Canonical Status and Literary Influence: A Comparative Study of Danish Novels from the Modern Breakthrough (1870–1900)

Pascale Feldkamp, Alie Lassche, Jan Kostkan, Márton Kardos,
Kenneth Enevoldsen, Katrine Baunvig and **Kristoffer Nielbo**
Aarhus University
{pascale.moreira, kenneth.enevoldsen}@cc.au.dk,
{a.w.lassche, jan.kostkan, martonkardos, baunvig, kln}@cas.au.dk

Abstract

We examine the relationship between the canonization of Danish novels and their textual innovation and influence, taking the Danish Modern Breakthrough era (1870–1900) as a case study. We evaluate whether canonical novels introduced a significant textual novelty in their time, and explore their influence on the overall literary trend of the period. By analyzing the positions of canonical versus non-canonical novels in semantic space, we seek to better understand the link between a novel's canonical status and its literary impact. Additionally, we examine the overall diversification of Modern Breakthrough novels during this significant period of rising literary readership. We find that canonical novels stand out from both the historical novel genre and non-canonical novels of the period. Our findings on diversification within and across groups indicate that the novels now regarded as canonical served as literary trendsetters of their time. To ensure reproducibility, code and raw data are available at https://github.com/centre-for-humanities-computing/memo-canonical-novels.

1 Introduction

At the beginning of the 21st century, the Danish government published an Educational Canon (*Undervisningskanon*) and a Cultural Canon (*Kulturkanon*) in an attempt to improve knowledge of Danish literature among the population, and to standardize school syllabi (Harbild et al., 2004). Both canons were met with criticism, and the canon debate flared up in full force – a development similar to, e.g., the Netherlands and Germany (Mai, 2016; Smid, 2022). Much of the criticism in Denmark was of the canon being unrepresentative and lacking diversity, including but one female author at the time (Fibiger, 2004).

The Danish canon debate echoes a central issue in literary scholarship, dating at least to the 'canon wars' of the 1980s (Guillory, 1995; Witt, 2000): the critique that the canon is a top-down, contemporary construct that does not accurately reflect works' historical significance, their impact on readers, or the breadth of literary production of a period. Still, advocates for the 'canon' being a meaningful term[1] argue that canonical works exhibit an enduring literary value and distinguish themselves by their lasting influence and innovation at the textual level (Bloom, 1995; Van Peer, 2008).

In this paper, we examine two hypotheses reflecting these polar stances on the canon: that canonical works are top-down and present day constructs, so that they would not stand out textually from their contemporaries (H1); and that canonical works distinguish themselves by textual novelty and literary influence, in which case we would expect them to show an impact on their literary field (H2).

To gauge whether books that are (today) considered canonical exhibit these distinguishing traits – textual innovation and literary impact and whether this resonates in what was published after – we compare canonical novels to novels that did not make it into the different constructions of 'canon' in a case study of Danish novels written in the Modern Breakthrough era. This late 19th century period is particularly suited as a case study of canon/non-canon dynamics for three reasons: *1)* It allows us to examine the status of canonical works within their historical context. *2)* It allows us to work with a complete corpus of the literary production of a time period – albeit limited. This is particularly significant because attempts to address issues of canon representativity often face the challenge of not catching the 'dark numbers' of literary production – the extensive numbers of titles forgotten or overlooked (i.e., the 'great unread' (Moretti, 2000)).

[1] Many literary scholars argue the opposite, that the canon should be rethought or revised (von Hallberg, 1983), and that terms like 'classics' belong to the "precritical era of criticism itself" (Guillory, 1995).

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 140–155

Lastly, *3)* the Modern Breakthrough era is a period of significant political and cultural upheaval, where we would expect to see literary innovation.

The Modern Breakthrough (1870-1900) – in Danish, *det Moderne Gennembrud* – marked a significant shift towards realism and naturalism, diverging from the romantic and idealistic styles that characterized the preceding period (D'Amico, 2016). Spearheaded by the influential critic Georg Brandes,[2] this era brought as much cultural as social change (Bjerring-Hansen and Wilkens, 2023). Literature of the period was unprecedented in emphasizing social issues, individualism, and a scientific approach, advocating for art to reflect and critique society (Mai, 2022). Moreover, the period saw a great rise in the number of literary publications (Bjerring-Hansen and Jelsbak, 2010), as well as an incline in previously underrepresented voices: journalists, teachers, and female authors published more (Bjerring-Hansen and Wilkens, 2023).

The current study focuses exclusively on Modern Breakthrough novels, which is not just a methodological choice to increase comparability,[3] but also recognises the novel's central position in the literary field of the late 19th–20th century. In this period, the novel expanded its reach to a broad and diverse readership, whereas poetry largely catered to a limited elite audience (Levine, 2008; Bjerring-Hansen and Jelsbak, 2010). This democratization of literature, coupled with the novel's generic capacity to reflect the complexities of a rapidly changing society[4] – social, political, and personal – renders it the period's most dynamic and malleable genre. It is a genre in which we expect much – even short-term – development in this period. As both a popular and prestigious genre, the novel also reflects the evolving tensions between canonical authority and popular appeal, making it an unparalleled document for tracing literary influence and its directionality in the period.

This paper is structured as following: Section 2 contains a discussion of related work on canonicity, literary innovation and influence, and the Danish literary context. In Section 3, we describe the dataset and annotations. Our methodological pipeline is described in Section 4, and includes the creation of document embeddings using both a multilingual model and TF/IDF, clustering methods for validating embeddings, and measuring diachronic change to explore how the canon and non-canon evolve over time. The results are presented in Section 5, followed by a discussion in Section 6. We finish with concluding remarks in Section 7, and a discussion of the limitations of this study (Section 8).

2 Related Work

2.1 Textual profile of canonical works

The discussion about canon has been torn between two extreme poles, where canonicity is either seen as something conferred 'from above' or as signaling the excellence of particular works 'from below' in terms of text-intrinsic features (Bloom, 1995). Recent studies show a nuanced take on the debate: while they show that text-extrinsic features[5] might be good predictors of canonicity (Brottrager et al., 2021), canonical works also appear to have a unique textual profile compared to non-canonical works (Barré et al., 2023; Brottrager et al., 2021; Porter, 2018). Beyond the binary distinction (canon/non-canon), canonical works exhibit textual profiles different from other types of excellence categories in literature, e.g., bestselling or prize-winning novels (Bizzoni et al., 2024; Wu et al., 2024). They have been found to have a denser nominal style (Wu, 2023) and lower readability, elicit higher LLM perplexity, and show more unpredictable sentimental dynamics (Bizzoni et al., 2024).

The axis along which canonical works are analyzed could be termed 'stylistic difficulty'. Here, traditional linguistic metrics and information theory have been employed to show that texts with greater literary prestige tend to exhibit higher levels of reading difficulty (Algee-Hewitt et al., 2016; Bizzoni et al., 2024; Wu et al., 2024), than more 'popular' works of literature which use a more accessible language, and find a broader audience (Bizzoni et al., 2023). However, few studies go beyond features of linguistic and stylistic complexity in

[2]Brandes' first Copenhagen lecture of the series "Hovedstrømninger" (1871), and the publication of J.P. Jacobsen's *Mogens* (1872) are often pinpointed as the start of the Modern Breakthrough (Bjerring-Hansen and Rasmussen, 2023).

