DOI: https://doi.org/10.5281/zenodo.14041674

A METHODOLOGICAL APPROACH TO LITERARY SELECTION FOR STUDENTS' CAPACITIES USING COMPUTATIONAL LINGUISTICS

Sattarova Sapura Beknazarovna
Computer Science department
Urgench State University,
sprsattarova@gmail.com

ABSTRACT

Selecting literary works that align with students' intellectual capacities is essential in education but often proves to be a complex task. Traditional text selection methods can be subjective and labor-intensive, resulting in mismatches between the reading difficulty of texts and students' abilities. This paper introduces a methodological approach that leverages computational linguistics (CL) and Natural Language Processing (NLP) techniques to enhance the selection process. By utilizing algorithms such as cosine similarity, TF-IDF, and Word2Vec, literary works are analyzed for linguistic complexity and matched to students based on their intellectual profiles. The study follows a three-step methodology: assessing students' intellectual capacity, analyzing linguistic features of texts, and employing machine learning algorithms for optimal selection. The results indicate that this approach increases student engagement and improves reading comprehension, making the text-matching process more efficient and accurate. Ultimately, the findings highlight the potential of CL methods to foster personalized learning by providing students with intellectually stimulating texts.

Keywords*: Literary selection, reading culture, natural language processing, text similarity algorithms, intellectual capacity, educational development, literary works, corpus linguistics.*

INTRODUCTION

Selecting literary works that are appropriate for students' intellectual capacities is a fundamental aspect of the education system. The right materials can stimulate cognitive development, enhance reading comprehension, and sustain student

engagement. In particular, literary texts serve as powerful tools for fostering critical thinking, emotional intelligence, and cultural awareness. However, the process of selecting suitable literary works is often subjective and labor-intensive for educators. It requires a nuanced understanding of the complexity of texts and the individual capabilities of each student. As a result, many students may encounter texts that are either too challenging or too simplistic, hindering their ability to engage meaningfully with literature. The traditional methods of literary selection typically rely on educators' experiences, intuition, and familiarity with the curriculum. While these methods may work well in some cases, they can lead to inconsistencies and biases, ultimately compromising the learning experience. Consequently, there is an urgent need for more systematic and data-driven approaches to text selection that take into account the diverse intellectual capacities of students.

Advancements in **Computational Linguistics (CL)** and **Natural Language Processing (NLP)** provide promising solutions to these challenges. By leveraging NLP techniques, educators can automate and enhance the text selection process. Methods such as text analysis, machine learning, and algorithms like **cosine similarity**, **TF-IDF**, and **Word2Vec** enable the creation of models that can evaluate literary works based on various linguistic and thematic characteristics. These algorithms can analyze a vast corpus of literary texts, assessing their complexity, vocabulary richness, and readability levels. For instance, cosine similarity measures the angle between vectors representing students' reading abilities and the linguistic features of texts, helping to identify the best matches. Meanwhile, TF-IDF evaluates the importance of specific terms within a text, facilitating the selection of materials that introduce students to new vocabulary and concepts at an appropriate level of challenge. Word2Vec can capture semantic relationships among words, allowing for a deeper understanding of the themes and contexts present in the literature.

This paper explores methods for improving the selection of literary works suitable for students' intellectual capacities through the application of computational linguistics techniques. The aim is to provide educators with a more efficient, accurate, and

personalized approach to literary selection. By employing these advanced methodologies, we can not only streamline the text selection process but also enhance the overall quality of literary education. The findings of this research have significant implications for educational practices, as they offer a pathway to better match students with texts that foster their academic growth and personal development. In an era where technology increasingly influences learning environments, integrating computational methods into literary selection can pave the way for more effective and engaging educational experiences.

LITERATURE REVIEW

Numerous studies by both local and international researchers have focused on developing educational resources and enhancing school textbooks. However, there has been little research on creating a corpus and selecting literary books aimed at evaluating the potential for increasing the knowledge of secondary school students in the Uzbek language.

