Patterns of Canon: A Multilingual Network Study

Judith Brottrager[1], Jean Barré[2], Yuri Bizzoni[3], and Pascale Feldkamp Moreira[3]

[1] Technical University of Darmstadt, Darmstadt, Germany
[2] LaTTiCe-CNRS, École Normale Supérieure - PSL University, Paris, France
[3] Center for Humanities Computing, Aarhus University, Aarhus, Denmark

Abstract

This study examines whether canonical literature exhibits consistent structural signatures in networks of textual similarity across languages. Using four diachronic corpora of prose fiction (5,000 texts, 17th–21st century), we construct time sensitive similarity networks, with edges weighted by textual proximity. Canonical status—derived from multiple markers of canonization—is analyzed in relation to various centrality measures. We find that global metrics such as betweenness and degree centrality show limited association with canonization, while sub-cluster centrality offers a more reliable signal. Our analysis also reveals temporal shifts: in some periods, canonical status aligns with structural centrality, while in others the correlation weakens or even reverses, with canonization favoring more distinctive or peripheral works.

Keywords: canonicity, network analysis, computational literary studies, multilingual text analysis, multilingual stylistics, multilingual stylometry

1 Introduction

Highly canonized texts are central to the way we teach, tell, and write both literature and literary history. They feature in school and university curricula, are well-explored subjects of literary studies, and often become part of the reference system of national cultures. While it is not necessary that canonized texts share aesthetic or poetic characteristics,[1] they are part of what Gadamer calls a "living cultural tradition" [25, p. 161]: evolving through interpretation and re-interpretation across generations, they are maintained in cultural memory and kept accessible for new generations of readers. Evaluative acts [28] ensure this cultural accessibility—such as being "[r]epeatedly cited and recited, translated, taught and imitated, and thoroughly enmeshed in the network of intertextuality" [40, p. 53]. Such repetition and reinforcement can also be assumed to have a concrete influence on literary historical developments. Because canonized texts remain part of what Assmann describes as the "active working memory" [3, p. 106] of culture, it is plausible that their textual characteristics are copied and reproduced, leading to patterns of text similarity that place them in central positions.

If culturally central texts are in fact structurally central, this centralities can be modeled and examined using network models of text similarities: these networks encode similarity relationships between texts as a system of nodes—representing the texts themselves—and edges—the formalized similarity between a pair of nodes. By building on these pairwise similarities, relationships between texts can be described and contextualized at both local and global levels, with the advantage that underlying similarities can be analyzed across languages despite surface-level linguistic differences. As a result, the networks can act as cross-linguistic and cross-cultural comparative frameworks, comparing texts based on the network positions they are assigned.

Judith Brottrager, Jean Barré, Yuri Bizzoni, and Pascale Feldkamp Moreira. "Patterns of Canon: A Multilingual Network Study." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 59–77. https://doi.org/10.63744/Y0dv4ooACREY.

[1] Recent research has found similarities between canonical texts based on more complex features [8; 48].

As we are interested in trans-national patterns of canonicity, we use four corpora of national literatures to explore *whether cultural centrality is reflected in structural centrality*. For this, we operationalize intra-corpus similarities based on low-level textual features, ranging from type-token ratio, readability, and compressibility to average word and sentence length, to model the relationships between corpus texts as similarity networks. Using these network models, the structural centrality of each text, measured by various centrality metrics, can then be compared to its centrality in the canon. We represent a text's status in the canon by scores that summarize markers of canonization processes, such as inclusion in literary historiography and presence in schools and universities [14]. Further exploring the network structures using temporal filtering and clustering algorithms—in our case, Infomap clustering [20]—we look for clusters with higher or lower correlations between canonization scores and centralities to add more localized perspectives to a possible global link between a text's status in the canon and its position in the network. By introducing temporal constraints, we can also model the relationship between cultural and structural elements over time, identifying temporal patterns in national canons.

With this methodological framework, we imply and test a relationship between degrees of canonization and textual features over multiple national literary traditions. There are at least two reasons to suspect such a correlation. On the one hand, canon formation may stabilize certain stylistic conventions—through imitation, teaching, and institutional reinforcement—so that canonical works come to define linguistic norms detectable in low-level features such as sentence length or lexical diversity at the literary level. On the other hand, the relationship may work in the opposite direction: texts with exceptional stylistic efficacy or evocative power may endure because such qualities attract critical and cultural attention over time. In both cases, measurable surface traits can act as faint but persistent traces of literary value attribution as it is transmitted, negotiated, and reinforced over generations.

2 Background and Related Work

Canon formation and the problem of the 'Great Unread' [16] have been central to the earliest contributions of what is now called Computational Literary Studies (CLS) [31; 34; 44]. Building on this foundation, numerous computational studies have sought to operationalize concepts such as canon, prestige, popularity, and literary quality through quantitative data analysis. Many adopt a binary view of the canon, using inclusions in curated lists such as literary histories, prize archives, or educational syllabi as a basis for classification [1; 4; 5; 15; 22; 48]. Others rely on indicators of prestige, such as the number of literary prizes or presence in literary encyclopedias [45], as well as academic attention measured through references in journals or bibliographies [37; 44; 45]. Reprint frequency, translations, and inclusion in central text collections are also used to assess popularity and cultural reach [1; 35], while more direct reader perception has been captured through surveys [18], and platforms such as Goodreads, which can act as proxies for broad audience engagement [37]. While binary approaches make it easy to compare two well-defined groups, other models highlight *gradation*, for instance by quantifying the degree of scholarly attention [17] or measuring representation across multiple canon sources [14]. Our approach falls into this latter category of operational strategies: using different markers of canonization (see Section 3), we model canonicity as a continuous variable rather than a binary label.

To capture how this value distributes across literary corpora, we turn to network-based methods that have already been used to model character interaction graphs—from Moretti's analysis of Hamlet [34, pp. 211–240] to the large-scale infrastructure of DraCor [23]. These studies quantify interaction frequency [43], information flow [2], and character roles based on network centrality [7]. Beyond drama, co-occurrence networks have been used to model the co-appearance of poems in anthologies [29; 30] or link topics in topic models via shared terms [26, pp. 177–188]. In stylometry, Eder [19] introduced networks as a way of visualizing text similarities based on distance

metrics; this approach has since been extended to capture thematic and stylistic patterns beyond authorship attribution [27; 36; 46].

Because networks model similarity at scale and introduce an additional level of abstraction beyond surface linguistic features, they provide a framework for comparative analysis. By defining text similarities in terms of positions within an interconnected system, network models can help address a key methodological challenge in CLS: enabling meaningful comparisons across linguistic and national boundaries [47]. Unlike approaches that rely on raw linguistic forms, network-based methods focus on relational patterns, avoiding difficulties inherent in cross-linguistic comparison. Formalized character networks in drama texts [23], for example, facilitate comparisons of dramatic structure across different national traditions, languages, and historical periods. Stylometric studies have also used network visualizations and similarity metrics to examine multilingualism and translation across national literatures [27; 38; 39], showing that network models can operationalize cross-linguistic comparison at scale.

3 Data

3.1 Corpora

For the implementation of our cross-linguistic comparison, we use four different corpora of European languages and literary traditions. These corpora, summarized in Table 1, vary in size, temporal scope, and conception: the English and German corpora cover the period from 1688 to 1914—the long eighteenth and nineteenth centuries—, the Danish corpus focuses on the late nineteenth century, and the French corpus spans from 1811 to 2020.

Language	Period	# Texts	Markers of canonization
Danish	1870–1900	839	National canon lists, Danish encyclopedia presence, entry length, intra-lexical references, title mentions
English	1688–1914	679	Complete/collected works editions by authors, student editions, mentions in literary history, university reading lists
French	1811–2020	2,961	student editions, school examination lists, university syllabi, literary awards, complete/collected works editions by author
German	1688–1914	571	Complete/collected works editions by authors, student editions, mentions in literary history, university reading lists

Table 1: Overview of corpora and markers of canonization

In addition to these differences in coverage, the corpora have been created according to diverging compilation strategies: The Danish corpus, covering a shorter time span of thirty years, is a near-complete collection comprising almost all Danish novels—from highly canonized to unknown—from the period 1870–1899, known as the 'Modern Breakthrough'.[2] The French corpus originates from the *ANR Chapitres* project [32] and comprises 2,961 digitized novels from the early nineteenth to the early twenty-first century. Although it spans a broad temporal range, it is

[2] The Danish MiMe-MeMo corpus was compiled by Jens Bjerring-Hansen, Philip Diderichsen, Dorte Haltrup, and Nanna Emilie Dam Jørgensen, based on the Danish national bibliography—Dansk Bogfortegnelse. It indexes all publications from 1830 onward and includes novels by Norwegian authors published in Denmark. Non-narrative prose and other genres (e.g., short story collections) were excluded. For details, see [9]. Version 1.1, used in this study, is available at: `https://huggingface.co/datasets/MiMe-MeMo/Corpus-v1.1`.

