Textometric and comparative analysis of the bilingual corpus of the journal *Infotheca*

UDC 811.163.41'322.2:811.111'322.2
DOI 10.18485/infotheca.2026.26.1.2

ABSTRACT: This paper presents a textometric and comparative analysis of the bilingual corpus of the scientific journal *Infotheca*, which comprises parallel texts in Serbian and English. The corpus was extracted from the digital library *Bibliša*, with metadata linked to Wikidata. Particular attention is given to the analysis of the Serbian and English subcorpora using textometric methods, including frequency analysis, keyword analysis, collocation analysis, and topic modeling. Following the individual analyses, a comparative analysis was conducted with the aim of identifying differences and similarities in the lexical characteristics of the two languages. The results show that, although the texts are translation equivalents, there are notable differences in term distribution, indicating the influence of linguistic and scientific conventions. The paper contributes to the development of methodologies for analyzing bilingual corpora in the fields of digital humanities and language technologies.

KEYWORDS: bilingual corpus, textometrics, Infotheca, parallel texts, comparative analysis, Wikidata, digital libraries.

PAPER SUBMITTED: 04 September 2020
PAPER ACCEPTED: 25 November 2020

Ranka Stanković
ORCID 0000-0001-5123-6273
ranka.stankovic@rgf.bg.ac.rs

University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

1 Introduction

Bilingual corpora play a significant role in contemporary linguistics and digital humanities, as they enable systematic comparison of linguistic structures, vocabulary, and language patterns across different languages. As parallel collections of texts, they represent a valuable resource for studying translation

strategies, terminology, and linguistic variation. In the context of language technologies, bilingual corpora serve as a foundation for the development and evaluation of tools such as machine translation, automatic annotation, and language modeling. At the same time, in digital humanities research, they facilitate a deeper understanding of cultural and scientific developments through multilingual sources, bridging quantitative methods with interpretative approaches.

The journal *Infotheca* is a relevant and reliable source of data for analysis in the fields of library science, information science, and language technologies. As a bilingual scientific journal, it includes papers in both Serbian and English, enabling the construction of a parallel corpus suitable for comparative research. Structured metadata and the availability of full-text articles further enhance its usability in textometric and corpus-based analyses, making it a valuable resource within digital humanities research.

Stanković et al. (2012) presented the *Bibliša* system, a tool designed to enhance the search of large collections of TMX (Translation Memory eXchange) documents generated through the parallelization of bilingual texts from multilingual digital libraries. Its functionality was tested on a collection of TMX documents derived from bilingual articles of the journal *Infotheca*. The continuous development of the system has introduced new capabilities, transforming the platform into a repository of parallel texts from various collections. It should be emphasized that *Infotheca* was the first collection integrated into Bibliša and remains the one that is regularly updated with new content. Specifically, upon the publication of an issue in its online version on the OJS platform[1], the articles are parallelized, enriched with metadata, and subsequently integrated into Bibliša and the Wikidata knowledge base (Stanković and Davidović 2021; Andonovski 2026).

The textometric approach has long been applied as an effective method for corpus analysis across various fields of the humanities and social sciences. By combining lexicometric and statistical methods with advanced corpus technologies, textometry enables non-linear, both quantitative and qualitative analysis of digital textual resources.

The TXM software platform (Heiden 2010) provides a comprehensive environment for textometric analysis, supporting a wide range of statistical computations and graphical visualizations of results[2] (Pincemin, Heiden, and Mazuet 2022).

1. http://infoteka.bg.ac.rs/ojs/
2. TXM - textometric platform

The SrpELTeC corpus has served as a testbed for various textometric studies (Jaćimović 2019; Krstev 2021; Stanković, Krstev, and Vitas 2024), and, following successful applications to literary texts, the focus has shifted toward the analysis of scientific texts. In this paper, the potential of the textometric approach is illustrated within the TXM environment through the analysis of the bilingual corpus of the journal *Infotheca*. Particular attention is devoted to the extraction and comparison of the Serbian and English subcorpora, as well as to the application of various textometric methods, including frequency analysis, specificity analysis, collocation analysis, and temporal progression. The obtained results are further interpreted through appropriate visualizations, enabling a clearer understanding of developmental trends and the thematic structure of the corpus.

The remainder of the paper is organized as follows: Section 2 presents the corpus development, while Section 3 provides the frequency analysis of nouns, verbs, adjectives, and multiword terms of the adjective-noun type. The textometric analysis is presented in Section 4, which includes the analysis of progression and topic specificity across periods. Topic modeling is discussed in Section 5, followed by the discussion in Section 6 and the conclusion in Section 7.

2 Corpus Development

Digital objects within the *Bibliša* system are described using structured metadata tailored to the representation of scientific journals and their contents (Stanković et al. 2012). These metadata include information about individual journal issues as well as the articles they contain. At the issue level, basic bibliographic data are recorded, such as the identification number, volume, issue number, month, and year of publication.

At the article level, metadata are organized bilingually, in Serbian and English, and include information about authors, title, article type, page range, abstract, keywords, and the availability of full text, with the option of recording translator information where applicable. Such structured metadata enable systematic representation, retrieval, and the analysis of the bilingual corpus (Stanković, Obradović, and Trtovac 2012; Stanković et al. 2016).

The *Bibliša* database is implemented within the MongoDB[3] environment, enabling flexible data management and organization. The system supports

3. https://www.mongodb.com/

data export in various formats, such as TXT and XML, with the option to select different levels of detail.

For the purposes of this analysis, a subset of metadata was extracted, including the article identifier, issue number, year of publication, title, and information about the first author. In addition, textual versions of the articles in Serbian and English were used, allowing for both individual and comparative analysis.

After extraction, the corpora srInfoteka and enInfoteka were imported into the TXM environment, within which integrated models for morphosyntactic annotation and lemmatization were applied. For the automatic identification of part-of-speech categories and lemmas, TXM relies on the TreeTagger tool[4] (Schmid 1994), which is a widely used standard in corpus processing and supports tagging across multiple languages.

