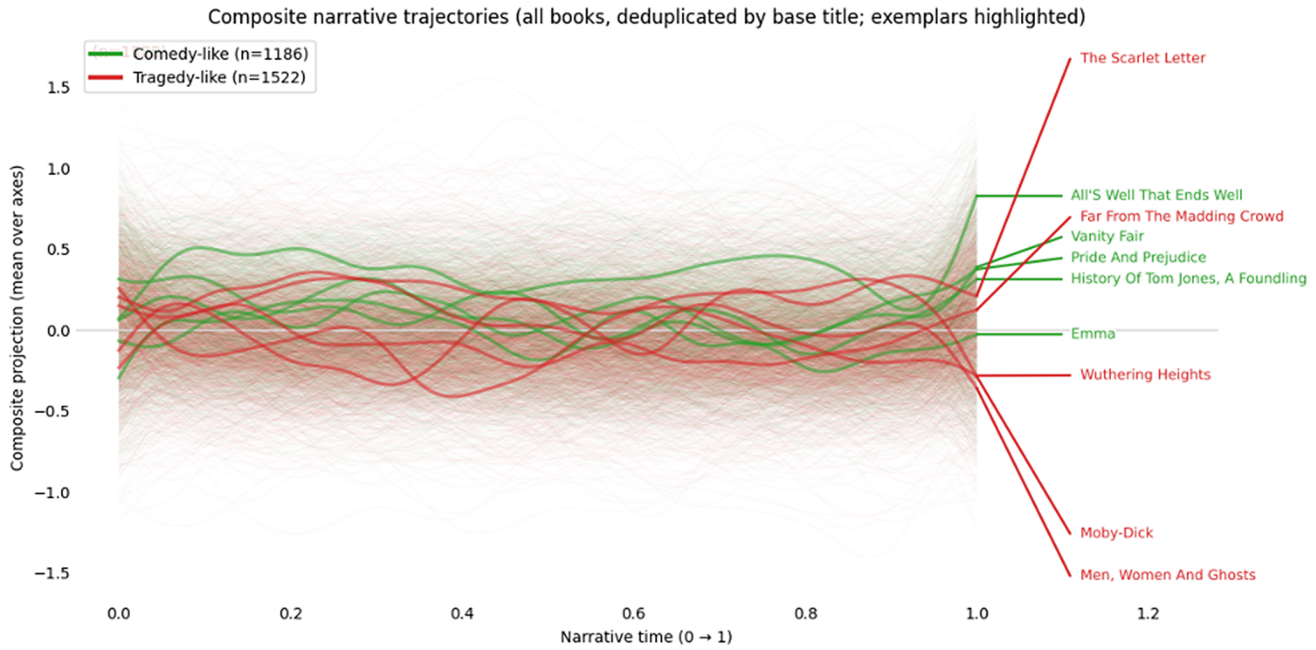


Figure 6. *Composite semantic* **axis genre** plot with labeled examples, using cosine similarity (per arc) to genre labels. Composite arcs are formed through concatenating semantic axis projection vectors and *z*-scoring. Comedies and tragedies are the same as in Figure 5, except with the addition of a small set of exemplars from Figure 4.