[3]This choice is naturally also restricted by data availability, as a corpus spanning the whole population (i.e., covering the 'dark numbers' of literary production is a rare resource.

[4]The ability to reflect social reality is an often highlighted generic trait, as in the seminal *The Rise of the Novel* of Watt (2001), describing the novel as 'truth to individual experience'. Similarly, Armstrong (1987) suggested that the 19th century novel reflected societal upheaval but was also an important instrument of change in bringing the middle class to light.

[5]I.e., cultural, political or market traits, as in Wang et al. (2019).

examining the canon, although some have shown that both sentiment and semantic profiles may be good predictors of more popular literature (Maharjan et al., 2018; Bizzoni et al., 2024).

Since literature is clearly a multidimensional phenomenon, we ideally take all these textual levels into account when we try to grasp the difference between canon and non-canon. Therefore, we make use of document embeddings, which are able to capture text characteristics at various levels, including stylistics and semantics (Wang et al., 2023; Terreau et al., 2024; Reimers and Gurevych, 2019).

Moreover, while textual metrics are generally used to predict a modern label (e.g., what has been shown to sell well/has become canon), few studies have looked into the dynamics of the literary field within the period itself (Brottrager et al., 2022). Although authors like Henrik Pontoppidan are regarded as canonical and influential today, it is uncertain whether their exceptional status was equally recognized by their contemporaries. Thus, the issue of canonicity is closely tied to the concepts of intertextuality and literary influence, which have traditionally focused on how individual authors were shaped by their predecessors (Bloom, 1975, 2011; Bassnett, 2007). If canonicity can be viewed as a marker of reception, we must consider how latent this reception was. In the context of this study, we are interested in the direction of literary influence, specifically whether books that become canonical influence subsequent novels or if they adapt to the overall novel production.

2.2 Canon and popular literature in the Modern Breakthrough

Reading audiences grew significantly during the Modern Breakthrough, and a more differentiated selection of literature became available to more and better readers than before (see also the increase of novels published in Appendix A) (Bjerring-Hansen and Jelsbak, 2010; Hertel Hans, 1983).[6] In the period, we also see an intellectual disdain for the 'popular novel', what Brandes spoke of as 'døgnlitteratur' (ephemeral literature, or literature 'of the day') (Brandes, 1877). *Døgnlitteratur* included, for example, the historical novel such as by Walter Scott, who had an enormous influence on the Danish and European literary field in the time preceding the Modern Breakthrough (Munch-Petersen Erland, 1978; Lukács, 1964).[7] As a match for the popularity of translated Scott novels (Munch-Petersen Erland, 1978), in the Danish context, especially B.S. Ingemann should be foregrounded. Ingemann had a diverse audience – from sailors to the (Sorø) academy – and received the same disdain from the intellectual elite in the period of the Modern Breakthrough as Scott (Bjerring-Hansen and Rasmussen, 2023). While the scorn of the popular novel was itself not a new phenomenon – also present in the reception of Ingemann (Martinsen, 2012) – it was in the Modern Breakthrough accompanied by a decline in the historical novel genre (Bjerring-Hansen and Rasmussen, 2023), and a rise in what Bjerring-Hansen and Wilkens (2023) have broadly called the 'realist novel', pitching the two types of novels starkly against each other (Bjerring-Hansen and Rasmussen, 2023).

However, this polarization within the genre and the dynamics of trends and innovation in the novel of this period are less explored – a period where the appearance of the Modernists in Danish literature coincided with the decline of the previously very viable popular genre of the historical novel at a time at which the demand for popular literature was on the rise.

3 Data

Our dataset consists of a collection of 838 original Danish and Norwegian novels (1870-1900), with connected metadata, e.g., number of pages, book prizes, and publishing house.[8] Previously, Bjerring-Hansen and Rasmussen (2023) tagged the corpus for whether a work is a historical novel or not. The corpus consists of all original first-edition novels published by Danish publishers in the period.[9]

As we sought to examine the relationship be-

[6] Beyond the growth of novel readership in the period, which Bjerring-Hansen and Wilkens (2023) call a 'reading explosion', Danish daily press also saw a great increase in this period, from 36 newspapers in 1847 to 156 in 1914 (Bjerring-Hansen and Wilkens, 2023).

[7] Moretti has also shown how the historical novel à la Scott gained a predominant position in the literary field 1740-1840, marginalizing older genres (Moretti, 2007).

[8] All novels, including the ones written by Norwegian authors, were published in the Danish language and at Danish publishing houses.

[9] The MiMe-MeMo corpus was compiled by Jens Bjerring-Hansen, Philip Diderichsen, Dorte Haltrup, and Nanna Emilie Dam Jørgensen, based on the Danish book index (*Dansk Bogfortegnelse*). It indexes all publications (1830-), including novels by Norwegian authors at Danish publishers. Creators excluded everything not novels (e.g., short story collections). For details, see Bjerring-Hansen et al. (2022). Version 1.1 (used in the present study) is accessible at: `https://huggingface.co/datasets/MiMe-MeMo/Corpus-v1.1`.

tween today's canonized novels from the Modern Breakthrough and the overall production of the period, we added a tag that informs us about the canonicity of the work. To compare the canons defined by a government-designated committee – which do not include Norwegian authors – with a canon that we assume to be created from a literary expert point of view (and less driven by a political agenda), we create a second canon that includes novels that are listed and mentioned in the lemma 'det moderne gennembruds litteratur' of the encyclopedia *Den Store Danske*.[10] We thus added the following tags to the novels in our corpus:

- **CE Canon**: Cultural/Educational Canon, referring to novels whose titles are included in the Cultural Canon, or whose author is included in the Educational Canon.
- **LEX Canon**: Lexicon Canon, referring to novels that were not included in the Educational Canon and Cultural Canon, but whose author is mentioned in the lists of novels and novellas in the 'det moderne gennembruds litteratur' lemma of *Den Store Danske*.
- **E Canon**: Extended Canon, referring to all novels that are included in CE Canon and/or LEX Canon.[11]
- **Other**: Other, referring to the novels that are neither tagged as historical, nor included in one of the canons.

Statistics of the corpus and every category can be found in Table 1.[12]

	titles	*authors*
Corpus	838	371
Cultural/Educational Canon	36	6
Lexicon Canon	110	19
Extended Canon	114	21
Historical Novels	65 (8)	19 (4)
Other	667	335

Table 1: Statistics on the corpus. Note that there is overlap between the categories: there are titles that are both in the Cultural/Educational Canon and the Lexicon Canon. The numbers between brackets in the Historical Novels category refer to titles that are tagged as a historical novel, but also included in one of the canons.

	$\overline{x}$ **SEB**	**Historical**	**SoI**
Number of Datasets →	24	1	25
Models ↓			
m-e5-large-instruct	**66.65**	**40.10**	**53.38**
m-e5-large	<u>60.69</u>	27.66	44.18
DFM-large	55.14	35.13	<u>45.14</u>
MeMo-BERT	36.85	<u>35.38</u>	36.12

Table 2: The performance of encoder models on the Scandinavian Embedding Benchmark (SEB) tasks and on the custom historical task. The Score of Interest (SoI) reflects the model's average score across tasks. The highest score is in bold, and the second highest is underlined.

