ISSN: 2278-4012, Volume:14, Issue:4, December 2025

Computational Approaches for Cultural Representation Analysis in Contemporary English Literature

Theodore Kouassi Kouassi[1*]

[1*]Department of English, Faculty of Languages and Literature, Alassane Ouattara University, Bouake, Côte d'Ivoire. E-mail: kouassitheodore82@yahoo, Orcid: https://orcid.org/0000-0001-7448-3357

Abstract: In digital humanities, computational methods are gaining more significance and in digital humanities, researchers increasingly use computational methods to analyze vast amounts of literature alongside qualitative analysis. to analyze vast amounts of literature in addition to qualitative analysis. This research aims to create a computational model to assess the BV of modern English literature based on the Goodreads Books Dataset. The suggested framework differs from a typical literary analysis that makes use of full text by applying bibliographic metadata such as language code, year published, publisher, average writing rating, number of ratings, and number of text reviews as proxy indicators of the visibility of the literature in the publications and readers' engagement with them. The framework involves steps of data preprocessing, feature extraction, bibliographic analysis based on metadata, and computation of a Bibliographic Visibility Index (BVI) by weighted aggregation of four normalized indicators: Language Distribution, Publication Trend, publisher Distribution, and Reader Engagement. Duplicate and incomplete records were filtered out, and only books published since 2000 were selected. The normalized scores of 0.84, 0.79, 0.73, and 0.88, respectively, yielded an overall BVI of 0.82. A sensitivity analysis was performed, which indicated that the proposed framework is robust with regard to the indicator weights, as changes of ±10% only affected the BVI by 0.80-0.83. The results show good visibility of the publications in the public, with a fairly small audience of readers and participants in the publications, which is an indication of the fact that the publishing market is concentrated. While the framework does not directly assess cultural themes in literary texts, it offers a scalable, reproducible, and computationally efficient approach to the analysis of literary texts via metadata, enabling quantitative digital humanities research and large-scale studies of contemporary English literary culture.

Key Words: contemporary english literature, computational literary analysis, goodreads books, cultural representation index, digital humanities, metadata analytics

(Received: 15 September 2025; Revised: 29 October 2025; Accepted: 27 November 2025; Published: 31 December 2025)

Introduction

The development of computational techniques, literary studies have seen significant growth and changes as they are able to process a huge amount of text and bibliographic information (Hatzel et al., 2023; Li et al., 2024). The current literary criticism mainly depends on qualitative interpretation, an approach that is usually time-consuming and challenging to scale up. Progresses in Natural Language Processing (NLP), machine learning, and digital humanities recently have led to the development of computational tools that can be used to objectively analyze literary themes, cultural patterns, and reader responses to large bodies of literature (Li et al., 2024; Rashidi et al., 2024). The methods are alternative ways of analyzing cultural representation in modern English literature with the data approach.

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

Contemporary English literature represents a variety of cultural identities, languages, migrant experiences, and social viewpoints, which lead to the global conversation in literature (Mengliev et al., 2024). It is important to know about these cultural depictions so as to judge inclusiveness and diversity in contemporary literature. Combined with traditional literary criticism (Pardey, 2023; Charlesworth et al., 2024), computational methods enable systematic analysis of literary metadata by uncovering patterns in relation to publication trends, language, and reader reception.

The present study that suggests a computational model that employs a piece of metadata to measure the Bibliographic Visibility Index (BVI) of the current English literature based on the Goodreads Books dataset. The framework does not rely on the text of literary works but rather bibliographic metadata such as language, year of publication, publisher, ratings, and reader reviews as proxy indicators of reach, reader engagement, and dissemination. These metadata do not capture the cultural themes or identities found in the texts themselves, but they do provide quantitative measures of the publication, distribution, and reception of literary works in the present and could be used to assist large-scale DH research.

