The *SERBOVERB* Language Resource and Its Multifunctionality

Saša Marjanović

University of Belgrade – Faculty of Philology,
Studentski trg 3, Belgrade 11 000, Serbia
sasa.marjanovic@fil.bg.ac.rs

Abstract

Serbian verb inflection is known for its complexity and unpredictability, posing a challenge for L2 Serbian speakers. Existing dictionaries are not well-suited to address the needs of L2 speakers. To overcome these challenges, the author presents *SerboVerb*, an electronic resource and application that offers a dynamic approach to processing Serbian verb inflection. *SerboVerb* includes a conjugation, dictionary, and gamification module, and offers paradigms for more than 34,000 verbs. The resource has been developed through a research project between the University of Toulouse Jean Jaurès (France) and the University of Belgrade, Faculty of Philology (Serbia). The author describes the structure and multifunctionality of *SerboVerb*, highlighting its potential to provide a more accessible and user-friendly resource for L2 Serbian speakers seeking to resolve their communication problems and improve their language skills. By offering a multifunctional and comprehensive approach to Serbian verb inflection, *SerboVerb* represents a significant step forward in electronic lexicography.

Keywords: *SerboVerb*; verb inflection; Serbian; resource; dictionary

1. Introduction

Serbian verb inflection is quite complex. The paradigm of the average Serbian verb in the active voice includes hundreds of inflected forms (*cf.* Krstev, 1997; Tošović, 2012). The relationship between these inflected forms and their basic (lemma) form — which is conventionally used to represent the entire verb paradigm — is only predictable in a small number of inflectional classes (Jelaska, 2005; Marjanović, 2016b). Hence, mastering Serbian verb inflection can be quite challenging for average L2 Serbian speakers (*cf.* Krajišnik, 2011; Babić, 2021). The task is rendered even more difficult by the fact that some inflected forms are hard to match to their lemma form. The existing Serbian dictionaries, both mono- and bilingual, where L2 speakers might search for an inflection information, are not well tailored to the needs of average L2 speakers: they list verbs generally only in the lemma form, while the forms relevant for establishing the entire paradigm (*cf.* Marković, 2014) are very often lacking (Marjanović, 2016a). Although there are different ways to process Serbian verb inflection in printed dictionaries to satisfy all the prototypical communication-related and cognitive needs of target users (see Marjanović, 2016a & 2016b), we believe that the most appropriate and up-to-date solution is found in electronic lexicography, in the form of an electronic conjugator.

This paper provides an overview and evaluation of currently available Serbian language conjugators in Section 2. Since these resources have some limitations and a new one is needed, Section 3 examines the existing inflection lexicons developed for Serbian language processing that could serve as a starting point for a new conjugator. Section 4 introduces *SerboVerb*, an innovative linguistic resource and its application, designed for Serbian L2 speakers. Developed as part of a research collaboration between the University of Toulouse – Jean Jaurès, France, and the University of Belgrade, Serbia, *SerboVerb* processes Serbian verb inflection dynamically, eliminating the limitations of static paper-based resources. The resource is accessible for free via a website and mobile app for Android and iOS. The paper emphasizes the potential of electronic lexicography to overcome traditional resource limitations and better meet the needs of L2 speakers. In addition, this paper details the structure of *SerboVerb*, highlighting its exhaustiveness, simplicity, and availability in processing verb inflection. Section 5 delves into its various functionalities. The paper concludes by outlining future plans in Section 6 and providing closing remarks in Section 7. Overall, the paper aims to showcase the multifunctionality of *SerboVerb* as a valuable language resource for learners of Serbian.

2. Previous Resources

SerboVerb is not the sole Serbian conjugator intended for human use, nor is it the first. To the best of our knowledge, several such electronic tools have been developed since the 1990s. Section 2.1 of this paper provides a chronological review of existing conjugators, while Section 2.2 offers a comparative evaluation of their strengths and weaknesses.

2.1 Existing Conjugators

The first Serbian conjugator was developed by a private company *Lexicom* (https://lexicom.rs) based in Belgrade (Serbia). However, there is no accompanying technical or scientific documentation related to this resource, so it is unclear how extensive the resource is and how many verbs it processes. The resource was freely searchable through the company's website (*cf.* Marjanović, 2016a), but is no longer accessible. The verb paradigm was presented in a tabular format. It is worth noting that, while Serbian can be written using both the Cyrillic and Latin alphabets, the verb lemma search and display in this particular case were exclusively limited to the Latin script.

The *Grammatical Dictionary of Serbian* is the second conjugator, a linguistic resource created by the private company *Srbosoft* from Obrenovac (Serbia), which offers a range of Serbian language lexicographic resources, mostly retro-digitized from previous paper editions. The resource is available on the company’s website (http://srpskijezik.com). It has been available online since the end of 2017 and can be searched with an annual subscription. While there is no documentation for the resource either, it contains approximately 117,296 lemmas, which would include around 20,000 verb paradigms. The database can be searched by lemma using both Cyrillic and Latin alphabets, but the paradigm output is exclusively in the Cyrillic alphabet. The output is presented in plain text format, showing one tense at a time. To access other tenses, users need to click on the corresponding tab. However, it's important to note that the paradigm display presents inflected forms in a tabular format, numbered from 1 to 6. This means that the third person plural is listed as the sixth person. This sequencing might lead to potential confusion among users. The resource also provides accent markings for all inflected forms, allowing the user to obtain information about the pronunciation of each form.

The *Verbix* conjugator (https://www.verbix.com) is the third conjugator available for Serbian and provides access to conjugators for over one hundred languages. Users can search the verb database by entering any form of the verb without creating an account. However, unlike the previous two resources, *Verbix* can only be searched in Cyrillic script, and the output of the verb paradigm is also only in Cyrillic. The resource includes both simple and compound forms, but does not provide verb participles nor verb adverbs. There are typographical and encoding errors, as well as frequent instances of uncorrected inflected forms, which may compromise its reliability. However, the advantage of this conjugator is its more accessible paradigm layout. Additionally, 20 randomly selected verbs belonging to the same inflectional class are listed in the lemma form, prompting users to consider the similarities and differences between the paradigms of related verbs.

