A Survey of Resources and Methods for Natural Language Processing of Serbian Language

Ulfeta Marovac[1†], Aldina Avdić[1*†] and Nikola Milošević[2]

[1]Department of Technical and Technological Sciences, State University of Novi Pazar, Vuka Karadžića 9, Novi Pazar, 36300, Serbia.

[2*]Research and Development, Bayer Pharmaceuticals, Müllerstrasse 178, Berlin, 13353, Germany.

*Corresponding author(s). E-mail(s): apljaskovic@np.ac.rs;
Contributing authors: umarovac@np.ac.rs;
nikola.milosevic@bayer.com;
†These authors contributed equally to this work.

Abstract

The Serbian language is a Slavic language spoken by over 12 million speakers and well understood by over 15 million people. In the area of natural language processing, it can be considered a low-resourced language. Also, Serbian is considered a high-inflectional language. The combination of many word inflections and low availability of language resources makes natural language processing of Serbian challenging. Nevertheless, over the past three decades, there have been a number of initiatives to develop resources and methods for natural language processing of Serbian, ranging from developing a corpus of free text from books and the internet, annotated corpora for classification and named entity recognition tasks to various methods and models performing these tasks. In this paper, we review the initiatives, resources, methods, and their availability.

Keywords: natural language processing, text mining, language resources, Serbian language

1 Introduction

The Serbian language is a south Slavic language currently actively spoken by about 12 million people worldwide. It is one of four mutually intelligible varieties of pluricentric language called Serbo-Croatian (other varieties include Croatian, Bosnian, and Montenegrin). Serbo-Croatian languages are morphologically rich (Delić et al., 2010), containing many inflections of words, due to three genders, seven cases for nouns, and seven tenses for verbs, whose inflections are followed by other parts of speech and word types, as well as twelve sound changes occurring in word inflections (Klajn, 2005). Serbian is also the only European language that is formally digraphic and whose speakers are functionally digraphic, using both Cyrillic and Latin alphabets (Magner, 2001). The majority of Serbian speakers live in Serbia (6,330,919 based on 2011 census[1]), but a significant number of speakers also live in Montenegro, Bosnia and Herzegovina, Croatia, Macedonia, Slovenia, Albania, Hungary, Austria, Sweden, Germany, and other countries. The Serbian language is an official language in Serbia, Bosnia and Herzegovina, Montenegro, while it is recognized as a minority language in Croatia, Macedonia, Romania, Hungary, Slovakia, and the Czech Republic. Variants of the Serbo-Croatian language (Serbian, Croatian, Bosnian, and Montenegrin) are spoken by about 19 million people, and therefore the importance of these languages are quite significant (Eberhard et al., 2022).

Natural language processing is a branch of artificial intelligence that examines methods to analyze, process and ultimately make natural languages understandable for computers (Reshamwala et al., 2013). Therefore, the field is addressing many challenges related to human/natural languages. Even though a majority of work in the field is predominantly done on the English language, there has been also work on other languages.

High morphological complexity, variety of word inflection, and relatively low amount of resources available for Serbian and Serbo-Croatian pose a challenge for natural language processing and language technologies. The morphological richness of Serbo-Croatian makes it particularly interesting for examining how natural language processing methods perform on languages with a variety of inflection and how to efficiently handle word inflection in morphologically rich and low-resource languages. In sense of language technologies and natural language processing, Serbian cannot be viewed in isolation, as differences between Serbian, Croatian, Bosnian, and Montenegrin are small, and often approaches developed for one of these variants would perform well on others. Despite these challenges, there have been several initiatives, organizations, and significant academic work performed to address some of the specific challenges in Serbian. A number of resources and corpora for syntactic analysis, classification, and named entity recognition were developed, as well as a number of approaches for document analysis, classification, semantic similarity, and even analysis of rhetorical figures such as similes.

[1] https://data.stat.gov.rs/Home/Result/3102010401?languageCode=en-US

The development of digital lexical resources is an important and strategic task for every language and should have national priority. The results of natural language processing are dependent on the quality and volume of available digital resources, as well as the availability and comprehensiveness of tools for processing digital resources (Nenadić, 2004). Our goal is to collect available resources and methods for processing textual data in the Serbian language, describe them, and identify shortcomings that can be advanced and expanded with the most needed resources according to the development trends of NLP. To the best of our knowledge, this is the first review of NLP resources and methods for Serbian of newer date and scope.

A brief history of NLP resources and method development in Serbia

The development of the first digital corpora in the Balkans started shortly after the development of the first digital corpora in the world, and it was started by the psychologist Djordje Kostić, in 1957. with the goal of developing language technologies for speech recognition and machine translation from the Serbo-Croatian language. This corpus was developed until 1962, however, it was not digitally processed, so the first digital corpus was published in Zagreb in 1967. This corpus contained the epic Osman by Ivan Gundulić prepared by Željko Bujas. Development of corpora and corpus linguistics in the western Balkans in the period between 1950 and 1990 is presented by Dobrić (2012). Language resources and tools that were mainly developed at the Faculty of Mathematics, University of Belgrade, until the year 2003, have been previously reviewed (Vitas et al., 2003b,a). During the project called META-NET in 2012, the analysis of the language resources for 23 official languages of the European Union was done, and as a part of white pages was published book "The Serbian language in the digital age" (Vitas et al., 2012). The Regional Linguistic Data Initiative (ReLDI) project has made a significant contribution to promoting the relevance and importance of open language resources for Serbian and related languages (Samardžić et al., 2015). The open and freely available language resources for processing the Serbian language, developed within the ReLDI project or independently built, are briefly presented by Batanović et al. (2020).

In this paper, we aim to review corpora, resources, methods, models, and tools that were developed over time for the Serbian language. We intentionally limit this review to the Serbian language only. While we agree that some approaches do work as well on related languages in the Serbo-Croatian group of languages, there are still small differences between them, that would make evaluation and comparison of the resources and methods challenging.

2 Review methodology

To cover all authors who deal with natural language processing in Serbian, we started with the National Repository of Dissertations in Serbia (https://nardus.mpn.gov.rs/). By using keywords such as natural language

processing, NLP, text mining, text data processing, computational linguistics, electronic dictionaries, corpora, sentiment analysis, emotional analysis, text classification, lexical resources, and other synonyms and related terms, we identified dissertations that contain these keywords. From the most relevant dissertations, those that deal with natural language processing in Serbian were selected. Additionally, a set of dissertations were identified by searching for known NLP scientists, supervisors, and groups at Serbian universities. Dissertations were identified by searching for known scientists acting as a supervisor or a member of the thesis committee. A total of 29 dissertations in the field of natural language processing were selected. By analyzing the dissertations and references cited in them, we identified 316 papers indicating NLP for Serbian. Further searches were conducted on Google Scholar for prominent authors (or author groups) and selected topics.