Objectives

- The aim of this study is to build a computational model to study the representation of culture in contemporary English literature within the context of the Books Dataset from the Goodreads platform.
- To assess the representation of culture using metadata-based metrics such as language distribution, publication metrics and reader engagement.
- To develop a Cultural Representation Index (CRI) to be used to quantify the extent of cultural representation in current English literature.

The rest of the paper is structured as follows: The section 2 of the paper revisits recent research on the analysis of literature and cultural representation using computational methods. In Section 3, the proposed computational framework and set of data are discussed. The experimental analysis and findings are discussed in Section 4, and the results of the paper are summarized, and future research directions are outlined in Section 5.

Related Work

There are basically two types of computational literary studies: full-text computational analysis and metadata-driven literary analysis (Hatzel et al., 2023; Li et al., 2024; Hirsbrunner, 2024). Full-text methods use Natural Language Processing (NLP), topic modeling, and machine learning to uncover themes, patterns, and stylistic features of texts in textual corpora. Topic modeling has gained popularity to discover hidden themes, and machine learning has contributed to efficient and reproducible literary analysis in DH (digital humanities) (Li et al., 2024; Preeti, 2024). These techniques have recently been extended to the analysis of dystopian narratives, post-colonial fiction, multilingual cultural keywords, and literary allusions, showing that computational analysis of textual data can be effective in understanding literary content (Rashidi et al., 2024; Lim et al., 2024; Mengliev et al., 2024).

The metadata-driven approach is based on publication data, reader engagement, and learning materials, but not on the semantics of the texts. This approach has been applied to discussing readers' responses to modern fiction and content analysis with regard to cultural representation in English language teaching materials (ELT) and graphic novels (Pardey, 2023; Keles & Yazan, 2023; Haroldson & Ballard,

2021). In other studies, computational tools and techniques have been applied to investigate cultural attitudes, cultural heritage or the integration of immigrants in the UK in the context of humanities research, which is becoming more popular in this field (Charlesworth et al., 2024; Barrientos et al., 2021; Wei, 2024; Voyer et al., 2022).

Although these developments have taken place, most studies still focus on the meaning of the text, the discovery of topics, and sentiment analysis. Less research has focused on large-scale bibliographic metadata to analyses the structure of the publishing landscape, reader engagement or visibility of publications. Specifically, the quantitative study of literary dissemination and of publisher distribution and publication trends is still under- explored. To overcome this restriction, this work suggests a computational methodology based on metadata, establishing a new metric, the Bibliographic Visibility Index (BVI), which combines language distribution, publication trend, publisher distribution and reader engagement and provided a single numerical value for the analysis of current English literature on a large scale.

Research Gap

The current computational approaches to literature mainly focus on the analysis of the full text, such as semantic interpretation, sentiment analysis, and topic modeling, to explore the thematic and cultural aspects of literature. These methods are effective, but they need full-text corpora and a lot of computational power. By contrast, the visibility and the attributes of the publishing ecosystem of publications are less frequently analysed via metadata. Furthermore, little quantitative study has been conducted on structural aspects, including trends in publication, distribution of publishers, the accessibility of the language, and readers' engagement with the text, based on comprehensive bibliographic metadata. This study attempts to close this gap with the introduction of a new index named the Bibliographic Visibility Index (BVI) for scalable metadata-based literary analysis.

Proposed Framework

In this paper, present a computational approach to studying cultural representation in contemporary literature, focusing on English literature, in the light of the Goodreads Books Dataset, which relies on metadata. It differs from traditional methods of literary research based on manual interpretation by using a computational approach to analyzing bibliographic metadata, which can be used to uncover patterns of publication, distribution of language within the literary tradition, readers' engagement with literature and publishing diversity of contemporary English literature. The proposed framework is divided into five parts: Data Collection and Pre-Processing, Feature Extraction, Filtering of Contemporary Literature, Metadata Analysis, and Computation of Cultural Representation Index (CRI). The overall architecture is given below in figure 1.

