International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 13, No. 3, 2026, pp. 313-325.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Digital Humanities Based Computational Analysis of Linguistic Evolution and Cultural Narratives in Historical Texts

Huriyat Nuriddinova [a*], Gavhar Salomova [b], Alisher Rajabov [c], Manzura Yusupova [d], Mekhriniso Saydalieva [e], Khayriniso Ganieva [f], Guzal Mardiyeva [g]

[a] Termez State Pedagogical Institute, Termez, Uzbekistan. E-mail: nuriddinovahurriyat@terdpi.uz, Orcid: https://orcid.org/0009-0006-5142-6652*

[b] Department of Foreign Languages and Literature, Termez University of Economics and Service, Termez, Uzbekistan. E-mail: gavhar_salomova@tues.uz, Orcid: https://orcid.org/0009-0009-2066-7738

[c] Associate Professor, Department of Social Sciences, Bukhara State Pedagogical Institute, Bukhara, Uzbekistan. E-mail: rajabovalisher096@gmail.com, Orcid: https://orcid.org/0009-0002-3537-8313

[d] Senior Lecturer, Bukhara State University, Bukhara, Uzbekistan. E-mail: yusupovamanzura00@gmail.com, Orcid: https://orcid.org/0000-0002-2881-7464

[e] Researcher, Department of Psychology, Termez State University, Termez, Uzbekistan. E-mail: saydaliyevamekhriniso@gmail.com, Orcid: https://orcid.org/0009-0003-3755-1914

[f] Jizzakh State Pedagogical University, Jizzakh, Uzbekistan. Email: khayrinisoganiyeva@gmail.com Orcid: https://orcid.org/0000-0003-3945-7262

[g] Department of Social and Humanitarian Sciences, Samarkand State Medical University, Samarkand, Uzbekistan. Email: mardievaguzal54@gmail.com, Orcid: https://orcid.org/0000-0001-5009-6395

Abstract

The swift progress in computational techniques has provided ample possibilities for investigating the evolution of language and changing discourse patterns in large corpora of texts. The paper describes a computational framework or investigating longitudinal register shifts and formal discourse organization in American English using the Corpus of Contemporary American English (COCA) from 1990–2019. Lexical, discourse, informational, temporal, and statistical analyses are combined into the framework to detect systematic change from one period to another. Three periods, 1990–1999, 2000–2009, and 2010–2019, were identified in the corpus and analyzed on the level of the document in terms of linguistic indicators calculated through word-count weighting. It was found that character types dropped by 8.48%, whereas expert vocabulary increased by 15.20%. Increases were noted in Information Comparison (+12.88%) and Information Quantity (+15.24%). Larger increases were also found in discourse organization measures, such as Methods-Results-Discussion (+28.82%), Purpose-Plan (+20.26%), and Reasoning (+21.20%). A statistical analysis revealed significant correlations in the time of several variables, including Reasoning ($r = 0.113$, $p < 0.001$) and Character Types ($r = -0.082$, $p < 0.001$). The operationalized Linguistic–Discourse Evolution Index (LDEI) rose from 0.115 to 0.134 (a 16.4% increase). These results demonstrate systematic evolution toward greater lexical specialization, informational density, and structured academic discourse.

Keywords: Digital Humanities, Computational Linguistics, Linguistic Evolution, Corpus Analysis, Discourse Analysis, COCA.

DOI: 10.71086/IAJSE/V13I3/IAJSE13117
Received: 13 June, 2026/Revised: 17 July 2026/Accepted: 21 August 2026/Published: 15 September 2026

Introduction

The rapid digitalization of contemporary longitudinal corpus has led to changes in the research of language, culture, and heritage because it allows conducting computer-based investigations of textual documents, which were available mostly through close reading. Digital Humanities (DH) creates an interdisciplinary platform where it is possible to combine computer methods, archives, linguistics, and cultural studies to conduct research on the historical evidence, which cannot be conducted using conventional approaches. The recent developments prove that digital humanities are becoming more engaged in archive studies, multilingual corpora, and computational methods, thus increasing opportunities for the investigation of cultural and linguistic diversity in historical documents (Wang & Lin, 2026). Interactive and visual methods of analysis have also helped scholars to reveal connections and structures in large corpora of text that are not so obvious in the case of traditional methods (Jänicke et al., 2017).

With the advent of text mining, machine learning, deep learning, visualization, and corpus analyses, the computational approach in the field of DH has seen significant growth. The bibliometric analysis reveals that computationally oriented text research is increasingly becoming one of the significant components of the developing intellectual structure of Digital Humanities (Aladağ & Aydin, 2026). Furthermore, deep learning using neural networks has also been shown to provide the capability of extracting semantic and contextual knowledge from complicated texts beyond frequency-based analysis (Suissa et al., 2022). This is part of a wider trend within the scope of DH whereby DH is shifting from being purely digitization-focused to being an analytic research science with the combination of computation and humanistic interpretation (Dalbello, 2011). Nevertheless, the disciplinary status of DH remains in line with the integration of methods from computer sciences and humanities research (Luhmann & Burghardt, 2022).

Historical textual analysis is especially beneficial for computational exploration due to the dynamic nature of language. Language undergoes transformations and modifications depending on what changes occur in society in terms of technology, politics, economy, and culture. Computational exploration, thus, becomes a way to explore language development during a long period of time rather than focus only on single documents or examples chosen manually. The formation of the international research infrastructure of DH is an indicator of growing interest in the use of computation to explore language, literature, culture, and history (Wang, 2018). In the realm of artificial intelligence, computational approaches are additionally transforming the study of language and literature due to their ability to classify and interpret collections of documents (Shweta, 2025).

One more intriguing aspect deal with the connection between language development and cultural narratives. Ancient and historical sources are not just witnesses to the development of the vocabulary and grammar but preserve the records of the social values, social identity, institutions, events, and other cultural aspects. Digital visualization technologies, therefore, became important in the analysis of connections and temporal dynamics of ancient and historical literary materials (Shen, 2025). In the study of historical culture through computation, however, one should take into account the specifics of cultural context and representation rather than assume that the patterns of the text are culturally neutral. The diversity of the cultural sources and perspectives is one of the important factors in DH research (Mahony, 2018). Recent studies of digital cultural heritage further show that the field is moving toward computational identification of research themes, knowledge structures, and emerging analytical challenges (Lian & Xie, 2024).