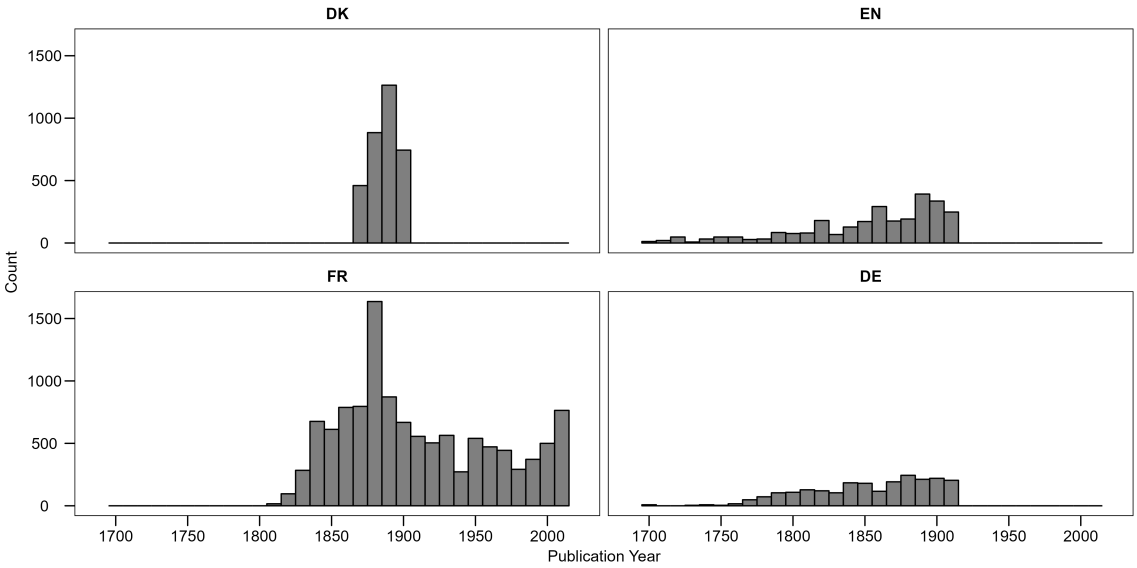

Figure 1: Distribution of publication years for the four corpora.

somewhat skewed toward the late nineteenth century—especially the 1880s, which account for nearly 10% of the texts (see Figure 1 for the distribution of corpus texts over time). The corpus is mainly based on digitized novels available online[3] in the French National Library, reflecting the historical processes of selection, publication, and preservation. The English and German corpora, by contrast, were compiled using literary historiography as a guiding principle. Texts were included based on their appearance in literary historical sources, with such inclusion treated as a proxy for perceived relevance within literary historiographical discourse [13].

3.2 Operationalizing Canonicity as a Spectrum

The four corpora used in this study originate from different research projects and thus differ, in addition to size and scope, in their approaches to operationalizing canonicity. An important aspect of our methodological framework is thus the alignment of these different indicators so that, despite their variety, they can be treated as equivalent measures of canonization. Following Heydebrand and Winko [28, p. 222], we understand evaluative acts as practices that sustain a literary canon through publication, critical editions, teaching, literary histories, and engagement by later authors. Within this framework, the various markers in each corpus thus serve as proxies for the same underlying concept.

For the **English** and **German corpora**, canon markers are based on four criteria: availability of student editions, inclusion in literary histories, publication of complete or collected works, and presence on university reading lists, capturing both academic recognition and institutional endorsement. In the **French corpus**, a more granular approach distinguishes author-level and novel-level markers, including student editions, school examinations, university syllabi, literary awards, and collected works. The **Danish corpus** relies on six proxy indicators combining institutional and expert recognition: whether the author has a dedicated page in the Danish encyclopedia *Den Store Danske*,[4] the word count of that page, whether the title is mentioned on it, how often the author is referenced on other pages dedicated to other authors, and two binary indicators of canonization: mention in the lemma on ‘det moderne gennembruds litteratur’ (the Modern Breakthrough)

[3] See `https://gallica.bnf.fr`

[4] `https://denstoredanske.lex.dk/det_moderne_gennembruds_litteratur`

and inclusion in the Danish Educational Canon (*Undervisningskanon*) and Cultural Canon (*Kulturkanon*).[5]

To allow cross-linguistic comparison, all indicators were normalized to a 0–1 scale, producing a composite canonization score. A score of 0 indicates no canonical evidence, 1 the strongest recognition. For the English and German corpora, this relies on logistic regression trained on minimum and maximum values [12]. The French spectrum combines three layers—frequency in syllabi or collections, author status in prestigious series such as the *Pléiade*, and major literary prizes—weighted with institutional inclusion strongest, author reputation intermediate, and prizes a smaller effect. For the Danish corpus, numeric features were normalized, binary features binarized, and all equally weighted to yield a final score between 0 and 1.

4 Methods

4.1 Textual Features

To formalize stylistic comparison across diverse literary corpora, we propose a computational framework that uses low-level linguistic features to establish a baseline for analyzing similarities across languages and historical periods. By focusing on simple, interpretable features, we create standardized text profiles that facilitate meaningful comparisons within and between corpora and enable the construction of temporally informed networks, reflecting the historical dynamics of literary influence and reception. We deliberately avoid raw word-frequency features: although powerful, they are less interpretable, closely tied to topical content, and sensitive to orthographic, morphological, and lexical variation, making cross-linguistic comparison harder. In contrast, features such as readability, lexical diversity, and compressibility are more abstract, and reflect structural or rhythmic aspects of writing—dimensions that could link back to stylistic conventions, endurance, and canon formation.

Feature	Description	Abbreviation
Type-token ratio	Sliding window (100 tokens) lexical diversity	TTR
Readability	Flesch score or equivalent per language	READ
Compressibility	Approximation of formulaicity or redundancy	COMP
Average sentence length	Mean number of words per sentence	ASL
Average word length	Mean number of characters per word	AWL

Table 2: Text-level linguistic features

To implement this framework, we extracted a set of text-level linguistic features that are comparable across languages and corpora, including type-token ratio (TTR), readability, compressibility, and other basic structural measures (see Table 2 for an overview and Table 4 in the Appendix for implementation details and interpretation). Such features might relate to canonicity, as texts that balance accessibility with stylistic distinctiveness often endure both as models of form and as vehicles of lasting aesthetic engagement [8]. All features were computed individually for each text: TTR was computed using a 100-token sliding window, and readability was assessed with established metrics such as the Flesch score or language-specific equivalents. Each work was then represented as a vector in a shared feature space, enabling the computation of pairwise similarities within each corpus. Features were normalized to a 0–1 range using min-max scaling.[6] Each text

[5] The Danish encyclopedia lemmas are authored by leading scholars. Unlike the institutional canon, the encyclopedia also lists Norwegian and Swedish authors who published extensively in Denmark during this period.

[6] From `Scikit-learn`: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`

was then represented as a vector in a shared feature space, and pairwise cosine similarities between these normalized vectors were computed to generate similarity matrices reflecting stylistic relationships between texts.

4.2 Networks

We used the resulting similarity matrices to construct fully connected text similarity networks, where nodes represent individual texts and edge weights indicate the degree of similarity between them. To incorporate temporal constraints and better model the historical conditions of literary influence and reception, we applied a time-sensitive filtering procedure before the actual network analysis. In these filtered networks, each text connects only to its three most similar predecessors—i.e., the three most similar texts published before it—based on cosine similarity in the normalized feature space. This filtering ensures that the networks adhere to the temporal directionality of literary development, excludes anachronistic links to later texts, and reduces network density while preserving meaningful structural relationships.