In addition to the three conjugators for the Serbian language, a Croatian conjugator called *Croatian Morphological Lexicon* (hereafter referred to as *CML*) (http://hml.ffzg.hr) has also been available since 2005 (Tadić & Fulgosi, 2003; Tadić, 2005; *cf.* Ljubešić et al., 2016). The relevance of Croatian conjugators to this paper lies in the fact that Croatian and Serbian are standardized micro-languages that are part of the same macro-language system. They share the same inflectional patterns and have a significant overlap in their lexical systems. Access to this *CML* conjugator requires an account approved by the author. However, it is not possible to reliably present the resource as access to it was not obtained at the time of writing. Based on literature (Tadić & Fulgosi, 2003), the resource contains about 36,000 lemmas, of which 7,735 are verbs, with two types of searches possible: by lemma and by any inflected form. The results of the searches are not hyperlinked, meaning that the user cannot access the complete paradigm of the selected lemma from an inflected form without conducting a new search. The first version of the conjugator listed the inflected forms alphabetically (Tadić, 2003), while the second version grouped them into traditionally organized paradigms (Tadić & Fulgosi, 2003). Additionally, in both versions, the inflected forms were tagged with a morphosyntactic code.

Finally, the *Croatian Language Portal* (hereafter referred to as *CLP*) (https://hjp.znanje.hr), a combination of retro-digitalized previously published Croatian monolingual dictionaries, includes a conjugator that provides users with the complete conjugation of 12,011 out of 15,699 Croatian verbs. However, the paradigms generated by the Portal's conjugator do not always match the data provided in the traditional morphological blocks of the dictionary entries, resulting in inconsistencies in the data presented to users. It is worth noting that the morphological block, which forms an essential component of the *CLP* dictionary entry, includes only the relevant inflection data necessary to establish the complete paradigm of a verb. In some instances, the conjugator offers only a single paradigm for a verb, disregarding the possibility of multiple potential paradigms as indicated by the data in the morphological block. Furthermore, the paradigms provided by the conjugator do not indicate any competing forms within corresponding tenses, further eroding the overall credibility of the *CLP* as a reliable language resource.

2.2 Comparative Evaluation

While the conjugators mentioned earlier can be useful for L2 Serbian speakers, each of them has its own limitations that electronic lexicographic resources should strive to overcome (*cf.* Tarp, 2008; Tarp, 2012; Lew, 2012; Grønvik & Smith Ore, 2013; Simonsen, 2014; Simonsen, 2015). These limitations concern the following eight points: *availability*, *access*, *content*, *scope*, *reliability*, *updating*, *searchability*, and *display*.

Regarding the first point, it can be concluded that all conjugators are available except for the first one (*Lexicom*), which, to the best of our knowledge, cannot be accessed for unknown reasons. Previously, *Lexicom* was open and available for free search without an account, much like *Verbix* and *CLP*. However, to search the *Srbosoft* conjugator and *CML*, users need to create an account, which is then verified by administrators before use. Unlike the others, *Srbosoft* conjugator access is not free and requires an annual subscription. Therefore, only two conjugators (*Verbix* and *CLP*) are currently available for completely open access.

In terms of content, Serbian and Croatian conjugators can be divided as follows: the first group (*Lexicom*, *Srbosoft*, and *Verbix*), exclusively includes verbs pertaining to the Ekavian variety. Conversely, the second group (*CML* and *CLP*), only contains Ijekavian verbs. These variations are a result of the different diatopic reflexes of the Old Slavonic sound *yat*. Consequently, a single verb that previously had the *yat* sound can now have two standard variants: the Ekavian — *e* — variant (e.g., *deliti*) and the Ijekavian — *ije* or *je* — variant (e.g., *dijeliti*), which are marked by areal distinctions. While it is expected for the Standard Croatian to include only Ijekavian forms, it is not justified for Serbian conjugators, as the Standard Serbian encompasses both Ekavian and Ijekavian variants. As a result, Serbian conjugators may not be helpful to users in need of inflection data on Ijekavian verb forms. Additionally, although

Serbian can be written in both Cyrillic and Latin scripts, all Serbian conjugators are available in only one script, with *Lexicom* and *Verbix* in Latin and *Srbosoft* in Cyrillic. Croatian conjugators use the Latin script, as it is the only script of the Croatian standard. This can pose a challenge for users who are not proficient in both alphabets.

When it comes to the scope of these resources, there are noticeable differences. Regarding the number of lemma, *Srbosoft* has the highest number of verbs (around 20,000), followed by *CLP* (12,011) and *CML* (7,735). Data on the number of verbs for *Lexicom* and *Verbix* is not available, but a random search of fifty verbs on *Verbix* reveals that even the most common verbs are missing. As only *CLP* and *Verbix* are freely accessible, it can be inferred that *CLP* has the most comprehensive coverage, but as it is a Croatian resource, it does not include verbs unique to the Serbian standard. Nonetheless, *CLP* is also the most inclusive concerning the number of inflected forms it encompasses, incorporating all simple and compound inflected forms. Conversely, *Verbix* excludes non-finite verb forms, while *Srbosoft* does not provide a paradigm for compound forms. There is no information on the data for *Lexicom* and *CML*, but as their resources were primarily created for *NLP* purposes, it is likely that these conjugators exclude compound forms.

Furthermore, *Verbix* contains many spelling, encoding, and material errors, while the other conjugators are reliable. However, this would not be a problem if the *Verbix* database were regularly updated and errors corrected. Unfortunately, this is probably not the case. It is also not clear whether any of these resources are regularly expanded with new verbs.

When it comes to searchability, all conjugators allow searching by entering the corresponding verb in the alphabet in which the verbs are stored in the database. Only the *Srbosoft* conjugator enables alphabet-insensitive search, which means that the user does not have to use Cyrillic script in the search field, but the search results will still be in Cyrillic. This can be convenient for users who do not have Cyrillic keyboards. *Verbix* and *CLP* offer the most flexibility regarding the linguistic form that can be entered in the search field. Users of these conjugators can enter any form of the verb in the search field, not just the lemma form, as is the case when searching with *Lexicom* and *Srbosoft* conjugators. On the other hand, *CML* is somewhere in between: this conjugator allows searching both by lemma and by inflected forms, but in the case of searching for an inflected form, the user is informed of its morphosyntactic description and directed to its lemma, which they need to search again if they require the entire verb paradigm.

It should be noted that while morphosyntactic identification of the searched form in *CML* is very useful, the MULTEXT-East format in which this description is encoded may be difficult for average users to decode. On the other hand, *Verbix* and *CLP* do not provide identification of the searched inflected form, but only display the full paradigm. It should also be pointed out that *Verbix* offers an autocomplete option when entering the verb in the search field, which saves time needed for typing the rest of the word.

Regarding paradigm display, most conjugators list inflected forms in a row, one after the other, with each tense being named. The *Lexicom*, *Srbosoft*, *CML*, and *CLP* conjugators provide numbered forms for persons. With all, except *Srbosoft*, the numbering follows the traditional description of three singular and three plural persons, meaning that singular and plural are numbered separately. In the *Srbosoft* conjugator, however, all persons are numbered continuously, with plural forms being numbered 4–6. This may be confusing for users accustomed to the traditional didactic description of forms and verb paradigm presentation.