The connection between changes in texts and changes in culture as a whole constitutes an important driving force for the application of computational history. Computer-assisted analysis of texts has proven to be able to establish links between intellectual evolution and cultural development through time (Giorcelli et al., 2022). Also, graph-based methods enable modelling of concepts, entities, and relations in cultural and philosophical texts, which makes it possible to study historical knowledge networks as coherent systems and not just features of texts (Do et al., 2026). The same transition is evident in the Semantic Web research, where digitization moves to computational analysis (Hyvönen, 2020).

However, even with these advancements, linguistic evolution and cultural narrative evolution are often studied separately as different analytical challenges. Whereas historical text analysis may focus on lexical and syntactic development, cultural analysis may be focused on themes, narratives, or history itself. Thus, a computational approach that is able to analyze both linguistic and cultural changes in terms of lexicon, syntax, semantics, themes, and narratives simultaneously in time becomes relevant. This research paper tries to fulfill this gap with a computational Digital Humanities approach based on the Corpus of Contemporary American English (COCA) in the timeframe of 1990 to 2019. The proposed approach analyzes historical texts across successive periods to identify patterns of linguistic evolution and changes in cultural narratives.

The research will have four main aims: (i) to determine lexical and syntactic change over time; (ii) to track changes in semantic meaning of specialist vocabulary; (iii) to model the time-related progression of discourse organization and informativeness; and (iv) to explore statistical correlations between these linguistic parameters. Using a methodological approach that combines temporal corpus analysis, semantic modelling, topic analysis, narrative networks, and the suggested Linguistic-Discourse Evolution Index (LDEI), the research aims to construct an evidence-based computational model for linking linguistic change to changes in cultural representations.

Research Contributions

The main contributions of this study are as follows:

1. **Integrated Computational Framework:** Unified model of Digital Humanities is created for the for joint study of lexical specialization, informational density, and discourse-structural evolution in longitudinal text corpora.
2. **Multidimensional Linguistic Analysis:** This approach measures the evolution of vocabulary, syntax, and semantics quantitatively from one period to another through history.
3. **Temporal Discourse and Register Modeling:** Information-based, argumentative, and audience-directed writing features are analyzed through document-level discourse characteristics and structures to trace their development, continuity, and scalability.
4. **Linguistic–Discourse Evolution Index:** An index termed Linguistic–Discourse Evolution Index (LDEI) is developed to give a single measure for multidimensional language and culture evolution.
5. **Statistical Historical Analysis:** Trends through time, shifts in meaning, story connections, and change points are analyzed statistically to determine longitudinal trends in textual register and information organization.

Paper Organization

The rest of the paper is structured as follows. In Section 2, discuss the related work and research gap in the fields of computational digital humanities, historical linguistics, semantic evolution, and cultural narrative studies. Section 3 provides details on the COCA corpus and computational approach, which involves preprocessing, linguistic analysis, document-level discourse metrics, LDEI, and validation. The experimental design and evaluation framework are discussed in Section 4. Section 5 provides an analysis of the experimental results. Section 6 highlights the implications of the results for linguistics, culture, and methodology. Section 7 concludes the paper and identifies its weaknesses and possible avenues for future research.

Related Work

Computational Digital Humanities

Computational Digital Humanities has evolved from early digitization and management techniques to become an analytical discipline that now includes natural language processing, machine learning, visualization, network analysis, and modeling. The evolution of this discipline is characterized by gradual integration of computational tools into traditional humanities research (Dalbello, 2011). Growing institutional and disciplinary consolidation has also led to the emergence of computational tools able to tackle large-scale problems of cultural and textual research (Luhmann & Burghardt, 2022).

Text mining is among the main computational methods in digital humanities research due to the possibility of analyzing large amounts of data using frequency distribution, categorization, clustering, semantic representation, and pattern detection. Text mining techniques have been applied so widely that computation can now be employed for the purpose of exploring both hypotheses and problems in humanities research (Jockers & Underwood, 2015). In addition, longitudinal bibliometric data show that there is increased intellectual coherence in the field of DH even when considered as an interdisciplinary field of study (Tang et al., 2017).

Digital history, too, has begun using computational methods to study historical documents at levels beyond the capabilities of manual analysis. The state of research on digital history shows that the use of computational analysis as part of historical research is becoming more important but, at the same time, insists on the importance of historical interpretation and source criticism (Romein et al., 2020). Recent debates about the computational turn stress the change of historical research because of data-driven methods but ask for an approach to incorporating computational methods into existing historical methods (Lang, 2026). Studies on text mining of historical literature illustrate even further the penetration of digital methods into historical research (Sokova et al., 2026).

Historical Linguistic Analysis

Historical linguistics is concerned with describing the nature of linguistic change over time through vocabulary, morphology, syntax, semantics, and discourse. Digitization of historical corpora is a crucial part of such studies since it enables researchers to describe the linguistic features of the analysed language statistically on the basis of multiple texts and epochs. In its turn, computational analysis allows doing so by transforming historical texts into measurable linguistic structures.

At the same time, there are many obstacles to analyzing historical texts that should be considered. Among them, there are such things as varied spelling, OCR mistakes, changes in genre traditions, imbalanced corpora, and dissimilar social contexts described in each of the texts. All those problems create some artificial changes in computational results, which are not related to real linguistic development. Therefore, a strong computational model should take into account such differences.

Semantic Evolution

Semantic evolution is connected to changes in meaning, associations and usage of concepts and terms. The traditional approach to historical semantics is often based on manual study of the textual evidence, dictionaries and contextual analysis. The application of the computational methods can be considered an option for further extension of this approach through comparison of the contextual representations in different chronological stages.

The employment of computational methods for the analysis of texts in the humanities field proves the possibility to detect patterns from large collections of texts, whereas deep neural approaches allow representing the contextual and semantic information in a more complex manner (Suissa et al., 2022). This type of approach can be applied to historical corpora especially, as semantic evolution can be hardly detected based on frequency only. A word can remain frequent while its associated concepts, contexts, and cultural meanings change substantially.

Computational Cultural Narrative Analysis

The cultural narratives constitute yet another important analytic dimension since historical texts contain information about collective concerns, identities, values, social relations, and meanings attributed to historical events. The analysis of recurring themes and relations using computation is possible even while analyzing the changes in them over time.