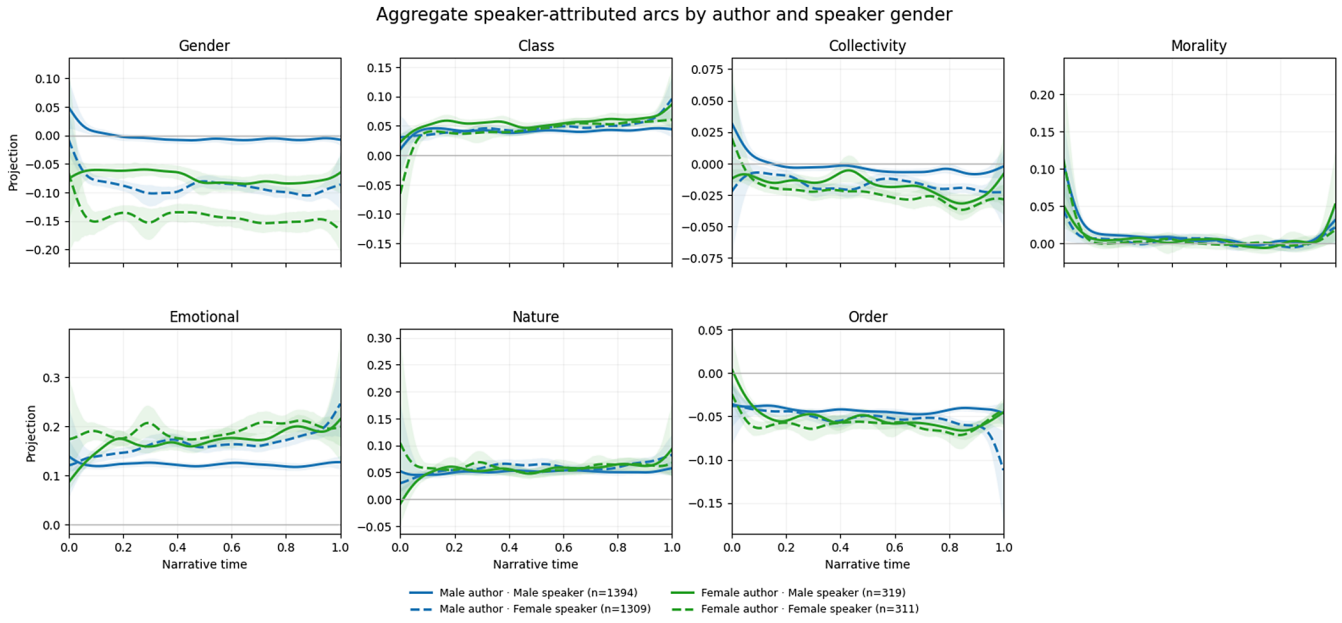


Figure 7. Author–speaker gender projections.

on nineteenth century reworking of masculine discourses). Potentially consistent with feminist narratological theory, female character dialogue also tends to be represented slightly (although not markedly) higher on individuality than collectivity, congruent with the influential claim in feminist narratology that female-coded characters were historically structurally positioned and voiced outside of a male-coded social collective (while often functioning to subvert or critique it). These findings are corpus-level and do not, on their own, establish claims from structuralist and feminist narratology, but they appear interpretively compatible with long-influential claims in feminist narratology. Individual text variance is quite high especially on constructs of interest above, indicating that writers were not strictly constrained by these trajectories at the individual book level, but that they also may represent possible field positions (mean absolute deviation gender = 0.101, class = 0.060, morality = 0.045, collectivity = 0.037, emotional = 0.116, nature = 0.072, and order = 0.053).

The aim of these analyses is to explore, as articulated in postclassical gender critiques of classical narratology, how gender standpoint might have informed patterned departures from formal narratological expectations shaped within a gendered literary field, particularly as women novelists became increasingly prominent