We reviewed the dissertations and papers we identified, excluding those that did not pertain to natural language processing in Serbian, and classified them based on their topic and date of publication. In this review, we follow this classification, with each section covering a broad area of natural language processing. The content in each section is primarily arranged in chronological order.

3 Corpora

A corpus is a set of machine-readable texts representing a sample of a language or text type. Corpora can be classified based on their parameters such as medium, scope (size), domain, purpose, period, source, method of annotation, number of languages involved, etc. (Vitas et al., 2003b). Given that corpora can include texts in one or more languages, they are divided into monolingual and multilingual corpora. According to this classification, we will present the corpora of the Serbian language.

3.1 Monolingual corpora

The Diachronic Corpus of the Serbo-Croatian Language (DCSCL, Table 1) of Professor Kostić digitized in 2003 contains texts from the period from the 12th to the 20th century, divided into five-time samples. The corpus comprises 11 million words that have been manually annotated with lemmas and information on various morphological categories such as gender, number, case, person, tense, and more (Kostić, 2014). In 1981, the NLP group at the Faculty of Mathematics (NLP_MATF) initiated the development of a corpus for the contemporary Serbian language. The first version of this corpus, named *"The Untagged Corpus of Contemporary Serbian Language"* (UCCSL, Table 1), was created in 2003. This corpus contains literature published during or after the 20th century and lacks any annotations. Subsequently, the corpus was enhanced by incorporating bibliographic information into the corpus texts, and this new version was called *"SrpKor2003"* (SrpKor2003, Table 1) (Krstev and Vitas, 2005; Utvić, 2014).

Most of the monolingual corpora have morphosyntactic annotation and bibliographic information about the corpus texts. Morphosyntactic annotation is a linguistic annotation that adds tags to the token: type of speech (Part of Speech Tagging), canonical form or lemma (lemmatization), and morphological word categories. By expanding SrpKor2003, a new version of the corpus of contemporary Serbian *"SrpKor2013"* (SrpKor2013, Table 1) was created, which contains literary and artistic texts by Serbian writers in the 20th and 21st centuries, as well as scientific texts, administrative texts, and general texts. *The Corpus of Contemporary Serbian* contains bibliographic data and it has been automatically morphosyntactically annotated (with part-of-speech and lemmas). It contains more than 122 million words. Its subset *"Lematized Corpus of the Modern Serbian Language"* (SrpLemKor, Table 1) contains 3.7 million corpus words. Both corpora are available with registration under a license (Popović, 2010; Utvić, 2011).

Among the available corpora of the Serbian/Serbo-Croatian language at the Faculty of Mathematics of the University of Belgrade[2], there are also the following monolingual corpora. *Henning's Corpus of Serbo-Croatian* (HennC, Table 1) consists of approximately 700,000 words of Serbo-Croatian. The texts are taken from modern Yugoslav fiction and all Serbo-Croatian-speaking areas are represented (Serbia, Croatia, Montenegro, and Bosnia-Herzegovina) (Corpora etc, 1992). *The Untagged Corpus of Vuk's Folk Proverbs* (UnVukC, Table 1), containing folk proverbs along with Vuk's comments on them (Krstev, 1997). Besides this corpus, Vuk's collection of similes has been augmented by employing grammatical rules, machine learning, and manual review. As a result, a corpus of contemporary similes containing 852 similes was developed (VukSimC, Table 1)[3] (Milosević and Nenadić, 2016, 2018). *Electoral Crisis 2000* corpus, which includes the entire webcasts of the daily newspaper "Politika" from September 10th to October 20th, 2000, and the *Labeled corpus of the Serbian language*, which consists of texts with a minimal set of structural labels, lack the detailed information on size and are available on the same source[4].

There are smaller corpora that have been collected mostly for specific domains (medicine, law, etc.) and particular purposes (name entities recognition, semantic similarity, etc.). Among them are the corpora *MRCOR1* and *MRCOR2* (Table 1) consisting of medical reports reviewed from 32 medical centers in Serbia. The primary data set contains 2212 medical reports with a diagnosis of measles. The other dataset consists of 2000 medical reports with ten different types of diagnoses. Medical and non-medical terms are manually marked in the medical reports. For each medical report is assigned a diagnosis code (Avdić et al., 2020). A corpus (DMRC, Table 1) of 100 discharge lists and 50 reports from doctors from the Faculty of Dentistry at the University of Belgrade was used to evaluate the system's effectiveness in automatically analyzing temporal expressions of medical narrative texts. Previously, the texts

[2]http://www.korpus.matf.bg.ac.rs/prezentacija/korpusi.html
[3]https://ezbirka.starisloveni.com
[4]http://www.korpus.matf.bg.ac.rs/prezentacija/korpusi.html