For Serbian, models adapted to the language and aligned with the Universal Dependencies (UD) tagset[5] were used, whose training and description are provided in (Utvić 2011; Stanković, Škorić, and Šandrih Todorović 2022). This approach enables consistent and standardized annotation of part-of-speech categories, which is particularly important for subsequent textometric analysis. For English, the Penn Treebank tagset[6] was employed, one of the most widely used standards for part-of-speech annotation in English. It provides fine-grained grammatical distinctions and is broadly applied in corpus linguistics and natural language processing.

Given that different tagsets were used for Serbian and English (UD and Penn Treebank), it was necessary to harmonize them to enable direct comparative analysis. To this end, Penn Treebank tags were mapped to the corresponding categories of the UD tagset, which represents a universal standard for morphosyntactic annotation. The mapping process relied on existing conversion schemes and linguistic correspondences between the two systems, with adjustments in cases where a one-to-one mapping was not possible. This normalization enabled a consistent view of part-of-speech categories across both subcorpora and a more reliable comparison of their grammatical characteristics.

The subcorpora each consist of 196 articles and 30,062 aligned segments (mostly sentences). The English subcorpus enInfoteka contains 871,253 tokens, while the Serbian subcorpus srInfoteka contains 775,669 tokens,

4. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
5. https://universaldependencies.org/u/pos/
6. https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf

making the English subcorpus approximately 12% larger in terms of token count.

When considering only words (excluding punctuation), srInfoteka contains 653,215 words, whereas enInfoteka contains 751,185, meaning that the English subcorpus is about 15% larger. In terms of distinct tokens, srInfoteka has 68,631, while enInfoteka has 40,358, indicating that Serbian has approximately 70% more distinct tokens, primarily due to its richer inflectional morphology. However, the number of distinct lemmas is also higher in srInfoteka by about 16%, with 36,640 lemmas compared to 31,373 in enInfoteka, despite the English corpus being slightly larger overall.

Lexical richness can be measured in various ways. One of the basic measures is the Type–Token Ratio (TTR), which represents the ratio between the number of distinct words (types) and the total number of tokens. Since it is dependent on corpus size, comparisons are most appropriate for corpora of the same or similar size. For srInfoteka, this ratio is 0.09, while for enInfoteka it is 0.05, indicating greater lexical richness in the Serbian subcorpus. Another measure, Root TTR, represents the ratio between the number of distinct words and the square root of the total number of tokens. For srInfoteka, this value is 77.93, while for enInfoteka it is 43.24, further confirming the higher lexical diversity of the Serbian subcorpus.

The chart in Figure 1 presents the distribution of parts of speech according to UD categories (UPOS) in the Serbian (sr) and English (en) subcorpora, revealing both similarities and clear differences. The NOUN category is dominant in both languages, with a slightly higher frequency in the Serbian corpus, indicating a somewhat greater nominal density. Similarly, punctuation (PUNCT) is highly represented and relatively balanced across the two subcorpora, reflecting the structural characteristics of scientific discourse.

Significant differences can be observed in certain grammatical categories. In the English corpus, determiners (DET) and adpositions (ADP) are considerably more frequent, which is consistent with the analytic nature of the English language. In contrast, Serbian shows lower frequency of these categories, as syntactic relations are often expressed morphologically, primarily through case inflection.

Verbs (VERB) are slightly more frequent in English, which may indicate a higher degree of explicitness in predicate structure, while adjectives (ADJ) and adverbs (ADV) are relatively balanced across both corpora. Conjunctions (CCONJ, SCONJ) exhibit moderate differences, with English showing a greater tendency toward their use.

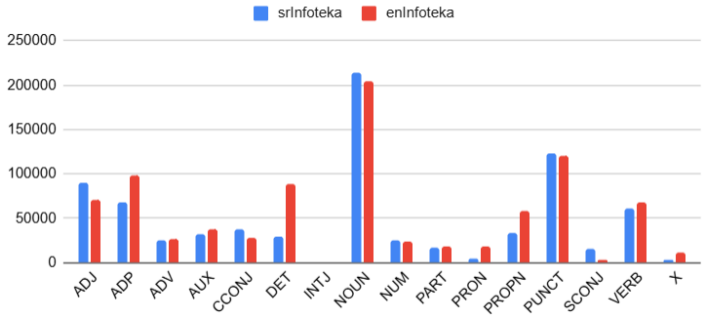

Figure 1. Number of tokens by part-of-speech categories in the Serbian and English subcorpora.

A particularly notable difference is the higher frequency of proper nouns (PROPN) in the English corpus, which may be attributed to differences in naming or translation conventions. Overall, the observed differences reflect the typological characteristics of Serbian (inflectional) and English (analytic) languages, as well as the specific features of scientific style in both.

It can be concluded that, although the INFOTHECA corpus consists of parallel texts, differences in categories such as DET, VERB, and ADP reflect underlying typological and stylistic distinctions between Serbian and English.

3 Frequency Analysis

3.1 Noun Frequencies

The analysis of noun frequencies in the parallel INFOTHECA corpus was conducted using appropriate CQL queries for the English and Serbian subcorpora. In the English part of the corpus, using the query `[enpos="NN.*"]`, a total of 204,262 noun occurrences were identified, comprising 16,396 distinct forms and 12,392 lemmas.

Figure 2 presents the distribution of the most frequent nouns in the English and Serbian subcorpora. In the English corpus, the most frequent nouns include *library*, *language*, *text*, *information*, and *system*, indicating the dominance of terminology related to library science, language technologies, and text processing. In addition, nouns such as *corpus*, *project*, *user*, *word*, and *document* frequently occur, reflecting the technological and user-oriented aspects of the analyzed texts.

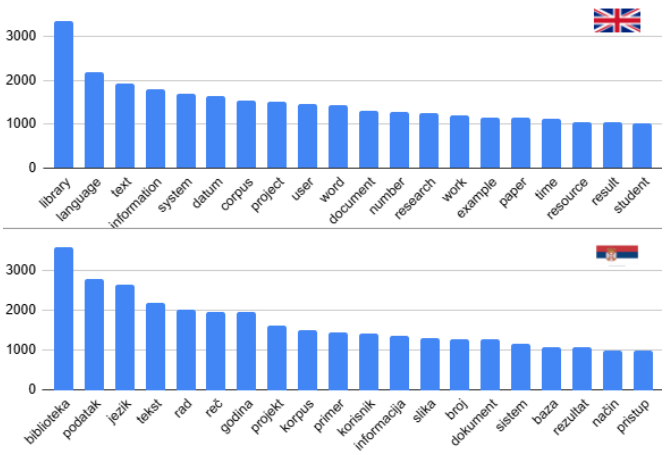

Figure 2. Twenty most frequent nouns in the subcorpora.