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

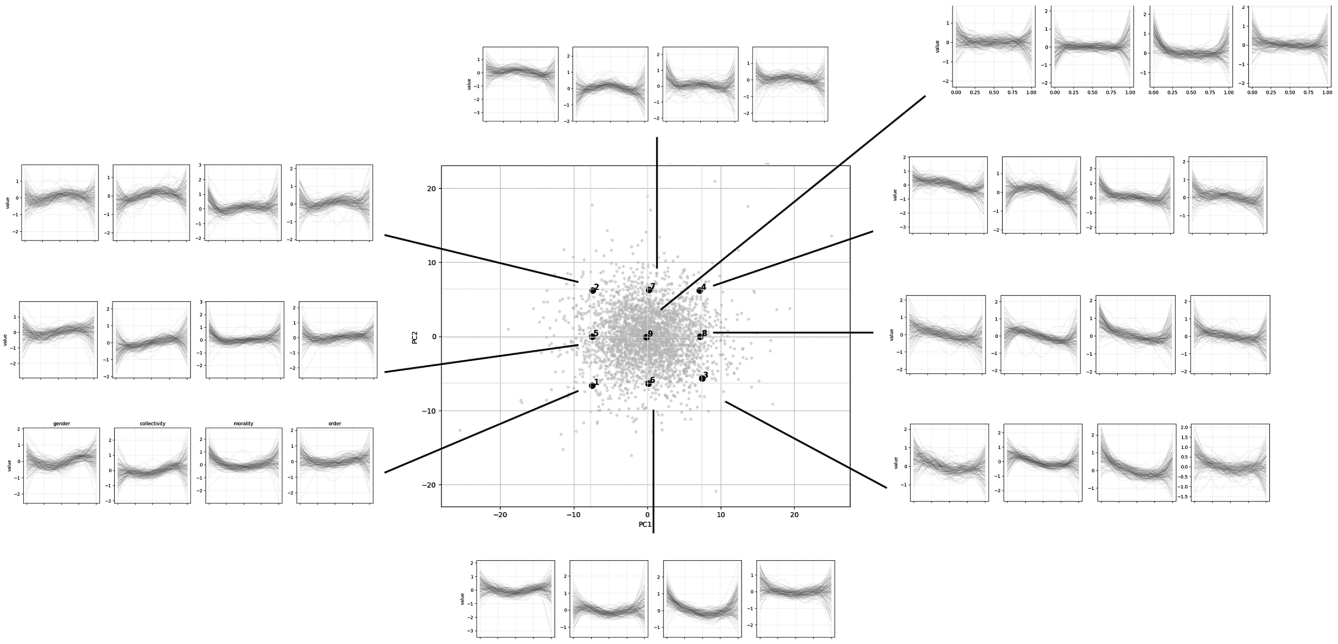


Figure 8. Recovering narrative curve types using PCA on sequence trajectory projections of four axes (gender, collectivity, morality, and order). We construct vectors composed of four projection values per chunk, reduce with PCA, and find that PC1 captures ascending to descending curve shapes, and PC2 captures u-shaped to inverted u-shaped curve shapes, consistent with functional PCA analysis. We can then analyze how narratives distribute within this space of curve types. Recovered curve shapes visually deteriorate with null model tests, reported in the Supplementary Material, along with a variant with all seven axes.

over the nineteenth century. Feminist narratology has broadly emphasized how the male gaze (de Lauretis 1987) and patriarchal norms (Gilbert and Gubar 2020) constrained the kinds of narrative roles historically expected of female characters and the stylistic expectations of women writers, often placing female characters in marginal or subversive roles. Miller (1980) explicitly emphasized how genre forms promoted male-coded themes, such as coming of age. Lanser (1992) highlighted how narrative perspectives, especially those speaking collectively, were historically male-coded. Warhol (1989) argues by contrast that female-coded narrative speech was often articulated directly or personally.

We can interpret the trajectory plots as reinforcing some of these arguments. For instance, women writers, in positioning female characters as progressively breaking from normative collectivity – seen here on the collectivity axes – might cohere with Gilbert and Gubar's (2020) argument that female characters were frequently positioned in subversive roles formally opposed to a masculine-coded notion of the universal collective (Lanser 1992). This is only one possibility, however, suggested by the aggregate curves, not an evinced finding, and significant close reading and supporting analysis would be required to further interrogate whether this particular trend is motivated by this postclassical narratological argument or not.

Relational trajectory space analysis

So far, we have interpreted aggregate trends; however, our model is also conducive to a relational interpretation of narrative form if we conceptualize it in terms inspired by (though not technically identical with) functional data analysis. In the above visualization (Figure 8), we run PCA on trajectory chunks for the four axis projections (morality, collectivity, gender, and class) after z-scoring and find that PCA separates a space of functional curve forms – PC2 (the y-axis) distinguishes inverted u-shape and u-shaped curvature (top to bottom) and PC1 distinguishes ascending versus descending curvature (left to right).[8] Each cutout plots 150 books with 0.25 Gaussian smoothing, although these shapes are retained in unsmoothed curves, presented in the Supplementary Material. PCA with z-scoring yields the space of curve shapes above separating raw curve shapes while removing differences in amplitude and starting absolute position; PCA on sequences without z-scoring incorporates wave amplitude and absolute starting position. We can then analyze separation in the z-scored shape space or in its non-z-scored variant for various variables.

When we test this on Shakespearean genre using a reduced axis set expected to distinguish genre (collectivity, morality, order, and nature), we find using linear discriminant analysis that the z-scored PCA-reduced sequences accurately separate texts by genre (Figure 9). Based on the PCA results, comedies tend to follow descending and u-shaped arcs while tragedies tend to follow ascending, inverted u-shape, and stationary arcs on the four semantic axes. These are composite results of the four axes, but they reflect u-shaped collectivity curves for comedy and possibly inverted u-shape collectivity curves for tragedy, ascending nature curves for tragedy, possible inverted u-shape order curves for tragedy, and u-shaped morality curves for comedy. This is consistent again with the earlier treatment of, for example, Frye's genre, wherein comedies move in u-shape away from and toward collective and moral integration, and where tragedy involves shifts away from earlier moral and collective integration.

We next test whether our PCA curve space distinguishes other variables such as time period. For time period, we test the model's

[8] That is, we treat each text as a vector $v \in \Re^{nT}$. This yields a new feature space distributing the four axis projections through time per text. One could generalize this approach onto other axes of interests, or the embedding values themselves.