On the other hand, unlike the aforementioned conjugators, the forms in *Verbix* are not numbered at all, which reduces the paradigm's readability. *Srbosoft* compensates for this shortcoming with a better paradigm view: only one verb tense is shown at a time, while others are selected by clicking on a tab above the verb forms. In other conjugators, the paradigm view is uninterrupted, and the user must scroll down to find data not immediately visible.

Finally, it can be said that conjugators do not provide information on the meaning of verbs. This is expected given that conjugators only offer data on verb inflection. However, in the era of linked resources, it is regrettable that the presented conjugators cannot be used with other tools. The only exception is *CLP*, which provides, for each entry, a description from integrated Croatian monolingual dictionaries, but it should be noticed that its data does not always match the data provided by morphological blocks. Therefore, there is a need for a new conjugator that would address all the shortcomings mentioned in this evaluation.

3. Related Resources

The starting point in the development of a new conjugator can be the use of the outcomes of Serbian and Croatian language processing. The first results date back to the 1990s, but they were not available for long. Tadić's conjugator, mentioned above, is based on the author's *Croatian Morphological Lexicon*, which has been available through META-share since 2012. It consists of entries in triples format: first, the inflected form is listed, followed by the lemma, and finally, the morphosyntactic description encoded according to the MULTEXT-East recommendations (Tadić, 2003; Tadić & Fulgosi, 2003). However, this lexicon is based on entries from a medium-size one-volume Croatian dictionary, which limits its coverage of Serbian standard vocabulary due to its focus on the most frequent Croatian words.

At the same time, a more extensive resource called the *Serbian Morphological Dictionaries* (*SrpMD*) was developed in the DELA format, relying on UNITEX systems (Krstev, 1997; 2008). This resource consists of several text files, including one containing simple-word lemmas (DELAS), one for multi-word lemmas (DELAC), and two files for inflected forms of simple and complex words, respectively (DELAF and DELACF), generated automatically (see Stanković et al., 2018). The lemma lexicon includes entries in lemma form, their corresponding POS category, and the label of a finite-state transducer, which allows for the unambiguous production of all inflected forms and their morphosyntactic properties. The inflected form lexicons include entries in inflected form, their lemma, and their morphosyntactic properties. The lemma lexicon also often includes a series of markers that indicate features of the entry or indicate the type of feature and specify its value. The resource includes both Ekavian and Ijekavian word forms of the Serbian language and is encoded in ASCII to neutralize the difference between Cyrillic and Latin characters. The number of entries in *SrpMD* is constantly increasing, and according to the literature, its size has grown significantly over the years. The initial version of the simple word lexicon DELAS comprised 6,569 lemmas, with 1,884 of them being verbs (Krstev, 1997). Ten years later, the lexicon expanded to include 84,607 lemmas, of which 15,907 were verbs (Krstev, 2008). Presently, the lexicon contains a total of 205,003 lemmas, with 21,159 of them classified as verbs (Rujević, 2022: 32). Development of this resource was initially carried out through the *WS4LR* application interface, which was later upgraded and renamed to *LeXimir* (Stanković et al., 2018). Although this resource is indexed on Meta-Share, it is only available to a limited group of users upon request, and other researchers — unfortunately — cannot use or distribute it for either commercial or non-commercial purposes (see Ljubešić et al., 2016; Miletic, 2017 & 2018).

Another noteworthy lexicon for Serbian language processing is the accentual-morphological lexicon developed for the *AlphaNum* speech synthesizer (Sečujski & Delić, 2011). This lexicon contains entries with information about the lemma, encoded accentual configurations, and morphosyntactic properties. As of 2011, it contained around 100,000 lemmas, with ongoing additions facilitated by the *ARecnik* user interface. The interface enables manual entry of new words or automatic input from connected text files. Based on the entered data, the program generates inflected forms, morphosyntactic properties, and accentual configurations. However, this lexicon is not available for download.

According to published references (Tošović, 2012 & 2014), significant efforts were made between 2008 and 2015 to carry out morphological annotation of inflected and uninflected words in Serbian, Croatian, and Bosnian. The project aimed to establish the minimum number of rules required to generate the maximum and complete system of inflected forms using the *MorfoGenerator* system. The project covered 30,030 verbs out of 112,000 words, using 378 out of 822 rules to generate inflected forms for each verb. The resulting lexicon, *Gralis-MorfoGenerator*, was used for morphosyntactic annotation of texts in the multilingual *Gralis* corpus. Regrettably, the manually verified inflected form paradigms and the *MorfoGenerator* tool, which were intended to be publicly available, are not currently accessible for search or download from any repository. Furthermore, the webpage cited in the papers is no longer reachable.

The first freely available morphological lexicon of the Serbian language, *Wikimorph-sr*, was derived by parsing the pages of the Serbo-Croatian version of *Wiktionary* based on a dump from October 2, 2015 (Miletic, 2017). The primary purpose of the lexicon was to enable multilayered annotation of Serbian texts in the multilingual parallel corpus *ParCoLab* (Miletic et al., 2017). It was supplemented with a list of entries extracted from a previous manually POS-tagged Serbian texts. The lexicon is in triples format, in accordance with MULTEXT-East recommendations, and contains 117,445 lemmas, including 11,299 verbs. Its coverage was tested on three contemporary Serbian novels, consisting of around 150,000 tokens, or 28,980 unique word forms, of which over 50% appear only once. The lexicon was found to cover 72% of word forms in these novels, which increases to around 80% for words that appear more than 10 times. The author notes that this result may be higher if a larger sample of texts were tested, but also suggests that the lexicon should be manually supplemented.

SrLex (Ljubešić et al., 2016) is another open-source lexicon that was created alongside the Croatian lexicon *hrLex*. These lexicons were built by expanding a publicly available lexicon from the *Apertium* machine translation system, which contained 10,183 lemmas assigned to 413 inflectional patterns. To identify missing words, the *hrWaC* and *srWaC* corpora were searched by frequency. A team of six linguists then used a graphical interface to review the missing Croatian words. They could either accept one of the automatically predicted lemma and inflectional pattern candidates or flag the word as not belonging to any of the predicted candidates. The process was repeated six times to improve coverage. The Serbian data was processed in just two rounds due to the significant lexical overlap with Croatian. As a result of the expansion, the Serbian lexicon (*srLex*) contains 105,358 lemmas, with an increase in the number of verb patterns from 167 in the original *Apertium* lexicon to 568 in *srLex*. The lexicon is freely available in both MULTEXT-East and Universal Dependencies formats.