The literary process involves not only an authorial subject who generates the artistic text, but another subject—the reader—who must perceive the text as artistic. Each subject is governed in part by the socially shared artistic norms of his time and place. And since "high" literature because of its written form is open to history literary works of art are perceived against the background of artistic norms different from those which gave birth to them [1]. The article [2] describes the goals and tasks of teaching Uzbek literature in Uzbek language classes in Russian-system schools, the problems of defining the educational content and ways to solve them. Also, selecting high-level artistic works from the new Uzbek literature, including referring to the works of famous writers, lexical and grammatical difficulties that arouse respect and respect in the representatives of Uzbek literature with their works, striving to get samples of their own works. This article [3] examines the theoretical and practical aspects of the Uzbek electronic corpus as a linguistic tool in computational linguistics. It details the corpus's functional capabilities, development experiences, and conceptual architecture, including its structure, linguistic annotation, metadata, and management. The platform

supports linguistic analysis in computational linguistics and NLP, with ongoing development of Uzbek language technology. The article [4] discusses the basic principles of corpus linguistics, a new field in applied Uzbek linguistics, as well as the process of designing and constructing corpus. Examples of achievements of world linguistics in the creation of corpus resources in the Uzbek language are given. Practical linguistic experience proves how to set up the first stage of corpus linguistics. The research [5] examines the automatic categorization of Ottoman texts using transliterated versions of the Latin alphabet. The authors use the method of applying Naive Bayes and Support vector machines to features such as word frequency and token length. The results of the statistical analysis will help to recommend style markers and methods for future work. This article [6] explores the concept of corpora as essential tools in linguistic analysis, defined as collections of texts representing a language (Tognini-Bonelli, 2001). It examines how corpus linguistics, leveraging computer technology, goes beyond mere methodology to integrate data gathering with theoretical insights, offering a new perspective on language.

The field of semantic analysis of any text is crucial in computational linguistics, where the focus is on improving the processing of text corpora through algorithms. In this research work [7], a keyword search algorithm for Kazakh language texts was developed, in which a reference dictionary was created using the Porter stemmer tool. The method used in the paper covers semantic affinity and vector marking for neural network learning. The advantage of this approach is that it automates text analysis and has potential applications in online student performance assessment. The novelty of the algorithm is its use of neural networks for Kazakh text and elimination of shortcomings in local computational linguistics research. This article examines the development issues of the Uzbek WordNet modeled after the Turkish WordNet [8]. In another article [9], a study on the detection of stop words is carried out using the "School Corpus" as an example, and the detection of stop words in Uzbek texts is summarized through automatic analysis of previous research. It is important to create an educational corpus suitable for the intellectual potential of primary school students to improve the quality

of education. Educational materials that are not appropriate for the age and intellectual potential of students can diminish their interest in learning.

In a different article [10], research is conducted on the creation of educational materials suitable for primary school students based on a corpus developed from 35 Uzbek primary school textbooks. In this paper [11], the educational corpus, which is a fragment of the Uzbek National Corpus taken from school textbooks and dictionaries, is studied. The paper details the factors, principles, models, and systems involved in the development of this corpus. In this study, the authors of [12] proposed a method for evaluating the list of stop words, tested in Uzbek but adapted to similar agglutinative languages, showing acceptable accuracy in automatic detection. Machine transliteration means transferring written words of one language from one alphabet to another, preserving their meaning and pronunciation. This article [13] presents a transliteration tool between three Uzbek scripts: Old Cyrillic, Official Latin, and New Latin. Annotated corpora development is very important in NLP, especially for resource-poor languages like Uzbek.

The paper [14] addresses automatic text summarization, focusing on two main approaches: summarizing with equivalent words and extracting key sentences. It introduces a text summarization model using the TF-IDF algorithm to automatically summarize texts in Uzbek. The model identifies unique words, calculates sentence weight, and utilizes the n-gram model to enhance summarization. The paper [15] tackles the scarcity of such resources by creating a novel POS and syntactic tag set for Uzbek. This article [16] examines the application of the Jaccard similarity method to the creation of appropriate reading lists for high school students. To achieve this, a corpus is created based on high-quality literature textbooks, and this corpus is compared with literary works. Books with the highest similarity results are recommended for reading. The problem was fully addressed using literature textbooks for 5th-11th grade students and literary works in the Uzbek language.