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

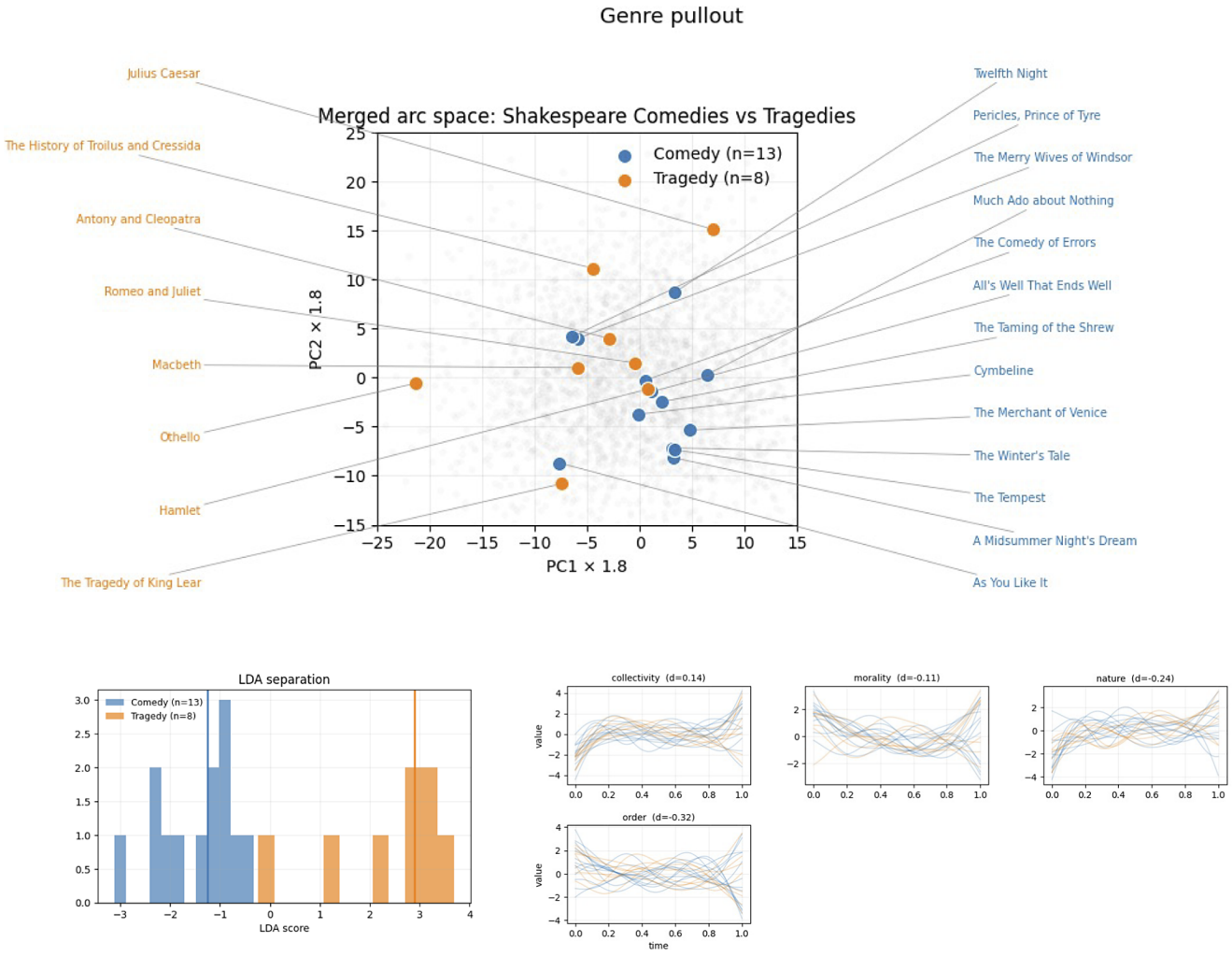


Figure 9. Distribution of genre (Shakespearean) in PCA curve space. The model, without supervision, appears to separate comedy and tragedy into ascending/u-shaped versus descending/inverted u-shaped or stationary arcs. It also locates "problem plays," romances, or genre ambiguous texts at the boundary of the two. Axis effect sizes (Cohen's *d*) are collectivity = 0.135, morality = −0.113, nature = −0.244, and order = −0.319.

ability to distinguish the early and late 1800s, and we report in Figure 10 both non-z-scored (top) and z-scored results (bottom). We use all seven semantic axes to construct the PCA space. Again, z-scoring isolated curve shapes only, whereas non-z-scored results also separate data by absolute semantic projection scores (affecting starting separation and curve magnitudes). For time period trends, by contrast to genre, we find separation only for the non-z-scored plot. This suggests weak evidence in our model for changing plot trajectory shapes through time, but modest evidence for semantic shifts (on both absolute semantic drift and curve magnitude).

Discussion

We discuss what our results suggest as a prototype framework for embedding sequence analysis as well as implications of this technique more broadly for computational narratology. We also briefly synthesize our substantive results, interpreting their implications for narrative theory.