These directed, weighted networks form the basis for calculating structural centrality. Four standard centrality measures were computed for each text: *Indegree* counts the number of incoming edges, reflecting how often a text resembles later works stylistically. *PageRank* [11] estimates influence recursively, identifying texts widely connected to other central texts and reflecting hierarchical patterns of stylistic centrality rather than simple connectivity. *Betweenness* measures how frequently a text appears on the shortest paths between other texts, highlighting its role as a stylistic bridge. *Closeness*, defined as the inverse average shortest path distance to all other reachable nodes, indicates a text's overall accessibility within the network. Together, these measures estimate each text's structural position within its corpus-specific network of stylistic predecessors, covering different aspects of centrality.

Metric	Compact definition
Indegree	Count of incoming edges; raw family resemblance
PageRank	Recursive weight of a node based on the importance of its neighbors
Betweenness	Share of all shortest paths that traverse the node; measures brokerage
Closeness	Inverse of mean shortest-path length to all others; captures reachability speed

Table 3: The four centrality measures used in the network analysis

To identify broader stylistic structures within each corpus, we applied the Infomap algorithm for community detection [20] to time-filtered similarity networks. This algorithm partitions the network into communities by modeling the flow of information and optimizing the description length of a random walker's movements. While applicable to both directed and undirected networks, we retained weighted and directed edges to preserve the asymmetry and magnitude of stylistic similarity. The resulting clusters represent cohesive groupings of texts with high internal stylistic similarity compared to the rest of the corpus.

5 Results

To assess the relationship between structural centrality and canonization, we computed Spearman correlations between each centrality measure and the canonization scores, separately for each language (Table 3). Across all four corpora, the correlations are weak. The **English** data shows the clearest pattern, with all four measures—indegree, PageRank, betweenness, and closeness—positively and significantly correlated with canonization ($\rho = 0.14$–0.17, all $p < 0.001$). However, even here, the effect sizes are small. The **French** corpus displays similarly small coefficients

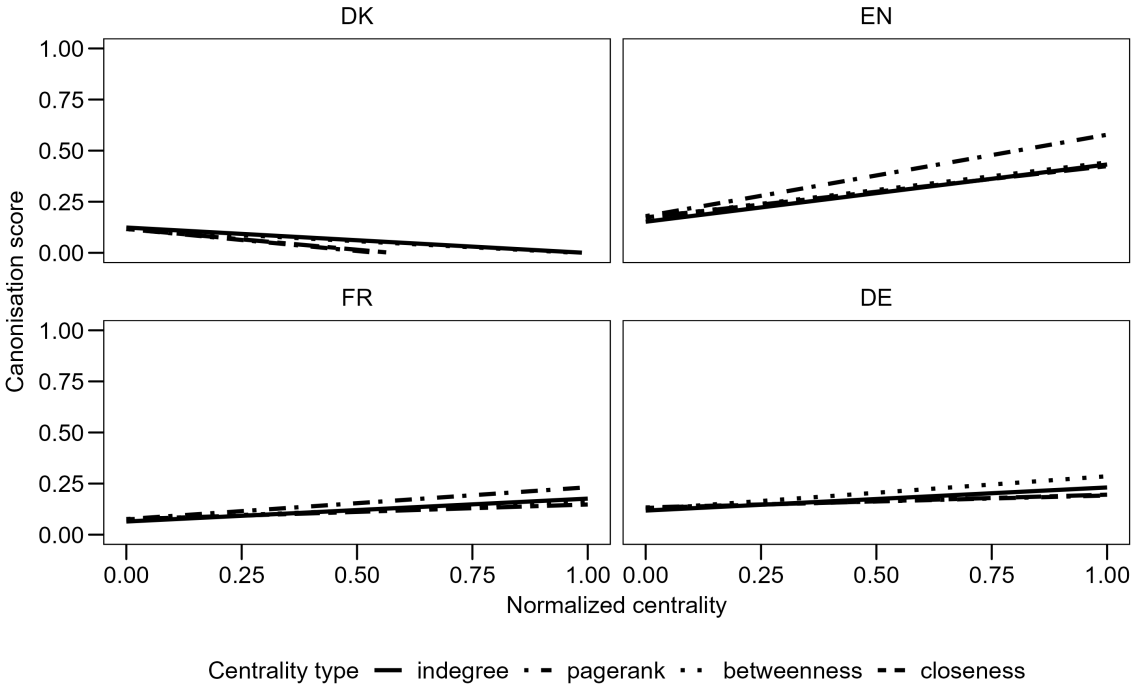

Figure 2: Relationship between normalized centrality measures and canonization scores by language. Linear fits summarize overall correlations per language.

($\rho \approx 0.07$–0.08), despite their statistical significance. For **German**, correlations are positive but very small and not statistically significant ($\rho = 0.05$–0.07, all $p > 0.05$). In the **Danish** corpus, correlations are close to zero or slightly negative ($\rho = -0.06$– -0.09), with some reaching statistical significance due to the large sample size rather than effect strength. Overall, these results suggest that structural centrality, as derived from basic stylistic similarity, has at most a very weak association with canonization.

To detect finer-grained variation that may be masked on the corpus level, we examined correlations within Infomap clusters (see Figure 2). In the **Danish network**, the 35 clusters show substantial heterogeneity. While most exhibit weak and statistically insignificant correlations, several stand out. Cluster 12 displays moderate positive correlations across all centrality measures, with closeness ($\rho = 0.41$, $p = 0.023$), PageRank ($\rho = 0.41$, $p = 0.021$) and indegree ($\rho = 0.38$, $p = 0.037$) reaching significance. This suggests that, within this stylistic grouping, more central texts are more likely to be canonized. In fact, novels with higher canonization scores, such as *Uden Midtpunkt* (1878) and *Smaafolk* (1880) by Schandorph, *Fru Marie Grubbe* (1885) by Jacobsen and *I Sabinerbjergene* (1871) by Bergsøe belong to the most central nodes in this cluster, suggesting a convergence of structural and cultural centrality. In contrast, Cluster 3 shows moderate negative correlations for all measures, with indegree ($\rho = -0.28$, $p = 0.0048$) and closeness ($\rho = -0.30$, $p = 0.035$) significant at the 0.05 level, indicating the opposite tendency: The cluster's center is dominated by titles with the lowest possible canonization score of 0, as, for example, Brosbøll's *Tranens Varsel* (1870) and *Viben Peter* (1875).

In the **English corpus**, correlations between centrality and canonization vary across the 37 Infomap clusters. Cluster 0 shows consistently positive and significant correlations across all centrality measures (indegree $\rho = 0.28$, $p = 0.033$; PageRank $\rho = 0.34$, $p = 0.008$; betweenness $\rho = 0.35$, $p = 0.006$; closeness $\rho = 0.34$, $p = 0.008$), suggesting that central texts in this stylistic community are more canonized. Its most central nodes include Dickens' works (*A Christmas*