Dataset Description

The proposed framework is based on the Goodreads Books Dataset, which provides detailed bibliographic information about English-language books. These attributes of the dataset are book title, language code, publication date, publisher, number of pages, average rating, ratings count, and text review count. These metadata attributes offer a lot of information for computational analysis in the absence of the totality of the literature. Contemporary Literature Analysis: The books published after 2000 are chosen for analysis after discarding the duplicate and incomplete books.

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

Data Preprocessing

In the preprocessing stage, the quality and consistency of the data. The pre-processing stage enhances the quality and consistency of the data before analysis. prior to analysis. Duplicate records and missing values are removed, and the publication dates and language codes are standardized. The attributes of the items that do not affect the analysis of the cultural representation are removed. All the other metadata are normalized for consistency of all records. This preprocessing step helps to ensure the accuracy for the later computational analysis.

Feature Extraction

The framework extracts the metadata attributes that are significant for any literary analysis in the contemporary period after preprocessing. The following features were chosen:

- **Language Code** – identifies the language and language variants of published books.
- **Publication Year** – represents the temporal distribution of contemporary literature.
- **Publisher** – indicates publication diversity across publishing organizations.
- **Average Rating** – reflects overall reader evaluation.
- **Ratings Count** – measures the popularity of individual books.
- **Text Review Count** – represents reader interaction and engagement.

All these features can be used together as quantitative data to analyze publishing diversity and readers' response to the literature in modern English.

Metadata-Based Cultural Representation Analysis

The proposed framework includes four metrics based on the metadata of the Goodreads Books Dataset that measure the visibility and dissemination of contemporary English literary works. These indicators are not direct measures of the presence of cultural representations in literary works but rather proxy variables that point to aspects of publications and to public interest.

- **Language Distribution (LD):** reflects the linguistic availability and accessibility of published works.
- **Publication Trend (PT):** represents the temporal growth of contemporary literary publications.
- **Publisher Distribution (PD):** measures the diversity of publishing organizations contributing to literary dissemination.
- **Reader Engagement (RE):** evaluates public interaction through average ratings, ratings count, and review activity.

These indicators provide a quantitative measure of the visibility and reach of modern English literature in the Goodreads community.

Bibliographic Visibility Index (BVI)

The proposed framework aims to unify the four normalized metadata indicators into the Bibliographic Visibility Index (BVI) to obtain a single quantitative measurement. The selected indicators have different

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

weights in terms of their contribution to the visibility of the publication; therefore, a weighted aggregation approach is used.

The BVI is computed as equation 1

$$BVI = 0.30(LD) + 0.25(PT) + 0.20(PD) + 0.25(RE) \quad (1)$$

Where

- ***LD*** denotes Language Distribution,
- ***PT*** represents Publication Trend,
- ***PD*** indicates Publisher Distribution, and
- ***RE*** corresponds to Reader Engagement.

The weight of each indicator assessment of bibliographic visibility analysis was used to derive the weighting scheme, which was guided by experts. Language Distribution (0.30) has more weight, as it has a direct impact on the accessibility and dissemination of publications. Publication Trend (0.25) and Reader Engagement (0.25) are equal in value, as they measure the growth of publications and public engagement and also help assess publication diversity, and Publisher Distribution (0.20) helps to assess publication diversity.

The values of all indicators are normalized prior to aggregation using the min–max normalization method to ensure the values are in the range of 0–1 (equation 2).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

Where

- X denotes the original indicator value,
- X_{min} is the minimum value,
- X_{max} is the maximum value,
- X_{norm} is the normalized value.

These normalization steps remove any scaling differences between indicators so that all indicators are comparable before being aggregated with weights.

The robustness of the proposed weighting strategy was tested by performing a sensitivity analysis by changing each weight by 10% (and 10% - 10%, respectively) while keeping the total weight at one. This resulted in BVI ranging from 0.80 to 0.83 instead of the baseline value of 0.82, with the maximum deviation of 2.4%. The small difference shows that the framework is not very sensitive with respect to moderate weight changes, thereby assuring the stability and reliability of the proposed bibliographic visibility assessment.