Natural Language Processing (NLP) technologies have revolutionized various domains by enabling machines to understand, interpret, and generate human language

data. However, for languages with limited digital resources and complex linguistic structures, such as Uzbek, NLP faces unique challenges. This paper [17] delves into the specific challenges encountered in NLP for Uzbek, focusing on lemmatization, stemming, sound recognition, and semantic analysis. In today's globalized world, providing quality education to students is one of the urgent tasks of the education system. This article [18] details how to create a model of the solar system using Python's graphical capabilities. This practice increases students' interest in graphic programming, and by visualizing the planets, students' interest and outlook on space science expand. In this article [19], the problem of creating a terminological dictionary for the chapter on the Delphi programming language, based on the textbook for academic lyceum students in Informatics and Information Technology, is considered. These articles [20, 21, 22] provide information about the importance of forming a culture of reading in students, creating the TF-IDF process for Uzbek literary texts, and what needs to be done in this regard. The paper [23] examines the development of a reading culture in students through the selection of literary works that match their intellectual abilities. Focusing on 5th-grade students, the study presents a literary corpus model built using Natural Language Processing (NLP) techniques to ensure appropriate text selection. The novelty lies in the creation of a corpus with statistical data on authors and literary types, which enables the application of text similarity algorithms. This approach enhances the selection process, contributing to the improvement of students' reading culture in school education.

METHODOLOGY

This study applied a range of CL and NLP methods to improve the selection of literary works suitable for students' intellectual capacities. The process involved assessing students' reading abilities, analyzing the linguistic complexity of literary works, and using machine learning algorithms to match students with appropriate texts.

1. Assessing Students' Intellectual Capacity

The first step was to evaluate the intellectual and cognitive capabilities of students using standardized reading comprehension tests. These tests were designed to measure:

- **Reading fluency**: The ability to read texts smoothly and at a comfortable pace.
- **Vocabulary knowledge**: The extent of a student's vocabulary, including understanding of complex or specialized terms.
- **Syntactic comprehension**: The ability to understand sentence structures and grammatical patterns.
- **Semantic processing**: The depth of understanding of text meaning and context.

The results of these assessments were used to create detailed intellectual profiles for each student, which captured their reading level, comprehension skills, and learning preferences.

2. Linguistic Analysis of Literary Texts

A corpus of literary texts was compiled for analysis, and the following linguistic features were extracted using NLP tools:

- **Syntactic Complexity**: The syntactic structure of sentences was analyzed, including sentence length, clause usage, and grammatical complexity. Longer, multi-clause sentences were considered more complex, while simpler sentences were deemed easier to comprehend.
- **Vocabulary Richness**: The variety of words in each text was evaluated using measures such as **Type-Token Ratio (TTR)** and word frequency distributions. Texts with a higher proportion of rare or specialized vocabulary were considered more advanced.
- **Readability Scores**: To determine the readability of each text, indices such as the **Flesch-Kincaid Grade Level** and the **Gunning Fog Index** were calculated. These scores indicate the educational level needed to comprehend a text easily.
- **Semantic Depth**: Word embeddings, generated using the **Word2Vec** model, captured the meaning of words based on their context in the text. This enabled the analysis of thematic complexity and the depth of ideas presented in each literary work.

3. Book Selection Algorithms

Once both student profiles and text features were prepared, various algorithms were employed to match students with suitable literary texts. These algorithms helped determine which texts best aligned with a student's intellectual profile, balancing text complexity and thematic depth with the student's comprehension skills.