While word embeddings are routinely used in natural language processing, with contextual embeddings becoming increasingly popular since their introduction in 2017, embeddings have been less frequently used in time-series and sequence analysis. Our work most directly extends diachronic word embedding modeling, which traces how individual words change over time series, from corpus-level analysis to the analysis of individual text sequences. This article's approach demonstrates that chunking embedded text and then pooling chunks to form sequence trajectories appears to generate interpretable, meaningful narrative sequence representations and that analyzing these representations in aggregate tracks known historical trends in narrative structural form.

Sequential time-series applications of language embeddings represent a midrange technique positioned between aggregate corpus statistics, which reduce books to a single data point, and classification and generative models, which are comparably more opaque. The fact that sequences, moreover, appear to encode coherent trends at the scale of ~3,000 books implies that representational learning may capture organization at scales rarely analyzed by common natural language benchmarks.

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

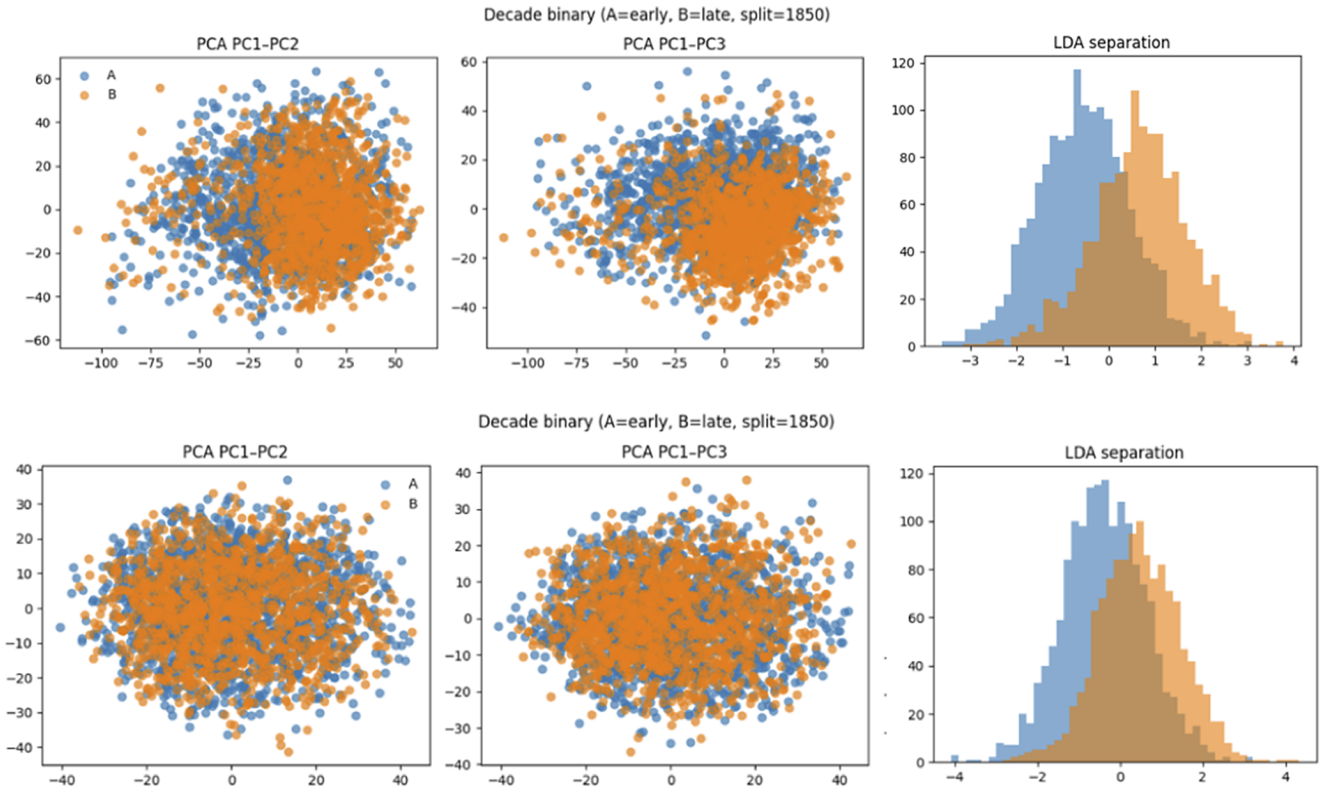


Figure 10. Plot trajectories separated by time yield similar distributions without z-scoring (top) and with z-scoring (bottom). Effect sizes (Cohen's d) for PCA without z-scoring: gender = 0.101, class = −0.229, collectivity = 0.287, morality = 0.745, emotional = −0.099, nature = −0.272, order = 0.560; effect sizes for PCA with z-scoring: gender = 0.060, class = 0.068, collectivity = −0.033, morality = 0.066, emotional = 0.031, nature = −0.002, and order = 0.098.