Table 1 Monolingual Serbian corpora

Corpus label	Text type	Number of unit	Annotation [1]	Reference
DCSCL	general	11 000 000 words	S, L, M;MA	(Kostić, 2014)
UCCSL	literary	22 200 000 words	U	(Krstev and Vitas, 2005; Utvić, 2014)
SrpKor2003	literary	22 200 000 words	B; MA	(Krstev and Vitas, 2005; Utvić, 2014)
SrpKor2013	literary-artistic scientific	122 000 000 words	B,L,PoS; MA,AA	(Popović, 2010; Utvić, 2011)
SrpLemKor	literary-artistic scientific	3 700 000 words	B,L,PoS; MA,AA	(Popović, 2010; Utvić, 2011)
HennC	literary	728 952 words	B; MA	(Corpora etc, 1992)
UnVukC	literary	6919 proverbs	U	(Krstev, 1997)
VukSimC	literary	852 similes	U	(Milosević and Nenadić, 2016, 2018)
MRCOR1	medical reports	2212 reports	MT,NMT; MA	(Avdić, 2021)
MRCOR2	medical reports	2000 reports	MT,NMT; MA	(Avdić, 2021)
DMRC	medical reports	150 reports	U	(Jaćimović et al., 2015)
LAWC	text of laws	59167 texts	S; MA	(Petrović and Stanković, 2019)
LTC	text of laws	7981446 tokens	S,NE; MA	(Vasiljević, 2015)
ATC	newspapers scientific	200 000 words	M; MA	(Sečujski and Delić, 2008)
SrpNEval	news	89425 words	NE; AA, MA	(Krstev et al., 2012)
NormTagNER	social network	89 425 words	M,NE;MA, AA	(Miličević and Ljubešić, 2016)
SETimes.SR	news	86 726 tokens	S, L, PoS, SD, NE; MA	(Batanović et al., 2018; Batanović V et al., 2018)
paraphrase.sr	news	1194 pairs	SS; MA	(Batanović et al., 2011)
STS.news.sr	news	1192 pairs	SS; MA	(Batanović et al., 2018)
SrELTeC	novel	5 263 071 words	S, L, PoS, NE; AA	(Stanković et al., 2021)
SrpELTeC-gold	literary	330 119 tokens	NE; AA,MA	(Todorović et al., 2021)
SrpKor4Tagging	literary	342 803 tokens	PoS, L; AA	(Stanković et al., 2020)
RudKorP	academic text	2 340 000 words	PoS, L; AA	(Utvić et al., 2019)
TorlakKor	culture interview transcript	498021 tokens	MS,A,L; MA,AA	(Vuković, 2020)
COPA-SR	question answer	1000 premises	P; MA	(Ljubešić et al., 2022)
CorFoA	biographical interviews transcripts	171552 tokens	MS, L; AA	(Lemmenmeier-Batinić et al., 2021)
MLNews	news	659084 tokens	MS, L; AA	(Bogetić and Batanović, 2020a)
MLN-COM	news comments	878482 tokens	MS, L; AA	(Bogetić and Batanović, 2020b)
srWaC	web text	554627647 tokens	MS, L, SD; AA	(Ljubešić and Klubička, 2016; Ljubešić and Klubička, 2016)
CorLeg	laws	5 files	U	(Bogdanović and Tošić, 2022)

[a]Annotation target: U - unannotated, MS - morphosyntactic, L - lemmatization, M - morphological categories, PoS - Part of Speech Tagging, S - structural, NE - named entity, UK - unknown annotation, A - accentuation, B - bibliographic, SD - syntactic dependencies, P - plausibility, MT - medical terms, NMT - non-medical terms, SS - semantic similarity; Annotation type: AA - automated annotation MA - manually annotated

had been automatically de-identified, but the time expressions had not been changed (Jaćimović et al., 2015). The *LAWC* (Table 1) set of data includes a collection of 1120 texts of laws, segmented into a total of 59167 texts of individual articles of law (Petrović and Stanković, 2019). The corpus *LTC* (Table 1) consists of legal texts in electronic form that are available on the website of the National Assembly of the Republic of Serbia. The laws passed by the end of May 2014 contain 681 texts (Vasiljević, 2015).

AlfaNum which deals with automatic speech recognition (ASR), has built its resource *AlfaNum Text Corpus* (ATC, Table 1), characterized by morphological categories and accentuation and contains approximately 200,000 words (Sečujski and Delić, 2008). *The Named Entities Evaluation Corpus for Serbian* (SrpNEval, Table 1) consists of 2000 short news Serbian daily newspapers from 2005 and 2006. Both the Cyrillic and Latin official scripts for the Serbian language are used in the corpus (Krstev et al., 2012). *ReLDI-NormTagNER-sr 2.1* (NormTagNER, Table 1) is a manually annotated corpus of Serbian tweets for evaluation of tokenization, sentence segmentation, word normalization, morphosyntactic tagging, lemmatization, and named entities recognition of non-standard Serbian language (Miličević and Ljubešić, 2016). *SETimes.SR* (SETimes.SR, Table 1) is a reference training corpus of Serbian, which has been annotated on multiple levels. The texts in SETimes.SR were obtained from the multilingual parallel corpus *SETimes* (SETimes, Table 2), which is a collection of news articles from the now-defunct Southeast European Times news portal (Batanović et al., 2018), (Batanović V et al., 2018). Sentences from online press sources were collected for *The Serbian Corpus of Paraphrases* (paraphrase.sr, Table 2). A binary similarity score was manually assigned to each pair of sentences, indicating whether the sentences in the pair are semantically similar enough to be considered close paraphrases (Batanović et al., 2011). Another corpus for determining semantic similarity, *The Serbian Corpus of Short News Texts* - (STS.news.sr, Table 2), consists of 1192 pairs of sentences in Serbian that were collected from news sources on the internet (Batanović et al., 2018).

Old Serbian novels from the 1840s to the 1920s are collected in *SrELTeC* (SrELTeC, Table 1) and have been digitally preserved as part of the COST action CA16204 (Stanković et al., 2021). *ELTeC*'s section for Serbian contains 120 novels (Odebrecht et al., 2021). The novels have structural annotations, and sentence splitting, words are POS-tagged, lemmatized and seven classes of named entities are annotated. Some of the other resources available through the ELG[5] portal are *SrpELTeC-gold*, *SrpKor4Tagging*, and *RudKorP* (Table 1) . The corpus for training the recognition of named entities SrpELTeC-gold is a sub-corpus of the literary corpus of the Serbian language, marked with named entities by the SrpNER(Krstev et al., 2014) system (Todorović et al., 2021). The SrpKor4Tagging corpus was formed by combining literary and administrative texts in the Serbian language (Stanković et al., 2020). The RudKorP corpus contains texts in the field of mining and processing of mineral raw

[5]https://live.european-language-grid.eu/

materials, created at the University of Belgrade, Faculty of Mining and Geology (Utvić et al., 2019). There are several more corpora available through the Clarin.si [6] platform, which are shown at the bottom of Table 1. *TorlakKor* is a corpus of transcripts of interviews with the local population of Timok (an area in southeastern Serbia) (Vuković, 2020). The *COPA-SR* dataset (*Choice of Plausible Alternatives in Serbian*) is a translation of the English *COPA* dataset (Ljubešić et al., 2022). *CorFoA* is a corpus of Serbian forms of the address containing transcripts of biographical interviews with 19 participants (Lemmenmeier-Batinić et al., 2021). *MLNews* is a comprehensive corpus of news articles that are Serbian language-related. It is complemented with a separate corpus of citizens' online comments on the news articles, available as *MLN-COM* (Bogetić and Batanović, 2020a,b). The web corpus of the Serbian language *srWaC* was built by crawling the .rs top-level domain for Serbia in 2014 (Ljubešić and Klubička, 2016; Ljubešić and Klubička, 2016). *CorLeg* is a corpus of legislation texts of the Republic of Serbia which was created using a large number of Serbian Legislation texts gathered from the official website [7] (Bogdanović and Tošić, 2022).