- **Cosine Similarity**: This algorithm measures the similarity between a student's intellectual profile and the linguistic features of a text. Both were represented as vectors, and the cosine of the angle between these vectors was calculated to assess similarity. A higher cosine similarity score (closer to 1) indicated that the text was a good match for the student's reading ability. This approach ensured that the complexity of vocabulary and sentence structures in a text were aligned with the student's cognitive level.
- **TF-IDF (Term Frequency-Inverse Document Frequency)**: TF-IDF was used to weigh the importance of certain terms in the texts. Terms frequently used in a specific text but less common across other texts were given higher importance, helping identify key themes and vocabulary that may challenge or interest the student. This allowed the system to prioritize texts that introduced new or specialized terms appropriate for the student's learning progress.
- **Word2Vec Embeddings**: To analyze the semantic meaning and thematic complexity of the texts, the Word2Vec model was applied. This model represents words as vectors based on their context in the text, capturing subtle semantic relationships. Using Word2Vec, the system was able to assess how well the themes and content of a text aligned with the student's intellectual and interest profile.
- **Clustering Algorithms**: Clustering methods, such as **k-means clustering**, were used to group texts with similar linguistic features. This allowed for efficient organization of the texts into different difficulty levels and themes. Students were then matched to the cluster that best aligned with their intellectual profile, ensuring they received texts appropriate for their reading abilities and interests.

4. Evaluation Process and Feedback Loop

After the initial text selection using the algorithms mentioned above, educators reviewed the matches to ensure that the selected texts were suitable for each student. The feedback from teachers was then used to refine the algorithmic models, making the system more adaptive and responsive to the changing needs of students over time.

RESULTS

The application of CL and NLP methods to the selection of literary works yielded several important results:

1. **Improved Student Engagement**: Students who received texts tailored to their intellectual levels demonstrated greater engagement with the material. They were more likely to finish the reading assignments and reported higher levels of interest in the texts.
2. **Better Reading Comprehension**: Matching the complexity of the texts to students' reading abilities resulted in significant improvements in reading comprehension. Test scores showed that students had a deeper understanding of the texts, with increased ability to analyze and interpret literary themes.
3. **Efficient and Accurate Text Matching**: The use of algorithms like cosine similarity and TF-IDF significantly reduced the time needed to match students with suitable texts. Additionally, the accuracy of the matches was much higher than manual selection methods, ensuring that students consistently received texts that were challenging but within their intellectual reach.
4. **Scalability**: The system's ability to automatically analyze and classify large corpora of literary works allowed educators to scale the process to a larger number of students, making personalized learning more feasible even in larger classrooms.

CONCLUSION

This paper demonstrates that using computational linguistics and NLP methods significantly improves the process of selecting literary works that are suitable for students' intellectual capacities. By leveraging algorithms like cosine similarity, TF-IDF, and Word2Vec, the methodology ensures that students are matched with texts that are challenging, engaging, and intellectually stimulating.The integration of machine learning and feedback loops further enhances the adaptability of the system, allowing for continuous improvement based on real-world performance data. As CL and NLP technologies continue to evolve, the potential for even more refined and personalized educational tools will grow, ultimately benefiting students and educators alike.

Future work could explore deeper integrations of artificial intelligence (AI) and advanced NLP techniques, potentially incorporating adaptive learning systems that dynamically adjust text difficulty as students progress in their intellectual development.

REFERENCES:

1. Vodička, Felix. "The Concretization of the Literary Work". The Prague School: Selected Writings, 1929-1946, edited by Peter Steiner, New York, USA: University of Texas Press, 1982, pp. 103-134
2. Niyozmetova Roza Khasanovna. (2023). SELECTION OF WORKS BASED ON NEW UZBEK LITERATURE: PROBLEMS AND SOLUTIONS. World Bulletin of Social Sciences, 22, 170-173.
3. Abdurakhmonova N. et al. Uzbek electronic corpus as a tool for linguistic analysis //Компьютерная обработка тюркских языков. TURKLANG 2022. – 2022. – С. 231-240.
4. Vosiljonov, Azizbek. "Basic Theoretical Principles of Corpus Linguistics." Academicia Globe, vol. 3, no. 02, 2022, pp. 173-175, doi:10.17605/OSF.IO/36RWP.
5. Can E. F. et al. Automatic categorization of ottoman literary texts by poet and time period //Computer and Information Sciences II: 26th International Symposium on Computer and Information Sciences. – Springer London, 2012. – С. 51-57.
6. Say B. et al. Development of a corpus and a treebank for present-day written Turkish //Proceedings of the eleventh international conference of Turkish linguistics. – Eastern Mediterranean University, 2002. – С. 183-192
7. Akanova, A. et al. "Development of the Algorithm of Keyword Search in the Kazakh Language Text Corpus." 2019.
8. A. Madatov, D. J. Khujamov, and B. R. Boltayev, “Creating of the Uzbek WordNet based on Turkish WordNet,” in AIP Conference Proceedings, 2022. doi: 10.1063/5.0089532.
9. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” Data Brief, vol. 43, 2022, doi: 10.1016/j.dib.2022.108351
10. Madatov, K. A., and Sattarova, S. "Creation of a Corpus for Determining the Intellectual Potential of Primary School Students." *2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM)*, Altai, Russian Federation, 2024, pp. 2420-2423. doi:10.1109/EDM61683.2024.10615103
11. Abjalova, M., Adalı, E., Iskandarov, O. "Educational Corpus of the Uzbek Language and Its Opportunities." *2023 8th International Conference on Computer Science and Engineering (UBMK)*, IEEE, 2023, pp. 590-594.
12. Madatov, K., Bekchanov, S., and Vičič, J. "Accuracy of the Uzbek Stop Words Detection: A Case Study on ‘School Corpus.’" *CEUR Workshop Proceedings*, 2022.

13. Salaev, U., Kuriyozov, E., and Gómez-Rodríguez, C. "A Machine Transliteration Tool Between Uzbek Alphabets." CEUR Workshop Proceedings, vol. 3315, 2022, pp. 42–50.
14. Madatov, K. A., and Bekchanov, S. K. "The Algorithm of Uzbek Text Summarizer." *2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM)*, Altai, Russian Federation, 2024, pp. 2430-2433. doi:10.1109/EDM61683.2024.10615191
15. Sharipov, M., Mattiev, J., Sobirov, J., and Baltayev, R. "Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language." *CEUR Workshop Proceedings*, vol. 3315, 2022, pp. 93–98.
16. Мадатов Х., Саттарова С. Using the Jaccard similarity method for recommendation system of books //Общество и инновации. – 2024. – Т. 5. – №. 1. – С. 59-69
17. Sattarova S. Advancing natural language processing in uzbek: challenges and solutions: Advancing natural language processing in uzbek: challenges and solutions //MODERN PROBLEMS AND PROSPECTS OF APPLIED MATHEMATICS. – 2024. – Т. 1. – №. 01.
18. Khodjinazarovna B. F., Kamaliddinovich S. A., Beknazarovna S. S. Visualizing the solar system using python and its importance in education //International journal of advanced research in education, technology and management. – 2023. – Т. 2. – №. 6.
19. Sattarova S. B., Bekchanova F. X., Shermetov A. K. Terminologik lug‘at yaratish texnologiyasi va uning ta’lim tizimidagi ahamiyati //Academic research in educational sciences. – 2023. – Т. 4. – №. 5. – С. 422-434.
20. Beknazarovna S. THE IMPORTANCE OF ELECTRONIC CATALOGS IN THE DEVELOPMENT OF READING CULTURE //ILM SARCHASHMALARI (2). – 2024. – С. 193-197.
21. Madatov X. A., Sattarova S. B. YOSHLARDA KITOBXONLIK MADANIYATINI RIVOJLANTIRISHNING ASOSIY OMILLARI //Educational Research in Universal Sciences. – 2023. – Т. 2. – №. 17. – С. 1017-1025
22. Madatov, Khabibulla, and Sapura Sattarova. "Vectorization of Uzbek Texts Using the TF-IDF Vectorizer Method." O‘zMU XABARLARI, vol. 11, 2023, pp. 177-180. ISSN 2181-7324.
23. Sattarova, S. B. "Developing an Uzbek Literature Corpus: Enhancing Literary Selection for 5th-Grade Education." *Science and Innovation International Scientific Journal*, vol. 3, no. 9, 2024, pp. 4-13.