Integration of Serbian Language Version of Wikipedia into Educational Systems and the Advancement of Language Technologies

Scientific paper DOI: 10.18485/judig.2025.1.ch26

Nebojša Ratković[1], 0000-0001-5983-9642

Abstract

The Serbian-language version of Wikipedia is one of the most comprehensive and accessible knowledge resources in the digital age. Its integration into educational systems can modernize teaching, enhance digital and language literacy, and support the development of language technologies. This paper explores both the opportunities and challenges of using Wikipedia in formal education, with particular emphasis on the creation of digital corpora and their role in advancing natural language processing and related technologies. It also discusses strategic steps for embedding Wikipedia into curricula, including teacher training, the adaptation of teaching programs, and the development of digital platforms.

The first part of the paper examines Wikipedia as a teaching tool and its capacity to promote critical thinking, research skills, and digital literacy. Examples from Serbian schools and universities illustrate how Wikipedia-based assignments improve student motivation and contribute to the availability of quality Serbian-language content online.

The second part focuses on the technical dimension, demonstrating how educational contributions enrich digital corpora, enabling the development of tools such as machine translation, grammar correction, and speech recognition. The systematic use of Wikipedia in education thus strengthens the digital infrastructure of the Serbian language and ensures its relevance in the evolving technological landscape.

Keywords: Wikipedia, education, digital corpora, language technologies, Serbian language, curriculum development

[1] Wikimedia Serbia, Education Program Manager, E-mail: nebojsa.ratkovic@vikimedija.org

1. Introduction

In the contemporary digital era, education is inseparably linked to the availability and quality of online resources. Wikipedia has become one of the most frequently consulted sources of knowledge worldwide (Pavlović et al., 2017), and its Serbian-language version represents a valuable repository of information that is widely accessible, constantly updated, and created collaboratively by a large community of volunteers. While the use of Wikipedia in informal learning has been well documented (Soler-Adillon et al., 2018), its systematic integration into formal educational systems remains insufficiently explored in Serbia. Recognizing its potential to modernize teaching practices, improve digital and language literacy, and stimulate the advancement of language technologies, this paper argues for a structured approach to incorporating Wikipedia into schools and universities.

The integration of Serbian Wikipedia into education aligns with global trends that emphasize open educational resources, participatory learning, and the democratization of knowledge (Stakić et al., 2021). For students, engaging with Wikipedia offers an opportunity to develop critical thinking, research skills, and collaborative competencies, all of which are increasingly necessary in knowledge-based societies. For teachers, it provides a dynamic tool that can enrich traditional teaching materials, facilitate innovative pedagogical methods, and contribute to the formation of a more interactive classroom environment. At the same time, Wikipedia's digital content can serve as a foundation for the development of linguistic corpora that are essential for natural language processing, machine translation, and other language technologies (Marovac et al., 2023; Leipzig Corpora Collection, 2021).

Despite these opportunities, the use of Wikipedia in Serbian classrooms is still perceived with skepticism, largely due to concerns about the accuracy of content, insufficient institutional support, and a lack of teacher training (Pavlović et al., 2017). This paper therefore seeks to examine both the opportunities and challenges of such integration. It will present examples of successful initiatives, highlight technical aspects related to the creation of digital corpora, and propose strategic recommendations for educational policy. By analyzing the dual role of Wikipedia as both an educational tool and a linguistic resource, the paper aims to demonstrate its relevance for modern education and its broader contribution to the digital advancement of the Serbian language (Krstev & Stanković, 2022).

2. Wikipedia as an Educational Tool

The Serbian-language version of Wikipedia has gradually moved beyond being merely a popular online reference and has become an increasingly valuable educational tool. Unlike static textbooks, Wikipedia is dynamic, continuously updated, and enriched by a diverse community of contributors. Its adaptability and openness make it especially suitable for integration into classrooms, where it can serve not only as a source of information but also as a platform for skill development, collaborative learning, and active knowledge creation (Pavlović et al., 2017).

One of the most significant benefits of using Wikipedia in education is its potential to enhance students' critical thinking and research competencies. Articles on Wikipedia are expected to cite reliable sources, and this requirement exposes students to the process of verifying information and evaluating the credibility of references. Instead of passively consuming information, they learn to question, analyze, and compare sources. Through editing, students also gain firsthand experience with academic debate and peer review, since their contributions are publicly visible and subject to scrutiny by the community. This visibility encourages responsibility, accuracy, and a higher level of academic engagement (Soler-Adillon et al., 2018).