3.2 Multilingual corpora

Multilingual corpora are a particular type of corpus that contains texts written in multiple languages. Parallel corpora include both the original texts and their translations into one or more other languages presented in such a way that their logical structure is explicitly connected at the document, chapter, paragraph, sentence, or word level. Table 2 shows multilingual corpora containing original texts or translations in the Serbian language. One of the early attempts to develop multilingual corpora is the creation of an alignment corpus of Plato's *"Republic"* containing translations into 21 languages, including Serbian. The corpus has been annotated at the sentence level and has been utilized for both tool development and automated alignment (Krstev and Vitas, 2011). The multilingual language resources and tools for extracting information from the language corpora of CEE languages (Central and Eastern European Languages), called MULTEXT-East[8] were created as part of the project Multext. The book "1984" is included in this parallel and sentence-aligned corpora, *Multext-East Corpora* (G. Orwell's "1984")(G.O.1984, Table 2), along with translations into several other languages. Krstev and Vitas (2011) created a translation of this novel into Serbian and a morphosyntactic annotation in the MULTEXT-East format, for which they had previously developed a specification for the Serbian language. The parallel corpus *Verne80days* (Table 2) contains the French original and 17 translations of Jules Verne's novel "Around the World in 80 Days". The alignment was performed on the subsentence level for each language (Vitas et al., 2008). *The Serbian-French Corpus* (SrpFranKor, Table 2), which consists of 31 subsentence-aligned texts

[6] https://www.clarin.si
[7] https://www.pravno-informacioni-sistem.rs/
[8] http://nl.ijs.si/ME/

that were originally written in French and then translated into Serbian and vice versa, is the first bilingual corpus in the Serbian language (Vitas and Krstev, 2006; Vitas et al., 2006). *ParCoLab* (Table 2) is a parallel online searchable corpus consisting of sentence-alignment texts in French, Serbian, English, Spanish, and Occitan. Each of these languages is at the same time a source language and a translation language (Balvet et al., 2014).

The Serbian-English Corpus (SrpEngKo, Table 2) is the second bilingual collection. It consists of English source texts aligned with their translations into Serbian, and visa-versa, as well as several aligned English and Serbian translations of literary texts originally written in French (Krstev and Vitas, 2011). The corpus *SETimes* is based on the articles posted on the news website SETimes.com. Bulgarian, Bosnian, Greek, English, Croatian, Macedonian, Romanian, Albanian, and Serbian are among the ten languages in which the news is available. Part of the SETimes, sub-corpus *BALKANTIMES* was used for the expansion of SrpEngKo (Batanović et al., 2018). Parallel texts from the fields of law, business, education, and health care are also added to SrpEngKor, resulting in the creation of the sub-corpus *Serbian-English Law Finance Education and Health* (SELFEH, Table 2). Almost 150 parallel texts make up SELFEH, which was utilized in term extraction and machine translation research as well as to test various taggers for the Serbian language (Utvić, 2011). Another Serbian-English corpus is *srenWaC* (Table 2), which consists of sentence-aligned parallel texts pulled from the .rs top-level domain (Ljubešić et al., 2016). In addition to the SrpFranKo and SrpEngKo bilingual corpora, a similar corpus was created for the German (SrpNemKor, Table 2). It contains 48,004 translated pairs of literary texts in Serbian and German, which are aligned to the sentence level. Available tools for annotation of named entities in texts in both languages as well as tools for terminology extraction were applied to the prepared parallel corpus (Andonovski et al., 2019; Andonovski, 2020).

Additionally, there are multilingual parallel corpora, some of which are displayed at the table's end (Table 2). *OpenSubtitles* is a database with about 4 million sentence-level translations of movies and television shows in more than 62 different languages (Tiedemann, 2012). *The Bosnian, Croatian, and Serbian Web Corpora* (BsHrSrWaC, Table 2) are top-level-domain web corpora. They were used to create a method for separating similar languages that is based on unigram language modeling on the crawled data only (Ljubešić and Klubička, 2014). The Twitter user dataset (Twitter-HBS, Table 2) consists of tweets and their language tag (Bosnian, Croatian, Montenegrin, or Serbian). The main goal of creating this corpus is discrimination between closely related languages at the level of Twitter users (Ljubešić and Rupnik, 2022). The *PE2rr* corpus includes source language texts from many fields, as well as automatically produced translations into a number of morphologically rich languages, post-edited versions of those texts, and error annotations of the post-edit processes that were carried out. This corpus contains texts in Spanish, German, Serbian,

Table 2 Parallel Serbian corpora

Corpus label	Text type	Number of unit	Annotation[1]	Languages	Reference
Plato's Republic	philosophical text	21 text translation	S, A; MA, AA	multilingual	(Vitas et al., 1998)
G.O.1984	general	100000 words	L, M, PoS, A;AA, MA	multilingual	(Krstev and Vitas, 2011)
Verne80days	literary	32 aligned texts	L, M, PoS, A; AA,MA	multilingual	(Vitas et al., 2008)
SrpFranKor	literary	1738752 words	S, A; AA,MA	Serbian, French	(Vitas and Krstev, 2006; Vitas et al., 2006)
ParCoLab	general	32 000 000 words	S, A; AA,MA	multilingual	(Balvet et al., 2014)
SrpEngKo	general	4.420.711 words	S,A; AA,MA	Serbian, English	(Krstev and Vitas, 2011)
SETimes	news	86 726 tokens	L,M,PoS; MA	multilingual	(Batanović et al., 2018)
SELFEH	law, finance, education, and health	2 000 000 words	L,M,PoS, A;AA, MA	Serbian, English	(Utvić, 2011)
srenWaC 1.0	web text	23139804 words	A;AA	Serbian, English	(Ljubešić et al., 2016)
SrpNemKor	literary	1 657 329 words	S,A; AA,MA	Serbian, German	(Andonovski et al., 2019; Andonovski, 2020)
OpenSubtitles	film translations	2 793 243 tokens	S,A; AA,MA	multilingual	(Tiedemann, 2012)
BsHrSrWaC	web text	894 000 000 tokens	M,L,S,D; AA,MA	multilingual	(Ljubešić and Klubička, 2014)
Twitter-HBS	social network	390268 texts	S; AA,MA	multilingual	(Ljubešić and Rupnik, 2022)
PE2rr	general	43938 words	S,ER,A; AA,MA	multilingual	(Popović and Arčan, 2016)
BERTić-data	general	8387681518 words	S; MA	multilingual	(Ljubešić and Lauc, 2021)
CLASSLA-Wiki	general	486258862 tokens	LN; AA	multilingual	(Ljubešić et al., 2021)