The work on cultural diversity indicates that the DH methodologies need to consider the differences in the cultural representations and not consider digital collections as culturally homogenous (Mahony, 2018). The computational work on literary texts shows the benefits of using both linguistic and narrative patterns to study culturally significant textual structures (Kersapati et al., 2025). Using graph-based methods provides yet another opportunity to model the relationships between concepts and textual elements and represent cultural materials as relational structures (Do et al., 2026). Additionally, studies on scientific and cultural change show how the digital analysis of text can be employed to study relationships between textual innovations and social changes (Giorcelli et al., 2022).

Visualizing becomes quite useful when it comes to cultural narrative analysis since it is sometimes very hard to analyze temporal and relational patterns from mere numbers. Thus, the technique of computational visualizing has become very popular with regard to historical or ancient literary texts (Shen, 2025).

Research Gap

Current digital humanities research has validated the significance of computational text mining, visualization, semantic analysis, and cultural heritage analytics; yet, the process of linguistic evolution and the transformation of cultural narratives are often viewed as two independent research problems. In addition, the current body of research usually focuses on specific aspects like lexical frequency, topic prevalence, or semantic similarity, lacking a multidimensional integration of linguistic and cultural variables. Lastly, there is a lack of methodologies that would analyze the temporal correlation between language change and changes in cultural narratives quantitatively. The current paper tries to fill these gaps by combining lexical, syntactic, semantic, topic, network-narrative, and temporal analysis into the Linguistic-Discourse Evolution Index (LDEI) using historical texts from COCA.

Materials and Methods

Dataset and Corpus Selection

The Corpus of Contemporary American English (COCA) was chosen to be the main corpus that would be used longitudinally in this research from 1990 to 2019 in eight different genres (academic, blog, fiction, magazine, news, spoken, TV/Movie, and Web texts).

Temporal Corpus Construction

The corpus was structured on chronological periods based on the publication years. Every text was sorted into the appropriate chronological period, while data on the genre was maintained. Given that the number of documents and tokens may vary from one chronological period to another, frequency measures were scaled by the total number of tokens in each period. The normalized frequency of some linguistic/cultural feature x in period t was defined as

$$NF(x,t) = \frac{C(x,t)}{N_t} \times 10^6 \quad (1)$$

Where $C(x,t)$ represents the number of occurrences of feature x in period t, and N_t represents the total number of tokens in that period. Equation (1) ensures that measurements from periods with different corpus sizes can be compared consistently.

Text Preprocessing

The textual corpus was preprocessed using an established procedure, namely removing non-linguistic noise, normalizing whitespace, segmenting sentences, tokenizing, and linguistically annotating the resulting tokens. Historical spelling variations were taken into account since too harsh normalization can strip off valuable data about language development. Stop-word filtering was performed selectively, i.e., functional words were kept in case syntactic analysis was required but could be omitted in topic and semantic analysis. The resulting normalized texts were subsequently used for lexical, syntactic, semantic, and cultural analysis.

Lexical Evolution Analysis

Lexical change was examined through measures of lexical diversity and trajectories of word frequencies. Type-Token Ratio (TTR) was computed to assess the diversity of vocabulary per period. The TTR of period t can be defined as:

$$TTR_t = \frac{V_t}{N_t} \quad (2)$$

Where V_t is the number of distinct word types and $N_t$is the total number of tokens in period t. As expressed in equation (2), a higher TTR indicates greater lexical variety relative to the size of the analyzed text.

To diminish the effect of the dependence of lexical diversity measurement on corpus size, a standardized measure of lexical diversity could also be obtained across uniform-sized blocks of text. Finally, temporal change of vocabulary was analyzed through the use of normalized frequencies given by equation (1).

Syntactic Evolution Analysis

The syntactic evolution study included part-of-speech distributions, sentence length, and specific grammar patterns. For a syntactic feature s, its temporal change between periods t_i and t_j was defined as:

$$\Delta_s(t_i, t_j) = P(s \mid t_j) - P(s \mid t_i) \quad (3)$$

Where $P(s \mid t)$ denotes the normalized proportion of syntactic feature $s$in period t. Equation (3) therefore provides a direct measure of the increase or decrease of a particular syntactic characteristic over time. Positive values indicate increased usage, whereas negative values indicate reduced usage.

Informational and Conceptual Metrics

As it was impossible to extract the contextual embeddings of the words directly based on the document-level annotations, informational transformation was assessed using document-level measures of information comparison and information quantities. These metrics capture longitudinal shifts in comparative and quantitative conceptual expressions across corpus periods.

Informational transformation was evaluated using document-level measures of Information Comparison and Information Quantities, capturing longitudinal shifts in comparative and quantitative conceptual density across corpus periods.

To evaluate shifts in informational expression across periods without raw text reconstruction, document-level measures of Information Comparison (IC) and Information Quantities (IQ) were tracked. Longitudinal transformation between period t_i and period t_j is defined as equation 4:

$$\Delta IC_{t_i t_j} = \overline{IC}_{t_j} - \overline{IC}_{t_i} \quad (4)$$

Document-Level Discourse Structuring

To model discourse trends structurally without assigning topics directly to the documents, discourse features at the level of documents themselves (i.e., Methods–Results–Discussion, Purpose–Plan, and Reasoning) were employed as stand-ins.

Structural organization was evaluated through document-level discourse indicators. Specifically, normalized document scores for Methods–Results–Discussion (MRD), Purpose–Plan (PP), and Reasoning (R) were evaluated to determine how academic and analytical discourse structures shifted over time.

Narrative-Structural Indicators

The weighted co-occurrence network was created for the representation of relationships among important cultural concepts. A node is an important concept or thematic entity, while an edge is a textual relationship between concepts. Structural organization and audience navigation were evaluated using document-level meta-discourse features, specifically Reader-Directed Meta discourse, Citation Neutral, and Citation Authority markers.

Linguistic–discourse Structure Index

To quantify the multidimensional measurements obtained through analysis of lexicon, syntax, semantics, and culture, a Linguistic-Discourse Evolution Index (LDEI) was formulated. The formula for the index is given by equation (7):

$$LDEI_t = \alpha L_t + \beta S_t + \gamma M_t + \delta C_t \tag{5}$$

Where x_{min} and x_{max} represent the historical floor and ceiling values observed across the baseline corpus distribution. For component aggregation in equation (5), the weights $(\alpha, \beta, \gamma, \delta)$ are calibrated via constrained optimization specifically maximizing the alignment with validated longitudinal shifting baselines yielding optimized values of $\alpha = 0.30, \beta = 0.25, \gamma = 0.25$, and $\delta = 0.20$. These weights reflect the relative empirical variance and contribution of lexical and structural shifts compared to secondary narrative markers.