Our analysis remains preliminary and requires further experimentation, although the results presented here are promising and imply reasonable methodological robustness. Our use of contextual sampling from Project Gutenberg for axis construction is intended to extend axis construction to contextual embedding models applied to historical texts. While this is theoretically validated as an extension of prior work such as Lucy, Tadimeti, and Bamman (2022), changing corpora or anchoring methods could change results. One limitation of this approach is that sentence transformers are trained on contemporary corpora, so while our semantic axes use historical examples, embedding representations may still project contemporary linguistic associations onto the historical corpus (a possibility we cannot rule out here). There are comparably few historical embedding models, and we think that the use of historically grounded semantic axes with sentence transformers partly obviates this limitation. However, this mismatch remains one of the stronger limitations today on applying contemporary natural language processing models to historical texts. A more comprehensive future validation could further consider stabilization on larger and multiple corpora and models. It could also consider fine-tuning embeddings, though fine-tuning can also erase or confound semantic associations.

As a contribution to narrative theory, our findings primarily complement and draw on well-known ideas from postclassical narratological theory. We provisionally capture widely theorized trends in the rise of the novel, such as shifts toward individualism and realism, as well as trends in the emergence and development of genre as it realizes itself in novel form. We also capture some evidence for trends developed by feminist narratology when segmenting texts by gender.

We note with caution some of the unique inferential limitations and caveats of our modeling assumptions. As noted earlier, chunking and embedding texts operate at a somewhat coarse level of granularity, although sentence transformers are designed to represent sequences uniquely. When we project sequences onto semantic axes, we nonetheless collapse local semantic variance into an average representation. Resampling and smoothing further collapse local semantic fluctuations into broader patterns. This is intended to isolate semantic tendencies that persist from local noise, though it risks collapsing some local variance (which can often be quite narratively rich), or dialectical constructs that signify by oscillating between semantic values.

The analysis of semantic axes is arguably one of the most widely used methods in NLP in the social sciences (with meaningful use in cultural analytics), and linear axes can frequently recover interpretable directions from natural language representations. They are not the only available technique for model interpretability or classification today, however, and they possess known limitations such as fragility to seed word choice (Antoniak and Mimno 2021). We engage them because they are interpretable, well tested, straightforward to validate, and computationally lightweight to integrate with our broader framework. Our validation results indicate that the selected axis seeds recovered meaningful, corpus-specific semantic directions from the sentence transformer model's representational space.

One of the largest analytic concerns, however, with this kind of approach is that it might synthesize trends while blurring local semantic distinctions. We believe that our statistical robustness tests – especially time-shuffled null permutations and smoothing and resampling sensitivity analyses – generally point away from

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press

this possibility, though the risk remains inherent to this kind of analysis. Statistical variance observed in aggregate should be read as a floor rather than a ceiling on possible semantic trends and associations embedded in the corpus and further contextualized using close reading when applied in practice in order to mitigate these limitations.

Our corpus exhibits some noise – both due to Gutenberg metadata mismatches that allow nonfiction works to fall through and due to the inclusion of volumes, collected works, and anthologies such as poetry collections. It also includes introductions and nonfictional sections of variable length. Rather than manually correct for these in this article, we retain them and rely on a uniform book-length inclusion criterion – this generally yields a conservative estimate of fictional trends in our narrative trajectory curves. These issues likely introduce balanced noise rather than systematic bias, and our time shuffle null results strongly suggest that they do not fundamentally degrade the overall representations.

Where our technique may contain potential for its most novel contribution is in its capacity to analyze directly the constituent elements of narrative form – plot temporality – at large, aggregate scales rather than through generalization from micro-scale close reading. Structuralist theory and historical analysis, in conjunction with close reading, triangulates well between macro-historical narrative trends and their micro-formal articulation. However, the aggregate analysis of the constituent elements of narrative form – the conditional, patterned representation of content in time, order, or sequence rendered in the voices of specific characters – specifically captures the spirit of empirical aggregate analysis for which classical and postclassical narratological theory has repeatedly called, but for which statistical tools have been historically incomplete.

While our technique cannot mirror the hermeneutic density of close reading (nor does it aspire to), it can provide a potentially new way to interpret in the space between individual text and corpus-level pattern. In practice, for instance, we observed when reading individual text samples that our approach highlighted historically specific trends for several under-read or non-canonical sources, allowing corpus-scale analyses that may offer the potential to situate and elevate rather than reduce away constituent, marginal, or oppositional representations.