In the Serbian subcorpus, 214,129 noun occurrences were recorded, comprising 24,996 distinct forms and 13,583 lemmas. The most frequent nouns include *biblioteka*, *podatak*, *jezik*, *tekst*, and *rad*, indicating a similar thematic core to that of the English part of the corpus. Among the frequently occurring nouns are also *reč*, *godina*, *projekt*, *korpus*, *primer*, and *korisnik*, as well as *informacija*, *slika*, *broj*, *dokument*, *sistem*, *baza*, *rezultat*, *način*, and *pristup*.

The comparison reveals a high degree of overlap in the key terms across the two subcorpora, which is expected for parallel texts. However, differences in the distribution of certain lexemes can also be observed. The English corpus shows a slightly higher concentration of technically oriented terms, while the Serbian corpus encompasses a broader range of more general and context-dependent nouns. These differences may be attributed both to language-specific characteristics and to varying translation and stylistic choices.

3.2 Verb Frequencies

In the Serbian subcorpus, using the query `[srpos="VERB"]`, a total of 61,121 verb occurrences were identified, comprising 11,272 distinct forms and 3,982 lemmas. In the English subcorpus, using the query `[enpos="VV.*"]`, 67,891 verb occurrences were found, with 5,433 distinct forms and 2,328 lemmas. The results indicate a greater diversity of forms in Serbian.

Figure 3 shows the distribution of the most frequent verbs in the English and Serbian subcorpora. In both languages, general, high-frequency verbs dominate, reflecting their broad use in scientific discourse. In the English corpus, the most frequent verb is *use*, which appears significantly more often than others, followed by verbs such as *do*, *make*, *include*, and *create*. These verbs reflect a characteristic feature of English scientific style, emphasizing the explicit description of procedures and methods.

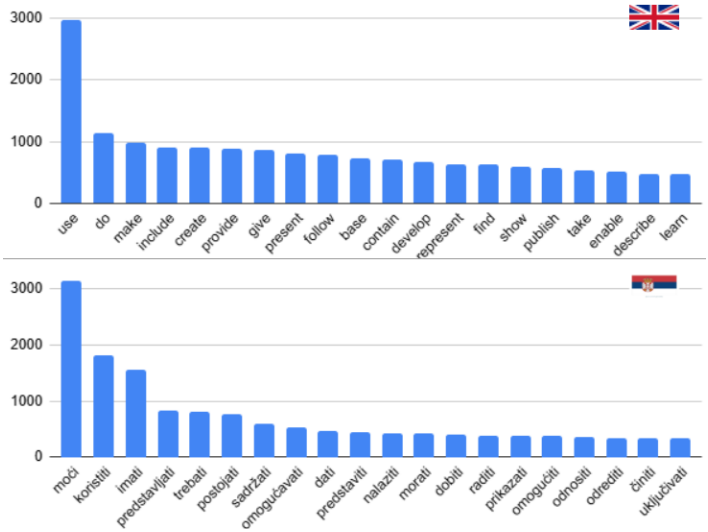

Figure 3. Twenty most frequent verbs in the subcorpora.

In the Serbian corpus, the most frequent verb is *moći* ('can'), indicating a more frequent use of modal constructions and the expression of possibility. It is followed by verbs such as *koristiti* ('use'), *imati* ('have'), *predstavljati* ('represent'), and *trebati* ('need/should'), which are also typical of academic style.

The comparison highlights differences in discourse strategies: English tends toward more direct and action-oriented expression, while Serbian more frequently employs modal and descriptive constructions. These differences may be attributed to typological characteristics of the languages, as well as to translation choices within the parallel corpus.

3.3 Adjective Frequencies

In the English subcorpus, using the query `[enpos="JJ.*"]`, a total of 70,582 adjective occurrences were identified, comprising 6,259 distinct forms and 5,400 lemmas. In the Serbian subcorpus, using the query `[srpos="ADJ"]`, 89,363 adjective occurrences were found, with 20,537 distinct forms and 9,153 lemmas. These results also indicate a significantly greater diversity of forms in Serbian.

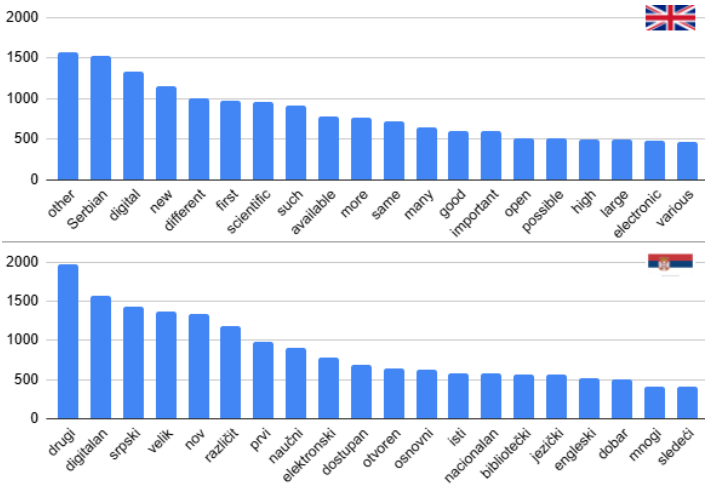

Figure 4. Twenty most frequent adjectives in the subcorpora.

Figure 4 shows the distribution of the most frequent adjectives in the English and Serbian subcorpora. In the English part, general and highly frequent adjectives such as *other*, *Serbian*, *digital*, *new*, and *different* dominate, reflecting their broad use in scientific discourse. In particular, *Serbian* and *digital* stand out, indicating the thematic focus of the corpus on the Serbian language, digital resources, and related processes. Among the frequently occurring adjectives are also *scientific*, *such*, *available*, *same*, and *many*, while adjectives such as *important*, *possible*, *large*, and *various* reflect a tendency toward generalization and evaluative expression in academic style.