In a study by Miletic (2018), the last two lexicons were mutually compared. It was shown that *Wikimorph-sr* contains only 21% of the entries found in *srLex*, while *srLex* contains 41% of the entries from *Wikimorph-sr*. Although the first finding is not surprising, the latter is less expected. Therefore, these resources were integrated into a single resource called *ParCoLex*, to assess whether their combined use could provide better coverage of *ParCoLab* text samples. The assessment used a sample of 16,389 tokens, corresponding to 6,301 unique inflected forms. The results showed that *srLex* provided better coverage than *Wikimorph-sr*, with 94% coverage of tokens compared to 73% for *Wikimorph-sr*, and 93% coverage of unique inflected forms compared to 63% for *Wikimorph-sr*. However, the newly integrated *ParCoLex* outperformed both resources, achieving 98% coverage for all tokens and 95% coverage for unique inflected forms. With its largest number of lemmas (157,886, including 14,562 verbs), *ParCoLex* can serve as a valuable resource for researchers and developers working on Serbian language-related projects, such as *SerboVerb* (presented in next sections), since it offers a comprehensive and relatively reliable source of morphological information.

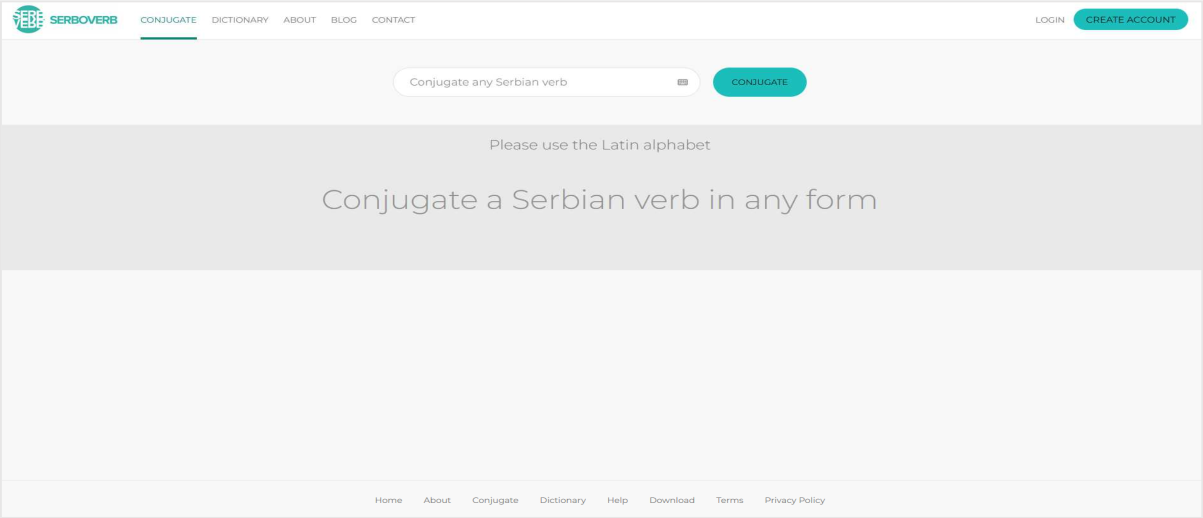
4. The *SerboVerb* Language Resource

In response to the limitations of existing conjugators for Serbian (and Croatian), as discussed in Section 2, a project was launched in 2017 at the University of Toulouse - Jean Jaurès (France) to develop a new, comprehensive, and multifunctional conjugator for Serbian, which was named *SerboVerb*. The project aimed to create an electronic resource that could be easily searched through a user-friendly application, taking into account the availability of an extensive morphological lexicon for non-commercial use (as discussed in Section 3).

The development of the *SerboVerb* application was funded by the Research Valorization Unit of the University of Toulouse – Jean Jaurès (France) and Toulouse Tech Transfert, a French company dedicated to promoting local research results through technology transfer. The development of the *SerboVerb* resource began in 2018 and has been ongoing since then. It is being carried out by an expert group consisting of linguists, lexicographers, and NLP researchers from the University Toulouse – Jean Jaurès and the University of Belgrade, Faculty of Philology (Serbia). This expert group had already established an intensive collaborative relationship in the field of NLP (*cf.* Miletic et al., 2017). External collaborators were also involved, including volunteers from both universities.

The entire resource is hosted on servers provided by Huma-Num, the French digital infrastructure supported by the CNRS (the French National Center for Scientific Research). It can be accessed for free via the website (https://serboverb.com), as well as through a mobile app available for Android and iOS operating systems, which can be downloaded from the Google Play Store and the App Store, respectively. The web application also serves as a resource management system. Figure 1 shows the homepage of the web application.

In order to enhance the overall functionality of the resource, a complex verb database, including their inflection paradigms and foreign languages equivalents, was implemented into the application, along with additional external educational materials. Consequently, the *SerboVerb* application now comprises three modules: a conjugation module, a dictionary module, and a gamification module, which will be presented in the following subsections.



Please use the Latin alphabet

Conjugate a Serbian verb in any form

Figure 1: Homepage of the web-based *SerboVerb* application

4.1 Conjugation Module

The Conjugation module is a part of the application that enables users to search the verb database and display the inflectional paradigms of the searched verbs. The *SerboVerb* database was created based on the *ParCoLex* morphological lexicon (see Section 3). The lexicon was converted from a text file in MULTEXT-East format to XML using the P5 schema of the Text Encoding Initiative (TEI). Additionally, since the *ParCoLex* lexicon exclusively stored simple verb forms such as present, imperative, synthetic future, aorist, imperfect, active and passive participle forms, as well as present and past participle forms, active compound forms — including perfect tense, analytical future, future II, conditional, and pluperfect forms — were automatically generated. Passive compound forms were not included. When generating these compound forms, special attention was given to include forms that usually occur in context with a subject, as well as forms when the subject is omitted. However, it was noticed that some relatively common verbs were missing from this extensive inflection database of 14,562 verbs, so work on expanding the *SerboVerb* database began immediately.

The expansion work was carried out in four phases. In the first phase, all the verbs in their lemma form were extracted from the *ParCoLex* lexicon and arranged in tabular form. Then, all the verbs were extracted from the *Reverse Dictionary of the Serbian Language* (Nikolić, 2000) and compared with the list of verbs in the lexicon. Any verbs missing were retained, and merged with the first list. Since reflexive verbs in *ParCoLex* do not contain the reflexive particle *se*, while the *Reverse Dictionary* includes reflexive verbs in their lemma form, merging these two lists enabled the identification of existing reflexive verbs in the *SerboVerb* database. Additionally, a small number of verbs were manually added from other specialized paper lexicographic resources, primarily slang dictionaries, dictionaries of neologisms and anglicisms.