At the same time, we recognize caution is warranted in inferring strong conceptual generalization from our model's outputs. Many of the constructs we engage – morality, individualism, reason, order, and disorder – are analytically complex and historically situational, if not notoriously evasive. Our model is on the other hand intentionally constructed as simple and parsimonious. This juxtaposition is partly deliberate – our aim is to inscribe as few assumptions as possible into our modeling strategy to test the degree to which conceptual generalities appear in aggregate as well as how they vary in particular expression. We intentionally include some fairly abstract concepts to test the potential range of this type of approach. This framework does not and cannot resolve them in their full, polyvalent expression – it sheds a certain kind of aggregate and general light on macro-level patterns that are often hard to capture through direct reading. Interpreting their meaning, however, is intentionally left as an exercise to the reader.

Conclusion

In conclusion, this article demonstrates a straightforward method for analyzing narrative sequences by embedding books as sequences of chunked text, projecting chunks onto semantic axes, and aggregating book-level sequences into corpus-wide sequences. We find that this approach is highly robust to initial statistical tests, such as within-book time shuffle nulls, and to chunking, resampling, and smoothing parameter variation. While there is more work to be done to validate our approach, we suggest that it shows promise for the analysis of sequential text data in general and narrative structure in particular. More broadly, the approach suggests a path toward modeling how meaning develops over time within texts rather than only how it is distributed across them.

Future work might consider alternatives to prespecified semantic axes, such as methods for inducing corpus-specific semantic dimensions, or adapting the method to alternative kinds of narratives (such as narratives embedded in speeches, scientific discourse, or online communication). One extension, for example, could apply it to video transcripts, to complement similar time-series multimodal models.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/chr.2026.10035.

Data availability statement. All data for this project were drawn from the english_gutenberg dataset available on the HuggingFace data repository using the sampling strategy described in the article. Texts are in the public domain and free to use for research under Project Gutenberg's license. To avoid duplicating the dataset, we do not include it here. For the out-of-domain ground truth tests, we include the sentence pairs in the Supplementary Material.

Acknowledgements. I am very grateful to the anonymous peer reviewers, whose suggestions helped considerably to strengthen this essay. Some code for the analysis was initially generated with the assistance of a language model and independently modified and verified by the author. The author takes full responsibility for the results and analysis.

Author contributions. Conceptualization: W.R.; Data curation: W.R.; Investigation: W.R.; Methodology: W.R.; Software: W.R.; Writing: W.R.

Funding statement. The author declares that no specific funding has been received for this article.

Competing interests. The author declares none.

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Al-Laith, Ali, Daniel Hershcovich, Jens Bjerring-Hansen, Jakob Parby, Aidan Conroy, and Timothy Tangherlini. 2024. "Noise, Novels, Numbers: A Framework for Detecting and Categorizing Noise in Danish and Norwegian Literature." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, edited by **Yaser Al-Onaizan, Mohit Bansal, Yun-Nung Chen**, 3344–54. Miami, Florida: Association for Computational Linguistics.

An, Jisun, Haewoon Kwak, and Yong-Yeol Ahn. 2018. "SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics beyond Sentiment." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by **Iryna Gurevych, Yusuke Miyao**, 2450–61. Melbourne, Australia: Association for Computational Linguistics.

Antoniak, Maria, and David Mimno. 2021. "Bad Seeds: Evaluating Lexical Methods for Bias Measurement." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, edited by **Chengqing Zong, Fei Xia, Wenjie Li, Roberto Navigli**, 1889–904. Online: Association for Computational Linguistics.

Bakhtin, Mikhail M. 2010. *The Dialogic Imagination: Four Essays*, Vol. 1. United States of America: University of Texas Press.

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. 2020. "*Longformer: The Long-Document Transformer*." Preprint, *arXiv:2004.05150*.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. "Quantifying and Reducing Stereotypes in Word Embeddings." Preprint, arXiv:1606.06121 [cs.CL].

Boyd, Ryan L., Kate G. Blackburn, and James W. Pennebaker. 2020. "The Narrative Arc: Revealing Core Narrative Structures through Text Analysis." *Science Advances* 6, no. 32: eaba2196.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356, no. 6334: 183–6.

Charlesworth, Tessa E. S., Aylin Caliskan, and Mahzarin R. Banaji. 2022. "Historical Representations of Social Groups across 200 Years of Word Embeddings from Google Books." *Proceedings of the National Academy of Sciences* 119, no. 28: e2121798119.

de Lauretis, Teresa. 1987. *Technologies of Gender: Essays on Theory, Film, and Fiction*. Bloomington, Indiana: Indiana University Press.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by **Jill Burstein, Christy Doran, Thamar Solorio**, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics.

Elsner, Micha. 2015. "Abstract Representations of Plot Structure." *Linguistic Issues in Language Technology* 12, 1–22.