The Serbian subcorpus exhibits a similar pattern, with the dominance of adjectives such as *drugi* ('other'), *digitalan* ('digital'), *srpski* ('Serbian'), *veliki* ('large'), and *nov* ('new'), with *digitalan* and *srpski* again indicat-

ing the thematic orientation of the corpus. Among the frequently occurring adjectives are *različit* ('different'), *prvi* ('first'), *naučni* ('scientific'), *elektronski* ('electronic'), and *dostupan* ('available'), which contribute to more precise description and classification of concepts. Domain-specific and stylistic features are further shaped by adjectives such as *otvoren* ('open'), *osnovni* ('basic'), *isti* ('same'), *nacionalan* ('national'), *bibliotečki* ('library-related'), and *jezički* ('linguistic'), while *dobar* ('good'), *mnogi* ('many'), and *sledeći* ('following') indicate elements of evaluation and textual organization.

Overall, both subcorpora are characterized by a combination of general and domain-relevant adjectives, which is typical of scientific discourse, with a high degree of thematic alignment between English and Serbian. This distribution confirms that adjectives in the corpus primarily serve the function of qualifying and specifying concepts.

3.4 Frequencies of Adjective–Noun Combinations

Among multiword terms, the most frequent pattern consists of an adjective followed by a noun, which plays a significant role in the specification and refinement of concepts. In the English subcorpus, the analysis of adjective–noun collocations using the query `[enpos="JJ.*"][enpos="NN.*"]` identified a total of 47,813 occurrences. These collocations comprise 26,731 distinct forms and 23,901 lemmas. The results indicate a high diversity of adjective–noun combinations, reflecting the richness of nominal phrases in English scientific discourse.

Figure 5 (top) presents the distribution of the most frequent domain-specific terms in the analyzed corpus. It can be observed that terms such as *digital library*, *Serbian language*, and *scientific research* dominate, indicating a clear thematic orientation toward digital libraries, language resources, and scientific research. Notably frequent are also terms such as *natural language*, *open access*, and *cultural heritage*, which reflect the broader context of applying language technologies within digital humanities. Among the frequent expressions are also *metadata*, *digital object*, *personal name*, and *foreign language*, highlighting the importance of data structuring and multilingualism in the corpus. The presence of terms such as *parallel corpus*, *social medium*, and *high education* further confirms the interdisciplinary nature of the analyzed texts.

Figure 5 (bottom) presents the distribution of the most frequent adjective–noun collocations in the Serbian subcorpus. The dominance of combinations such as *srpski jezik* ('Serbian language'), *univerzitetska biblioteka*

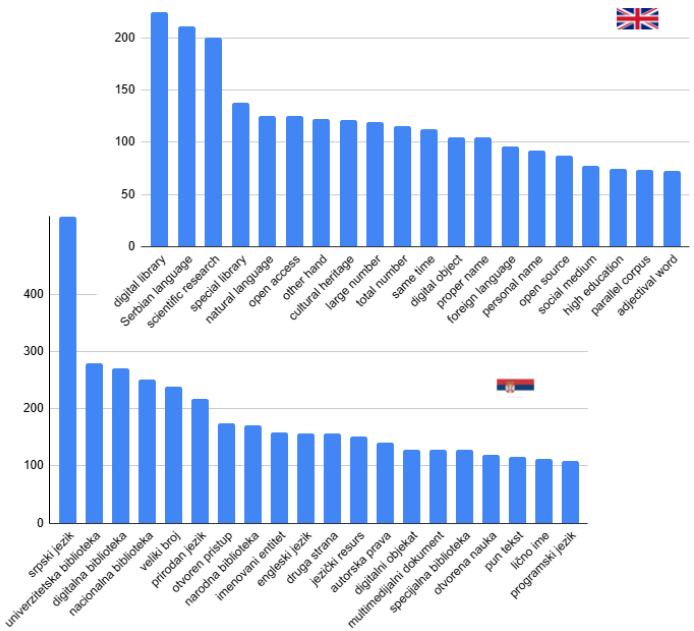

Figure 5. Twenty most frequent adjective–noun combinations.

('university library'), *digitalna biblioteka* ('digital library'), and *nacionalna biblioteka* ('national library') can be observed, indicating a clear thematic orientation of the corpus toward language, library science, and digital resources.

Among the frequent collocations are also *veliki broj* ('large number'), *prirodan jezik* ('natural language'), *otvoren pristup* ('open access'), *narodna biblioteka* ('public library'), and *imenovani entitet* ('named entity'), reflecting key concepts in the fields of natural language processing and information management. The presence of collocations such as *engleski jezik* ('English language'), *druga strana* ('other side'), *jezički resurs* ('language resource'), and *autorska prava* ('copyright') points to the multilingual and legal context of the analyzed texts. Furthermore, collocations such as *digitalni objekat* ('digital object'), *multimedijski dokument* ('multimedia document'), *specijalna biblioteka* ('special library'), *otvorena nauka* ('open science'), *pun tekst* ('full text'), *lično ime* ('personal name'), and *programski jezik* ('programming language') confirm the interdisciplinary nature of the corpus. Overall,

the distribution of adjective–noun combinations shows that the Serbian subcorpus is characterized by a rich and thematically coherent use of nominal phrases, which play a key role in the precise naming and specification of concepts in scientific discourse.

A significant difference in frequency can be observed for the term *srpski jezik* (536) and its translation equivalent *Serbian language* (211). This is because in English texts the noun *language* is often omitted, as it is implicitly understood. This can be seen in Figure 6, which presents concordances from the parallel INFOTEKA corpus obtained via the Noske platform[7] - NoSketch Engine (Kilgarriff et al. 2014).

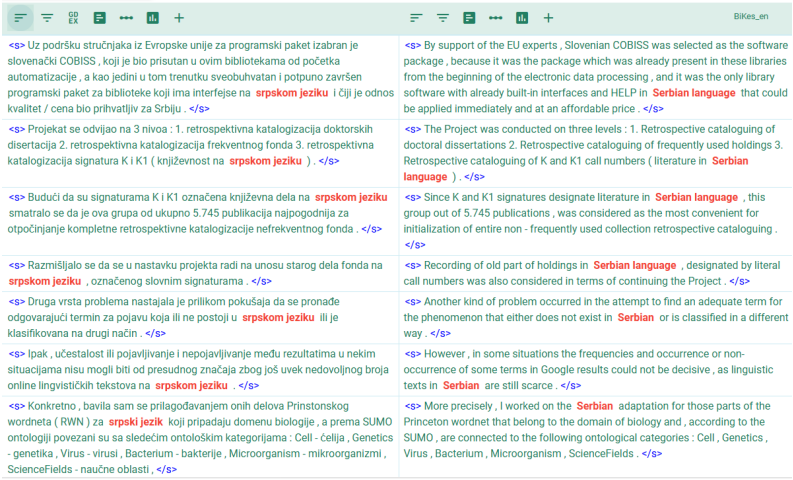

Figure 6. Concordances of the INFOTHEKA subcorpus on the Noske platform.