Finally, as all Serbian lexicographic sources are based on relatively outdated material, all missing verbs found in the *srWac* and *hrWac* corpora (Ljubešić & Klubička, 2014) were automatically extracted. The resulting list included 34,049 verbs. In the second phase, the verbs were annotated. Two annotators worked on this task, which lasted for six months. Firstly, based on the existing linguistic descriptions, 121 inflectional patterns were identified. Then, for each verb, a manual tag was assigned to indicate its membership to one of these patterns. In cases where a verb could also have a paradigm according to another inflectional pattern, an additional tag was assigned. However, the patterns did not include imperfect tense forms. For each verb in the database, a verb aspect was also indicated to mark the absence of certain verb forms (e.g. imperfect tense forms for perfective verbs, aorist tense forms for imperfective verbs). Each verb associated with either the Ekavian or Ijekavian variety was annotated with a distinct tag, and its corresponding counterpart in the opposite variety was added. Similarly, a subset of approximately 16,000 most frequent verbs and a subset of 1,844 core Serbian verbs (*cf.* Section 4.2) were specifically tagged. The lists of these most frequent and core verbs were published in the form of a paper-based conjugation dictionary for the needs of Serbian L2 speakers (Marjanović & Radosavljević, 2019). However, the entire *SerboVerb* database has not been made available for distribution.

In the third phase, rules for generating verb paradigms belonging to the most frequent and productive inflectional patterns were developed, as the inflectional patterns were designed to allow for the creation of rules for unambiguous generation of the complete verb paradigm. Simple and compound inflected forms were generated for all verbs that follow productive inflectional classes, which were missing from the database generated based on the *ParCoLex* lexicon. The imperfect tense forms were generated using a separate set of rules. The newly generated forms were added to the *SerboVerb* database at the end of 2018. The source element in the XML structure of the *SerboVerb* database provided clear indication of the *ParCoLex* paradigms and the newly added verbs, as well as their generated inflected forms.

During the fourth and final phase, the manual verification of newly generated inflected forms began in the spring of 2019. The verb paradigms formed on the basis of data from the *ParCoLex* lexicon were immediately published and have since been accessible to end-users. Initially, these forms were not subject to verification, as the creation of the *srLex* resource, which formed the basis of *ParCoLex*, involved linguists who verified the verb lemmas and their predicted paradigms (*cf.* Ljubešić et al., 2016; see Section 3). However, within the *SerboVerb* application, these verbs are internally labelled as unchecked. This label does not imply that the paradigms of these verbs are entirely accurate, nor does it mean that they will remain unchecked. The decision was made to prioritize the verification of the newly generated forms to speed up the process of verifying the entire database. As a result, the verification of the paradigms of these verbs will be conducted after the verification of the newly generated forms. Additionally, special attention is given to verifying the imperfect tense and passive participle forms of these verbs, as the imperfect tense forms of some verbs were not generated simultaneously with the other inflected forms. Furthermore, the transitivity of some verbs was not marked in the manually annotated database, necessitating thorough verification of the resulting paradigms.

The accuracy verification of the forms is carried out in rounds, which are organized once a year. Each round covers 4,000 verbs and is conducted in two stages, with each stage lasting four months. In the first stage, a group of 10 trained and experienced native speakers of Serbian receive a batch of verbs and, following detailed instructions provided by the *SerboVerb* team, verify, correct, and supplement their paradigms. If there are no errors in the generated verb paradigm, the collaborator marks the verb with an appropriate flag. If a collaborator encounters a problem or has a doubt about a particular inflected form, they flag it for further review. In the second stage, the *SerboVerb* team coordinators provide additional verification. They publish verified verbs that are ready for publication and simultaneously review, correct, and supplement verb paradigms for which collaborators had doubts. At the time of writing this paper, 20,158 verbs have been reviewed. The remaining verbs will be reviewed in the following rounds.

4.2 Dictionary Module

The Dictionary module is a component of the application used to search and display the multilingual dictionary database of the *SerboVerb* language resource. The database is also structured in XML format according to TEI Guidelines, since it is merged with the *SerboVerb* inflection database. It can be searched in the same way as the conjugation module database (see Section 4.1). In the dictionary module, users can enter a verb lemma or any inflected form of the verb, and receive a bilingual dictionary description of the desired Serbian verb in one of the 36 available languages (both European and non-European).

The dictionary description contains one or more senses introduced by a gloss, marked with one or more labels, followed by one or more equivalents, each of which may also contain one or more labels, and finally, one or more translated examples. Therefore, this is a dictionary description in which Serbian is the source language, and other languages are the target languages (TL). Users can choose the TL they need for the first dictionary look-up, and that language will remain as the default language for subsequent searches in the dictionary module.

The development of the multilingual database started in autumn 2022. During the first phase, basic equivalents were added for a list of 1,844 core Serbian verbs (previously mentioned in Section 4.1), extracted from the annotated *SerboVerb* database. These verbs are representative enough for most L2 speakers up to level B2 (Upper Intermediate level) according to the Common European Framework of Reference for Languages. The selection criteria for these verbs are not discussed in this paper. Currently, the entries for core verbs have basic equivalents in Albanian, English, French, German, Portuguese, Russian, Spanish, and Ukrainian. However, equivalents for Czech, Danish, Italian, Norwegian, Polish, Slovak, Swedish, and Turkish are still being added. Insertion of equivalents in Bulgarian, Greek, Hungarian, Macedonian, Romanian, and Slovene started in April 2023. Equivalents for other languages such as Arabic, Chinese, Dutch, Estonian, Farsi, Finnish, Hebrew, Japanese, Korean, Latvian, Lithuanian, Romani, Rusyn, and Swahili are being prepared for autumn 2023. The insertion of equivalents is carried out by a team of collaborators who possess a minimum proficiency level of C1 (Advanced level) in the respective languages. Each group comprises one to four members, and their work lasts for up to four months. Once the equivalents have been entered for all languages, the coordinators of the *SerboVerb* team plan to conduct a manual cross-check of all entries to ensure that the dictionary module is consistent across all languages.

4.3 Gamification Module

The Gamification module is designed to provide an interactive way for L2 Serbian speakers to learn, practice, and improve their verb inflection skills. Development of the module began in autumn 2022 and is currently ongoing. The initial content was created by the *SerboVerb* team, and external collaborators with expertise in teaching Serbian as an L2 or heritage language have been engaged to prepare additional education materal. This material is expected to be added to the module in the near future, further enhancing its value as a learning tool.