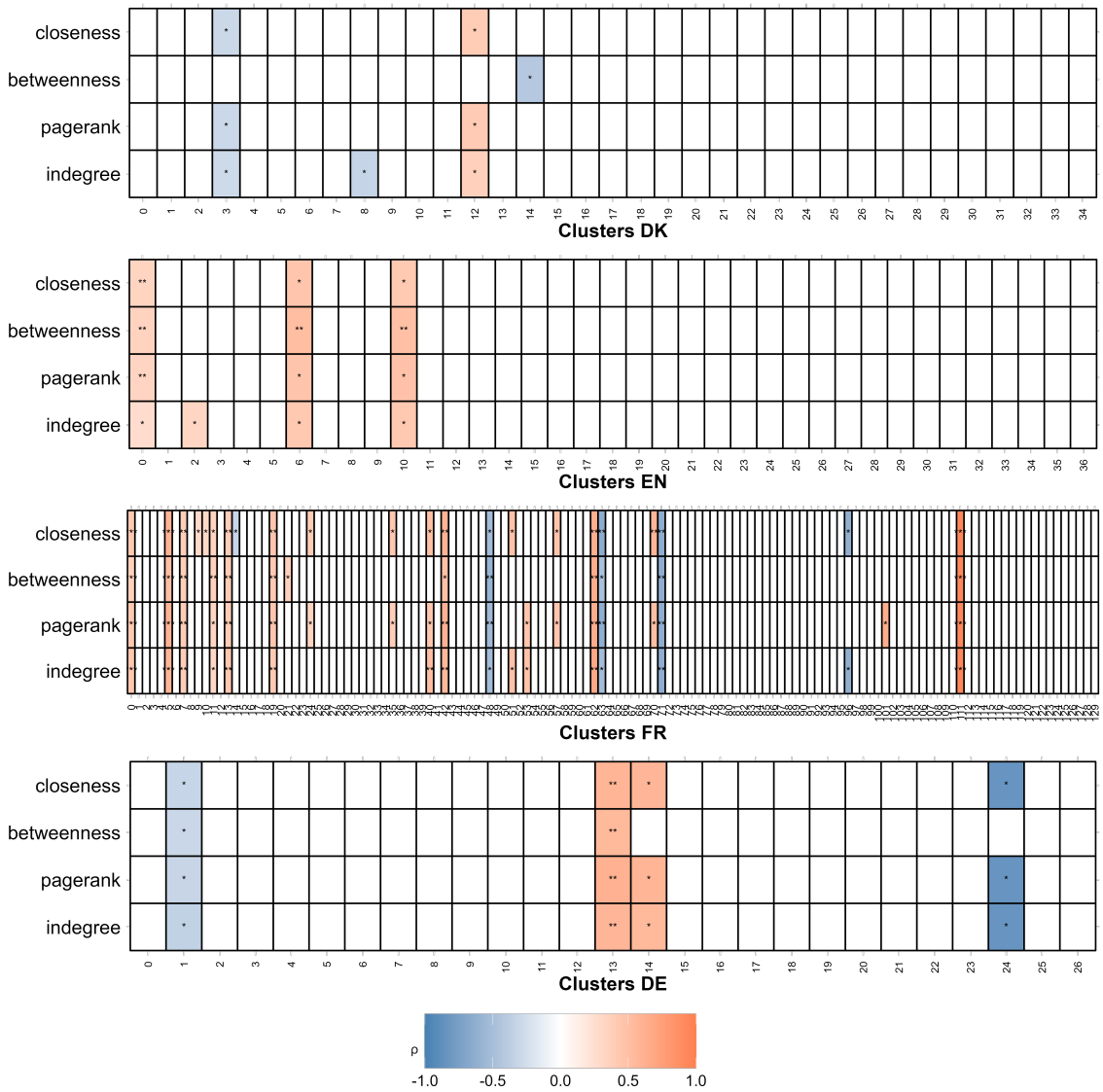

Figure 3: Correlation between canonization scores and centrality measures across clusters for four languages. Only significant correlations are colored (blue = negative, red = positive; −1 to 1), with asterisks indicating significance levels (p < 0.05, ** p < 0.01, *** p < 0.001). Non-significant correlations are shown in white.

Carol, 1843; *The Cricket on the Hearth*, 1845) and Rymer's *The String of Pearls* (1847), reflecting the influence of Victorian publishing dynamics such as serialization and market competition. Cluster 6 exhibits even stronger correlations (indegree $\rho = 0.44$, $p = 0.020$; PageRank $\rho = 0.46$, $p = 0.013$; betweenness $\rho = 0.52$, $p = 0.004$; closeness $\rho = 0.45$, $p = 0.015$), centering on early satirical novels like Fielding's *Joseph Andrews* (1742), *Tom Jones* (1749), *Jonathan Wild* (1743), and Swift's *Gulliver's Travels* (1726). Cluster 10 similarly aligns network centrality with canonization (indegree $\rho = 0.44$, $p = 0.022$; PageRank $\rho = 0.48$, $p = 0.011$; betweenness $\rho = 0.50$, $p = 0.008$; closeness $\rho = 0.42$, $p = 0.027$), with central texts including Thackeray's *The Newcomes* (1855), Austen's *Pride and Prejudice* (1813), and Hardy's *Two on a Tower* (1882).

In the **French network**, several clusters show considerable variation in the strength and direction of correlations. Several show significant positive correlations: Cluster 0 displays moderate associations across all centrality measures ($\rho = 0.37$–0.44, all $p < 0.001$), as does Cluster 5 (e.g.,

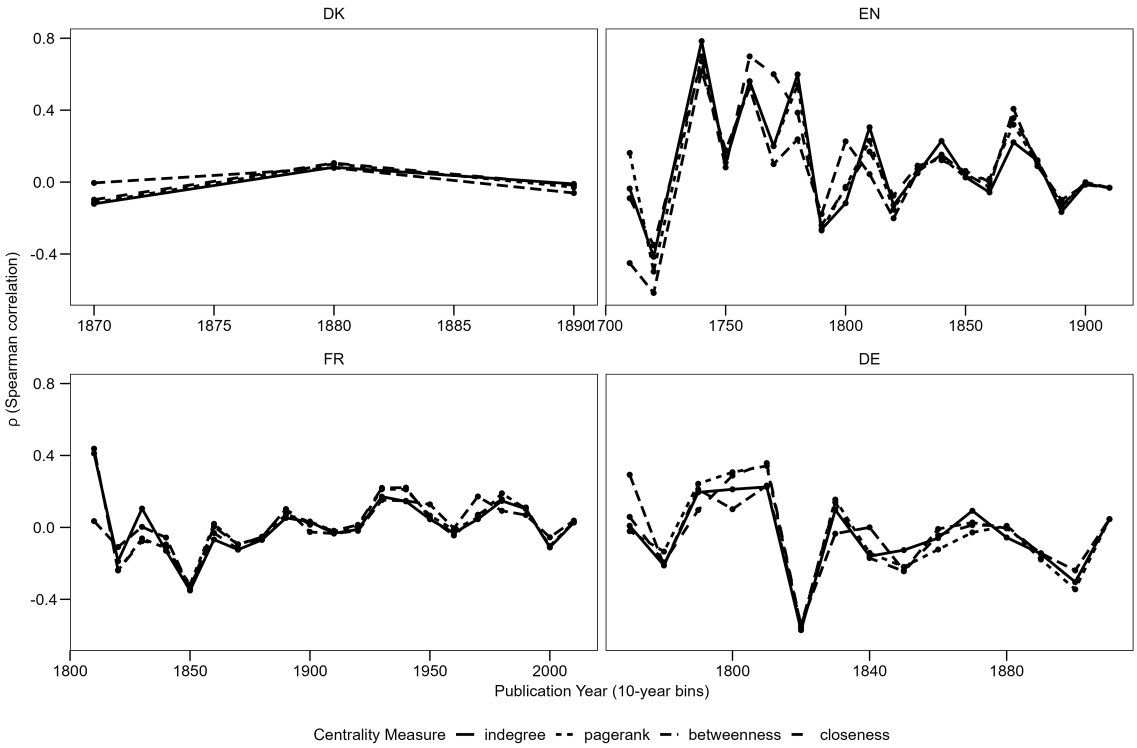

Figure 4: Temporal trends of Spearman's ρ between canonization scores and centrality measures in 10-year bins. Each subplot displays a single language, with line types indicating centrality metrics.

PageRank $\rho = 0.59$, $p = 9.6 \times 10^{-7}$; closeness $\rho = 0.59$, $p = 9.3 \times 10^{-7}$), and Cluster 7 ($\rho \approx 0.38$–0.39, $p \approx 0.004$). Clusters 19, 42, and 62 also show strong positive correlations across all centrality measures ($\rho = 0.42$–0.68, all $p < 0.01$). What these clusters have in common is that their central nodes represent texts by highly canonized authors: Balzac (Cluster 0), Dumas (Cluster 5), Saint-Exupery and Duras (Cluster 7), Verne and Flaubert (Cluster 19), and Sand (Cluster 42 and 62). Conversely, Cluster 48 and Cluster 71 show significant negative correlations (PageRank $\rho = -0.54$ to -0.64, $p < 0.02$), meaning that canonized texts are relatively peripheral in these stylistic communities. Central nodes include Féval's *Jean Diable* (1862), Sue's *Les Mystères de Paris* (1843), Boisgobey's *Le Crime de l'Opéra* (1878), and Madame de Stolz's *Valentine* (1875), typical of the feuilleton tradition of urban mysteries and melodramatic intrigue. Here, stylistic cohesion stems from formula and serialization rather than innovation, producing strong popular centers of influence that remain largely outside canonical prestige.