Practical examples from Serbia strongly support these claims. Many higher education institutions have partnered with Wikimedia Serbia to integrate Wikipedia editing into their curricula. At the Faculty of Organizational Sciences in Belgrade, for instance, students were assigned to create or expand articles relevant to their fields of study. This long-term cooperation at the Faculty of Organizational Sciences has resulted in the creation and improvement of several hundred articles, while engaging several hundred students across different courses over the years. Research shows that students who participate in such assignments demonstrate stronger motivation, higher levels of engagement, and greater awareness of academic integrity (Stakić et al., 2021). Similarly, workshops at the Faculty of Philology, carried out through long-term collaboration, engaged several hundred students across multiple courses and semesters. Their work resulted in the creation and improvement of several hundred articles, significantly enriching Serbian Wikipedia with high-quality linguistic content and advancing students' academic writing skills. Over the past decade (2015 to 2025), educational collaborations between Wikimedia Serbia and Serbian universities have included around 160 project cycles, engaging approximately 4,800 students who created or improved several thousand Wikipedia articles. These long-term activities have substantially contributed to the expansion of high quality Serbian language content and to the enhancement of digital and academic competencies among participants.

Additional initiatives at the Faculty of Economics in Belgrade and the Faculty of Contemporary Arts demonstrated how disciplines outside philology could benefit from such approaches. Students in economics contributed articles related to business, finance, and management, thereby expanding the availability of specialized content in Serbian. Students in the arts, on the other hand, engaged with cultural and artistic topics, where visual materials and multimedia enriched their contributions. In both cases, participants expressed pride in creating resources with lasting public value, noting that their academic work transcended the boundaries of the classroom and reached a broad online audience (Pavlović et al., 2017).

In secondary schools, pilot workshops illustrated how Wikipedia can foster digital and media literacy. Guided by trained facilitators, students explored the fundamentals of editing, citation, and neutrality. Teachers observed that such assignments helped pupils understand the importance of perspective and bias in historical and social narratives.

Another important dimension is teacher training. Experience from accredited professional seminars organized by Wikimedia Serbia has shown that educators are often initially skeptical about Wikipedia. Concerns usually revolve around reliability and the risk of plagiarism. However, once teachers are introduced to methods for using Wikipedia constructively, such as incorporating it into project-based learning, assigning article editing as coursework, or using talk pages to practice academic debate, they tend to recognize its pedagogical value (Pavlović et al., 2017). During the period 2015–2025, several dozen teachers participated in accredited seminars on integrating Wikipedia into education. While not all of them continued the practice, a number of individual teachers chose to incorporate Wikipedia-based assignments into their classrooms, thereby sustaining the collaboration and extending its impact. Teachers emphasized that Wikipedia-based activities encouraged student independence, creativity, and responsibility in ways traditional assignments often did not. Moreover, the open and collaborative nature of the platform aligned well with the principles of inclusive and participatory education promoted in Serbia's educational reforms (Krstev & Stanković, 2022).

These practices reveal that Wikipedia's role in education goes far beyond functioning as a simple reference. It becomes a didactic instrument that transforms learning into an interactive, student-centered process. Tasks based on Wikipedia stimulate collaboration, creativity, and digital responsibility while also enriching the body of Serbian-language knowledge available online. In doing so, educational institutions contribute to the visibility and vitality of the Serbian language in the digital sphere, ensuring that it remains relevant and accessible in an era increasingly dominated by global digital platforms (Marovac et al., 2023).

The integration of Wikipedia into formal education in Serbia requires a structured and strategic approach that goes beyond sporadic initiatives. While isolated examples from universities and secondary schools demonstrate the platform's potential, a more systematic model is necessary for long-term sustainability. Such integration involves aligning Wikipedia-related activities with curricula, providing adequate teacher training, and creating supportive digital infrastructures. These steps are crucial for transforming Wikipedia from an occasional classroom tool into an established component of Serbia's educational framework (Soler-Adillon et al., 2018).

2.1. Teacher Training, Professional Development and Curriculum Alignment

Teachers represent the key actors in integrating Wikipedia into schools and universities. Without adequate preparation, many remain hesitant to incorporate Wikipedia into their teaching practice, often citing concerns about accuracy, plagiarism, or assessment criteria (Pavlović et al., 2017). To address these issues, Wikimedia Serbia, in cooperation with educational institutions, has organized accredited professional development seminars. The introduction of accreditation was particularly significant, as it ensured that teachers' engagement with Wikipedia was formally recognized as part of their professional development. This alignment with national standards encouraged broader adoption, particularly among younger teachers eager to experiment with innovative methods. Teachers who later implemented Wikipedia-based assignments in their schools reported that students responded positively, perceiving such projects as more meaningful and motivating compared to traditional written essays (Stakić et al., 2021).