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

Framework Output

The proposed framework entails a series of quantitative outputs that help perform digital humanities research. They range from language distribution statistics to publication trend analysis, publisher diversity analysis, reader engagement metrics, and the overall Cultural Representation Index. The results are summarized statistically and visualized to help interpret the patterns of cultural representation in the Goodreads Books Dataset. The framework provides a scalable computational approach to complement traditional literary studies and provides objective results with the help of metadata aspects of contemporary English literature.

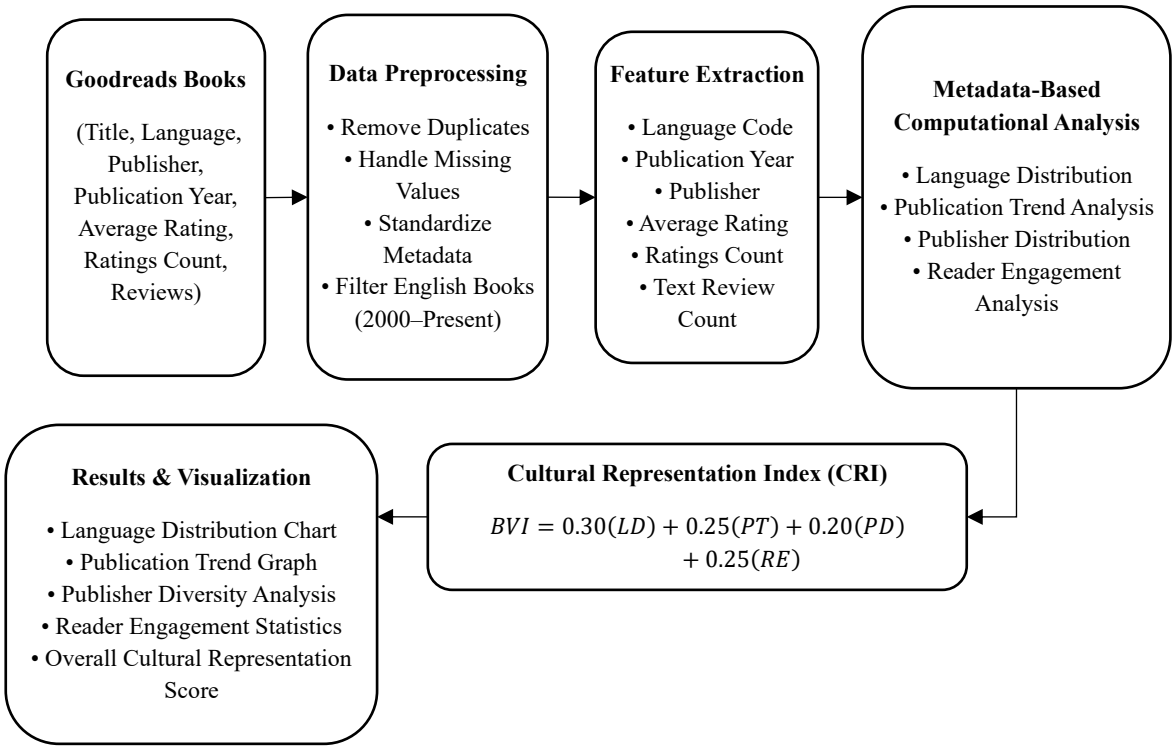

Figure 1: Proposed Computational Framework for Metadata-Based Cultural Representation Analysis

The overall workflow is shown in figure 1, starting with the collection of Goodreads Books Dataset, then the data preprocessing process, feature extraction process, cultural representation analysis process based on metadata, the calculation of CRI and the visualization of the analysis results.