Ethayarajh, Kawin, David Duvenaud, and Graeme Hirst. 2019. "Understanding Undesirable Word Embedding Associations." In *Proceedings of the 57th annual meeting of the association for computational linguistics*, edited by **Anna Korhonen, David Traum, Lluís Màrquez**, 1696–1705. Florence, Italy: Association for Computational Linguistics.

Frye, Northrop. 1957. *Anatomy of Criticism*. Princeton, New Jersey: Princeton University Press.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences* 115, no. 16: E3635–44.

Gilbert, Sandra M., and Susan Gubar. 2020. *The Madwoman in the Attic: The Woman Writer and the Nineteenth-Century Literary Imagination*. New Haven: Yale University Press.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. Berlin, Germany: Association for Computational Linguistics.

Jockers, Matthew, and Gabi Kirilloff. 2016. "Understanding Gender and Character Agency in the 19th-Century Novel." *Journal of Cultural Analytics* 2, no. 2: 1–26.

Jockers, Matthew L., and David Mimno. 2013. "Significant Themes in 19th-Century Literature." *Poetics* 41, no. 6: 750–69.

Jockers, Matthew L., and Rosamond Thalken. 2020. *Text Analysis with R*. Cham: Springer International Publishing.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review* 84, no. 5: 905–49.

Lanser, Susan S. 1992. *Fictions of Authority: Women Writers and Narrative Voice*. Ithica, New York: Cornell University Press.

Lenci, Alessandro, and Magnus Sahlgren. 2023. *Distributional Semantics*. New York: Cambridge University Press.

Lucy, Li, and David Bamman. 2021. "Characterizing English Variation across Social Media Communities with BERT." *Transactions of the Association for Computational Linguistics* 9: 538–56.

Lucy, Li, Divya Tadimeti, and David Bamman. 2022. "Discovering Differences in the Representation of People Using Contextualized Semantic Axes." In *Proceedings of the 2022 conference on empirical methods in natural language processing*, edited by **Yoav Goldberg, Zornitsa Kozareva, Yue Zhang**, 3477–3494. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Lukács, Georg. 1974. *The Theory of the Novel: A Historico-Philosophical Essay on the Forms of Great Epic Literature*. United States of America: MIT Press.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*, edited by **C.J. Burges and L. Bottou and M. Welling and Z. Ghahramani and K. Weinberger**, 26. Nevada.

Miller, Nancy K. 1980. *The Heroine's Text: Readings in the French and English Novel, 1722–1782*. New York: Columbia University Press.

Öhman, Emily, Yuri Bizzoni, Pedro F. Moreira, and Kristian Nielbo. 2024. "EmotionArcs: Emotion Arcs for 9,000 Literary Texts." In *Proceedings of the 8th Joint Sighum Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, edited by **Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva, Stan Szpakowicz**, 51–66. St. Julians, Malta: Association for Computational Linguistics.

Piper, A., So R. J., and Bamman D. 2021. "Narrative theory for computational narrative understanding." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 298–311.

Prince, G. 2012. *Narratology: The form and functioning of narrative (Vol. 108)*. Walter de Gruyter.

Prince, G. 2019. "Narratology." In *Oxford Research Encyclopedia of Literature*.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. "Learning Transferable Visual Models from Natural Language Supervision." In *Proceedings of the 38th International Conference on Machine Learning*, edited by **Marina Meila, Tong Zhang**, 8748–63. Virtual: PMLR.

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter S. Dodds. 2016. "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes." *EPJ Data Science* 5, no. 1: 1–12.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, edited by **Kentaro Inui, Jing Jiang, Vincent Ng, Xiaojun Wan**, 3982–92. Hong Kong, China: Association for Computational Linguistics.

Rockmore, Daniel, Jiayi Chen, Mohammad Javad Latifi Jebelli, Allen Riddell, and Harrison Stropkay. 2025. "On the Literary Landscapes of Vector Embeddings." *Computational Humanities Research* 1: e18.

Toubia, Olivier, Jonah Berger, and Jehoshua Eliashberg. 2021. "How Quantifying the Shape of Stories Predicts their Success." *Proceedings of the National Academy of Sciences* 118, no. 26: e2011695118.

Warhol, Robyn R. 1989. *Gendered Interventions: Narrative Discourse in the Victorian Novel*. New Jersey: Rutgers University Press.

Watt, I. 2001. *The rise of the novel*. Univ of California Press.

Zeng, Q., Jin M., and Voigt R. 2024. "ADAPTIVE AXES: A Pipeline for In-domain Social Stereotype Analysis." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15576–15593.

https://doi.org/10.1017/chr.2026.10035 Published online by Cambridge University Press