A systematic integration of Wikipedia requires its inclusion in official curricula. At the university level, this can be achieved by embedding Wikipedia editing projects into existing subjects such as language studies, history, sociology, or computer science. Several faculties have already experimented with such practices.

At the secondary school level, the potential lies in aligning Wikipedia activities with competencies outlined in Serbia's educational strategy, such as digital literacy, critical thinking, and teamwork. Students not only learned to evaluate multiple perspectives but also became active participants in knowledge production (Pavlović et al., 2017). By positioning Wikipedia-based tasks within the framework of project-based learning, schools can achieve curriculum goals while simultaneously contributing to the public pool of knowledge in Serbian.

However, curriculum alignment is still at an early stage. While individual teachers and faculties have adopted these methods, official documents such as the national Strategy for Education have yet to explicitly recognize Wikipedia as an educational resource. A future step would involve including Wikipedia and similar open educational resources in policy guidelines, ensuring consistency across schools and regions (Krstev & Stanković, 2022). Without this formal recognition, the use of Wikipedia risks remaining fragmented and dependent on the enthusiasm of individual educators rather than being embedded in systemic practice.

Despite the opportunities, integrating Wikipedia into the educational system faces challenges. The first obstacle is skepticism among educators, many of whom continue to perceive Wikipedia as unreliable (Pavlović et al., 2017). Although research and practice show that Wikipedia articles are generally as accurate as traditional encyclopedias, these perceptions persist and hinder wider adoption. Another barrier lies in the lack of official policy support—without explicit recommendations in strategic documents, teachers often fear that incorporating Wikipedia might not be formally recognized as valid educational practice (Krstev & Stanković, 2022).

Technical challenges also remain. Many schools, especially in rural areas, still lack adequate digital infrastructure, and teachers are not always trained to guide students in advanced digital assignments. Furthermore, issues such as plagiarism and improper citation require continuous monitoring and guidance. Addressing these challenges requires a multi-layered approach that combines teacher support, policy frameworks, and technological investment (Marovac et al., 2023).

3. Wikipedia and Language Technologies

The development of language technologies relies heavily on the availability of high-quality digital corpora. In this respect, Wikipedia represents an invaluable resource. Its openly licensed content, standardized structure, and wide thematic coverage make it an ideal foundation for building linguistic corpora used in natural language processing (NLP), machine translation, speech recognition, and other language technologies.

For the Serbian language, which is relatively underrepresented in global digital resources compared to English or major European languages, integrating Wikipedia into the construction of corpora is not only useful but essential for ensuring its digital presence and competitiveness (Krstev & Stanković, 2022).

Wikipedia articles are particularly suitable for corpus creation because they adhere to consistent formatting, rely on references, and are regularly updated by a community of editors. Each article provides structured text, metadata, and hyperlinks that facilitate linguistic analysis. Researchers can extract balanced datasets of different genres, such as history, science, culture, and technology, within a single platform. Unlike news portals, which often focus on short-term topics, Wikipedia provides encyclopedic content that is more stable and comprehensive.

For Serbian, a corpus based on Wikipedia articles currently contains over 26 million tokens and 1.7 million sentences, compiled in the Leipzig Corpora Collection (Leipzig Corpora Collection, 2021). In addition, the Serbian web corpus includes approximately 894 million tokens, demonstrating the scale of resources available for computational processing (UBC Library, 2025). Academic institutions in Serbia have already experimented with using Wikipedia as a basis for compiling corpora aimed at developing spell-checkers, grammar correction tools, and lemmatization resources (Marovac et al., 2023). However, systematic efforts remain limited, and much of the potential is yet to be realized.

The educational use of Wikipedia directly contributes to the enrichment of such corpora. When students and teachers create new articles or expand existing ones, they produce large amounts of standardized text that is publicly available and digitally accessible. Each contribution strengthens the corpus and provides raw material for further linguistic analysis.

This process also enhances the representativeness of corpora. By integrating educational contributions, subject areas such as economics, sociology, and the arts—traditionally less covered in standard corpora, become part of the linguistic record. The impact is therefore twofold: students gain valuable academic and digital skills (Stakić et al., 2021), and language technologies benefit from richer, more diverse resources.

Although Wikipedia provides a rich foundation, several linguistic challenges complicate the creation of robust Serbian corpora. The Serbian language is morphologically complex, with rich inflection and case systems, which requires advanced annotation and lemmatization methods (Marovac et al., 2023). Furthermore, Serbian is digraphic, written in both Cyrillic and Latin alphabets, which necessitates normalization across scripts for effective computational processing. These issues demand additional preprocessing steps, such as script unification, morphological tagging, and syntactic parsing.