The educational material in the gamification module is presented as a series of learning games, with various types available (see Mihaljević & Hudeček, 2022), such as quizzes, drag-and-drop exercises, fill-in-the-blanks, find-the-match, puzzles, crosswords, memory games, and hangman games. External collaborators may also contribute unique games. All games contain at least two gamification elements, such as levels, scoring, leaderboard, and time limit. The educational material is classified according to the required language competencies in Serbian as an L2 needed to solve them and is marked accordingly. Users are provided with a score of their performance to boost motivation. Based on their performance, they are ranked against other users who have completed the same game. Additionally, some games have a time limit.

All of the educational material is prepared using open-access gamification platforms that are freely accessible. As a result, this module is the least consistent in terms of content and presentation. However, this is not a problem, as the involvement of different and numerous collaborators ensures a variety of approaches and a wider reach in the use of *SerboVerb* app and its resources.

5. Multifunctionality of *SerboVerb*

The differences among the Serbian conjugators discussed in Section 2 can significantly influence the user's experience. Hence, it was crucial to take these aspects into account when creating the *SerboVerb* application as they can greatly impact the efficiency and effectiveness of the end product. Moreover, comprehending the advantages and limitations of each conjugator could help the *SerboVerb* team develop an application that cater for user's specific requirements better.

As previously demonstrated in the literature (Tarp, 2008), according to Function Theory, users for whom a particular language is a foreign language (in this case, Serbian) may have a primary or secondary need for inflection information, which can be satisfied by seeking help from a dictionary in all extralexicographic situations, including communicative (receptive and productive) and cognitive ones. The following subsections illustrate how the *SerboVerb* application provides data based on which appropriate information can be derived in all three mentioned situations.

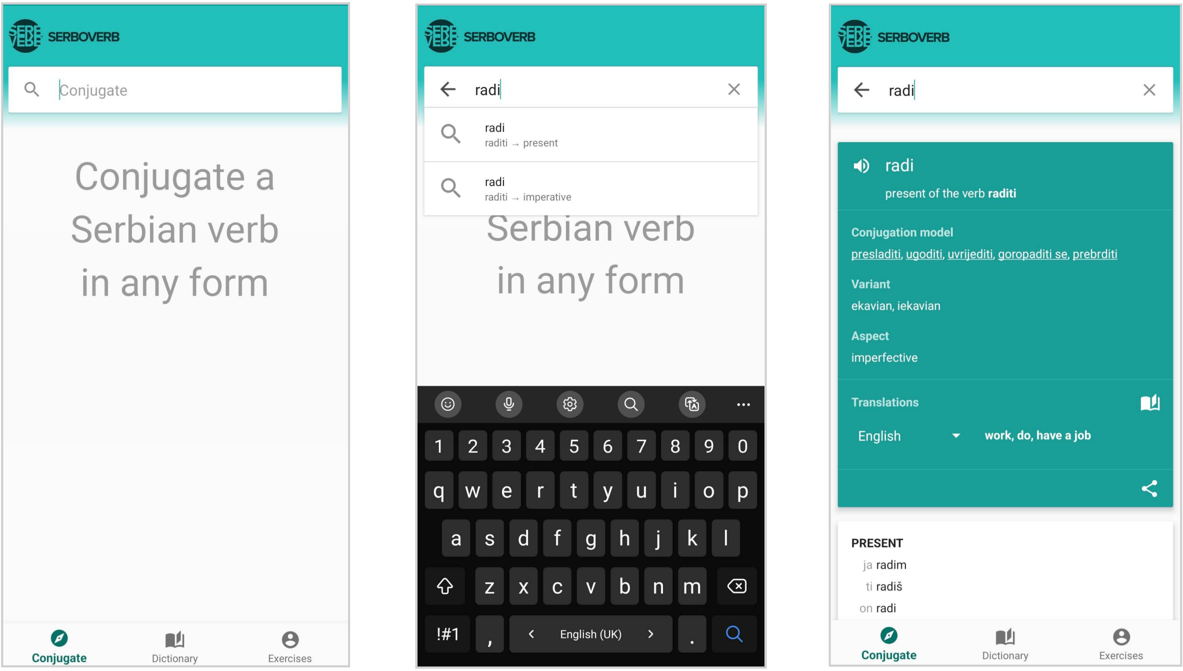
5.1 Receptive Functions

If an L2 Serbian speaker is not familiar with or unable to recognize a certain inflected form of a verb, they can search for it in the *SerboVerb* web-based or mobile application without creating an account and completely free of charge. Within the Conjugation module, the user can enter the unrecognized form in the search field (see Figure 2a). The searched form can be in its lemma or non-lemma form. Through the autocomplete feature (see Figure 2b), the application will suggest one or multiple possible results, along with a brief morphosyntactic identification of the form. This feature assists the user in identifying the tense in which the searched verb form is located within the written or spoken extralexicographic context where they first encountered the verb, and provides the corresponding result. By clicking on the appropriate form, the user can access the paradigm of the selected result (see Figure 2c).

The result page consists of two components: a shaded identification block (see Figure 2c) and a brighter paradigm block (see Figure 2c & 3a). The first block provides the user with more reception-relevant data: firstly, it identifies the searched inflected form by placing it in a specific tense from the verb paradigm; secondly, it indicates whether the verb is limited to Ekavian or Ijekavian areas or can be used in all varieties of Serbian standard language. If the usage is limited to a specific area, a cross-reference to the counterpart form is provided to the user. Then, the aspectual value of the verb is presented to the user. Finally, the identification block also provides basic equivalents for 1844 core Serbian verbs, which provide the lexical meaning of the searched verb and facilitate its reception. If the user needs a language that is not provided by default, they can select the appropriate language from the drop-down menu list (Figure 3b). If the user requires additional information (e.g., on the usage of the verb in context) to further understand its lexical meaning, they can click on the icon that opens the Dictionary module, which offers more data from the dictionary database. The second element in this result page provides the complete paradigm of the searched verb. By scrolling down, the user can locate the searched form within the full inflectional paradigm.