4 Methods

We developed a methodological pipeline consisting of the following steps:

1. Choosing embedding model. We test four embedding models to decide on the one best suited for our task and corpus. We test these using a weighted average between a historic clustering task and the Scandinavian Embedding Benchmark (SEB)[13] (Enevoldsen et al., 2024) to get a model performing well generally and across historical documents. The performance of the four models can be found in Table 2. For our models, we use the MeMo-BERT trained on Danish and Norwegian historical documents (Al-Laith et al., 2024), the best-performing Danish sentence encoder DFM-large (Enevoldsen et al., 2023) along with the two best-performing open-weight[14] models on SEB, m-e5-large as well as its prompt-based version m-e5-large-instruct (Wang et al., 2024b). Prompt-based models allow for adaptation of the embedding space depending on the use case and have been shown to improve performance significantly (Muennighoff et al., 2023; Enevoldsen et al., 2024; Wang et al., 2024b) as seen in Table 2 we also find this to be the case. For the prompt-based model, we used the instruction *"Identify the author of a given passage from historical Danish fiction"* for evaluation of the historical task. For readability,

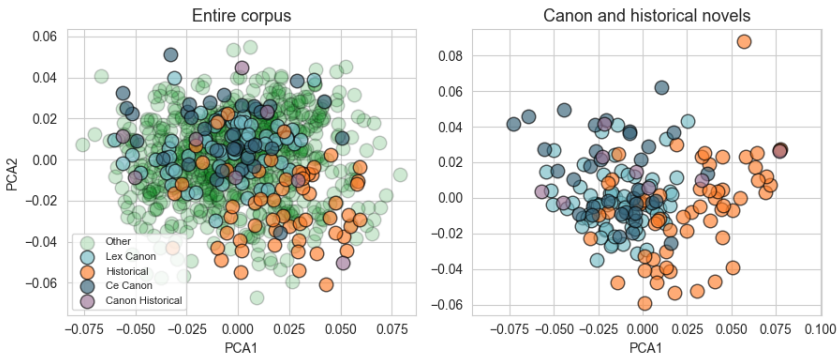
[10] See https://denstoredanske.lex.dk/det_moderne_gennembruds_litteratur. We are aware that the Educational and Cultural Canon and the *Den Store Danske* lemma include more genres than the novel. This paper focuses specifically on the canonical reputation of the novel.

[11] Note that some tags overlap, so that we tag as historical–canon in the following visualizations anything that was both tagged historical and was in either of the canons.

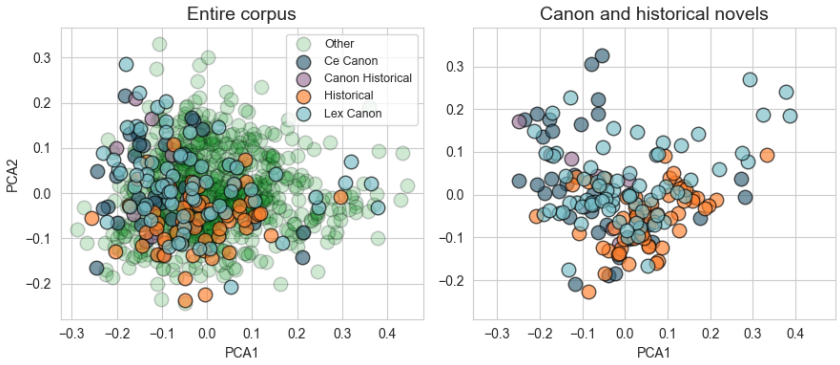
[12] An extended dataset (with added tags) is available on HuggingFace: https://huggingface.co/datasets/chcaa/memo-canonical-novels.

[13] We use the latest version of SEB (v0.13.6).

[14] We avoid using commercial APIs to ensure reproducibility.



(a) embeddings



(b) TF/IDF values

Figure 1: PCAs of the entire corpus (*left*) and the canonical and historical novels (*right*), based on embeddings and TF/IDF values. Note that canon and historical groups cluster more in the PCA based on embeddings.

we short the model names. For full model references along with revision, see Appendix F. We present the construction of the historic task in Appendix G. We continued with the best performing model, m-e5-large-instruct.

2. Creating document representations. We create two types of document representations: our main approach is creating *semantic embeddings*, while we use *lexical embeddings* to validate our semantic embeddings.

- **Semantic embeddings**. We slice each novel into chunks of the same size.[15] Afterwards, we create embeddings for every chunk with the m-e5-large-instruct model, using the same prompt as in the previous step. The average of all document embeddings of one novel is used as a representative embedding for that novel.

- **Lexical embeddings**. After pre-processing the documents (lowercasing, removing punctuation), we create a TF/IDF representation of each novel using sklearn (Pedregosa et al., 2011).

3. Clustering embeddings for method validation. We validate our method by clustering the obtained document embeddings using different measures and visualizations, including dendrograms (see Appendix B) and a PCA as implemented in sklearn (Pedregosa et al., 2011). We use PCA as it preserves the global structure of the embedding space.

4. Measuring diachronic change. We use intra- and inter-group (cosine) similarity to measure how the canon and non-canon evolve over time and how they influence each other.

5 Results

5.1 Validation of embeddings

We perform clustering methods on the two different types of embeddings to verify the novel distribution based on semantic and lexical features. The two PCAs in Fig. 1a are based on the semantic embeddings of the novels. The left PCA shows that overall, the novels that are tagged as canonical and/or historical (colored blue, purple, and orange)

[15] Since the maximum chunk size includes the length of the prompt, we use a chunk size of 512 - 87 = 425 tokens.

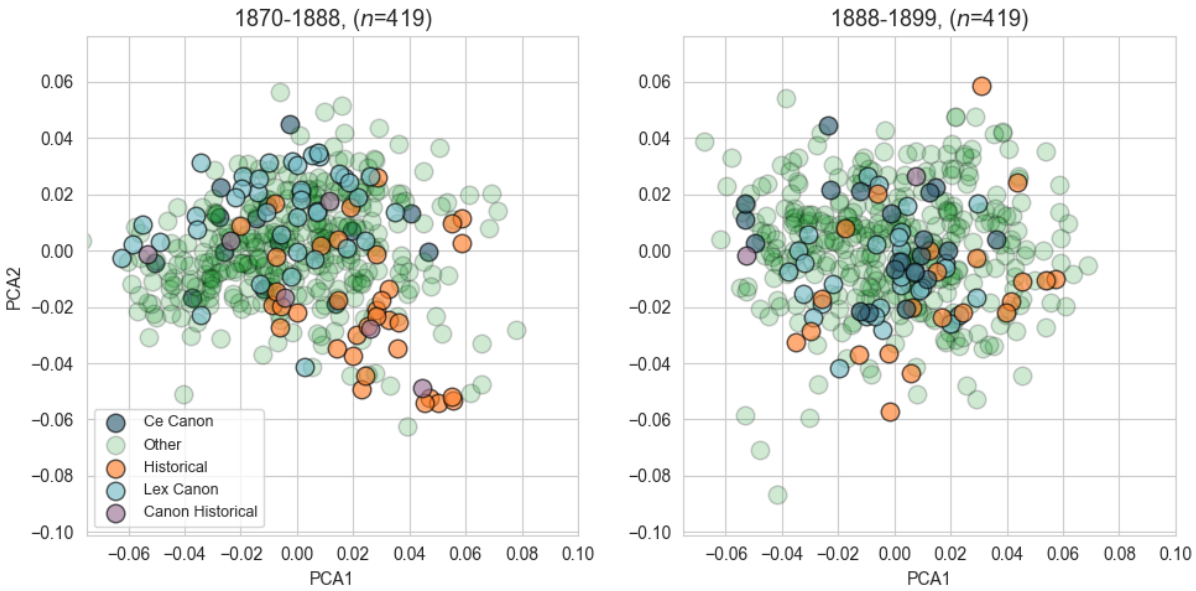


Figure 2: PCA's of (*left*) the first half of the corpus (1870-1888) and (*right*) the second half of the corpus (1888-1899), with novels ordered chronologically.