[a]Annotation target: U - unannotated, L - lema, MS - morphosyntactic, M - morphological categories, LN - linguistic, PoS - Part of Speech Tagging, UK - unknown annotation, S - structural, SD - syntactic dependencies, A - aligned, ER - error; Annotation type: AA - automated annotation MA - manually annotated

Slovene and English (Popović and Arčan, 2016). The *BERTić-data* text collection contains more than 8 billion tokens of mostly web-crawled text written in Bosnian, Croatian, Montenegrin, or Serbian. The collection was used to train the BERTić transformer model (Ljubešić and Lauc, 2021). The Wikipedia dumps of the Bosnian, Croatian, Macedonian, Montenegrin, Serbian, Serbo-Croatian, and Slovenian Wikipedias were collected in the comparable corpus *CLASSLA-Wikipedia* (CLASSLA-Wiki, Table 2). The linguistic annotation was performed with the classla package [9], on all levels available for a specific language (Ljubešić et al., 2021). Corpora for sentiment analysis are presented in a separate chapter.

4 Language resources

4.1 Dictionaries and terminologies

The term electronic dictionary considers the dictionary which is used for text processing. It consists of valuable information for solving problems of segmentation, morphological, and partly syntactic and semantic text processing (Vitas and Krstev, 2009). The automatic processing of text begins by analyzing individual words, which are the base units of the analyzed text. At times, individual words may not be the most appropriate base units for processing natural language. Therefore, there are two types of dictionaries: mono-lexemic, which consists of single words, and polylexemic which consists of multi-word units (Andonovski, 2020).

The international network of laboratories for computational linguistics, RELEX (Laporte, 2003), has created a model for building electronic morphological dictionaries that have been adopted by numerous organizations dealing with natural language processing. The *Unitex* [10] system works with electronic morphological dictionaries developed according to this model. These are dictionaries in DELA format (*Dictionnaires Electroniques du LADL - Laboratoire d'Automatique Documentaire et Linguistique*). In order to distinguish between monolexemic and polylexemic units, this electronic dictionary is organized into two separate subsystems: a dictionary of monolexemic units (DELAS - simple forms and DELAF - inflected forms) and a dictionary of polylexemic units (DELAC - compound forms, and DELACF - compound inflected forms).

Based on these models, within the Group for Language Technologies of the University of Belgrade, electronic morphological dictionaries of the Serbian language in Latin and Cyrillic (SrbMD) were built (Krstev, 1997; Vitas et al., 2003b,a; Krstev et al., 2006, 2010). According to (Mladenović, 2016), the SrpMD system currently contains 148,000 lemmas and over 1,000 final transducers that generate more than 5 million DELAF determinations. The tool Leximir (Stanković et al., 2011) is used as a dictionary management system. It is a multipurpose tool for supporting computational linguists in developing, maintaining, and exploiting e-dictionaries.

[9] https://pypi.org/project/classla/
[10] https://unitexgramlab.org/language-resources

The accentuation-morphological dictionary was created at the Faculty of Technical Sciences in Novi Sad and it contains over 4 million entries. It is used for context analysis within text-to-speech and automatic speech recognition systems for Serbian (Sečujski and Delić, 2008).

Ljubešić et al. (2015) presented *MWELex*, a multilingual lexical of Croatian, Slovene, and Serbian multi-word expressions (MWE) that were extracted from parsed corpora. The *srMWELex* lexicon v0.5 was automatically built during the short-term scientific mission inside the PARSEME COST action. It contains multi-word expression candidates extracted with the DepMWEx tool from the *srWaC* v1.0 web corpus. It consists of 22 290 entries and 3 273 369 multi-word units. The freely available morphological lexicon srLex is introduced in (Ljubešić et al., 2016). It is consisting of 105 359 lexemes and 5 327 361 (token, lemma, MSD) triples.

Miletic (2017) described the creation of a morphosyntactic e-dictionary for the Serbian language. It is derived from the Wiktionary edition for Serbo-Croatian, a manually POS-tagged corpus and specialized proposition list. This lexicon contains 1 226 638 million wordforms for 117 445 lemmas, corresponding to a total of 3 066 214 unique triples (wordform, lemma, MSD - morpho-syntactic description), and it is aimed for POS (part of speech) tagging and parsing tasks.

The DELAS-TOP and DELAS-PERS are dictionaries that respectively list geographic names and Serbian personal names (Krstev et al., 2008; Pavlović-Lažetić et al., 2004; Grass et al., 2002). The dictionary of geographic names DELA-TOP covers geographic concepts at the level of a high-school atlas (approximately 20.000 toponyms, oronyms, and hydronyms with their corresponding derivatives). The dictionary of personal names has been created from the list of the names of 1.7 million inhabitants of Belgrade as established in 1993. Based on this list, two dictionaries were constructed: DELA-FName for the first names, and DELA-LName for the last names (Vitas et al., 2003a).

The dictionary of librarianship and information sciences contains terminology (Kovačević et al., 2004) used in the theory and practice of librarianship, information sciences, and related fields in Serbian, English, and German. The online version of the dictionary currently contains: 40,000 definitions (approximately 14,000 in Serbian); 900 definitions or annotations of terms that are part of library standards; 2,300 acronyms of international and national organizations and institutions; 190 addresses of relevant websites[11].

The electronic geological dictionary (GeolISSTerm) is a specially prepared taxonomy of basic geological concepts and terms, and it is used for IT needs as an elementary resource in the formation of domains in the Geological Information System of Serbia (GeolISS)(Stanković et al., 2011).

(Vujičić-Stanković et al., 2014) extended the SrpMD by 636 entries of simple words and 612 entries of MWE (multi-word expressions) from the culinary domain.

[11] http://rbi.nb.rs/srlat/dict.html

Grljević (2016) provided several dictionaries for sentiment analysis in the field of education in her doctoral dissertation (sentiment words, domain-specific phrases, negation keywords, and stop words that are identified from the corpus). Negation signals, negative quantifiers, and particle intensifiers were added to the sentiment lexicon (Ljajić and Marovac, 2019). Similarly, for sentiment analysis, a domain-oriented stop words collection was created (Mladenović, 2016). In a separate chapter on sentiment analysis below, sentiment word lexicons and other lexicons used in sentiment analysis are described more.