In English, noun–noun constructions are frequently used where Serbian employs adjective–noun structures. This is illustrated by the case of *univerzitetske biblioteke* ('university libraries'), which ranks highly in the Serbian corpus but does not appear as such in English, where it is realized as a noun–noun construction (*university library*). A similar pattern can be observed in other examples: *jezičke tehnologije – language technology*, *maternji*

7. https://noske.jerteh.rs/ - an instance of the NoSketch Engine system maintained by the Language Resources and Technologies Society JeRTeh

jezik – mother tongue, *zdravstvena zaštita – health care*, and *obrazovni proces – education process*.

Overall, the distribution of terms shows that the corpus integrates topics from the fields of language technologies, library science, digital resources, and education, with a strong emphasis on natural language processing and digital content management.

4 Textometric Analysis

4.1 Progression

Progression in TXM represents the distribution and change in the frequency of selected linguistic units along a text or across a given corpus, enabling insight into their dynamics and development over time or within the structure of a document. Figure 7 shows the temporal progression of the terms *model*, *corpus*, and *dictionary/lexicon* in the Serbian subcorpus. It can be observed that the term *corpus* (blue line) has the highest overall frequency and exhibits a continuous increase throughout the entire period, with more pronounced growth around 2012, 2019, and from 2024 onward, indicating sustained interest in corpora and their use in research. The term *dictionary/lexicon* (green line) shows a stable and moderate increase, with a somewhat earlier rise compared to the other observed terms, and notable growth in 2024, reflecting the traditionally strong presence of lexical topics. However, its growth is more gradual and less dynamic compared to *corpus*. In contrast, the term *model* (red line) exhibits slower growth in the earlier period, followed by a noticeable acceleration in later stages, which may be associated with the development of modern machine learning methods and language models.

Overall, the progression indicates a shift in focus toward corpus-based and model-driven approaches, reflecting broader trends in the field of language technologies.

Figure 8 presents the temporal progression of selected concepts in the Serbian subcorpus, illustrating the development of thematic areas over the period covered by the corpus. Three groups of concepts were analyzed: *otvoren pristup/nauka* ('open access/science'), *jezički/leksički/lingvistički resurs* ('language/lexical/linguistic resource'), and *digitalan objekat/nauka* ('digital object/library').

It can be observed that digital objects and libraries (green line) exhibit a relatively stable growth in the period 2011–2023, with a sharp increase in 2017, reflecting the gradual digitization and development of digital library

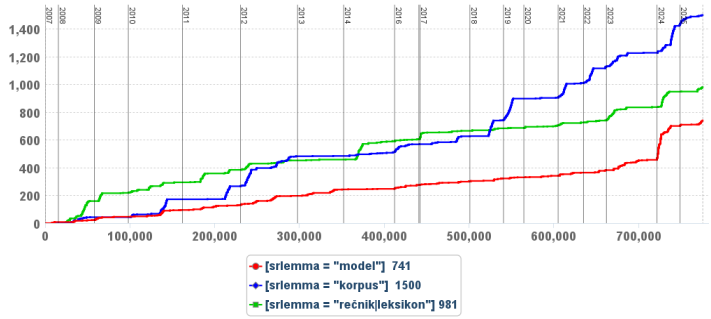

Figure 7. Progression of concepts related to language resources.

infrastructures. In 2023, a thematic issue of the journal was dedicated to the *eNauka* ('eScience') system, a publicly accessible information system that became operational in the first quarter of 2023.

Mentions of language resources (blue line) show noticeable peaks in 2008, 2012, and 2023, followed by further growth, indicating a strengthening interest in language technologies and resources in recent scientific discourse. In contrast, open science (red line) shows an initial increase in 2011, followed by more intensive growth only from 2018 onward. This trend reflects the expansion of open science initiatives and the increasing relevance of open access policies within the scientific community.

Overall, the progression indicates a clear development of infrastructural topics, language resources, and open science, with each of the observed categories reflecting different phases in the evolution of the digital scientific environment.

4.2 Topic Specificity by Periods

Specificity in TXM is a statistical measure that indicates the extent to which a given linguistic unit is overrepresented or underrepresented in a selected subset compared to a reference corpus, enabling the identification of characteristic terms for a given context or period (Heiden 2010). Figure 9 presents the specificity of selected terms (*korpus*, *rečnik*, *model*, *digitalizacija*, *anotacija*) across four time periods (2007–2025), with values expressed relative to the corpus average.

In the first period (2007–2011), dictionaries are mentioned more frequently than the corpus average, while other terms, particularly *digitalizacija*

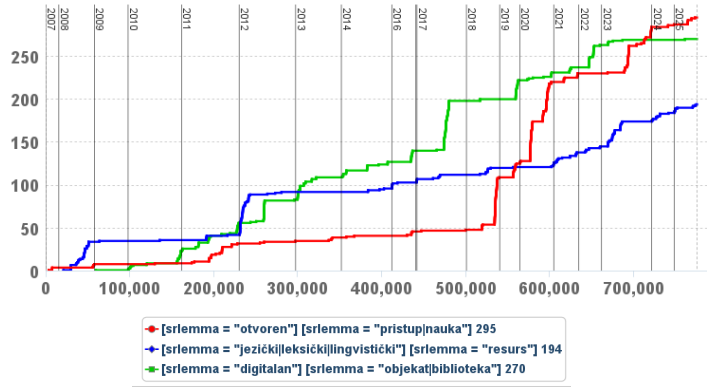

Figure 8. Progression of selected research topics.