largely overlap with the rest of the corpus (colored green). This tells us that, at first sight, their semantic style does not differ significantly from the overall literary production during the Modern Breakthrough. When we only look at the canonical and historical novels, as visualised in the right PCA, we see a distinction between the historical novels (orange) and the canonical novels. This can also be confirmed in the dendrogram in Appendix B (Fig. 6a), in which the clear orange cluster of historical novels at the right side of the plot suggests that they share similarity in semantic space.

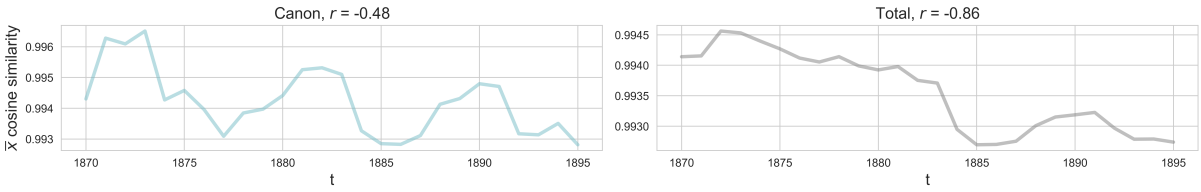
When we compare these figures with the ones based on the TF/IDF values, we see some interesting differences. The PCAs (Fig. 1b) do not show as clear clusters of canonical novels and historical novels, and the same goes for the dendrogram (Fig. 6b in Appendix B). The works of the earlier mentioned author Pontoppidan for example do not cluster on lexical style, and only parts of the historical novels cluster together, while the rest is spread out over the other branches. It suggests that our semantic embeddings go beyond lexical features. This is in line with previous results (Enevoldsen et al., 2024) indicating that document embeddings primarily capture semantics and, e.g., cannot differentiate between correct/incorrect word order.

5.2 Diachronic change

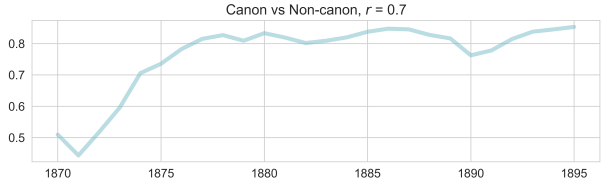
A diachronic comparison between semantic embeddings of the first 419 novels in our corpus (published between 1870 and 1888) and the latter 419 novels (published between 1888 and 1899) shows a difference over time in the relationship between canonical novels and the rest of the corpus (see Fig. 2). While the left PCA shows that canonical novels (colored light and dark blue) cluster on the upper-left side of the green cluster (with non-canonized novels) in the period 1870-1888, the canonized and non-canonized novels show much more overlap in the later period. It suggests that the early Modern Breakthrough works that became canonical differed semantically from the overall production of that time, while the later canonical works were more similar to the other novels of that time.

To verify this potential diachronic change, we compute the mean embedding of canonical novels and non-canon novels for each rolling window (window size of 4 years) across the 30-year period and plot the cosine similarity between the two groups for each window (Fig. 3b).[16] We see a ro-

[16]Due to the discrepancy in group sizes (e.g., canon vs. non-canon) and the overall skewed distribution of our corpus (see Fig. 5 in Appendix A), we used simulation methods to compare cosine similarity across time windows. For each window, we simulated 1,000 Gaussian distributions of cosine similarity for each group based on their respective means and standard deviations. The overall mean of these 1,000 runs was used for each group in the comparison. For intergroup comparisons, we employed the same approach by simulating the mean embedding of each group (1,000 runs per window) and then calculating the average cosine similarity between the groups' embeddings across runs for each window. To ensure



(a) Intra-group similarity across time, using the mean cosine similarity of books in a rolling window ($s = 4, step = 1$) over the years for each category – all books and our extended canon definition. Spearman's rho of the correlation between year and mean intragroup similarity (on top), $p < 0.01$. For each rolling window, group size > 2. Note that both groups tend to get more internally diverse across time, though canon books less so and less consistently.



(b) Inter-group similarity across time: The cosine similarity between the average canon and non-canon embedding per rolling window ($s = 4, step = 1$). Note that a correlation with time persists after 1875 (ρ .42).

Figure 3: Intra-group similarity and inter-group similarity across time.

bust positive correlation with time (Spearman's ρ $.7, p < 0.01$),[17] suggesting that canonical novels become increasingly similar to non-canonical novels. This positive correlation is partly due to the steep increase in inter-group similarity before 1875. Nevertheless, the correlation with time persists after that year (ρ .42). The flattening of the curve might suggest that certain subgenres continue to disappear (like the decline of the historical novel), or it might reflect how the entire literary field is more standardized from 1875 and onwards.

When we look at the intra-group similarity of all canonical works – using the same rolling window to extract intra-group cosine similarity over time – we see a decrease in similarity internally in the canon group, suggesting that the canon group becomes more internally diverse over time. The same trend, though slightly stronger, can be observed in the corpus as a whole (Fig. 3a).[18] Note that these correlations of intra- and inter-group similarity over time hold regardless of which model (among those tested) is used to create embeddings (see Table 3 in Appendix C.

To detect whether the canon moves towards the non-canon over time, or the other way around, we gauge the directionality of both groups. We split the corpus in two equal parts, in the same way as done for Fig. 2. The result consist of four subsets: early non-canonical novels, early canonical novels (both pre-1888), late non-canonical novels and late canonical novels (both post-1888). We have plotted all novels in one PCA (Fig. 4), using colors to distinguish between the four subsets. We fit the mean embedding of every subset to the same PCA. The resulting plot shows that the late non-canon has moved up in the direction of the early canon, suggesting that the novels that today have a canonical status, behaved as trendsetting novels at the time. In Appendix E, we include an alternative version of this PCA, using a rolling window size of 5 years (step 1) to show that the non-canon moves towards the canon.

6 Discussion

Seeking to validate what our embeddings capture, we compared canonical and historical novels both in terms of embedding space and lexical similarity (TF/IDF). The fact that these two groups of novels – canonical and historical – cluster differently at the level of embeddings is interesting for two reasons: firstly, it suggests a maintained coherence of the historical genre in this period, although it was in decline (Bjerring-Hansen and Rasmussen, 2023). Secondly, since this is not an equal comparison – one group being a genre and the other a category spanning a diversity of novels – we find that

results are not skewed because we assume a normal distribution, we also tested a bootstrap sampling, which yielded similar results.

[17] Due to our simulation approach, results may vary slightly for each run, so correlation coefficients with time should be taken as estimations rather than precise values.

[18] For further validation of the limited range of cosine similarity values in our study, see Figure 7b in Appendix D.