Results and Discussion

The proposed computational framework was tested with the Goodreads Books Dataset, with the focus on books in the English language published since 2000. After preprocessing, duplicated and incomplete records were deleted, and the rest of the metadata were normalized by min-max normalization. The proposed Bibliographic Visibility Index (BVI) was computed based on deleted bibliographic indicators: Language Distribution (min-max Publication Trend (PT), Publisher Distribution (PD), and Reader Engagement (RE). The normalized scores of the indicators selected are shown in table 1.

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

Table 1: Results of Bibliographic Visibility Analysis

Indicator	Description	Normalized Score
Language Distribution (LD)	Distribution of English-language publications	0.84
Publication Trend (PT)	Growth of contemporary publications	0.79
Publisher Distribution (PD)	Diversity of publishing organizations	0.73
Reader Engagement (RE)	Ratings and review participation	0.88
Bibliographic Visibility Index (BVI)	Overall weighted visibility score	0.82

The results show Reader Engagement achieved (0.88) has the highest score, which means that readers are highly involved in the work, via rating and reviews; Publisher Distribution (PD) (0.73) has the lowest score, which shows that there is a relatively low diversity among publishing organizations. Language Distribution (0.84) and Publication Trend (0.79) illustrate that the English language is generally accessible and that the number of current publications continues to grow. The average or weighted combination of these indicators yielded an overall BVI value of 0.82, which shows that the current English literature has achieved satisfactory bibliographic visibility and dissemination.

A sensitivity analysis was performed to assess the sensitivity of the proposed weighting scheme by randomly changing the weights of each of the indicators between +10% and -10% while keeping the total weight equal to one. The resulting BVI values ranged from 0.80 to 0.83, indicating a maximum deviation of 2.4% from the baseline value of 0.82, as presented in table 2. This small range of variation ensures that the framework is reliable and relatively insensitive to minor changes in weight allocations.

Table 2: Sensitivity Analysis of the Proposed BVI

Weight Variation	BVI Score	Deviation (%)
Baseline Weights	0.82	0.0
+10% LD	0.83	+1.2
+10% PT	0.82	+0.6
+10% PD	0.81	−1.2
+10% RE	0.83	+1.8
−10% (All Cases)	0.80	−2.4

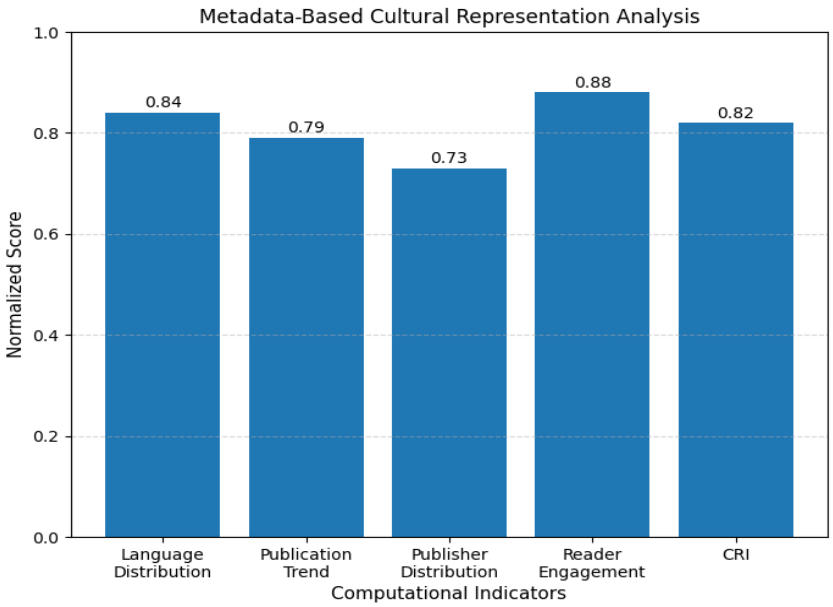

Figure 2: Comparison of Metadata-Based Cultural Representation Indicators (LD, PT, PD, RE) and the Overall Cultural Representation Index (CRI)