Another challenge lies in domain coverage. While Wikipedia articles are encyclopedic, certain fields (e.g., law, technical sciences) remain underrepresented, leading to gaps in corpus coverage. Supplementing Wikipedia with additional open sources, while still ensuring linguistic consistency, remains a task for future corpus-building projects (Krstev & Stanković, 2022).

Corpora derived from Wikipedia are already used in several technological applications relevant to Serbian:

- Natural Language Processing (NLP): Tools for tokenization, part-of-speech tagging, and lemmatization rely on corpora containing millions of words. The inclusion of Wikipedia texts allows these tools to be tested and improved on standardized material (Marovac et al., 2023).
- Machine Translation: Open datasets based on Wikipedia articles in multiple languages have been successfully used to improve translation engines such as Google Translate or DeepL. For Serbian, bilingual corpora derived from Wikipedia parallel articles with English and other South Slavic languages already contain aligned sentence pairs (Leipzig Corpora Collection, 2021).
- Lexical Databases and Semantic Networks: Dictionaries and ontologies can be enriched through Wikipedia entries, especially categories, interlinks, and metadata, which provide structured connections between concepts (Krstev & Stanković, 2022).

Each of these applications directly benefits from educational initiatives where students contribute content. The combination of pedagogy and technology thus produces a multiplier effect: education enhances language technologies, while language technologies in turn support more effective digital education.

4. International Comparisons and Societal and Cultural Benefits

Other countries provide valuable examples of how Wikipedia can be systematically used for corpus building and technology development. In Finland, Wikipedia content was incorporated into the national corpus project for Finnish, enabling progress in machine translation and AI applications for a language with a relatively small speaker base. In the Czech Republic, the Czech National Corpus uses Wikipedia extensively to enrich linguistic datasets, which has significantly improved their NLP tools and language-learning platforms. Similarly, Sweden integrated Swedish Wikipedia texts into large-scale corpora used in both academic research and commercial products (Soler-Adillon et al., 2018).

Beyond national initiatives, several international projects highlight the role of Wikipedia in cross-linguistic data development. For example, the Wikimedia Foundation's Content Translation Tool enables editors to translate articles between languages, simultaneously generating parallel corpora that can be used for machine translation training. European projects such as ELRC-SHARE and Europeana emphasize the digitization of cultural and linguistic heritage and often rely on Wikipedia as a multilingual source (Krstev & Stanković, 2022).

The role of Wikipedia in language technologies extends beyond technical applications. By enriching Serbian-language corpora, it contributes to preserving cultural heritage, supporting digital inclusion, and ensuring that Serbian remains a fully functional language in the digital age. High-quality NLP tools based on Wikipedia texts can facilitate inclusive education for students with disabilities (e.g., through text-to-speech applications), promote cross-cultural communication via improved translation tools, and enhance the accessibility of digital archives. In this way, Wikipedia strengthens not only the technological infrastructure of the language but also its cultural and social relevance (Marovac et al., 2023).

5. Discussion and Conclusion

The integration of Wikipedia into Serbian education and its role in advancing language technologies illustrate the dual potential of digital platforms: they serve both as tools for teaching and as resources for technological development. Research and educational practice consistently show that students perceive Wikipedia as a motivating and meaningful learning environment, particularly when their work becomes visible to the wider public. Teachers, although often initially skeptical due to concerns about reliability and plagiarism, frequently acknowledge its value once they receive appropriate training and methodological guidance.

From an educational perspective, integrating Wikipedia into curricula has proven to enhance critical thinking, digital literacy, and collaborative learning. It also contributes to the expansion of high-quality Serbian-language content online, aligning with global educational trends that emphasize open access, participatory learning, and interdisciplinary collaboration. However, these benefits can be fully realized only if institutions formally recognize Wikipedia as an educational resource and integrate it into official strategies and programs. Without such systemic support, the use of Wikipedia risks remaining fragmented and dependent on the initiative of individual educators.

From a technological perspective, Wikipedia represents a substantial linguistic and digital resource. Texts generated through educational projects contribute to the growth of digital corpora used for developing natural language processing tools, machine translation systems, and grammar and spell-checking applications. At the same time, challenges such as the morphological complexity of Serbian, the coexistence of Cyrillic and Latin scripts, and uneven thematic coverage call for continued collaboration between educators, linguists, and technology experts.

The broader implication of these findings is that education and language technologies should be viewed as mutually reinforcing domains. When embedded in formal education, Wikipedia functions as a catalyst for both pedagogical innovation and technological progress. Conversely, the development of advanced digital tools enhances its educational value, for instance through improved translation, text-to-speech, and search functionalities. This circular relationship ensures that each educational activity contributes to linguistic infrastructure, while every technological advancement further strengthens the educational process.