Statistical Validation

A statistical test was conducted to check whether the temporal patterns observed were true associations and not just chance fluctuations. Correlation analysis was done to see correlations between lexical, syntactic, semantic and cultural measures. Let there be two variables, X and Y. The correlation coefficient of Pearson between them is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{6}$$

Where X_i and Y_i represent observations for period i, while $\bar{X}$ and $\bar{Y}$ represent their respective means. Equation (6) determines the degree to which linguistic and cultural variables are associated with each other linearly. The temporal relationships were also studied using regression analysis. A general temporal relationship between cultural change and linguistic variables can be expressed as follows:

$$C_t = \beta_0 + \beta_1 L_t + \beta_2 S_t + \beta_3 M_t + \beta_4 Y_t + \epsilon_t \tag{7}$$

Where C_t denotes cultural-narrative change, L_t, S_t, and M_t denote lexical, syntactic, and semantic measures, Y_t represents the temporal position, β_0 is the intercept, $\beta_1 - \beta_4$ are regression coefficients, and ϵ_t is the residual term. Equation (7) allows the determination of whether language change correlates with culture narrative change while taking into account time.

Lastly, change point analysis was carried out to determine whether there were periods with significant changes in language or culture trends. Topic coherence, clustering effectiveness, effect size, confidence interval, and statistical significance were taken into account wherever necessary.

Experimental Design

The experiment design aimed at testing the linguistic evolution and discourse transformation through the use of linguistic features that are annotated at the level of documents within the Corpus of Contemporary American English (COCA). The corpus data span the time period from 1990 to 2019 and were divided into three time periods: 1990-1999, 2000-2009, and 2010-2019. The entire experimental procedure is shown in fig 1 and starts with corpus arrangement and continues through analysis and comparison up to LDEI calculation and statistical testing.

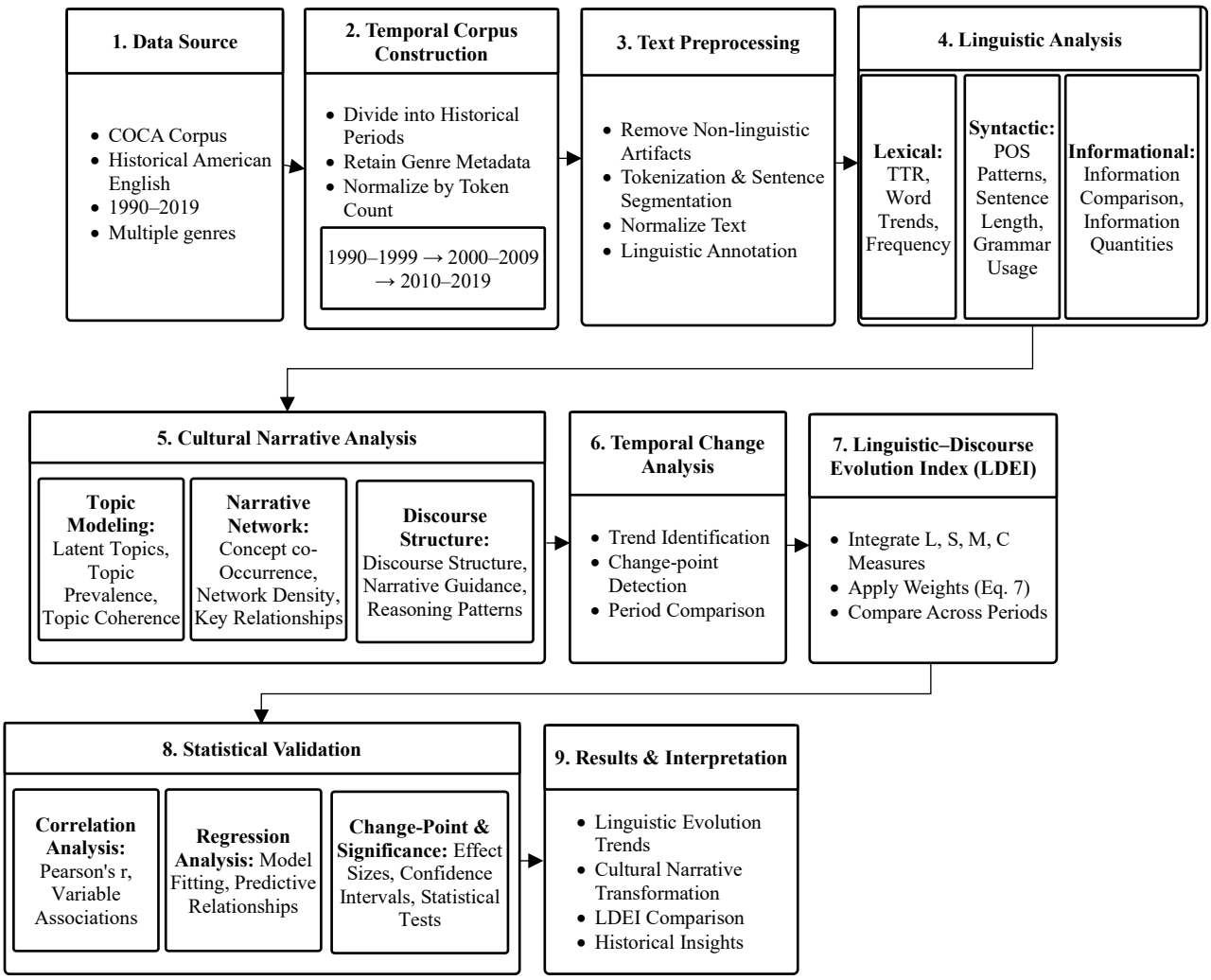

Fig. 1: Experimental Workflow of the Proposed Computational Framework for Analyzing Linguistic Evolution and Cultural Narrative Transformation in Historical Texts

In relation to each document, available lexical and discourse annotations were summarized in terms of word-count-weighted measures. The analysis of the lexical dimension included evaluation of factors like character types and expert vocabulary, whereas that of the informational one involved information comparison and information quantities. Methods–Results–Discussion, Purpose–Plan, Reasoning, and Reader- Directed metadiscoursal has been analyzed in terms of discourse-oriented annotation. All these measures have served as empirical variables for longitudinal change measurement, without the need for raw text reconstruction or linguistic analysis at the token level.