Avdić et al. (2020) created medical dictionaries for Serbian: names of diagnoses (7942 entries), diagnosis code (14194 entries), Latin names of the diagnosis (3794 entries), therapies (2232 drugs and 1317 ampoules, 2255 different terms), symptoms for the diagnosis of measles B05 (95 entries), specialties (41 entries), abbreviations from the medical domain. Non-medical dictionaries created in the same research are a set of negation symbols in the medical domain, places, and names.

Ostrogonac et al. (2020) created a domain vocabulary of jobs in Serbian. It has two versions, one of 40 thousand, one of 80 thousand words, and 30 thousand lemmas, and they are included in Python library *nlpheart*.

The Serbian stop word dictionary (SSW dictionary) contains 1241 different stop words for the Serbian language. It was created based on the grammar of the Serbian as well as by comparing with available sets of stop words for the Serbian language and a set of stop words for the Croatian language. SSW dictionary for the Serbian language contains words in different forms of their appearance. A word type label accompanies each word. The SSW dictionary is available as a CSV file - SSWdictionary.csv. The file contains two columns: word and label. The label describes the type of words: auxiliary verbs (V), pronouns (PRON), adverbs (ADV), prepositions (PREP), conjunctions (CONJ), exclamations (EXCL), particles (PART) and abbreviations (ABBR)(Marovac et al., 2021).

The SrHurtLex (Stanković et al., 2020) is a lexicon created for the detection of abusive words in Serbian. It is created using the lexical database Leximirka, the system of Serbian morphologic dictionaries SrpMD, and The Dictionary of Serbian Language (DS) (Vujanić, 2007), where the multi-word expressions labeled in dictionaries as augmentative, pejorative, derogatory, vulgar, etc. were collected.

4.2 Ontologies

The term "ontology" originates from philosophy and it represents science about existing concepts (types of things) and their relations (Vujičić-Stanković, 2016). In computer science, ontology is a structure that describes concepts, their relations, and existing constraints. Their purpose is the automatic sharing and reuse of knowledge between humans and computer, and between computers. Both parts which are included in sharing process have to have a certain level of understanding of the exchangeable information.

The hierarchy of classes is called taxonomy. Commonly, ontology describes terms and relationships between them for a particular domain.

The semantic network which describes proper names and their relations is developed during Prolex project (Krstev et al., 2007). It consists of 2000 proper names, mainly names of states and their capital cities.

The RudOnto is a terminological resource developed at the Faculty of Mining and Geology in Belgrade, and it is the reference resource for mining terminology in Serbian. It is managed by a terminological information system, and intended to produce the derived terminological resources in subfields of mining engineering, such as planning and management of exploitation, mine safety or mining equipment management (Stanković et al., 2011).

Tomašević (2018) developed a mining domain ontology *RuDokOnto* for the purpose of collecting, describing, and systematization of mining project documentation throughout the phases of the mining project's life cycle in a way that links other related ontologies.

RetFig is a linguistic domain, descriptive, formal ontology for rhetorical figures in Serbian and describes 98 figures (Mladenović and Mitrović, 2013).

4.3 Word networks

In traditional dictionaries, lexical concepts are alphabetically ordered and there is a definition for all possible meanings for each of them. In WordNet, all words expressing a concept are grouped together in a set of synonyms (synset - synonymous set). *Serbian WordNet (SerWN)* (Krstev et al., 2004; Koeva et al., 2008) is the lexical-semantic net for Serbian. Its development started within the project BalkanNet (Mladenović et al., 2020), and when it finished in 2004, it had 8000 synsets. After that, the development of WordNet continued, especially in biological, biomedical, psycho-linguistic, and gastronomical domains, etc. Its structure is basically the same as PWN (Princeton Word Net (Miller and Fellbaum, 2007)), and it is organized using nodes (synsets) and relations between them. Every word in synset is represented as an array of characters or literal, followed by the meaning of concrete literal in concrete synset. As a word can have multiple meanings, it can be part of multiple synsets.

According to Koeva et al. (2008), *SerWN* consists of 13612 synsets, 23139 literals, 18210 relations, 314 derived, and 83 derivatives.

Krstev et al. (2014) developed an ontology for the culinary domain in Serbian, and the Serbian Wordnet is enhanced with the synsets from this domain. This ontology is used for the determination of similarity between recipes and query expansion.

As a lexical resource, *SerWN* has been applied in multi-member lexical unit research (Krstev et al., 2010; Mladenović et al., 2014), text classification (Pavlovic-Lazetic and Graovac, 2010), the search of multilingual digital databases (Stanković et al., 2012), recognizing rhetorical figures (Mitrović et al., 2017; Mladenović and Mitrović, 2013; Mitrovic, 2014), analyzing feelings expressed in the text (Mitrović et al., 2015) and others.

Vujicic-Stankovic in created an ontology for the culinary domain, and expanded SrWN by 1,404 synsets from the culinary domain so it contains a total of 1,797 such synsets (Vujičić-Stanković et al., 2014; Vujičić-Stanković, 2016).

Universal Dependencies (UD) project[12] aims to develop cross-linguistically consistent treebank annotation for many languages, to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages while allowing language-specific extensions when necessary. As a part of this project, Serbian treebank is created, based on SETimes corpus (Samardžić et al., 2017).

5 Lexical and syntactic analysis methods

5.1 Transliteration and diacritic restoration

The tool for the automatic performing diacritic restoration of text which is potentially missing diacritics (e.g. transform "kuca" (dog) into "kuća" (house), if it is necessary) is described by Ljubešić et al. (2016). The accuracy of the tool is 99.5% on standard and 99.2% on nonstandard language.

Transliteration in Serbian is accommodated because each sound is a character. Characters map almost directly from Cyrillic to Latin, with exception of a few letters, that map from a single Cyrillic character to two Latin characters (e.g. њ -> nj, љ -> lj, or ђ -> dj). Systems for transliteration between Serbian Cyrillic and Latin alphabets exist since the 1950s (Matthews, 1952; Aurousseau, 1953; Gerych, 1965). Among newer tools for solving this problem is the Python package *nlpheart* (Ostrogonac et al., 2020), which has a possibility of conversion between the Cyrillic and Latin alphabet.