The **German corpus** again shows similar variation across its 27 clusters. Cluster 13 exhibits significant positive correlations across all centrality measures (e.g., PageRank $\rho = 0.61$, $p = 0.003$; closeness $\rho = 0.59$, $p = 0.005$), as does Cluster 14, with PageRank ($\rho = 0.58$, $p = 0.014$) and closeness ($\rho = 0.57$, $p = 0.016$) reaching significance, while betweenness is slightly lower ($\rho = 0.45$, $p = 0.073$). Both clusters have central nodes related to German Romanticism (with works by Brentano, Arnim, Hauff, and Eichendorff), but diverge in later developments: Cluster 13 incorporates more rural-Realist texts, while Cluster 14 bridges the transition from Romanticism to Poetic Realism. By contrast, Cluster 1 exhibits significant negative correlations across all centrality measures (e.g., indegree $\rho = -0.35$, $p = 0.020$; PageRank $\rho = -0.32$, $p = 0.032$), indicating that canonized texts in this subgroup tend to be less central. Cluster 24 also shows strong negative correlations (e.g., indegree, PageRank, and closeness $\rho = -0.85$, $p = 0.034$), but this is likely driven by the small number of nodes in the cluster (n=6).

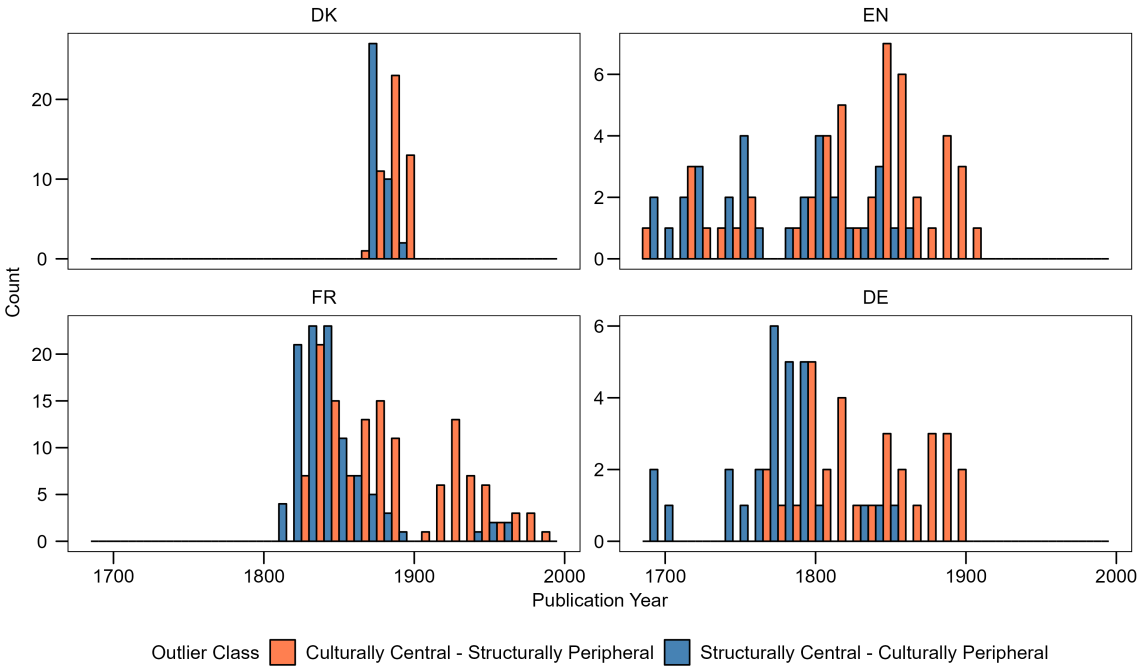

Figure 5: Distribution over time across four languages of texts that are outliers in at least one centrality metric, showing counts of texts that are structurally central but culturally peripheral (blue) and culturally central but structurally peripheral (coral). The histogram uses 10-year bins and is faceted by language.

Overall, these findings indicate that the relationship between structural and cultural centrality varies across stylistic communities, with some clusters showing clear alignment between network position and canonization, while others diverge. To examine temporal patterns, we computed Spearman's ρ between canonization scores and centrality measures within ten-year publication bins (Figure 3). The resulting trends fluctuate across all four corpora rather than following a consistent trajectory. In the **Danish corpus**, correlations are weakly positive in the 1880s, with earlier and later decades near zero or slightly negative. The **English corpus** exhibits strong positive correlations in the late eighteenth century (around 1740, 1760, 1780), which weaken by 1790 and during the early nineteenth century, remaining near zero for most of the Victorian period, with a brief resurgence around 1870 before declining again. Correlations in the **French corpus** fluctuate across the nineteenth and twentieth centuries, though less dramatically than in English; between 1930 and 1950, all four centrality measures—betweenness, PageRank, indegree, and closeness—show consistently positive associations, which weaken or turn slightly negative in later decades. The **German corpus** displays pronounced shifts between positive and negative correlations throughout the nineteenth and early twentieth centuries, with notable changes around 1850 and 1890, reflecting periods of both strong and weak alignment between network centrality and canonical status.

To further examine this variability, we analyzed the temporal distribution of outlier texts according to their cultural and structural centrality, grouped by decade (see Figure 4). We distinguish between two types of outliers: texts that are structurally central but culturally peripheral (shown in blue), and texts that are culturally central but structurally peripheral (shown in coral). Outliers were identified based on standardized residuals between a text's structural centrality and its canonization score, calculated for each language and centrality measure. We classified texts with residual z-scores greater than ±2 as outliers.[7] Across all four corpora, structurally central but culturally pe-

[7] See Figure 7 for the corresponding scatterplots of the distribution of canonization scores and centrality measures with

ripheral texts tend to cluster in particular decades and are followed, typically a few decades later, by increases in culturally central but structurally peripheral texts.

6 Discussion

Our results show that structural centrality within networks of stylistic similarity shows no sign of a strong correlation with canonization across the four corpora. This absence of a clear global relationship is itself a meaningful finding, indicating that stylistic embeddedness—as measured through centrality in a network of formal similarity—does not translate into cultural recognition in a systematic way. If canonization operated as a function of mere stylistic centrality, we would expect a stable and positive correlation across national contexts and over time. The weak global correlations instead suggest that the mechanisms driving canonization are more complex and heterogeneous, shaped by literary historical developments, institutional structures, and sociocultural dynamics that vary across time and space.

However, when we disassemble the networks and focus on stylistic clusters—i.e., communities of texts that share similar linguistic and formal features—stronger and more consistent relationships between structural and cultural centrality begin to emerge. This suggests that within certain stylistic environments, centrality may indeed reflect, or contribute to, canonization. The presence of local patterns within specific clusters indicates that smaller-scale groupings with coherent stylistic profiles mediate the relationship between formal and cultural values.

In addition to these local structures, our temporal analysis shows that the relationship between structural and cultural centrality has varied through history. Rather than a consistent trend, we find periods of closer alignment alternating with phases in which structural and cultural centrality display a weak, or inverse, association. We can conceptualize these shifts over time as a sort of canonical gravity: canonized texts can shape the stylistic landscape, affecting the structure of the similarity network because they act as stylistic role models. At certain moments, if a particular style or genre becomes dominant, this influence may become pronounced, resulting in a closer alignment between structural centrality and canonization. At other times, such as during periods of stylistic change or growing diversity, the connection may weaken or invert, with centrality becoming either an unreliable indicator of a text's cultural status, or holding a negative correlation with canonicity.

To identify such shifts, we examined how the correlation between structural and cultural centrality changes over time. This temporal approach highlights historical discontinuities in the relationship between canonical status and network position, periods in which the stylistic influence of canonical works diminishes and their centrality in the network weakens or even reverses. We applied formal change-point detection to the year-wise trajectories of Spearman's ρ, computed for each language–centrality pair using a five-year centered moving mean to smooth short-term fluctuations (see Figure 5). Significant changes in the mean or variance of these smoothed series were treated as breakpoints, signaling potential restructurings in the canonicity–centrality relationship.

Many of the identified change points align with major literary historical movements and broader cultural shifts across the four languages, reinforcing the idea that correlations between cultural and structural centrality are responsive to changing literary norms. In **Danish** literature, for example, the change points of 1879 and 1882 align with the early years of the Modern Breakthrough,[8] a movement led by critic Georg Brandes that championed Realism, Naturalism, and political engagement. A clear departure from earlier Romantic ideals marked this new period, which saw the rise of prominent figures like J.P. Jacobsen and Henrik Pontoppidan; it is thus plausible to see a higher correlation between canonization and centrality after a time of consolidation of new poetic and stylistic standards. However, due to the smaller time frame covered by the

highlighted outliers.

[8] The Modern Breakthrough corresponds to the period between the 1870s and the 1890s.