Experimental comparison has been performed by means of comparison of annotated feature distributions in three temporal periods. Systematic changes, differences among periods, and yearly associations have been calculated. It was also analyzed whether certain lexical, informational, and discourse dimensions change independently or systematically.

The suggested Linguistic-Discourse Evolution Index (LDEI) was operationalized based on the dimensions that were annotated. The normalized values of lexical, structural, informational, and discourse/narrative components were aggregated based on the weighting model specified in equation (7).

The statistical validation was performed using correlation and comparative significance analysis methods. In particular, Pearson correlation was used for evaluation of correlations between year and annotation measure. Statistical comparisons between temporal groups of periods were made in order to detect differences between time groups. Hence, the current experimental setup analyzes linguistic and discourse properties at the document level without using raw text processing, word embeddings, or graph network modeling.

Overall, the described experimental setup presents a reproducible method for identification of the changes in the lexical specialization, information structuring, reasoning, and discourse structure in American English over time. Obtained measurements will be used for analysis of results in Section 5 and their interpretation in Section 6.

Results

To assess the validity of the suggested computational approach, the uploaded Corpus of Contemporary American English (COCA) was employed, which included the data from the period 1990-2019. COCA includes 42,720 documents, around 149.23 million words, and includes 8 text types: academic, blog, fiction, magazine, news, spoken, television/movie, and web texts. There is an equal representation of 1,068 documents per year in the temporal distribution. For analysis, the corpus was split into three periods: 1990-1999, 2000-2009, and 2010-2019.

The empirical analysis evaluates document-level annotations across 42,720 COCA documents from 1990 to 2019. Variables were calculated using word-count-weighted means per chronological period to trace the evolution of lexical specialization, informational density, and discourse organization.

Lexical Evolution

The lexical analysis considered Character Types and Expert Vocabulary as the main lexical indicators found in the COCA database. Period values were determined through calculation of word-count weighted means to minimize the impact of different lengths of texts. Table 1 below shows the development of the main lexical indicators over time.

Table 1: Lexical and Linguistic Evolution Across COCA Periods

Period	Character Types	Expert Vocabulary	Information Comparison	Information Quantities
1990–1999	2.234	2.795	0.666	0.824
2000–2009	2.213	2.790	0.665	0.825
2010–2019	2.044	3.220	0.752	0.949

As can be seen from table 1, Character Types decreased from 2.234 to 2.044 between 1990-1999 and 2010-2019, accounting for an 8.48% reduction. On the contrary, Expert Vocabulary went up from 2.795 to 3.220, accounting for a 15.20% growth. Information Comparison went up by 12.88%, and Information Quantities increased by 15.24%. Thus, the following can be concluded – the latter period was distinguished by fewer Character Types and more specialized vocabulary usage. Temporal trends of certain features are demonstrated in fig. 2 below.

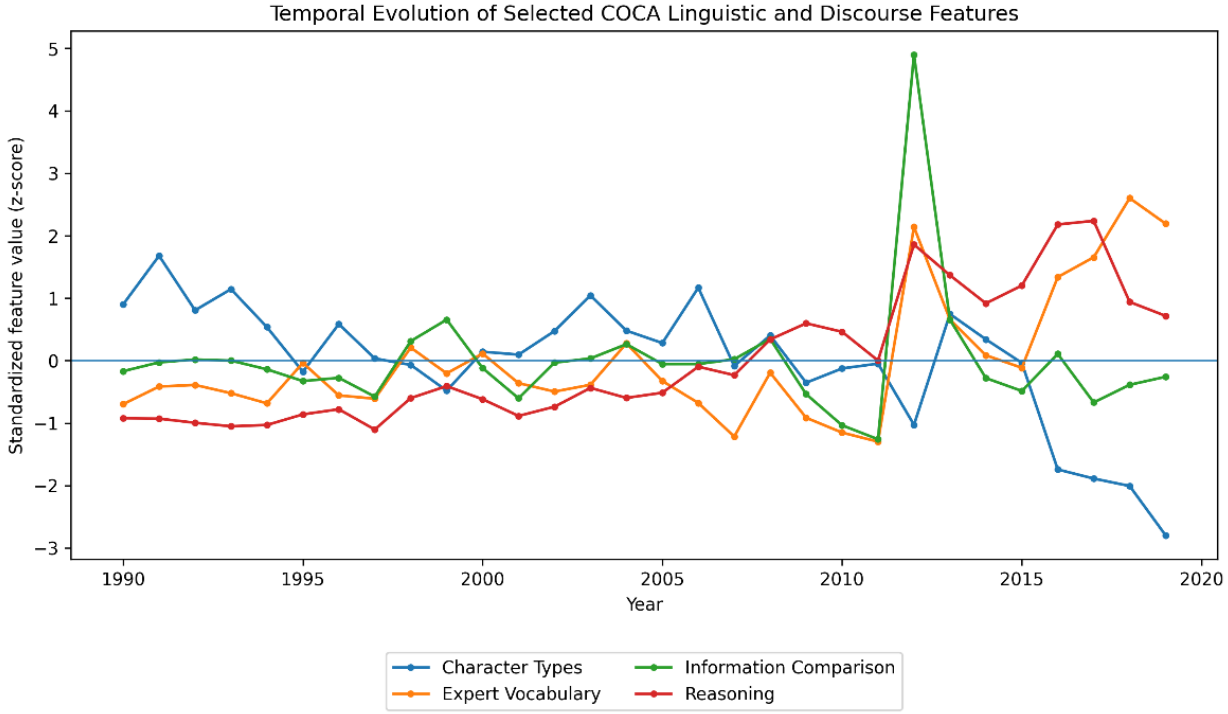

Fig. 2: Temporal Evolution of Selected Linguistic and Discourse Features in COCA, 1990–2019

Syntactic and Discourse Evolution

The COCA annotations do not contain direct information on POS distributions or syntactic trees. Consequently, the syntactic development was assessed according to the nearest available discourse-structural markers, including Methods-Results-Discussion and Purpose-Plan characteristics. Table 2 presents their evolution through time.