5.2 Tokenization and stemming

Sentence tokenization is the process of dividing the text into consisting sentences. Word tokenization's aim is to divide sentences into simple units, tokens, which are usually words, numbers, and punctuation marks. There are a number of multi-language tokenizers which have the ability to tokenize Serbian texts. The majority of these tools are available as Python modules, like Cutter(Graën et al., 2018), Spacy[13], CLASSLA and Reldi[14] tokenizers. Cutter tokenizer has a variant for online tokenization[15]. CLASSLA tokenizer is adapted Stanford NLP Python Library with improvements for specific languages - Fork of Stanza for Processing Slovenian, Croatian, Serbian, Macedonian and Bulgarian)[16]. Turanjin tokenizer for Serbian is available as a PHP library[17]. There is no precise information or comparison of the tokenization accuracy on Serbian documents.

[12] https://universaldependencies.org/introduction.html
[13] https://spacy.io/api/tokenizer
[14] https://github.com/clarinsi/reldi-tokeniser
[15] https://pub.cl.uzh.ch/projects/sparcling/cutter/current/
[16] https://pypi.org/project/classla/
[17] https://github.com/turanjanin/serbian-language-tools

Ostrogonac et al. (2020) present Python package *nlpheart* for text processing of Serbian that includes transliteration, tokenization, normalization, and automatic preparing for the application of machine learning models.

Stemming is a process of removing finishing letters of words, as derivation suffixes of words. The remaining part is a reduced form of the word called a stem. The stem differs from a dictionary form of the word (lemma). The first tool for stemming (in further text, stemmer) for Serbian is described by Kešelj and Šipka (2008), and it is rule-based (1000 rules) and its accuracy is 79%. Based on this stemmer, Milošević (2012b) created a new stemmer reducing the number of rules (180 rules) with an accuracy of 90%. Another solution can be found in literature, and it is created by S. Petković et al.[18] and it is based on Stemmer for Croatian (precision of 0.986 and recall of 0.961 (F1 0.973) for Croatian)[19]. There is no information about the stemming accuracy of this tool. Batanović et al. reimplemented the optimal and the greedy stemmers of Kešelj and Šipka (2008), improved the greedy algorithm proposed by Milošević (2012b), and reimplemented a stemmer for Croatian by Ljubešić & Pandžić, which is a refinement of the algorithm presented by Ljubešić et al. (2007), as a WEKA package (Holmes et al., 1994) – SCStemmers in (Batanović et al., 2016).

The stem is not a dictionary word form, it is the most common part of words with the same semantic meaning. So, in some normalization methods, n-gram analysis is used as a stemmer alternative. This means that a word could be normalized to a single sub-string of its letters whose size is n (tri-gram, tetra-gram etc.). The reason is that the n-gram analysis approach is language-independent, which means that it doesn't need any rules, lexicons, or corpora. Marovac et. al used n-gram analysis in the normalization of Serbian text (Marovac et al., 2012).

5.3 Lemmatization and Part-of-speech tagging

Lemmatization is a process that aims to determine the base morphological form of the word (lemma), which corresponds to a headword in a dictionary. This step in text mining is especially important for languages with rich inflectional morphology, such as Serbian. A given word can have multiple possible lemmas, and it depends on the context, so some lemmatizers use information obtained by POS or MSD tagging to achieve better accuracy.

There are a number of lemmatization approaches: rule-based, simple statistical-based methods, and machine learning-based methods (Akhmetov et al., 2020).

LemmaGen (Juršic et al., 2010) is a learning algorithm for the automatic generation of lemmatization rules in the form of a refined RDR (Ripple Down Rules) tree structure. It is compared with CST (Dalianis and Jongejan, 2006)

[18]Stefan Petković and Dragan Ivanović, Stemmer for Serbian language, 2019. https://snowballstem.org/algorithms/serbian/stemmer.html (accessed Apr 26, 2022)

[19]Ljubešić, Nikola. Pandžić, Ivan. Stemmer for Croatian, http://nlp.ffzg.hr/resources/tools/stemmer-for-croatian/

and RDR (Plisson et al., 2008) lemmatization algorithms and its lemmatization accuracy on Serbian corpora are given in 3.

BTagger[20] (Gesmundo and Samardzic, 2012) is a bidirectional tagger-lemmatizer tool that implements a lemmatization-as-tagging paradigm. Models are trained on the Serbian G.O.1984 corpus, reaching overall accuracies of 97.72% for lemmatization and 86.65% for MSD tagging.

Agić et al. (2013) tested hidden Markov model trigram taggers HunPos, lemmatization capable PurePos, TreeTagger, support vector machine tagger SVMTool, CST data-driven rule-based lemmatizer and BTagger on Serbian corpora and results are given in Table 3.

Table 3 Tools for normalization and POS tagging

Tool label	Application	Corpus	Accuracy
KeseljStemmer (Kešelj and Šipka, 2008)	stemmer	unknown	79.0 %
MilosevicStemmer (Milošević, 2012b)	stemmer	Politika	90.0%
LemmaGen (Juršic et al., 2010)	lemmatizer	Multext-East	up to 86.1%+-0.61%
CST (Juršic et al., 2010; Dalianis and Jongejan, 2006)	lemmatizer	Multext-East	64.0 %+-0.82%
RDR (Juršic et al., 2010; Plisson et al., 2008)	lemmatizer	Multext-East	63.8%+-0.80%
BTagger (Gesmundo and Samardzic, 2012)	lemmatizer	G.O.1984	97.73%
PurePOS (Agić et al., 2013)	lemmatizer	SETimes	86.63%
TnT tagger (Gesmundo and Samardzic, 2012)	POS tagger	G.O.1984	85.47%
BTagger (Gesmundo and Samardzic, 2012)	POS tagger	G.O.1984	86.65%
HunPOS (Agić et al., 2013)	POS tagger	SETimes	95.47%
CRF (Ljubešić et al., 2016)	POS Tagger	500k	97.86%
HunPOS (Agić et al., 2013)	MSD tagger	SETimes	87.11% (+lex 84.81%)
PurePOS (Agić et al., 2013)	MSD tagger	SETimes	74.4%
SVMTool (Agić et al., 2013)	MSD tagger	SETimes	84.99%
CRF (Ljubešić et al., 2016)	MSD tagger	500k	92.33%
Reldi-tagger (Ljubešić and Dobrovoljc, 2019)	MSD tagger	CLARIN.SI	92.03%
StanfordNLP (Ljubešić and Dobrovoljc, 2019)	MSD tagger	CLARIN.SI	95.23%

POS (part of speech) tagging is an NLP processing task where words in the text are annotated with corresponding grammatical categories (parts of speech: verb, noun, adjective, pronoun, etc.). POS tagging with more precise information about grammatical categories is MSD tagging (morphosyntactic tagging - tagging with morphosyntactic descriptions).