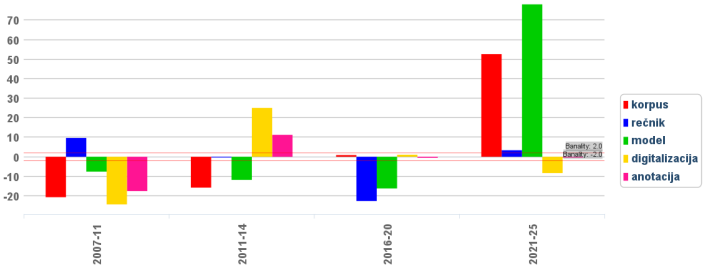

Figure 9. Specificity of selected research topics.

and *korpus*, show negative specificity, indicating their relatively low presence during this period. In the period 2011–2014, an increase in the specificity of *digitalizacija* and *anotacija* can be observed, reflecting a growing interest in digital resources and text processing. The period 2016–2020 is characterized by relative stabilization, with most terms showing values close to the average, while *rečnik* ('dictionary') and *model* exhibit negative specificity. For this period, notably positive specificity is observed for adjectives such as *normativan* ('normative'), *terminološki* ('terminological'), and *otvoren* ('open'), as well as for nouns such as *repozitorijum* ('repository') and *sintagma* ('syntagm').

The most pronounced changes occur in the period 2021–2025, where *model* and *korpus* ('corpus') show a strong increase in specificity, indicating the dominance of corpus creation and the use of language models. Notably,

these two terms attain positive specificity only in the final period. At the same time, *digitalizacija* ('digitalization') declines in importance, suggesting that it has become an underlying infrastructure rather than a central research topic.

Figure 10 presents the change in specificity of verb groups describing research activities across four time periods. In the early period (2007–2010), all groups exhibit negative or low specificity, indicating their limited presence. In the period 2011–2014, an increase in specificity can be observed for the group *kreirati-razvijati* ('create–develop'), as well as a moderate rise for *unaprediti-automatizovati* ('improve–automate'), while other groups remain less prominent. During the period 2016–2020, a decline in specificity is observed for most groups, particularly *unaprediti-automatizovati*, suggesting a temporary decrease in their relevance. The most significant changes occur in the period 2021–2025, where the group *obučavati* ('train') shows a strong increase in specificity, while *evaluirati-analizirati* ('evaluate–analyze') also rises, indicating a growing emphasis on describing language models and analytical approaches. These results reflect a shift from development and infrastructural activities toward methods based on machine learning and evaluation.

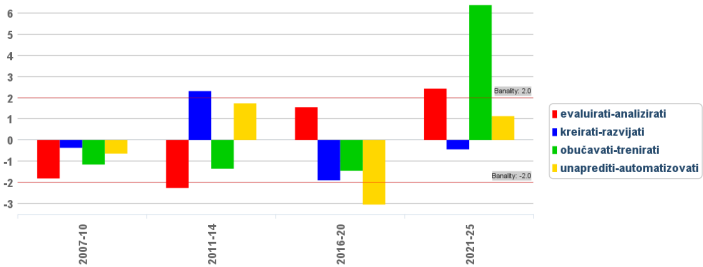

Figure 10. Specificity of selected research activities.

Overall, the results reveal a clear developmental trajectory: from an early phase focused on lexicography, through a phase of digitization and annotation, to a contemporary period dominated by corpus-based and model-driven approaches.

4.3 Comparison of Progression and Specificity

The comparison of progression and specificity for the terms *model.** and *katalog.** (‘catalog.*’) using the same CQL query reveals a clear shift in thematic focus over time.

The specificity analysis shows that the term *katalog.** (‘catalog.*’) is positively specific in the early period (2007–2010), indicating its relative prominence compared to the corpus average (Figure 11). In subsequent periods, its specificity declines and stabilizes around the average, followed by a decrease in the final period (2021–2025), suggesting that this concept is losing its central role in the discourse. In contrast, *model.** exhibits negative specificity in earlier phases but shows a markedly strong increase in the final period, indicating its growing thematic importance in the contemporary corpus and current research trends.

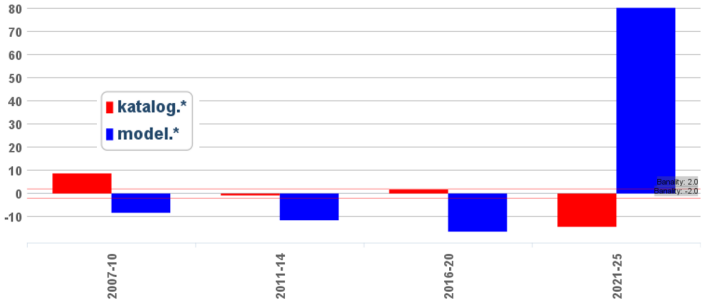

Figure 11. Specificity of selected research concepts.

This finding is consistent with the progression analysis, which shows that *model* exhibits relatively moderate and gradual growth until around 2023, followed by a sharp increase in frequency (Figure 12). In contrast, *katalog* (‘catalog’) and related terms (*katalogizacija* (‘cataloging’), *katalogizator* (‘cataloguer’)) display a stable but moderate increase without significant peaks up to 2021, after which they effectively stagnate (i.e., they occur very rarely), indicating a continuous but progressively less dominant role.

A combined consideration of these two aspects allows for a clearer interpretation: while progression reflects the absolute growth in the usage of terms, specificity analysis reveals their relative importance within the corpus. In this sense, although *katalog.** remains present throughout the entire

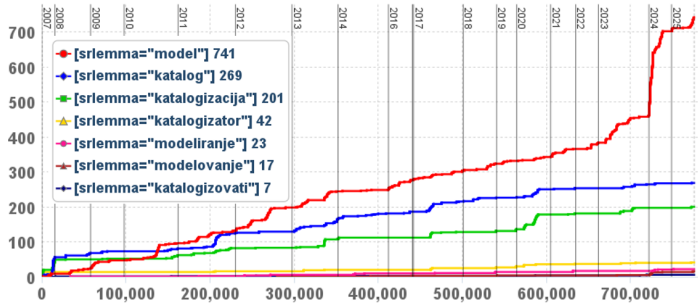

Figure 12. Progression of selected research concepts.

period, its relative importance declines due to the rapid growth of the term *model.**. This trend reflects a broader shift from traditional library and cataloging topics toward contemporary approaches based on models and machine learning.