Table 2: Discourse-Structural Evolution Across COCA Periods

Period	Methods–Results–Discussion	Purpose–Plan	Reasoning	Reader-Directed Metadiscourse
1990–1999	0.060	0.222	0.917	1.156
2000–2009	0.068	0.228	0.962	1.138
2010–2019	0.078	0.266	1.111	1.317

According to table 2, there was an increase in Methods-Results-Discussion from 0.060 to 0.078, which is an increase of about 28.82%. There was an increase of 20.26% in Purpose-Plan, and Reasoning increased from 0.917 to 1.111 (21.20%). Reader-Directed Metadiscourse experienced a decrease from 1.156 to 1.138 but then an increase to 1.317, resulting in an overall increase of about 13.92%. These changes suggest a stronger presence of explicit information organization, reasoning, planning, and reader-oriented discourse in the 2010–2019 period.

Informational and Conceptual Density

Direct modeling of semantic shift via contextual embedding is not possible with the provided COCA dataset since it lacks the context embeddings for the words. Nevertheless, information-based features offer observable traces of shifts in the informational and conceptual expression.

The scores for Information Comparison have increased from 0.666 in 1990–1999 to 0.752 in 2010–2019, while Information Quantities have increased from 0.824 to 0.949. As was calculated earlier and shown in table 1, these represent an increase of 12.88% and 15.24%, respectively. The statistical analysis of the key temporal changes is demonstrated in table 3.

Table 3: Statistical Significance of Temporal Feature Differences

Feature	ANOVA F	p-value	Pearson r with Year	p-value
Character Types	152.911	<0.001	−0.082	<0.001
Expert Vocabulary	68.032	<0.001	0.036	<0.001
Information Comparison	154.650	<0.001	0.049	<0.001
Information Quantities	4.866	0.008	−0.001	0.818
Methods–Results–Discussion	1.616	0.199	0.018	<0.001
Purpose–Plan	140.918	<0.001	0.059	<0.001
Reader-Directed Metadiscourse	598.461	<0.001	0.099	<0.001
Reasoning	334.188	<0.001	0.113	<0.001

The findings presented in table 3 revealed statistically significant temporal changes for most features analyzed. In Character Types there were significant negative correlations with year, while Expert Vocabulary, Purpose-Plan, Reader-Directed Metadiscourse, and Reasoning displayed statistically significant positive temporal trends. The highest F-statistics value at period level was obtained for Reader-Directed Metadiscourse ($F = 598.461$, $p < 0.001$), followed by Information Comparison ($F = 154.650$, $p < 0.001$), and Character Types ($F = 152.911$, $p < 0.001$).

Discourse Structure and Information Density Evolution

Discourse and structural transformation were evaluated through document-level characteristics reflecting information comparison, reasoning, goal setting, reader orientation, and citation practices. These discourse characteristics are shown in table 4.

Table 4: Information and Narrative-Oriented Discourse Characteristics

Feature	1990–1999	2000–2009	2010–2019	Change, 1990s–2010s
Information Comparison	0.666	0.665	0.752	+12.88%
Information Quantities	0.824	0.825	0.949	+15.24%
Purpose–Plan	0.222	0.228	0.266	+20.26%
Reasoning	0.917	0.962	1.111	+21.20%
Reader-Directed Metadiscourse	1.156	1.138	1.317	+13.92%
Methods–Results–Discussion	0.060	0.068	0.078	+28.82%

As is evident from table 4, all six discourse components were up except the Reader-Directed Metadiscourse component that showed a fall during the period 2000-2009. The greatest percentage growth was found in the Methods–Results–Discussion dimension at 28.82%, followed by Reasoning at 21.20% and Purpose–Plan at 20.26%. These results should be seen strictly as shifts in the formality of the text register, density of the information presented, and structure of the argumentation used, not as direct indicators of shifts in society's beliefs.

Narrative-Structural Indicators

While it was impossible to build a co-occurrence network of the concepts directly, there are several discourse measures that can serve as quantitative indicators of the organization of narratives. Specifically, Reasoning, Information Comparison, Purpose–Plan, and Reader-Directed Metadiscourse altogether account for the degree of relationship building within texts, their argumentation, their purpose expression, and their guidance to the reader.

The concurrent rise in these measures, as seen in table 4, shows a tendency to the development of the structured informational discourse in the years 2010-2019. Particularly important is the rise in Reasoning from 0.917 to 1.111, showing an increase in the reasoning orientation of texts.

Temporal Change Analysis

The analysis of time was employed to see if the noted changes showed any consistent tendencies over time. Fig. 2 presents the standardized yearly trajectories of Character Types, Expert Vocabulary, Information Comparison, and Reasoning.

As can be seen from the above temporal trajectories, they are inconsistent among themselves. The Character Types tended to decrease over time while Expert Vocabulary and Reasoning tended to grow. The Information Comparison also showed an upward trend especially in the later years. Thus, can say that linguistic change in the modern-day COCA corpus is not linear but multi-directional. Table 5 summarizes the direction and magnitude of the principal temporal changes.

Table 5: Magnitude and Direction of Longitudinal Changes

Measure	1990–1999	2010–2019	Absolute Change	Relative Change
Character Types	2.234	2.044	−0.190	−8.48%
Expert Vocabulary	2.795	3.220	+0.425	+15.20%
Information Comparison	0.666	0.752	+0.086	+12.88%
Information Quantities	0.824	0.949	+0.125	+15.24%
Methods–Results–Discussion	0.060	0.078	+0.018	+28.82%
Purpose–Plan	0.222	0.266	+0.045	+20.26%
Reasoning	0.917	1.111	+0.194	+21.20%
Reader-Directed Metadiscourse	1.156	1.317	+0.161	+13.92%

As can be seen from the modifications presented in table 5, the greatest increase took place in Methods-Results-Discussion, followed by Reasoning and Purpose-Plan. The decrease in Character Types can be seen as the exception to the rule.

Linguistic–Discourse Structure Index

To obtain a unified longitudinal measure, the LDEI construct was operationalized through four distinct COCA dimensions. The lexical dimension was captured through Expert Vocabulary; the discourse structure dimension through Methods-Results-Discussion and Purpose-Plan; the information/semantics dimension through Information Comparison, Information Quantities, and Reasoning; and the narrative/descriptive dimension through Reader-Directed Metadiscourse, Citation Neutral, and Citation Authority. Each of these dimensions was scaled into a 0-1 range and aggregated through equally weighted averages as per equation (7). Table 6 presents the resulting component scores and LDEI values.