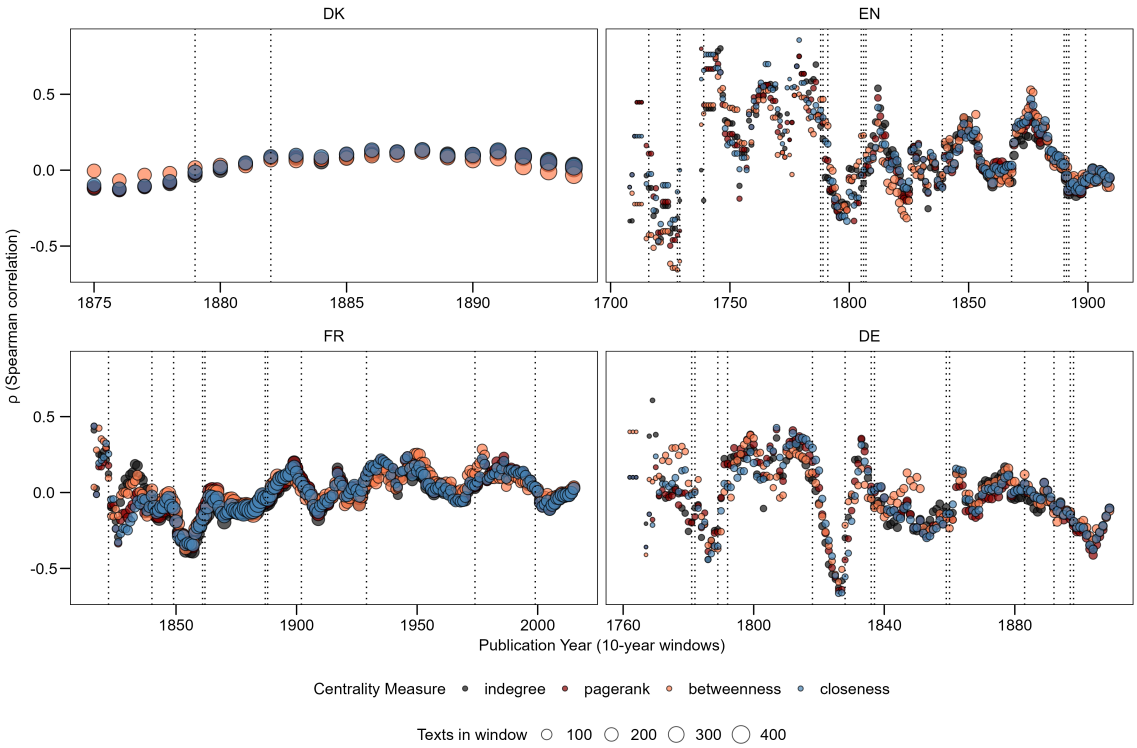

Figure 6: Change points in the correlation between centrality and canonization over time. Vertical dotted lines indicate detected change points in the correlation trend, suggesting shifts in the relationship between structural and cultural centrality.

Danish corpus and the smaller changes in correlations, these trends require more detailed study.

In the **English** corpus, structural inflection points are concentrated in four distinct periods. The first (1716–1739) aligns with the Augustan rise of prose, as writers like Defoe and Swift help expand early novelistic forms and shift network centrality. A second shift (1788–1807) spans the transition from the Enlightenment to Romanticism, with the French Revolution in 1789 introducing external pressure. The 1820s and 1830s mark a Victorian re-orientation, shaped in part by Dickens and new urban-industrial themes. A final phase (1868–1899) can be linked to the late Victorian turn toward aestheticism and early modernist tendencies, seen in the work of authors such as George Eliot and Thomas Hardy.

In the **French** time series, the detected breaks at 1822, 1862, 1887, 1902, 1929, 1974, and 1999 coincide with well-known literary reconfigurations rather than directly marking them. The early-to-mid-19th-century interval (≈ 1822–1862) overlaps the shift from high-Romantic modes toward emerging Realist/Naturalist prose (Hugo/Balzac through Flaubert). The late-19th-century breaks (≈ 1887, 1902) occur near the Symbolist/Decadent moment and Belle-Époque realignments (Moréas's 1886 manifesto; Zola's death in 1902), which shaped novelistic style and reception. The 1929 break aligns with the second Surrealist manifesto and the consolidation of interwar modernism, with existentialist prominence emerging later. Late-20th-century inflections (≈ 1974, 1999) correspond to post-1968/postmodern tendencies and the aftermath of the *Nouveau Roman*. The 1887 break, in particular, coincides with Jules Verne's prolific output (*Twenty Thousand Leagues Under the Seas*, 1870; *The Carpathian Castle*, 1892), exemplifying the rise of scientific adventure fiction and a peak of literary novelty [6].

In **German**, early change points between 1781 and 1792 coincide with the *Sturm und Drang* movement and the rise of Weimar Classicism, notably marked by Goethe's return from Italy in 1786

and the canonical consolidation of his and Schiller's works. Further change points in the early nineteenth century (1818, 1828, 1837) reflect the transition into the *Vormärz* period, characterized by politically engaged literature and the emergence of Young Germany. The later nineteenth-century cluster (1859, 1883, 1892, 1898) corresponds to the maturation of Realism and the ascendancy of Naturalism. These shifts suggest a gradual weakening of classical forms and a growing prominence of socially engaged prose.

In sum, our analysis shows that the relationship between stylistic centrality and canonization is neither fixed nor uniform. While there is no strong global correlation, the presence of consistent local patterns and historically specific alignments points to a more nuanced connection between form and cultural value. Canonization does not simply follow from structural embeddedness, but the two can align under certain stylistic and historical conditions. These findings suggest that stylistic similarity matters, but not in isolation: its role is shaped by changing literary norms, institutional frameworks, and broader cultural developments.

7 Conclusion

Our approach bridges literary history and network analysis to trace how cultural recognition relates to stylistic embeddedness in four national corpora—Danish, English, French, and German. Using time-sensitive similarity networks and multiple centrality measures, we explored how canonized texts occupy structural positions over time. Across all corpora, we found positive but limited correlations between cultural and structural centrality, with cluster-level patterns generally reinforcing these associations. Temporal analysis suggested an oscillating relationship between structural and cultural centrality: periods marked by structurally central but culturally peripheral texts are often followed, after a few decades, by an increase in culturally central but structurally peripheral texts. This pattern points to shifting dynamics in how stylistic embeddedness and institutional recognition interact over time. To better understand these shifts, we traced historical fluctuations in canonical gravity—moments when the cultural pull of canonized texts weakens, intensifies, or becomes uncoupled from their network positions.

This approach, however, comes with limitations. The structural centrality measures are derived from similarity networks filtered to connect each text only to its three most similar predecessors. This design ensures historical plausibility and reduces noise, but it also introduces a bias: periods with more available texts offer a larger pool of potential predecessors, increasing the chances for higher centrality. In other words, the temporal distribution of centrality scores may reflect corpus density rather than intrinsic stylistic distinctiveness. Moreover, our use of fixed ten-year time bins without down-sampling means that periods with a higher number of texts can exert disproportionate influence on both network structure and centrality metrics. These effects should be considered when interpreting historical trends, especially in analyses of long-term stylistic influence. Finally, we are only considering a limited set of textual features, both in scope and complexity: limited in scope because the feature set is relatively small, and limited in complexity because the features themselves are based on simple linguistic measures rather than deep semantic or syntactic representations.

Despite these constraints, the study provides insights into the formal structures underlying literary canons and offers a framework for comparing stylistic and cultural prominence across national traditions. The identification of canonization outliers opens avenues for reevaluating literary history, particularly for texts that have been structurally influential but culturally overlooked. In general, our analyses show that the relationship between cultural and structural centrality is not linear; instead, we have identified specific, localized, and temporally influenced patterns rather than more generalizable trends. This suggests that canon formation is shaped not just by stylistic prominence but also by broader cultural dynamics and historical contingencies. Future work could refine these models by incorporating richer features, experimenting with alternative network

constructions, and extending comparisons to other, ideally non-Western, national or multilingual corpora.

Acknowledgements

The authors thank the anonymous reviewers for their thoughtful comments and critiques.

References

[1] Algee-Hewitt, Mark, Allison, Sarah, Gemma, Marissa, Heuser, Ryan, Moretti, Franco, and Walser, Hannah. “Canon/Archive. Large-scale Dynamics in the Literary Field”. In: *Pamphlets of the Stanford Literary Lab* , no. 11 (2016). URL: https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

[2] Andresen, Melanie, Krautter, Benjamin, Pagel, Janis, and Reiter, Nils. “Knowledge Distribution in German Drama: An Annotated Corpus”. In: *Journal of Open Humanities Data 10* (Jan. 12, 2024), p. 6. DOI: 10.5334/johd.167.