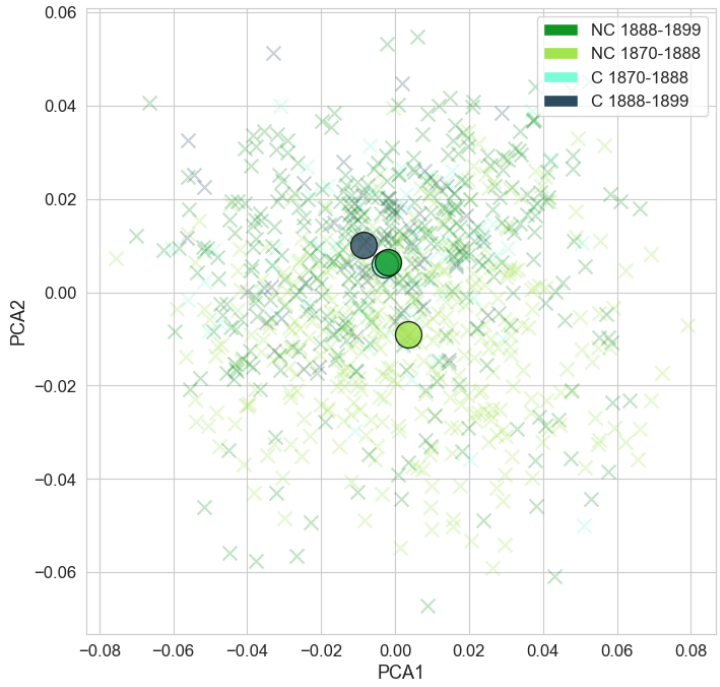


Figure 4: Positions of the mean embedding of the canon/non-canon groups over two time-periods (1870-1888, 1888-1899). Note that the later period non-canon mean seems to have moved closer to mean canon position. The PCA was fitted to all embeddings, then mean embeddings per group per period were fitted into the same PCA.

the clustering must suggest that our embeddings capture more than thematic elements, which more usually distinguish more popular genres (Moreira and Bizzoni, 2023).

Moreover, our clustering experiments show that different constructions of the canon, i.e., the more political, smaller canon, and the expert canon which includes non-Danish authors to a higher degree, do not seem to differ significantly: both versions include novels that stand out based on their semantic features. The textual features of the small selection of novels included in the Cultural Canon and the Educational Canon are not different from those of the novels that are included in what we call the Lexical Canon. This can be derived from the way in which novels from both groups have close proximity to each other and are mixed in the dendrograms (Figs. 6a and 6b).

For our main result, we show that the Modern Breakthrough novels that have a canonical status today show traits at the textual level which suggest their innovation and distinction from their non-canonical contemporaries, contra to our H1. This ties in with what is shown in earlier studies on the literary canon: canonical novels have text-intrinsic features that distinguish them from other novels (Brottrager et al., 2022; Bizzoni et al., 2024; Wu, 2023). Moreover, our results suggest that canonical novels behaved innovatively, introducing characteristics in the Modern Breakthrough era, which resonated in the literary production that came after – non-canonical novels in a sense tracing after the canonical novels. In a diachronic comparison, we see that non-canonical novels adapt to the canon, likely copying the innovative themes and style that these trendsetters introduced. This supports our H2. It makes the current study – to the best of our knowledge – the first that uses embeddings to show the relationship between canonical and non-canonical novels in terms of innovation and influence.

Furthermore, the decreasing intra-group similarity of both the canon and non-canon reflects the diversification of the literary field during the period, as outlined by (Bjerring-Hansen and Jelsbak, 2010; Hertel Hans, 1983). A larger variation of novels saw the light of day, and a more diverse selection of novels became canonical, reflected in canonical novels' increasing internal diversity. While this study confirms the textual innovation of canonical novels, it is possible that there are novels that show the same textual profile as the canon but that did not get canonical status. Future research should provide more insight into these potentially innovative

but today lesser known novels. Moreover, while we argue that the canonical status of novels might be related to their textual profile, we acknowledge that text-extrinsic features could also have played a role here. Future work could, therefore, explore the relationship between the canonical status of a novel and features such as the price of the book, and the publishing house – something the current dataset allows for.

A last note concerns the directionality of the canon and non-canon works, where our results suggest that works in the canon group may have acted as literary trendsetters. Further research, employing more sophisticated methods to gauge causality is needed to confirm our suggestion. Moreover, future work could compare the embedding spaces of canon and non-canon groups to the embedding space of non-fiction texts, as the latter may serve as a useful reference point for assessing the movement and direction of fiction works.

7 Conclusion

We have examined the relationship between the canonization of Danish novels and their textual innovation and influence, taking the Danish Modern Breakthrough era (1870–1900) as a case study. We created embeddings of the 838 novels in our corpus, and used a custom historic clustering task to decide on the best suited model for our task and corpus, which turned out to be the multilingual m-e5-large-instruct model. We validated our embeddings by creating a TF/IDF representation of each novel. Our results show that the embeddings capture semantic features and go beyond lexical features: historical novels and canonical novels cluster differently. Inter-group similarity shows that the similarity of canonical and non-canonical novels increases over time, while at the same time, intra-group similarity decreases, indicating that the canon group as well as the overall novel production becomes more internally diverse over time. We finally show that the non-canon moves towards the canon, suggesting that non-canonical novels adapt to the canon, possibly copying the innovative themes and style of these trendsetters.

8 Limitations

Prompts for embeddings: This work utilizes the prompt-based embedding model m-e5-large-instruct, and thus, it is likely that notably different results could have been obtained by changing the prompt. We examine this further in Appendix H.

Occurrence within training data: Canonical works are more likely to appear online or outside their original context due to their popularity. This could lead to differences in embeddings when using models trained on large web-based data sources simply because paragraphs from these novels may appear in varied contexts within the training data. However, we consider this influence to be minor, as historical novels likely represent only a small fraction of online discourse. This is especially the case for the multilingual embedding model used, where Danish likely comprises only a tiny fraction of the training data. Ideally, the training data should be examined to ensure this influence is not significant. However, this approach is often unfeasible, as pre-training data for these models is typically unavailable, and exploring it would require extensive computational resources. Additionally, the fact that historical canon has often been rewritten further complicates such efforts.

Canon definition: the concept of canonicity is inherently vague and subject to various interpretations. Our canon definition and our binary classification of canonical works may oversimplify a concept that may be better represented as a continuous variable (Brottrager et al., 2022). Our rationale in using two ideal classes (canon/non-canon) was to get an estimate of the difference between them, though it should be noted that the transition between them may be more fluid than it is represented here.

Acknowledgements

The authors of this paper were supported by grants from the Carlsberg Foundation (grant title: *The Golden Array of Danish Cultural Heritage*) and the Aarhus Universitets Forskningsfond (grant title: *Golden Imprints of Danish Cultural Heritage*). Part of the computation done for this project was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark.

References

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024. Development and evaluation of pre-trained language models for historical Danish and Norwegian literary texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources*

and Evaluation (LREC-COLING 2024), pages 4811–4819, Torino, Italia. ELRA and ICCL.

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2016. *Canon/Archive. Large-scale Dynamics in the Literary Field*. Stanford Literary Lab.