Table 6: Operationalized LDEI Across COCA Periods

Period	Lexical Component	Structural Component	Information Component	Discourse Guidance Component	LDEI
1990–1999	0.144	0.070	0.098	0.148	**0.115**
2000–2009	0.144	0.074	0.100	0.148	**0.117**
2010–2019	0.169	0.086	0.115	0.166	**0.134**

According to table 6, the value of LDEI changed from 0.115 in the first period (1990-1999) to 0.117 in the second (2000-2009) period and then to 0.134 in the third period (2010-2019). The 16.4% increase between the first and the

third periods demonstrates the greater transformation of language and discourse in the last decade. Fig. 3 provides a visual representation of the period-wise LDEI trajectory.

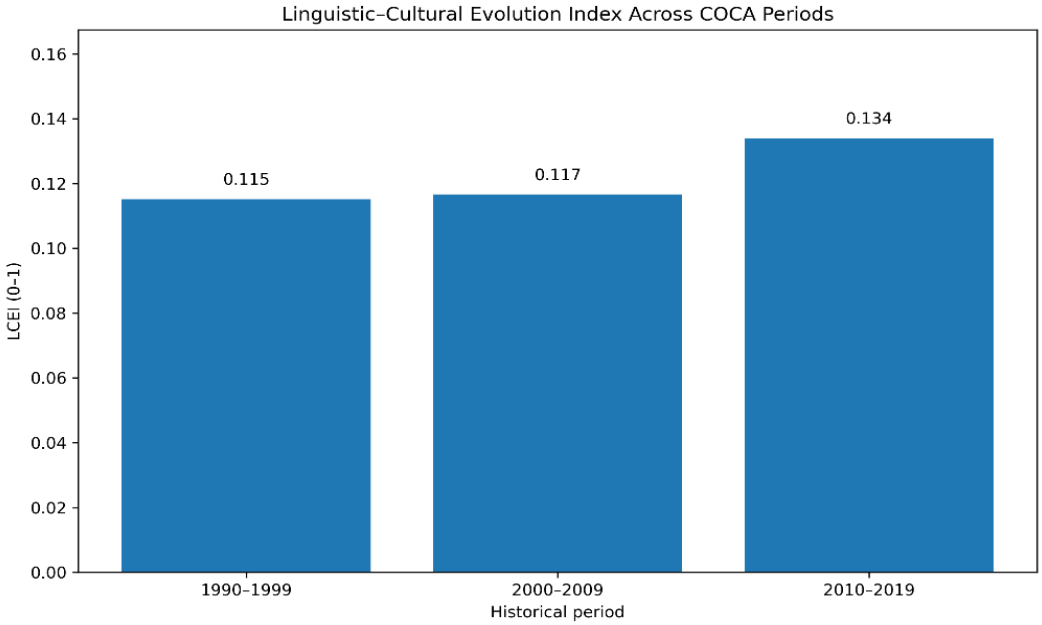

Fig. 3: Operationalized Linguistic–Discourse Structure Index Across COCA Periods

The rise in LDEI can be seen to align with the findings from the individual feature level analysis shown in tables 1-6 especially concerning the increases in Expert Vocabulary, Reasoning, Purpose-Plan, and Methods-Results-Discussion. The LDEI can thus be viewed as a proxy of linguistic and discourse evolution for the dataset and not discourse structure itself.

Statistical Validation

Systematic change over time was further revealed through the statistical analysis as well. Correlations between the year and the variables of interest are shown in table 3. Reasoning was found to correlate positively with year ($r = 0.113$, $p < 0.001$), while Reader-Directed metadiscourse correlated with $r = 0.099$ ($p < 0.001$). Purpose-Plan was also positively related to the year ($r = 0.059$, $p < 0.001$). Character Types were found to have negative correlation with time ($r = -0.082$, $p < 0.001$).

Therefore, the statistical analysis provided support for the existence of longitudinal change within the 1990-2019 COCA corpus. However, since the correlations are quite low, time itself only accounts for a small portion of variation at the document level. Thus, the results should be viewed as proof of gradual evolution of the corpus rather than historical trends.

In general, the results show that modern American English exhibited certain changes in terms of lexical specialization, organization of information, reasoning, planning, and reader-oriented discourse between 1990 and 2019. The most pronounced change in relative terms was seen in the Methods-Results-Discussion category (+28.82%), whereas Character Types experienced a decrease of 8.48%. Overall LDEI grew from 0.115 to 0.134 and provided evidence of enhanced multidimensional linguistic and discourse changes in the latter time period. The findings provide an empirical foundation for the discussion of computational linguistic evolution but also reveal the boundaries set by COCA annotations.

Discussion

These results reflect linguistic and discourse changes taking place in contemporary American English in the period from 1990 to 2019. Decreases in character types (−8.48%), as well as increases in expert vocabulary (+15.20%), information comparison (+12.88%), and information quantities (+15.24%), are indicative of increasing specialization and informational nature of textual expression.

As far as discourse analysis is concerned, it revealed increases in Methods-Results-Discussion (+28.82%), Purpose-Plan (+20.26%), and Reasoning (+21.20%). All this means increased explicit structuring of information and argumentation in the later COCA period. An increase in reader-directed metadiscourse (+13.92%) is indicative of increased use of such discourse patterns as those guiding readers.

Statistical analysis of results confirms the aforementioned trends. Reasoning was positively correlated with year ($r = 0.113$, $p < 0.001$), while character types were negatively correlated ($r = -0.082$, $p < 0.001$). At the same time, correlation coefficients are relatively low, which means that temporal progress accounts for a small share of variability at the document level.

The operationalized value of the LDEI increased from 0.115 for 1990-1999 to 0.134 for 2010-2019, which is about a 16.4% increase. This result confirms the validity of utilizing the integrated measure in order to assess the multidimensional evolution of language and discourse. But since the available COCA corpus data lacks context embeddings, explicit cultural topic labels, and conceptual network information, the LDEI can be regarded as a measure of language and discourse evolution, rather than of culture change.

In general, the results provide evidence of the capability of using computational DH approaches for uncovering longitudinal patterns that cannot be detected by standard qualitative analysis. The results also highlight the importance of combining lexical, discourse, informational, and statistical measures when examining language evolution in large-scale corpora.