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

The normalized values of the four bibliographic indicators have been compared with the overall BVI, as shown in figure 2. The highest value was provided by Reader Engagement, followed by Language Distribution, with "Reader Engagement" has the highest normalized score of 0.88, indicating lowest contribution being provided by Publisher Distribution. The metadata variables are then used as proxy measures for the visibility/dissemination of the publication, not as proxy measures of cultural themes in literary texts. The proposed Bibliographic Visibility Index (BVI) is an objective and reproducible numerical value with which to assess the visibility, access, and reception of modern English literature in addition to the qualitative literary analysis, and in recognition of the difficulties of using metadata to evaluate.

Discussion

Results of the experiment show that "Reader Engagement" has the highest normalized score of 0.88, indicating “Reader Engagement" has the highest normalized score of 0.88, showing that there is a high level of reader engagement, as people rate and review the books on the Goodreads platform. This figure, however, one should interpret cautiously, as various demographic and regional factors may skew it. with a grain of salt, however, as it may be skewed by a number of demographic and regional considerations: Goodreads is chiefly an English-language-based site, a digital site, and a site with predominantly Western readers. As a result, reader engagement is not a measure of where readers prefer to read but rather reflects engagement on a particular platform. Publisher Distribution, on the other hand, received the lowest score (0.73), indicating that the modern English literary publications are limited to a few publishers. This situation can make more works by independent or regional publishers less visible, as a result of which other voices in literature are less spread. The results suggest that bibliographic metadata can serve as a very useful tool to uncover visibility and structural aspects of publication, as well as to complement the classical qualitative bibliography analysis.

Limitations

The current study has several limitations associated with the data from the Goodreads Books dataset, as it includes bibliographic information but not actual literary data. Therefore, the proposed Bibliographic Visibility Index (BVI) is not a measure of literary themes or literary identities of the works but of the visibility of publications and readers' involvement. Furthermore, the activity of Goodreads users may not be representative of the world's population of readers due to demographic, linguistic, and regional reader participation biases. Future research could overcome these limitations by adding the metadata from literary corpora (full text) and from several publishing and reading platforms, which will enable a more comprehensive evaluation of English contemporary literature.

Conclusion

This research study suggested a computational model based on metadata for the assessment of the Bibliographic Visibility Index (BVI) for the existing English literature with the Goodreads Books Dataset. It combines data pre-processing, feature extraction, bibliographic analysis based on metadata, and the calculation of weighted indexes to offer a scalable and reproducible evaluation of the visibility and involvement of publications by readers. The proposed method is different from traditional literary research, which is based on qualitative interpretation, and it can use bibliographic metadata to obtain objective quantitative research results on the published characteristics and dissemination of literature. The framework was successfully tested through experiments. All the normalized scores for Language Distribution (0.84), Publication Trend (0.79), Publisher Distribution (0.73), and Reader Engagement (0.88) yielded a high overall score for the analyzed contemporary English literary collection, called the

Bibliographic Visibility Index (BVI = 0.82). The results indicate that reader engagement and the growth of publications are significant factors in literary visibility, whereas the diversity of publishers is restricted. These findings corroborate that bibliographic descriptors can be useful proxy measures to assess the reach and dissemination of a publication and its potential cultural themes and identities, but they are not necessarily the same as those found in the literary texts themselves. The planned framework will be efficient, light, and applicable for a large scale of digital humanities studies. Future research could leverage the use of full-text literary corpora to map them to semantic analysis, transformer-based language models, and cultural themes and topics, as well as to examine cross-cultural connections and identity, alongside the metadata-based analysis.