5.2 Productive Functions

In situations where an L2 Serbian speaker is not familiar with the inflectional paradigm of a certain verb, or is unsure about it, but needs it for text production purposes, they can search for the verb's inflectional paradigm in the Conjugation module. As in receptive situations (*cf.* Section 5.1), the search can be performed based on the form of the verb that the user first recalls. This can be either the lemma form or any inflected form. The search result page displays a shaded identification block and a brighter paradigm block. Unlike in receptive needs, where the identification block carries more informative weight, in productive needs, the primary importance of the data is in the paradigm block. In this block (see Figure 3a), the user scrolls down to search for the verb tenses that they require in the production situation. The verb tenses are arranged so that the most frequent ones in contemporary Serbian, and the ones that are first learned in Serbian L2 courses (present, imperative, perfect, and future tense), come first. Regarding the data in the paradigm, it should be noted that the user can also obtain information about all the compound tenses, as well as the paradigm of reflexive verbs, where forms have different word order depending on whether the subject is present or not. Furthermore, the graphical interface is designed to enable the user to quickly scroll through the paradigm, both up and down and left and right (especially when displaying forms for the appropriate gender). Moreover, in the identification block, the user can check whether the searched verb is used in the appropriate Ekavian or Ijekavian area and what aspectual value it carries. Then, if they need information about the use of the verb in context, they can switch to the Dictionary module (see Figure 3c).



radi

present of the verb **raditi**

Conjugation model

presladiti, ugoditi, uvrijediti, goropaditi se, prebrditi

Variant

ekavian, iekavian

Aspect

imperfective

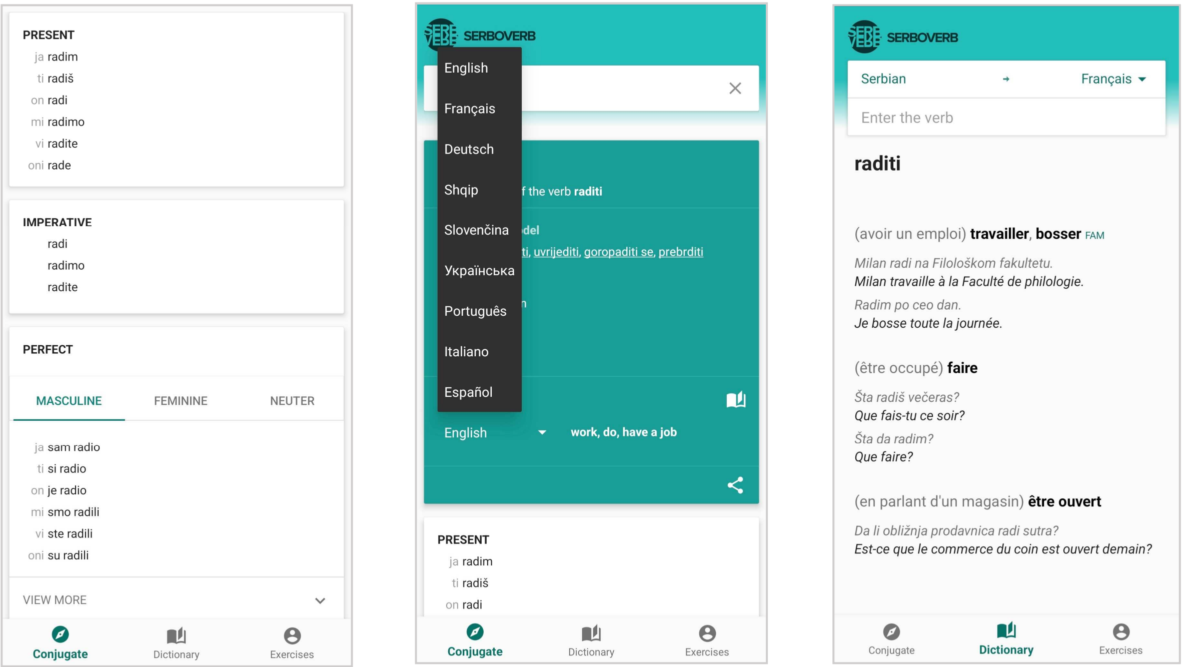
Translations

English — work, do, have a job

PRESENT

ja radim
ti radiš
on radi

Figure 2: The conjugation module in the Android app version:
a) the homepage, b) a search action, c) the conjugation result page



PRESENT

ja radim
ti radiš
on radi
mi radimo
vi radite
oni rade

IMPERATIVE

radi
radimo
radite

PERFECT

MASCULINE FEMININE NEUTER

ja sam radio
ti si radio
on je radio
mi smo radili
vi ste radili
oni su radili

PRESENT

ja radim
ti radiš
on radi

raditi

(avoir un emploi) **travailler**, **bosser** FAM

Milan radi na Filološkom fakultetu.
Milan travaille à la Faculté de philologie.

Radim po ceo dan.
Je bosse toute la journée.

(être occupé) **faire**

Šta radiš večeras?
Que fais-tu ce soir?

Šta da radim?
Que faire?

(en parlant d'un magasin) **être ouvert**

Da li obližnja prodavnica radi sutra?
Est-ce que le commerce du coin est ouvert demain?

Figure 3: The conjugation and the dictionary modules in the Android app version:
a) the conjugation result page, b) switching a language, c) dictionary view

5.3 Cognitive Functions

L2 Serbian speakers can use the *SerboVerb* application not only when they need to solve a communication–based problem but also in cognitive situations, where they want to independently confirm or acquire knowledge about the paradigms or inflected forms of certain verbs they are uncertain about. In such cases, the *SerboVerb* resource in the application can be searched in the same way as described in previous subsections (see Sections 5.1 & 5.2). An additional feature that is not relevant to the previous two functions is the cross-reference to five randomly selected verbs from the *SerboVerb* inflection database that belong to the same inflectional pattern. By comparing the paradigms of similar verbs, users can acquire and expand their knowledge of the conjugation properties of individual inflectional classes. Additionally, cognitive functions are satisfied through the use of a gamification module, described in Section 4.3.

6. Future Development

As stated in the previous sections, *SerboVerb* is an application and a language resource that are still in development. Further development is based on user needs, gathered through log file analysis and direct communication with users. So far, several needs have been identified for which both short-term and long-term plans have been made.

Despite the fact that the search field indicates that the verb database should be searched using Latin characters, it has been noticed that users occasionally search for verbs using Cyrillic alphabet. As a result, the short-term plan involves introducing an algorithm in the search field that instantly transliterates Cyrillic letters into Latin characters, enabling users to input forms in their preferred alphabet. Furthermore, the short-term plan entails conducting further verification of the inflection database to ensure that all verbs become available to users in the near future and that the data is as reliable as possible.

In the long term, the plan is to expand the multilingual database by adding examples for core verbs in their basic meanings, expressed in lexically simple and concise syntactic patterns, and translated into available languages. The gamification module will also receive regular updates with new content to cater for different types of users. Finally, a pronunciation module will be developed that enables users to hear the correct pronunciation of the searched form and other forms in the verb paradigm.