Looking ahead, several steps appear essential for maximizing the benefits of this integration. Policymakers should acknowledge Wikipedia as a legitimate and valuable educational tool within strategic documents. Teacher training programs need to be continuously expanded, ensuring that educators possess both technical and methodological competence. National corpus initiatives should systematically incorporate Wikipedia-based materials, supported by open access principles and inter-institutional cooperation. Furthermore, Serbia's participation in international projects that connect smaller languages to global digital infrastructures will ensure the linguistic equality and sustainability of the Serbian language in the digital era.

In conclusion, integrating the Serbian-language Wikipedia into education is not merely an innovation in pedagogy but a strategic investment in the language's digital future. By linking teaching and technology, Wikipedia provides a sustainable model for modernizing classrooms, empowering students, and enhancing language resources. Its dual role, as both a pedagogical tool and a linguistic resource, positions it as a cornerstone of twenty-first-century education in Serbia. Systematic implementation will ensure that the Serbian language remains vibrant, functional, and technologically supported within the global digital environment.

References

[1] Krstev, C., & Stanković, R. (2022). *Report on the Serbian Language* (Deliverable D1.35, European Language Equality project). European Language Equality. https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_35__Language_Report_Serbian_.pdf

[2] Leipzig Corpora Collection. (2021). *Serbian Wikipedia corpus based on material from 2021* [Dataset]. Leipzig Corpora Collection. https://corpora.wortschatz-leipzig.de/en?corpusId=srp_wikipedia_2021

[3] Marovac, U. A., Avdić, A. R., & Milošević, N. Lj. (2023). *A Survey of Resources and Methods for Natural Language Processing of Serbian Language*. arXiv preprint. https://doi.org/10.48550/arXiv.2304.05468

[4] Pavlović, D., Petrović, Z. S., & Mamutić, A. (2017). Wikipedia – From the Popular Source of Information to the Pedagogical Tool. *International Journal of Social Sciences & Educational Studies, 4*(3), 36–48. https://ijsses.tiu.edu.iq/index.php/ijsses/article/view/406

[5] Soler-Adillon, J., Pavlović, D., & Freixa, P. (2018). Wikipedia in Higher Education: Changes in Perceived Value through Content Contribution. *Comunicar, 26*(54), 39–48. https://doi.org/10.3916/C54-2018-04

[6] Stakić, Đ., Tasić, M., Stanković, M., & Bogdanović, M. (2021). Students' Attitudes Towards the Use of Wikipedia: A Teaching Tool and a Way to Modernize Teaching. *Área Abierta, 21*(2), 309–325. https://www.researchgate.net/publication/353110388_Students'_Attitudes_Towards_the_Use_of_Wikipedia_A_Teaching_Tool_and_a_Way_to_Modernize_Teaching

[7] UBC Library. (2025). *Serbian web corpus (srWaC): 894 million tokens* [Dataset]. University of British Columbia. https://guides.library.ubc.ca/c.php?g=306932&p=2051153

Интеграција Википедије на српском језику у образовни систем и унапређење језичких технологија

Небојша Ратковић

Сажетак

Интеграција Википедије на српском језику у образовни систем представља иновативан приступ који истовремено унапређује наставне методе и доприноси развоју језичких технологија. Рад анализира потенцијале и изазове овакве интеграције у три области: улогу Википедије као наставног алата, њено укључивање у наставне планове и програме, и значај за изградњу дигиталних корпуса и развој технологија обраде природног језика.

Први део рада показује да задаци засновани на Википедији подстичу критичко мишљење, развијају истраживачке вештине и дигиталну писменост ученика и студената. Примери из високог и средњег образовања у Србији потврђују да уређивање чланака повећава мотивацију ученика и њихов осећај одговорности, а истовремено обогаћује јавно доступан садржај на српском језику.

Други део наглашава важност стратешког приступа – обуке наставника, укључивања у наставне програме и развоја дигиталних платформи. Без институционалне подршке и званичног препознавања, примена Википедије у настави остаје ограничена на ентузијазам појединаца.

Трећи део указује на техничке аспекте и значај Википедије за изградњу корпуса неопходних за машинско превођење, алате за исправљање граматике, као и технологије препознавања и синтезе говора. Кроз образовне активности генерише се сталан прилив дигиталног текста, што обезбеђује да српски језик остане виталан и равноправан у дигиталном добу.

Кључне речи*:* Википедија, образовање, дигитални корпуси, језичке технологије, српски језик, наставни план и програм