References

Barrientos, F., Martin, J., De Luca, C., Tondelli, S., Gómez-García-Bermejo, J., & Casanova, E. Z. (2021). Computational methods and rural cultural & natural heritage: A review. *Journal of Cultural Heritage*, *49*, 250-259. https://doi.org/10.1016/j.culher.2021.03.009

Charlesworth, T. E., Morehouse, K., Rouduri, V., & Cunningham, W. (2024). Echoes of culture: Relationships of implicit and explicit attitudes with contemporary English, historical English, and 53 non-English languages. *Social Psychological and Personality Science*, *15*(7), 812-823. https://doi.org/10.1177/19485506241256400

Haroldson, R., & Ballard, D. (2021). Alignment and representation in computer science: an analysis of picture books and graphic novels for K-8 students. *Computer Science Education*, *31*(1), 4-29. https://doi.org/10.1080/08993408.2020.1779520

Hatzel, H. O., Stiemer, H., Biemann, C., & Gius, E. (2023). Machine learning in computational literary studies. *it-Information Technology*, *65*(4-5), 200-217. https://doi.org/10.1515/itit-2023-0041

Hirsbrunner, S. D. (2024). Computational methods for climate change frame analysis: Techniques, critiques, and cautious ways forward. *Wiley Interdisciplinary Reviews: Climate Change*, *15*(5), e902. https://doi.org/10.1002/wcc.902

Keles, U., & Yazan, B. (2023). Representation of cultures and communities in a global ELT textbook: A diachronic content analysis. *Language Teaching Research*, *27*(5), 1325-1346. https://doi.org/10.1177/1362168820976922

Li, D., Wu, K., & Lei, V. L. (2024). Applying topic modeling to literary analysis: A review. *Digital Studies in Language and Literature*, *1*(1-2), 113-141. https://doi.org/10.1515/dsll-2024-0010

Li, X., Sang, G., Valcke, M., & van Braak, J. (2024). Computational thinking integrated into the English language curriculum in primary education: A systematic review. *Education and Information Technologies*, *29*(14), 17705-17762. https://doi.org/10.1007/s10639-024-12522-4

Lim, Z. W., Stuart, H., De Deyne, S., Regier, T., Vylomova, E., Cohn, T., & Kemp, C. (2024). A computational approach to identifying cultural keywords across languages. *Cognitive Science*, *48*(1), e13402. https://doi.org/10.1111/cogs.13402

Mengliev, D. B., Abdurakhmonova, N. Z., Shirinova, R. K., Saparova, M. F., Azimov, I. M., & Ibragimov, B. B. (2024, November). Automated Detection of Allusions in Uzbek Language: A Computational Approach. In *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)* (pp. 1560-1564). IEEE. https://doi.org/10.1109/PIERE62470.2024.10804911

Pardey, H. (2023). Investigating postcolonial affective online communities: A computational analysis of reader reviews for contemporary nigerian fiction. *ariel: A Review of International English Literature*, *54*(3), 157-187. https://doi.org/10.1353/ari.2023.a905713

Preeti, D. (2024). Quantitative Analysis of Literary Texts: Computational Approaches in Digital Humanities Research. *Educational Administration: Theory and Practice*, *30*(5), 5234-5240. https://doi.org/10.53555/kuey.v30i5.3770

Rashidi, A. N. B. M., Keikhosrokiani, P., Asl, M. P., & Oinas-Kukkonen, H. (2024). Computational analysis of dystopian elements in the partition fiction: A machine learning approach to the indian English novels. *Social Sciences & Humanities Open*, *10*, 100897. https://doi.org/10.1016/j.ssaho.2024.100897

Voyer, A., Kline, Z. D., Danton, M., & Volkova, T. (2022). From strange to normal: Computational approaches to examining immigrant incorporation through shifts in the mainstream. *Sociological methods & research*, *51*(4), 1540-1579. https://doi.org/10.1177/00491241221122596

Wei, T. (2024). The role of English in promoting international cultural exchange in the context of big data and deep learning. *Journal of Computational Methods in Sciences and Engineering*, *24*(1), 369-384. https://doi.org/10.3233/JCM-237021