[3] Assmann, Aleida. “Canon and Archive”. In: *Cultural Memory Studies: An International and Interdisciplinary Handbook*, ed. by Astrid Erll, Ansgar Nünning, and Sara B. Young. Media and cultural memory ; Medien und kulturelle Erinnerung 8. Berlin; New York, NY: De Gruyter, 2008, pp. 97–107.

[4] Barré, Jean. “Latent Structures of Intertextuality in French Fiction”. In: *Proceedings of the Computational Humanities Research Conference 2024*. CHR 2024. Aarhus, 2024, pp. 21–26. URL: https://ceur-ws.org/Vol-3834/paper97.pdf.

[5] Barré, Jean, Camps, Jean-Baptiste, and Poibeau, Thierry. “Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature”. In: *Journal of Cultural Analytics 8*, no. 1 (2023), pp. 1–29. DOI: 10.22148/001c.88113.

[6] Barré, Jean and Poibeau, Thierry. “Beyond Canonicity: Modeling Canon/Archive Literary Change in French Fiction”. In: *Proceedings of the Computational Humanities Research Conference 2023*. 2023, pp. 814–830. URL: https://ceur-ws.org/Vol-3558/paper9925.pdf.

[7] Beine, Julia. “The Schemer Unmasked. Sketching a Digital Profile of the Scheming Slave in Roman Comedy”. In: *Journal of Computational Literary Studies 3*, no. 1 (2024), pp. 1–29. DOI: 10.48694/JCLS.3670.

[8] Bizzoni, Yuri, Feldkamp, Pascale, Lassen, Ida Marie, Jacobsen, Mia, Thomsen, Mads Rosendahl, and Nielbo, Kristoffer. “Good Books are Complex Matters: Gauging Complexity Profiles Across Diverse Categories of Perceived Literary Quality”. Apr. 2024. DOI: 10.48550/arXiv.2404.04022.

[9] Bjerring-Hansen, Jens, Kristensen-McLachlan, Ross Deans, Diderichsen, Philip, and Hansen, Dorte Haltrup. “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*. Vol. 3232. Uppsala, Sweden, 2022, pp. 177–186. URL: https://ceur-ws.org/Vol-3232/paper14.pdf.

[10] Björnsson, Carl Hugo. *Läsbarhet*. sv. Liber, 1968.

[11] Brin, Sergey and Page, Lawrence. “The Anatomy of a Large-Scale Hypertextual Web Search Engine”. In: *Computer Networks 30* (1998), pp. 107–117. URL: http://www-db.stanford.edu/~backrub/google.html.

[12] Brottrager, Judith. “From Canon to Score: Quantifying, Measuring, and Comparing Canonisation”. In: DH2025. Lisbon, 14-18 July 2025, 2025.

[13] Brottrager, Judith. “Unlocking the Archive: Exploring Literary History through Word Embeddings”. In: *Flows & Frictions: Intregrating Mixed Methods for Data-rich Research in Historical Media*. Gothenburg: University of Gothenburg Press.

[14] Brottrager, Judith, Stahl, Annina, and Arslan, Arda. “Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features”. In: *Proceedings of the Conference on Computational Humanities Research 2021*. CHR 2021. Amsterdam: CEUR Workshop Proceedings, 2021, pp. 195–205. URL: http://ceur-ws.org/Vol-2989/short_paper21.pdf.

[15] Calvo Tello, José. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press, Dec. 31, 2021. DOI: 10.1515/9783839459256.

[16] Cohen, Margaret. “Narratology in the Archive of Literature”. In: *Representations 108*, no. 1 (Nov. 1, 2009), pp. 51–75. DOI: 10.1525/rep.2009.108.1.51.

[17] Cranenburgh, Andreas van. *Canonizer*. The Hague: KB Lab, 2021.

[18] Dalen-Oskam, Karina van. *The Riddle of Literary Quality: A Computational Approach*. Amsterdam: Amsterdam University Press, 2023.

[19] Eder, Maciej. “Visualization in Stylometry: Cluster Analysis Using Networks”. In: *Digital Scholarship in the Humanities 32*, no. 1 (2017), pp. 50–64. DOI: 10.1093/llc/fqv061.

[20] Edler, Daniel, Holmgren, Anton, and Rosvall, Martin. “The MapEquation software package”. https://mapequation.org. 2025.

[21] Ehret, Katharina and Szmrecsanyi, Benedikt. “An information-theoretic approach to assess linguistic complexity”. en. In: *Complexity, Isolation, and Variation*. De Gruyter, July 2016, pp. 71–94. DOI: 10.1515/9783110348965-004.

[22] Feldkamp, Pascale, Lassche, Alie, Kostkan, Jan, Kardos, Márton, Enevoldsen, Kenneth, Baunvig, Katrine, and Nielbo, Kristoffer. “Canonical Status and Literary Influence: A Comparative Study of Danish Novels from the Modern Breakthrough (1870–1900)”. In: *Proceedings of the 4th international conference on natural language processing for digital humanities*, ed. by Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni. Miami, USA: Association for Computational Linguistics, Nov. 2024, pp. 140–155. URL: https://aclanthology.org/2024.nlp4dh-1.14.

[23] Fischer, Frank, Börner, Ingo, Göbel, Mathias, Hechtl, Angelika, Kittel, Christopher, Milling, Carsten, and Trilcke, Peer. *Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama*. 2019. DOI: 10.5281/ZENODO.4284002.

[24] Flesch, Rudolph. “A New Readability Yardstick”. In: *Journal of Applied Psychology 32* (1948), pp. 221–233.

[25] Gadamer, Hans-Georg, Weinsheimer, Joel, and Marshall, Donald G. *Truth and Method*. The Bloomsbury revelations series. London: Bloomsbury, 2013.

[26] Glawion, Anastasia. *Remembering World War II: A Mixed-Methods Exploration of Memory Practices on an Online Forum*. Digitale Literaturwissenschaft. Berlin, Heidelberg: Springer, 2023. DOI: 10.1007/978-3-662-66708-8.

[27] Herget, Katharina. *Reading at Scale. Eine Mixed-Methods-Analyse der „Deutschen Novellenschätze“*. Digitale Literaturwissenschaft. Berlin, Heidelberg: Springer, 2025. DOI: 10.1007/978-3-662-70346-5.

[28] Heydebrand, Renate von and Winko, Simone. *Einführung in die Wertung von Literatur: Systematik - Geschichte - Legitimation*. UTB für Wissenschaft Uni-Taschenbücher 1953. Paderborn: Schöningh, 1996.

[29] Houston, Natalie. "Topic Modeling the Nineteenth-Century Poetry Canon: English Poetry Reprinted in Anthologies". In: *Digital Humanities 2022*. DH2022. Tokyo, 2022, pp. 492–493.

[30] Houston, Natalie M. "Measuring Canonicity: A Network Analysis Approach to Poetry Anthologies". In: *Digital Humanities 2017 Conference Abstracts*. DH2017. Montréal, 2017, pp. 476–478.

[31] Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. Urbana, IL: University of Illinois Press, 2013.

[32] Leblond, Aude. "Corpus Chapitres". Version v1.0.0. 2022. DOI: 10.5281/zenodo.7446728.

[33] Liddle, Dallas. "Could Fiction Have an Information History? Statistical Probability and the Rise of the Novel". In: *Journal of Cultural Analytics* (2019). DOI: 10.22148/16.033.

[34] Moretti, Franco. *Distant Reading*. London, UK; New York, NY: Verso, 2013.

[35] "European Literary Text Collection (ELTeC)". COST Action Distant Reading for European Literary History, Apr. 2021. DOI: 10.5281/zenodo.4662444.

[36] Päpcke, Simon, Weitin, Thomas, Herget, Katharina, Glawion, Anastasia, and Brandes, Ulrik. "Stylometric Similarity in Literary Corpora: Non-Authorship Clustering and *Deutscher Novellenschatz*". In: *Digital Scholarship in the Humanities 38*, no. 1 (2023), pp. 277–295. DOI: 10.1093/llc/fqac039.

[37] Porter, J.D. "Popularity/Prestige". In: *Pamphlets of the Stanford Literary Lab* , no. 17 (2018). URL: https://litlab.stanford.edu/LiteraryLabPamphlet17.pdf.