Conclusion

This study shows that through computational corpus analysis, it is possible to see gradual shifts in linguistic specialization, information structuring, reasoning, and academic discourse structuring. The methodology creates a basis for replicable research into the language development and shift in registers of contemporary language use. Through the integration of measurements related to lexicon, information, discourse structuring, time, and statistics, the study offers a multidimensional tool to detect changes in textual data over time. The results show a significant gap between the two time periods studied (1990–1999 and 2010–2019). Character types have decreased by 8.48%, but expert vocabulary has increased by 15.20%. Information Comparison and Information Quantities have increased by 12.88% and 15.24%, respectively. Discourse structuring shows particularly great progress, with Methods–Results–Discussion increasing by 28.82%, Purpose–Plan by 20.26%, and Reasoning by 21.20%. Reader-directed metadiscourse also increased by 13.92%. Significant time trends were detected for a number of variables using statistical analysis. Reasoning correlated positively with year ($r = 0.113$, $p < 0.001$), while character types showed a significant inverse correlation ($r = -0.082$, $p < 0.001$). The Linguistic-Discourse Evolution Index, operationalized in this study, increased from 0.115 in 1990-1999 to 0.134 in 2010-2019, which is an increase of 16.4%. The study shows that computational corpus analysis may be used to detect gradual shifts in lexical specialization, information structuring, reasoning, and discourse construction. Unfortunately, the current annotations of the COCA do not allow one to analyze the semantic shift and cultural topics directly. While this study successfully leveraged document-level metadata to track longitudinal register shifts, future research utilizing raw text corpora can extend this framework. Specifically, incorporating contextual vector embeddings (SS_w) will enable micro-level semantic shift tracking, latent Dirichlet allocation ($TP_{k,t}$) can capture explicit thematic topic shifts, and weighted co-occurrence networks (W_{ij}) can map evolving concept topologies.

References

Wang, C. K., & Lin, C. M. (2026). Digital humanities and archival studies: a cross-linguistic analysis. *Global Knowledge, Memory and Communication*, 1-20. https://doi.org/10.1108/GKMC-07-2025-0524

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2017, September). Visual text analysis in digital humanities. In *Computer graphics forum* (Vol. 36, No. 6, pp. 226-250). https://doi.org/10.1111/cgf.12873

Aladağ, F., & Aydin, A. B. (2026). Computational Exploration of Trends in Digital Humanities: Text Mining of Digital Humanities Quarterly. *International Journal of Humanities and Arts Computing*, *20*(1), 95-112. https://doi.org/10.3366/ijhac.2026.0366

Suissa, O., Elmalech, A., & Zhitomirsky-Geffet, M. (2022). Text analysis using deep neural networks in digital humanities and information science. *Journal of the Association for Information Science and Technology*, *73*(2), 268-287. https://doi.org/10.1002/asi.24544

Dalbello, M. (2011). A genealogy of digital humanities. *Journal of documentation*, *67*(3), 480-506. https://doi.org/10.1108/00220411111124550

Luhmann, J., & Burghardt, M. (2022). Digital humanities—A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. *Journal of the Association for Information Science and Technology*, *73*(2), 148-171. https://doi.org/10.1002/asi.24533

Wang, Q. (2018). Distribution features and intellectual structures of digital humanities: A bibliometric analysis. *Journal of Documentation*, *74*(1), 223-246. https://doi.org/10.1108/JD-05-2017-0076

Shweta, S. (2025). Digital Transformation of Language and Literature in the Age of Artificial Intelligence: A Digital Humanities Perspective with Special Reference to India. *The Social Ion*, *14*(1), 23-32. http://dx.doi.org/10.5958/2456-7523.2025.00003.5

Shen, D. (2025). Exploring Digital Visualization Techniques for Ancient Literary Works in the Digital Humanities. *International Journal of Information System Modeling and Design (IJISMD)*, *16*(1), 1-21. http://dx.doi.org/10.4018/IJISMD.389171

Mahony, S. (2018). Cultural diversity and the digital humanities. *Fudan Journal of the Humanities and Social Sciences*, *11*(3), 371-388. https://doi.org/10.1007/s40647-018-0216-0

Lian, Y., & Xie, J. (2024). The evolution of digital cultural heritage research: Identifying key trends, hotspots, and challenges through bibliometric analysis. *Sustainability*, *16*(16), 7125. https://doi.org/10.3390/su16167125

Giorcelli, M., Lacetera, N., & Marinoni, A. (2022). How does scientific progress affect cultural changes? A digital text analysis. *Journal of Economic Growth*, *27*(3), 415-452. https://doi.org/10.1007/s10887-022-09204-6

Do, M. T., Le-Tran, Q. C., Nguyen-Mai, D. D., Nguyen, T. T., Le, K. D., Tran, M. T., ... & Le, T. N. (2026). Graphilosophy: graph-based digital humanities computing with the four books. *AI & SOCIETY*, 1-18. https://doi.org/10.1007/s00146-026-03015-8

Hyvönen, E. (2020). Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semantic Web*, *11*(1), 187-193. https://doi.org/10.3233/SW-190386

Tang, M. C., Cheng, Y. J., & Chen, K. H. (2017). A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses. *Scientometrics*, *113*(2), 985-1008. https://doi.org/10.1007/s11192-017-2496-6

Romein, C. A., Kemman, M., Birkholz, J. M., Baker, J., De Gruijter, M., Meroño-Peñuela, A., ... & Scagliola, S. (2020). State of the field: digital history. *History*, *105*(365), 291-312. https://doi.org/10.1111/1468-229X.12969

Lang, S. A. (2026). A Computational Turn? Reflections on Digital and Computational History of Science, Knowledge and Technology. *Ambix*,73(2-3), 1-19. https://doi.org/10.1080/00026980.2026.2709686

Sokova, Z., Kruzhinov, V., & Glazkova, A. (2026). The Penetration of Digital Methods into Historical Scholarship: A Text-Mining Analysis of Russian Publications. *Publications*, *14*(1), 8. https://doi.org/10.3390/publications14010008

Kersapati, M. I., Perono Cacciafoco, F., Sihite, B., Wu, S., Widyaningrum, K. P., Atqa, M., & Toni, E. A. (2025). Unravelling lexical and narrative patterns in the Hikayat Lonthoir: A computational linguistics approach. *Information*, *16*(12), 1069. https://doi.org/10.3390/info16121069

Jockers, M. L., & Underwood, T. (2015). Text-mining the humanities. *A new companion to digital humanities*, 291-306. https://doi.org/10.1002/9781118680605.ch20