5 Topic Modeling

Topic modeling was conducted on the English subcorpus of the journal, resulting in the identification of six thematic clusters that reflect different aspects of the analyzed corpus (Table 1). The topic modeling procedure, based on Latent Dirichlet Allocation (LDA), involves several key steps: text preprocessing, including cleaning, tokenization, and lemmatization, as well as the removal of stop words. The LDA model is then trained on the prepared data to extract latent topics within the corpus.

The first topic encompasses concepts related to library systems, scientific research, and digital infrastructure, while the second focuses on corpus linguistics and language resources, including text processing and translation. The third topic is associated with lexicographic aspects, such as dictionaries, word forms, and usage examples. The fourth topic relates to digital libraries and software projects, including platforms such as *Europeana*. The fifth topic covers the educational context, with a focus on students, teaching, and learning in digital environments. The sixth topic connects corpus and linguistic analyses with bibliographic and authorship aspects, including citation practices and text structure. The extracted topics point to the interdisciplinary nature of the corpus, and consequently of the journal itself, which integrates library science, language technologies, education, and digital humanities.

Topic	Keywords
Topic 1	library, system, data, research, scientific, user, university, digital, national, science
Topic 2	corpus, language, word, text, resource, serbian, model, used, document, translation
Topic 3	name, language, serbian, word, dictionary, example, used, form, figure, text
Topic 4	digital, library, project, software, material, europeana, data, computer, user, page
Topic 5	student, text, language, document, course, learning, school, web, programming, figure
Topic 6	word, language, serbian, citation, used, data, corpus, author, noun, number

Table 1. Topics extracted from the enInfoteka subcorpus.

The obtained results were further analyzed and interpreted using various visualizations, such as interactive topic displays, word clouds, and charts showing the distribution of topics across documents. Figure 13 presents the number of papers per topic, where one dominant topic is automatically assigned to each document using the LDA model.

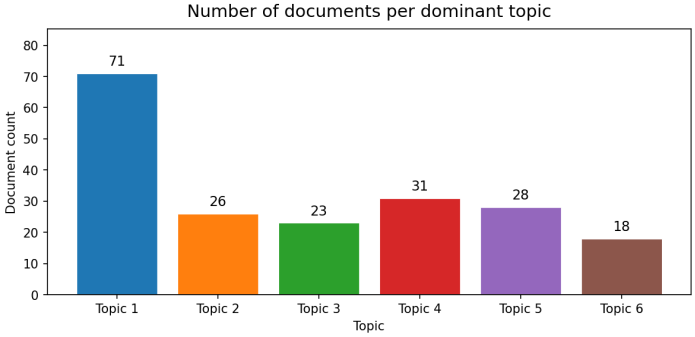

Figure 13. Distribution of papers across topics.

Figure 14 shows an interactive visualization of topic modeling results using the pyLDAvis tool (Sievert and Shirley 2014). The left panel represents the Intertopic Distance Map, obtained through multidimensional scaling,

where each circle corresponds to a topic. The size of each circle indicates the relative prevalence of the topic in the corpus, while the distance between circles reflects the degree of similarity between topics. It can be observed that some topics are closely related and partially overlapping, whereas others are more clearly separated.

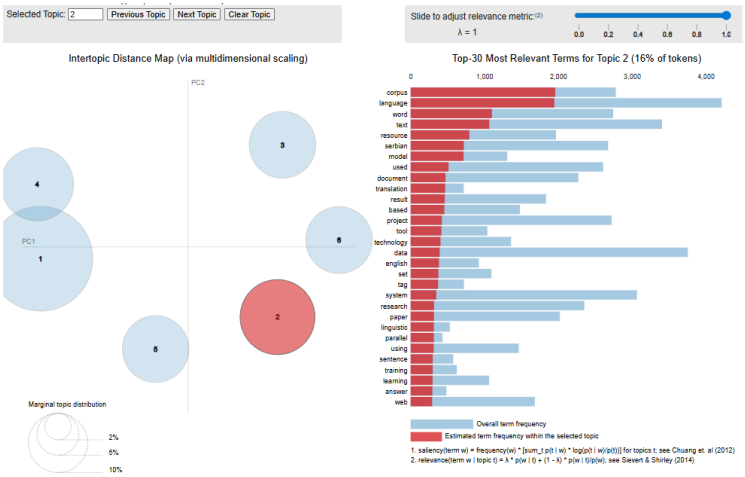

Figure 14. Interactive exploration of topics in the corpus.

The right panel displays the most relevant terms for the selected topic (Topic 2), which covers approximately 16% of the entire corpus. Among the key terms are *corpus*, *language*, *word*, *text*, *resource*, *Serbian*, and *model*, indicating that this topic is associated with corpus linguistics and language resources. The blue bars represent the overall frequency of terms in the corpus, while the red bars indicate their relevance within the selected topic (Sievert and Shirley 2014). More importantly, the red bars do not reflect only the absolute frequency of a word, but rather its relevance within the given topic, i.e., the estimated frequency of that word in documents associated with the topic. In other words, they show how characteristic a word is for the selected topic, rather than for the corpus as a whole. A term may be relevant to multiple topics, but with varying degrees of importance. Such visualizations facilitate the interpretation of topic structures and their characteristic vocabulary.

Figure 15 presents word clouds for the six identified topics, confirming the thematic diversity and interdisciplinary nature of the corpus.

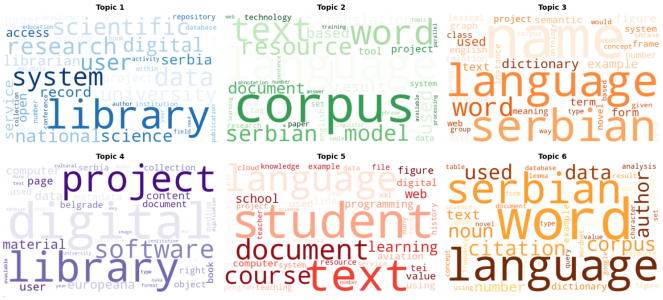

Figure 15. Word clouds by topic.

6 Discussion

The results of the textometric analysis of the bilingual Infoteka corpus reveal a high degree of thematic and lexical alignment between the Serbian and English subcorpora, which is expected given their parallel nature. At the level of frequency analysis, both subcorpora share a common core of concepts related to library science, language resources, and digital technologies, confirming the thematic consistency of the corpus.