7. Conclusion

This paper introduced an innovative language resource called *SerboVerb* and its accompanying application, which enable L2 Serbian speakers to effectively and dynamically meet all their needs related to verb inflection in various communicative and cognitive situations. As demonstrated, the application was designed to be freely and openly accessible, with a comprehensive database of verbs and their inflected forms, continuously updated and expanded, with flexible search capabilities, and an effective and highly readable graphical interface that presents a large amount of data in a clear manner. Additionally, the main inflection database is linked with other resources, such as dictionaries and educational content, further enhancing its utility. By relying on this more trustworthy tool than on previous conjugators, L2 Serbian speakers now have access to a valuable resource that includes Serbian, a language often considered low-resourced, thus enriching the electronic lexicographic landscape.

8. Acknowledgements

The paper was produced as part of the Scientific Research Program under a contract no. 200167, signed by the University of Belgrade, Faculty of Philology, and the Ministry of Science, Technological Development and Innovations of the Republic of Serbia.

9. References

Babić, B. (2021). *Unutarjezičke greške u nastavi srpskog jezika kao stranog.* Novi Sad: Filozofski fakultet.

Grønvik, O. & Smith Ore, Ch-E. (2013). What should the electronic dictionary do for you – and how? In I. Kosem et al. (eds.) *2013. Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.* Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 243–260.

Jelaska, Z. (2005). Glagolske vrste. In Z. Jelaska et. al. (eds.) *Hrvatski kao drugi i strani jezik.* Zagreb: Hrvatska sveučilišna naklada, pp. 170–185.

Krajišnik, V. (2011). Rječnik u nastavi srpskog kao stranog jezika. *Anali Filološkog fakulteta*, 23(2), pp. 245–258.

Krstev, C. (1997). *Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije.* PhD thesis. Beograd: Matematički fakultet.

Krstev, C. (2008). *Processing of Serbian: Automata, Texts and Electronic Dictionaries.* Belgrade: Faculty of Philology.

Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (eds.) *Electronic lexicography.* Oxford: University Press, pp. 343–362.

Ljubešić, N. & Klubička, F. (2014). {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In P*roceedings of the 9th Web as Corpus Workshop (WaC 9).* Gothenburg: Association for Computational Linguistics, pp. 29–35.

Ljubešić, N., Klubička, F., Agić, Ž. & Jazbec, I.-P. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* Portorož: European Language Resources

Association (ELRA), pp. 4264–4270.

Marjanović, S. & Radosavljević, N. (2019). *Srpski glagoli: Konjugacijski rečnik glagola srpskoga jezika.* Beograd: Klett.

Marjanović, S. (2016a). Glagolska fleksija u rečnicima. O recepcijskim i produkcijskim potrebama stranih korisnika. In V. Krajišnik et al. (eds.) *Srpski kao strani jezik u teoriji i praksi III.* Beograd: Filološki fakultet, pp. 261–277.

Marjanović, S. (2016b). Glagolska fleksija u dvojezičnom rečniku sa srpskim kao ciljnim jezikom. *Zbornik Matice srpske za filologiju i lingvistiku*, 59(2), pp. 109–128.

Marković, A. (2014). Gramatika u srpskim rečnicima. In R. Dragićević (ed.) *Savremena srpska leksikografija u teoriji i praksi.* Beograd: Filološki fakultet, pp. 69–91.

Mihaljević, J. & Hudeček, L. (2022). Model for developing educational games based on data from dictionary structure. *Studia lexicographica,* 16(30), pp. 111-133.

Miletic, A. (2017). Building a morphosyntactic lexicon for Serbian using Wiktionary. In *6e édition des Journées d'étude toulousaines : Les interfaces en Sciences du Langage. Actes des Journées d'études toulousaines 18 et 19 mai 2017.* Toulouse: Université Toulouse Jean Jaurès, pp. 30‒34.

Miletic, A. (2018). *Un treebank pour le serbe : constitution et exploitations.* PhD thesis. Toulouse: Université Toulouse – Jean Jaurès.

Miletic, A., Stosic, D. & Marjanović, S. (2017). ParCoLab: A Parallel Corpus for Serbian, French and English. In K. Ekštein & V. Matoušek (eds.) *Text, Speech, and Dialogue. 20th International Conference, TSD 2017, Prague, Czech Republic, August 27–31, 2017 Proceedings. Lecture Notes in Artificial Intelligence* 10415. Cham: Springer, pp. 156–164.

Nikolić, M. (2000). *Obratni rečnik srpskoga jezika.* Novi Sad – Beograd: Matica srpska, Institut za srpski jezik SANU, Palčić.

Rujević, B. (2022). *Rečnici u digitalnom dobu – informatička podrška za srpski jezik.* PhD thesis. Beograd: Filološki fakultet.

Sečujski, M. & Delić, V. (2011). *Automatska konverzija tekstualnih informacija u govor.* Beograd: Vojnotehnički institut.

Simonsen, H. K. (2014). Mobile Lexicography: A Survey of the Mobile User Situation. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus. 15-19 July 2014, Bolzano/Bozen*, pp. 249–261.

Simonsen, H. K. (2015). Mobile Lexicography: Let's Do it Right This Time! In I. Kosem et al. (eds.) *2015. Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom.* Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 84–104.

Stanković, R., Krstev, C., Lazić, B. & Škorić, M. (2018). Electronic Dictionaries – from File System to *lemon* Based Lexical Database. In *Proceedings of the 11th*

International Conference on Language Resources and Evaluation (LREC 2018) - *W23 6th Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018)*, Miyazaki, Japan, May 7-12, 2018.

Tadić, M. & Fulgosi, S. (2003). Building the Croatian Morphological Lexicon. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*. Budapest: Association for Computational Linguistics, pp. 41–45.

Tadić, M. (2003). *Jezične tehnologije i hrvatski jezik.* Zagreb: Ex Libris.

Tarp, S (2008). *Lexicography in the Borderland between Knowledge and the Non-Knowledge*. Tübingen: Niemeyer.

Tarp, S. (2012). Theoretical challenges in the transition from lexicographical p-works to e-tools. In S. Granger & M. Paquot (eds.) *Electronic lexicography.* Oxford: University Press, pp. 107–118.

Tošović, B. (2012). Morfogeneratorska tipologija glagola (na korpusu Rečnika srpskog jezika). *Slavistika*, 16, pp. 135–142.

Tošović, B. (2014). Automatsko kodiranje pomoću Morfogeneratora. *Slavistika*, 18, pp. 207–214.