Nancy Armstrong. 1987. *Desire and domestic fiction: a political history of the novel*, nachdr. edition. Oxford Univ. Press, New York.

Jean Barré, Jean-Baptiste Camps, and Thierry Poibeau. 2023. Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature. *Journal of Cultural Analytics*, 8(3).

Susan Bassnett. 2007. Influence and Intertextuality: A Reappraisal. *Forum for Modern Language Studies*, 43(2):134–146.

Yuri Bizzoni, Pascale Feldkamp, Ida Marie Lassen, Mia Jacobsen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024. Good Books are Complex Matters: Gauging Complexity Profiles Across Diverse Categories of Perceived Literary Quality.

Yuri Bizzoni, Pascale Moreira, Nicole Dwenger, Ida Lassen, Mads Thomsen, and Kristoffer Nielbo. 2023. Good reads and easy novels: Readability and literary quality in a corpus of US-published fiction. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 42–51, Tórshavn, Faroe Islands. University of Tartu Library.

Jens Bjerring-Hansen and Torben Jelsbak. 2010. *Boghistorie*, 1 edition. University Press, Aarhus.

Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. Mending Fractured Texts. A heuristic procedure for correcting OCR data: 6th Digital Humanities in the Nordic and Baltic Countries Conference, DHNB 2022. In *CEUR Workshop Proceedings*, volume 3232, pages 177–186, Uppsala, Sweden.

Jens Bjerring-Hansen and Sebastian Ørtoft Rasmussen. 2023. Litteratursociologi og kvantitative litteraturstudier: Den historiske roman i det moderne gennembrud som case. *Passage - Tidsskrift for litteratur og kritik*, 38(89):171–189.

Jens Bjerring-Hansen and Sebastian Ørtoft Rasmussen. 2023. Litteratursociologi og kvantitative litteraturstudier: Den historiske roman i det moderne gennembrud som case. *Passage - Tidsskrift for litteratur og kritik*, 38(89):171–189. Number: 89.

Jens Bjerring-Hansen and Matthew Wilkens. 2023. Deep distant reading: The rise of realism in Scandinavian literature as a case study. *Orbis Litterarum*, 78(5):335–352. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/oli.12396.

Harold Bloom. 1975. *The Anxiety of Influence: A Theory of Poetry*. Number 426 in A Galaxy Book ; GB. Oxford Univ. Press, London.

Harold Bloom. 1995. *The Western Canon: The Books and School of the Ages*, 1st riverhead ed edition. Riverhead Books, New York, NY.

Harold Bloom. 2011. *The Anatomy of Influence: Literature as a Way of Life*, 1 edition. Yale University Press, New Haven.

Georg Brandes. 1877. *Hovedstrømninger i det 19de Aarhundredes Litteratur : Forelæsninger holdte ved Kjøbenhavns Universitet 1871-1887. : Emigrantlitteraturen. 2. udgave*. Kjøbenhavn, Gyldendal.

Judith Brottrager, Annina Stahl, and Arda Arslan. 2021. Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features. In *CEUR Workshop Proceedings*, pages 195–205, Antwerp, Belgium. CEUR.

Judith Brottrager, Annina Stahl, Arda Arslan, Ulrik Brandes, and Thomas Weitin. 2022. Modeling and predicting literary reception. *Journal of Computational Literary Studies*, 1(1):1–27.

Giuliano D'Amico. 2016. Modern Breakthrough. In *Routledge Encyclopedia of Modernism*, 1 edition. Routledge, London.

Kenneth Enevoldsen, Lasse Hansen, Dan S. Nielsen, Rasmus A. F. Egebæk, Søren V. Holm, Martin C. Nielsen, Martin Bernstorff, Rasmus Larsen, Peter B. Jørgensen, Malte Højmark-Bertelsen, Peter B. Vahlstrup, Per Møldrup-Dalum, and Kristoffer Nielbo. 2023. Danish foundation models.

Kenneth Enevoldsen, Kardos Marton, Niklas Muennighoff, and Kristoffer L. Nielbo. 2024. The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding.

Johannes Fibiger. 2004. Kampen om litteraturhistorien. Om veje ind i og ud af litteraturen. *Bogens Verden. Tidsskrift for kultur og litteratur*, (6).

John Guillory. 1995. *Cultural Capital: The Problem of Literary Canon Formation*. University of Chicago Press.

Steen Harbild, Stefan Hermann, and Steen Lassen, editors. 2004. *Dansk litteraturs kanon: rapport fra kanonudvalget*, 1. udg. ; 1. opl edition. Undervisningsministeriets forlag, København.

Hertel Hans. 1983. *Den daglige bog: bøger, formidlere og læsere i Danmark gennem 500 år*. Forening for Boghaandværk, Kbh.

George Levine. 2008. *How to read the Victorian novel*. How to study literature. Blackwell Pub, Malden, MA.

György Lukács. 1964. *Probleme des Realismus*. Neuwied: Luchterhand.

Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018. Letting emotions flow: Success prediction by modeling the flow of emotions in books. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 259–265, New Orleans, Louisiana. Association for Computational Linguistics.

Anne-Marie Mai. 2016. Canons and Contemporary Danish Literature. *Folia Scandinavica Posnaniensia*, 19(1):109–132.

Anne-Marie Mai. 2022. *Danish Literature from 1000 to 1900*. Syddansk Universitetsforlag/University Press of Southern Denmark, Odense.

Lone Kølle Martinsen. 2012. Bondefrihed og andre verdensbilleder idehistoriske studier af b.s. ingemanns danmarkshistorie 1824-1836. *Temp - tidsskrift for historie*, 3(5):75–103.

Pascale Moreira and Yuri Bizzoni. 2023. Dimensions of quality: Contrasting stylistic vs. semantic features for modelling literary quality in 9,000 novels. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 739–747, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Franco Moretti. 2000. Conjectures on World Literature. *New Left Review*, (1):54–68.

Franco Moretti. 2007. *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, London New York.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Munch-Petersen Erland. 1978. *Romanens århundrede: studier i den masselæste oversatte roman i Danmark 1800-1870.* Ph.D. thesis, Forum, Kbh. Book Title: Romanens århundrede : studier i den masselæste oversatte roman i Danmark 1800-1870. ISBN: 9788755307179.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Jack Douglas Porter. 2018. Popularity/prestige. Technical report, Stanford Literary Lab.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Ally Smid. 2022. In de nieuwe literaire canon moet je de schrijfsters met een lampje zoeken. *Trouw*.

Enzo Terreau, Antoine Gourru, and Julien Velcin. 2024. Capturing Style in Author and Document Representation. ArXiv:2407.13358 [cs].

Willie Van Peer. 2008. Ideology or aesthetic quality? In Willie Van Peer, editor, *The Quality of Literature: Linguistic Studies in Literary Evaluation*, volume 4 of *Linguistic Approaches to Literature*, pages 17–29. John Benjamins Publishing Company, Amsterdam.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Robert von Hallberg. 1983. Editor's Introduction. *Critical Inquiry*, 10(1):iii–vi.

Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. Can Authorship Representation Learning Capture Stylistic Features? *Transactions of the Association for Computational Linguistics*, 11:1416–1431. Place: Cambridge, MA Publisher: MIT Press.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Xindi Wang, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási. 2019. Success in books: predicting book sales before publication. *EPJ Data Science*, 8(1):31.