However, the observed differences in the distribution of part-of-speech categories and lexical units point to the influence of typological characteristics of the languages. English, as an analytic language, exhibits a higher use of functional words, while Serbian, as an inflectional language, demonstrates greater morphological diversity, reflected in a higher number of distinct forms and a greater degree of lexical richness. The analysis of adjective–noun combinations indicates the stability of domain-specific terminology in both languages, with the presence of key concepts such as *digital library*, *natural language*, and *open access*, as well as their Serbian equivalents. At the same time, certain differences arise from translation conventions, such as the omission of the noun *language* in English.

The textometric analysis of temporal progression and specificity indicates a clear developmental trajectory of thematic interests. The early period is characterized by the dominance of lexicographic and library-related topics,

followed by a phase focused on digitization and annotation, while the contemporary period is marked by a strong increase in corpus-based approaches and language models. This trend is particularly evident in the analysis of the terms *model.** and *katalog.**, which demonstrates a shift in focus from traditional library activities toward modern methods based on machine learning.

The results of topic modeling further confirm the interdisciplinary nature of the corpus, which integrates the fields of library science, language technologies, education, and digital humanities. The identified topics show that the corpus is not thematically homogeneous, but rather encompasses multiple related yet distinct research directions.

7 Conclusion

This paper presented a textometric and comparative analysis of the bilingual Infoteka corpus, with the aim of examining the lexical, structural, and thematic characteristics of the Serbian and English subcorpora. By applying a range of methods, including frequency analysis, collocation analysis, temporal progression, specificity analysis, and topic modeling, comprehensive insights into the corpus were obtained.

The analysis has shown that, although the texts are parallel, there are differences influenced by the typological properties of the languages. The Serbian subcorpus is characterized by greater morphological and lexical diversity, while English demonstrates higher structural explicitness. At the same time, thematic analysis indicates a strong development of language technologies, with a shift in focus toward corpus-based approaches and language models in the contemporary period.

The results confirm that bilingual corpora represent a valuable resource for linguistic and interdisciplinary research, enabling both quantitative and qualitative analysis. A particular contribution of this study lies in the application of textometric methods to a domain-specific corpus, as well as in the integration of multiple analytical approaches.

Future research directions include the expansion of the corpus and the application of additional text processing and semantic analysis methods, in order to achieve a deeper understanding of the structure and evolution of scientific discourse.

Acknowledgment

We would like to express our gratitude for the corpus content to Prof. Dr. Cvetana Krstev, who served as editor of the journal *Infotheca* for 16 years, as well as to the many authors, reviewers, translators, and proofreaders. We also thank Dr. Jelena Andonovski, Dr. Biljana Rujević, and Dr. Aleksandra Tomašević for their work on corpus parallelization. This research was supported by the Science Fund of the Republic of Serbia under Project No. 7276, "Text Embeddings – Serbian Language Applications (TESLA)".

References

Andonovski, Jelena. 2026. "Infotheca: Journal for Digital Humanities - 2000-2026 -." *Infotheca – Journal for Digital Humanities* 26 (1): 87–118. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2026.26.1.4.

Heiden, Serge. 2010. "The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme." In *Proceedings of the 24th Pacific Asia conference on language, information and computation,* 389–398.

Jaćimović, Jelena. 2019. "Textometric Methods and the TXM Platform for Corpus Analysis and Visual Presentation." *Infotheca – Journal for Digital Humanities* 19 (1): 30–54. https://doi.org/10.18485/infotheca.2019.19.1.2.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. "The Sketch Engine: Ten Years on." *Lexicography* 1 (1): 7–36.

Krstev, Cvetana. 2021. "White as Snow, Black as Night – Similes in Old Serbian Literary Texts." *Infotheca - Journal for Digital Humanities* 21 (2): 119–135. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.6.

Pincemin, Bénédicte, Serge Heiden, and Franck Mazuet. 2022. "The textometric concept of active corpus." In *JADT 2022 Proceedings of the 16th International Conference on Statistical Analysis of Textual Data,* edited by Michelangelo Misuraca, Germana Scepi, and Maria Spano, II:691–698. Naples, Italy: VADISTAT - Per Simona Balbi, Univ. of Naples Federico II.

Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of the International Conference on New Methods in Language Processing.* Manchester, UK.

Sievert, Carson, and Kenneth Shirley. 2014. "LDAvis: A Method for Visualizing and Interpreting Topics." In *Proceedings of the workshop on interactive language learning, visualization, and interfaces,* 63–70.

Stanković, Ranka, and Lazar Davidović. 2021. "Infotheca (Q25460443) in Wikidata." *Infotheca - Journal for Digital Humanities* 21 (1): 87–98. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.1.5.

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac, and Miloš Utvić. 2012. "A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals." In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012),* edited by Nicoletta et al. Calzolari, 1710–1717. Istanbul, Turkey: European Language Resources Association (ELRA).

Stanković, Ranka, Cvetana Krstev, and Duško Vitas. 2024. "SrpELTeC: A Serbian Literary Corpus for Distant Reading." *Primerjalna književnost* 47 (2): 45–63. https://doi.org/10.3986/pkn.v47.i2.03.

Stanković, Ranka, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. 2016. "Keyword-based Search on Bilingual Digital Libraries." In *International KEYSTONE Conference on Semantic Keyword-Based Search on Structured Data Sources,* 112–123. Springer.

Stanković, Ranka, Ivan Obradović, and Aleksandra Trtovac. 2012. "An Approach to Development of Bilingual Lexical Resources." In *Proceedings of the Fifth Balkan Conference in Informatics (BCI 2012),* edited by Zoran Budimac, Mirjana Ivanović, and Miloš Radovanović, 101–104. Novi Sad, Serbia: Faculty of Sciences, Department of Mathematics / Informatics.

Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. "Parallel Bidirectionally Pretrained Taggers as Feature Generators." *Applied Sciences* 12 (10): 5028. https://doi.org/10.3390/app12105028.

Utvić, Miloš. 2011. "Annotating the corpus of contemporary Serbian." *Infotheca: Journal of informatics and librarianship* 12 (2): 36a–47a.