[38] Rybicki, Jan. "The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation". In: *Studies in Corpus Linguistics*, ed. by Michael P. Oakes and Meng Ji. Vol. 51. Amsterdam: John Benjamins Publishing Company, 2012, pp. 231–248. DOI: 10.1075/scl.51.09ryb.

[39] Schöch, Christof, Dudar, Julia, Fileva, Evgeniia, and Šeļa, Artjoms. "Multilingual Stylometry: The Influence of Language on the Performance of Authorship Attribution using Corpora from the European Literary Text Collection (ELTeC)". In: *Proceedings of the Computational Humanities Research Conference 2024*. CHR2024. 2024, pp. 386–408.

[40] Smith, Barbara Herrnstein. *Contingencies of Value: Alternative Perspectives for Critical Theory*. Cambridge, MA: Harvard University Press, 1988.

[41] Stajner, Sanja, Evans, Richard, Orasan, Constantin, and Mitkov, Ruslan. "What Can Readability Measures Really Tell Us About Text Complexity?" In: *Proceedings of Workshop on natural language processing for improving textual accessibility*. Istanbul, Turkey: Association for Computational Linguistics, 2012, pp. 14–22. URL: http://nlx-server.di.fc.ul.pt/~sanja/StajnerEtAl-12.pdf.

[42] Torruella, Joan and Capsada, Ramon. "Lexical Statistics and Tipological Structures: A Measure of Lexical Richness". In: *Procedia - Social and Behavioral Sciences 95* (2013), pp. 447–454. DOI: 10.1016/j.sbspro.2013.10.668.

[43] Trilcke, Peer, Ustinova, Evgeniya, Börner, Ingo, Fischer, Frank, and Milling, Carsten. “Detecting Small Worlds in a Corpus of Thousands of Theater Plays”. In: *Computational Drama Analysis*, ed. by Melanie Andresen and Nils Reiter. De Gruyter, June 17, 2024, pp. 7–34. DOI: 10.1515/9783111071824-002.

[44] Underwood, Ted and Sellers, Jordan. “The Longue Durée of Literary Prestige”. In: *Modern Language Quarterly 77*, no. 3 (2016), pp. 321–344. DOI: 10.1215/00267929-3570634.

[45] Verboord, Marc. “Classification of Authors by Literary Prestige”. In: *Poetics 31*, no. 3 (2003), pp. 259–281. DOI: 10.1016/S0304-422X(03)00037-8.

[46] Weitin, Thomas. *Digitale Literaturgeschichte*. Berlin, Heidelberg: Springer, 2021. DOI: 10.1007/978-3-662-63663-3_4.

[47] Wilkens, Matthew. “Digital Humanities and Its Application in the Study of Literature and Culture”. In: *Comparative Literature 67*, no. 1 (2015), pp. 11–20. DOI: 10.1215/00104124-2861911.

[48] Wu, Yaru, Moreira, Pascale Feldkamp, Nielbo, Kristoffer L., and Bizzoni, Yuri. “Perplexing Canon: A Study on GPT-Based Perplexity for Canonical and Non-Canonical Literary Works”. In: *Proceedings of LaTeCH-CLfL 2024*. Association for Computational Linguistics, 2024, pp. 172–184.

A Appendix

Code and data availability: https://github.com/crazyjeannot/MultiLingualCanon

A.1 Outliers

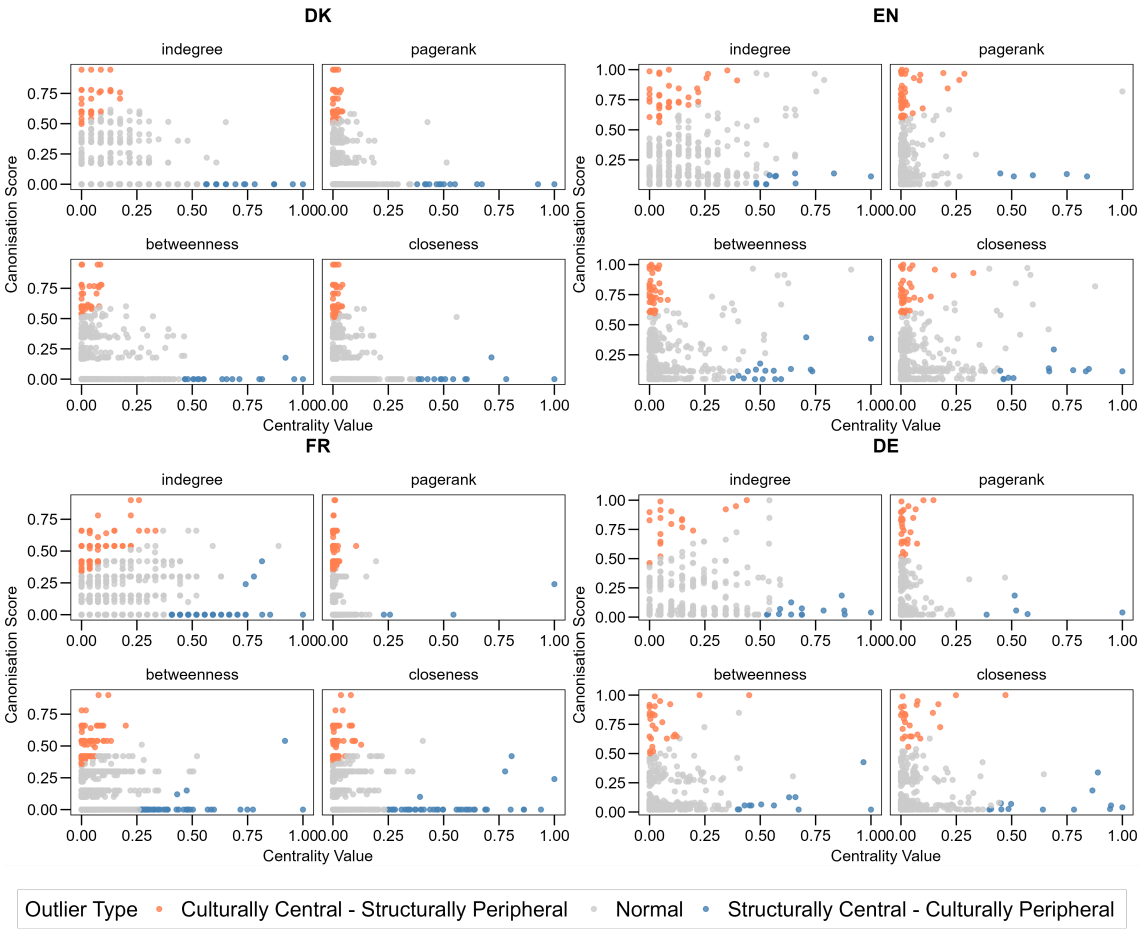

Figure 7: Scatterplots of canonization score versus centrality value by language and centrality type. Points are colored by outlier classification: *Structurally Central, Culturally Peripheral* (coral), *Culturally Central, Structurally Peripheral* (blue), and *Normal* (gray). Facets represent different centrality measures; panels represent languages.

A.2 Details on Textual Features

Feature	Abbreviation	Description
Type-token ratio	TTR	Measures lexical diversity by comparing the number of unique words (types) to the total word count (tokens), reflecting a text's vocabulary complexity [42]. We used Mean Segmental TTR (MSTTR), averaging type-token ratios across 100-word segments.
Readability	READ	Readability indexes estimate reading difficulty based variously on sentence length, syllable count, and word length/difficulty. We used the Flesch reading score [24]—a classic formula that remains widely used [41]—and equivalents in French and German. For Danish, we used RIX readability [10].
Compressibility	COMP	Compressibility estimates the degree of redundancy or formulaicity [21] in a text by measuring how much a compression algorithm can reduce its digital representation. High compressibility texts contain more repeated patterns and predictable structures, and less compressible texts display greater lexical and syntactic variation, implying higher informational complexity. Using Bzip2—an off-the-shelf file compressor—we calculated the compression ratio, i.e. file size divided by compressed size. Similar approaches appear in Dalen-Oskam [18]; Liddle [33]. To normalize, we averaged compression ratios applied to three random, non-overlapping 500-word chunks.
Average sentence length	ASL	Mean number of words per sentence in the whole text.
Average word length	AWL	Mean number of characters per word in the whole text.

Table 4: Text-level linguistic features with details on their calculation