Ian P. Watt. 2001. *The rise of the novel*. University of California Press, Berkeley.

Mary Ann Frese Witt. 2000. Are the Canon Wars Over? Rethinking Great Books. *The Comparatist*, 24:57–63.

Yara Wu. 2023. Predicting the Unpredictable. Using Language Models to Assess Literary Quality. Master's thesis, Uppsala University, Uppsala.

Yaru Wu, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo. 2024. Perplexing canon: A study on GPT-based perplexity of canonical and non-canonical literary works. In *Proceedings of the 8th Joint SIGHUM*

Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), pages 172–184, St. Julians, Malta. Association for Computational Linguistics.

A Distribution of titles

We see the distribution of titles per category in our corpus (1870-1900) in Figure 5.

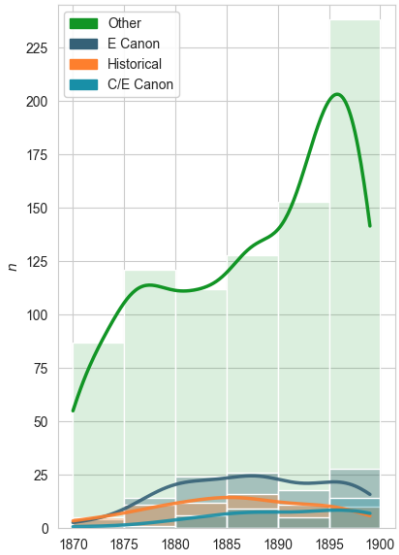


Figure 5: Distribution of titles per category in our corpus (1870-1900). This figure also reflects the actual incline of novels published in the period.

B Dendrograms

We see the dendrograms visualized in Figure 6.

C Intra and inter-group similarity

When calculating the similarity of books within groups over time (intra-group similarity), as well as the similarity between the canon and non-canon group over time (inter-group similarity), we used embeddings based on the m-e5-large-instruct model, using the prompt "Identify the author of a given passage from historical Danish fiction.". The direction of change of both intra- and inter-group similarity proved consistent when using embeddings based on other models. For reference, in Table 3, we show the correlation over time for the various models tested. In the same table, we also show the correlation when using unprocessed means of rolling windows over time versus when using means of simulated distributions, as described in Section 5.2 ($s = 4, step = 1$). Note, again, that correlations vary slightly for each time we run the rolling window with simulated means (1,000 simulations of Gaussian distributions per window), so that the correlation coefficient should be taken as an indication rather than an exact value.

D Cosine similarity ranges

It is clear that cosine similarities are very high in our analysis. As noted, we use pooled embeddings which may affect a higher cosine similarity due to information loss. However, cosine similarity values are also high when comparing embeddings of raw chunks. As noted in the model card, the m-e5-large-instruct, cosine similarity scores of embeddings produced with this distribute in a narrow (and high) range [19]. Developers note: "This is a known and expected behavior as we use a low temperature 0.01 for InfoNCE contrastive loss. For text embedding tasks like text retrieval or semantic similarity, what matters is the relative order of the scores instead of the absolute values, so this should not be an issue."

In Fig. 7, we show the distribution of cosine similarities for both raw and pooled embeddings for our corpus. Note that while the distribution of pooled embeddings does show a skew toward higher cosine similarity values, cosine scores of raw embedding chunks also exhibit a narrow range with a high mean. We therefore consider the very high cosine similarity scores an artefact of the model, rather than an effect of the pooling procedure per se.

E Canon directionality

We see the PCA with the mean embeddings of the canon/non-canon, plotted with a rolling window, visualized in Figure 8.

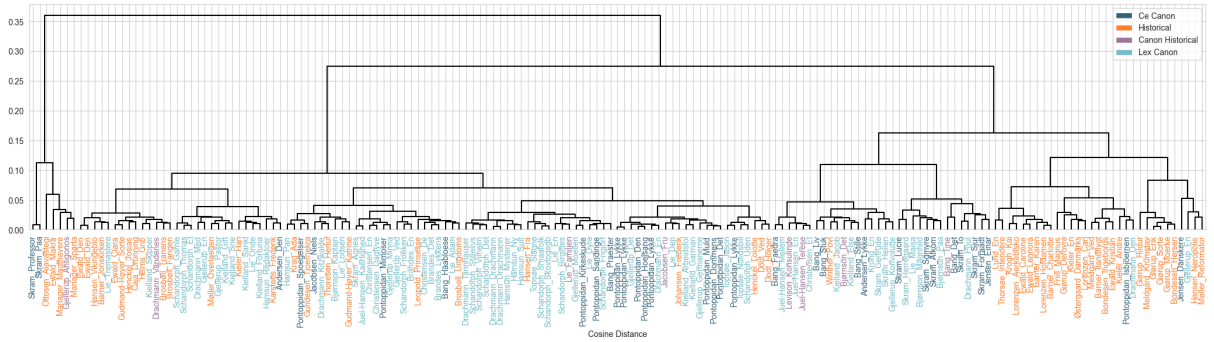
F Model References and Names

As many models are often updated, leading to a change in their output, we ensure reproducibility by specifying the revision IDs used in Table 4. The table also maps short-form model names used in the paper with their reference names as they appear on Huggingface.

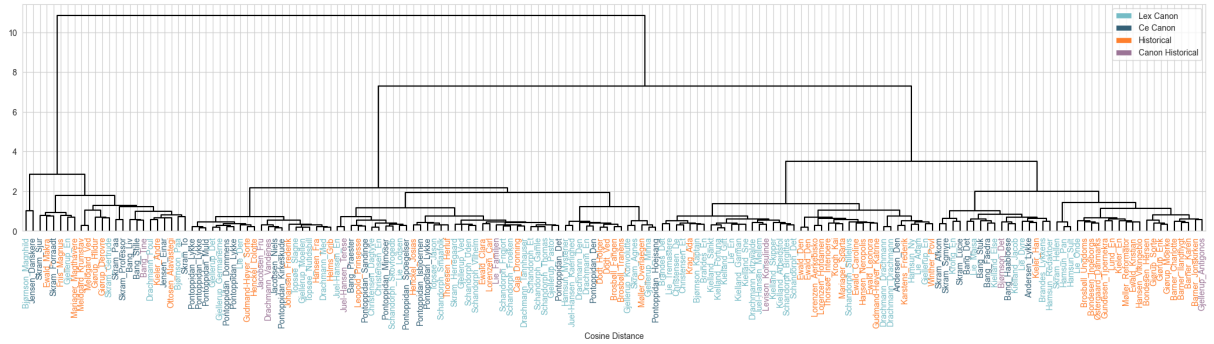
G Construction of Historic Evaluation task

Using the MiMe-MeMo corpus (Bjerring-Hansen et al., 2022), we similarly construct a clustering task as done in (Muennighoff et al., 2023; Enevoldsen et al., 2024). We down-sampled our corpus to 64 novels and took 32 chunks from each, adding up to 2,048 passages. The goal of the task is to see how well clusters of the embeddings correspond with the original authors. Clustering

[19] See FAQ, question 3, https://huggingface.co/intfloat/multilingual-e5-large-instruct



(a) embeddings



(b) TF/IDF values

Figure 6: Dendrograms based on cosine similarity of semantic embeddings (*a*) and TF/IDF (*b*). Dendrograms were calculated using Ward variance minimization, as implemented in SciPy v1.14.1 (Virtanen et al., 2020). Note that titles in purple are historical novels by – in our wider definition – canonical authors ($n = 171$).

	m-e5-large	m-e5-l-instruct (*identify*)	m-e5-l-instruct (*retrieve*)	DFM-large	MeMo-BERT
Canon	-0.61 (-0.82)	-0.48 (-0.62)	-0.59 (-0.73)	-0.68 (-0.83)	**-0.78** (-0.77)
Non-canon	-0.81 (-0.87)	**-0.87** (-0.90)	-0.81 (-0.84)	-0.76 (-0.75)	-0.77 (-0.81)
Total	-0.80 (-0.85)	**-0.86** (-0.91)	-0.78 (-0.82)	-0.76 (-0.79)	-0.72 (-0.76)
Canon/non-canon	0.67 (0.80)	**0.70** (0.74)	**0.70** (0.77)	0.62 (0.84)	0.68 (0.66)

Table 3: Correlation of intra and inter-group similarity over time using embeddings based on all models. Correlation over time based on the rolling windows' simulated means and correlation over time between actual values *in parenthesis*. Note that we show the results of the m-e5-large-instruct model when instructed with two different prompts, "retrieve" and "identify", see Table 4 in Appendix H for the full prompt, prompt 1 & 5. The strongest correlation is in bold, the second strongest is underlined. For all correlations, $p < 0.01$.

is performed using a K-means clustering of the authors of the passages. The performance is measured using V-scores similar to SEB (Enevoldsen et al., 2024). For the prompt-based model, we used the prompt *"Identify the author of a given passage from historical Danish fiction"*. To encourage future evaluations of historical Danish and Norwegian documents, we contribute our newly developed task to the Scandinavian Embedding Benchmark (SEB) in a pull request: https://github.com/KennethEnevoldsen/scandinavian-embedding-benchmark/pull/184.

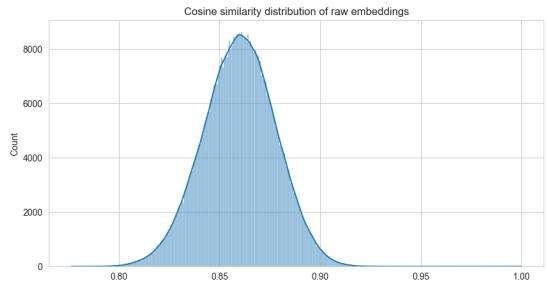
H Instruction prompts

When generating the prompt, we followed the format used in (Wang et al., 2024a), where instructions for all clustering tasks start with the word "Identify". We evaluated the performance of several versions of our final prompt on the custom historical task, which can be seen in Table 5.

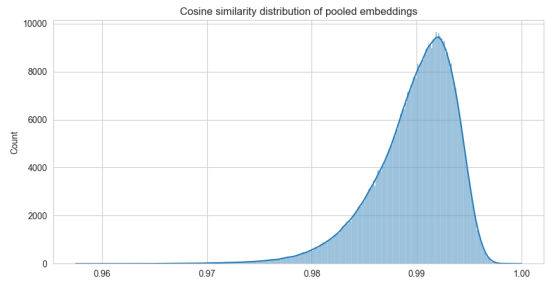
Specifics of the formulation do not seem to have a large impact on performance; Prompts 1 and 2 perform similarly. Performance drops with prompts 3 and 4, which instructs the model to perform a different task, than it is evaluated on (cluster on books instead of authors). Finally, using a differ-

Name in Paper	Reference Name and Revision
m-e5-large-instruct	intfloat/multilingual-e5-large-instruct `baa7be480a7de1539afce709c8f13f833a510e0a`
m-e5-large	intfloat/multilingual-e5-large `ab10c1a7f42e74530fe7ae5be82e6d4f11a719eb`
DFM-large	KennethEnevoldsen/dfm-sentence-encoder-large-exp2-no-lang-align `ec8293d8f447023de99d1e7fb79aa918d6258dc7`
MeMo-BERT	MiMe-MeMo/MeMo-BERT-03 `04cad875b848b56d9a76e80a031d60d66ae9cd02`

Table 4: Model names as their appear in the paper along with reference name as their appear of hugging face along with the revision ID.



(a) Cosine similarity range of raw embedding chunks. For this figure, we used the 20th chunk of each book and calculated cosine similarity between all pairs.



(b) Cosine similarity range of pooled embeddings of all books.

Figure 7: The range of cosine similarity scores for raw and pooled embeddings of the m-e5-large-instruct used for the main analysis, i.e., with prompt (1), see Table 5 in Appendix H.

ent task keyword (e.g. "Classify" instead of "Identify") has some impact on the performance, as can be seen in prompts 5 and 6. This is likely the result of the training procedure of m-e5-large-instruct, as the model learns to embed the text conditional on the task prompt. For example, with a task definition that asks the model to retrieve, the model is likely trying to find a good query vector that lands close to relevant documents in embedding space, instead of embedding similar documents close together as is the goal of a clustering task.

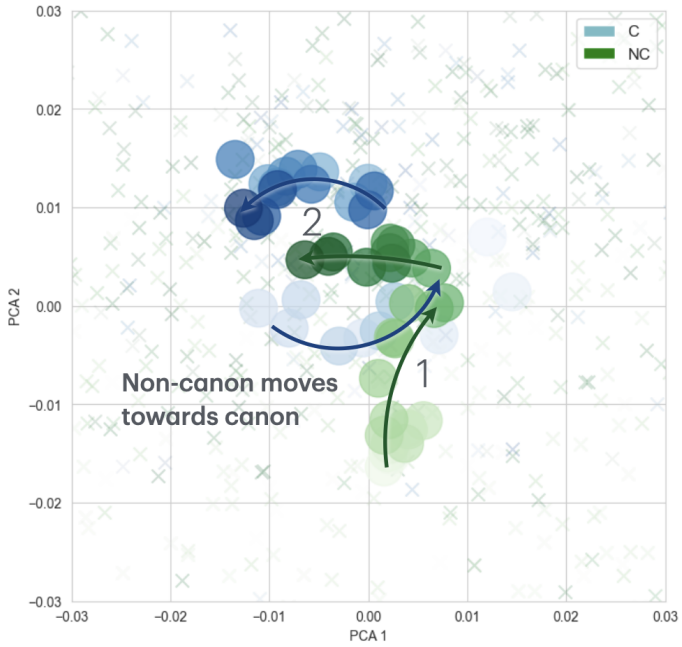


Figure 8: Positions of the mean embedding of the canon/non-canon, plotted with a rolling window size of 5 years, step 1. Time is indicated by the shading: darker colors are later in time. Note that the non-canon moves towards the canon. The PCA was fitted to all embeddings, then mean embeddings per group per window were fitted into the same PCA.

ID	Prompt	Historical
1	Identify the author of a given passage from historical Danish fiction	40.10
2	Identify the author of a specified passage taken from historical Danish literature	42.29
3	Identify which book from Danish historical fiction does the passage belong to	33.04
4	Identify the work from Danish historical fiction to which the provided passage belongs	34.35
5	Retrieve the author of a given passage from historical Danish fiction	42.56
6	Classify the author of a given passage from historical Danish fiction	46.43

Table 5: The performance of m-e5-large-instruct on SEB's custom historical task using different prompts.