Finite state automata used in the lexical and syntactic analysis, considering morpho-syntactic labels were described in (Krstev, 1997).

Sečujski and Kupusinac (Sečujski and Kupusinac) used HMM for morphosyntactic tagging on Alfanum and G.O.1984 corpora. The accuracy of annotation largely depends on the type of text and that some texts are more suitable for automatic annotation than others. For the AlphaNum corpus, an error of 18.44% was obtained, and for "1984" as much as 26.97%.

Popović (2008, 2010) evaluated five taggers (Tree Tagger, SVMTool, Brill – Rule Based Tagger, Trigrams'n'Tags and MXPOST) on three corpora

[20] https://github.com/agesmundo/BTagger

(*"Helsinške sveske br. 15, nacionalne manjine i pravo"*, Serbian Radio diffusion Law and materials from UNDP workshops, G.O.1984). TnT has shown the best performance, while Tree Tagger and SVMTool taggers have shown better performance in special cases.

The POS tagger for Serbian and Croatian based on CRF (conditional random fields) is described in (Ljubešić et al., 2016). It is trained on a manually annotated corpus of Croatian in combination with hrLex/srLex lexicons for each language. The set of morpho-syntactic labels used in the corpus is created according to instructions of the revised MULTEXT-East V5 set of labels for Croatian and Serbian. The accuracy of POS tagging for Serbian is 92.33% for MSD tagging and 97.86 for POS tagging.

The tools for tokenization, stemming, lemmatization, and POS and MSD tagging and their accuracy on Serbian corpora are shown In Table 3.

6 Classification

Text classification is a process of categorizing text data into predefined groups or categories based on its content. Text classification is often performed using supervised machine learning techniques.

Graovac (2014) proposed two methods for classifying text based on their content. The first method is based on the representation of a document as a profile containing a fixed number of n-grams of bytes that appear in the document, and a dissimilarity measure used to determine the class to which the document belongs. This method is language-independent and does not require any pre-processing of the text or prior knowledge of the content of the text or the language in which the text is written. The second method refers to the use of the information contained in the Serbian wordnet and the Serbian electronic dictionary.

Petrović proposes utilizing models and neural networks as a potential remedy to meet the demand for machine prediction of links or references within the text of newly enacted laws and other regulations (Petrović and Janičijević, 2019; Petrović, 2020). Training and validation of neural networks (RNN - Recurrent neural networks, CNN - convolutional neural networks, and HAN - hierarchical attention network model) are performed on a labeled data set, which is made by assigning to each segment of the text of the law (each article of the law) a corresponding label on the existence, or non-existence of a link or reference in that segment of the text. After that, the training procedure is based on a large set of data, which includes a collection of 1120 texts of laws, segmented into a total of 59 167 individual articles of law.

For all methods, the number of training parameters is reduced by over 99%.

6.1 Similarity

Marovac et al. (2013) proposed a method for similarity search of documents in Serbian. The searching query is represented as a word vector, as well as documents for search hing. The grouping of the documents is done using the

k-Means clustering algorithm, and keywords are extracted using TF and IDF features, and n-grams. The similarity values between query and documents are calculated using cosine measure, Jaccard's coefficient, or Euclidian distance. Furlan et al. (2013) proposed a new algorithm, called LInSTSS, which, when determining the semantic similarity of two short texts, also takes into account the specifics of the words these texts contain. The evaluation was carried out on a corpus of paraphrases for the Serbian language created in the same research. One solution of similarity search in e-government is described by Nikolić (2016) using the tool "Apache Lucene". Petrović and Stanković (2019) demonstrated how different preparation methods influence the calculation of text similarity.

Batanović (2020) presented the process of handling semantic tasks using statistical modeling and machine learning. The STS.news.sr is a corpus of news created and used for the task of semantic similarity where the similarity of news is annotated by score. Implementation is given in the library STSFineGrain (Java), available on GitHub. For semantic similarity, the combination of word alignment and the average of word vectors was used. The srWaC corpus (Web corpus of the Serbian language) is used for creating the word vectors. An evaluation of the effects of 3 different stemming techniques on text similarity for Serbian has been performed. Additionally, a new technique for calculating similarity was proposed called Part-of-Speech and Term Frequency weighted Short-Text Semantic Similarity.

6.2 Sentiment analysis

Sentiment analysis is the process of analyzing and deriving people's opinions, thoughts, and impressions regarding various topics, products, and services expressed in a part of the text. Sentiment analysis can be investigated on several levels: document level, sentence level, phrase level, and aspect level (Wankhade et al., 2022). For sentiment analysis, specific lexical resources are necessary, such as a dictionary of sentiment words, tools for processing negation, stylistic figures, and so on. One of the first tools for sentiment analysis at the sentence level for the Serbian language was given by Milošević (2012a). A binary classification of negative and positive sentiment was performed using the Naive Bayes(NB) algorithm. A steamer (Milošević, 2012b) that was designed for this purpose was used as part of the preprocessing. Stop words were eliminated, and negation processing was done by prefixing the word that follows the negation signal (words like no, none) with 'NE_'. The sentiment analyzer was created as a web tool and made available to the public[21].

Maximum entropy (ME), support vector machine (SVM), and NB machine learning methods were used to analyze tweet sentiment (Jolić, 2015). Procedures are offered to minimize the noise in these messages to increase accuracy. They achieved the best accuracy with the ME method of 80.5% using unigrams; however, when applying unigrams and bigrams, negation and phrases were also considered, increasing accuracy to 82.7%.

[21] https://inspiratron.org/SerbianSentiment.php

Mladenović et al. (2016) chose a hybrid approach that uses a dictionary of sentiments extended by morphological forms using a morphological dictionary SrbMD and synonyms using Serbian WordNet to reduce the disadvantages of using stemmers in morphologically rich languages. A sentiment dictionary was created (Mladenović, 2016), containing 1053 expressions (and 10704 inflectional forms) classified into 24 emotion categories, and augmented with synonyms and phrases. SentiWordNet has been integrated with Serbian WordNet to provide sentiment tags to the synsets from Serbian WordNet. A total of 4044 synsets were marked. An additional sentiment dictionary with 971 inflectional forms was created using these synsets. Using the TF-IDF approach, 577 (1428 inflectional forms) of the most frequent words from the 122 million-word corpus of the contemporary Serbian language SrpKor2013 were used to construct the list of stop words. A domain-oriented collection of stop words with 1372 inflectional forms were generated using the TF approach. The method was trained on a news set (TrN, Table 4) with two topics: "bad news" and "good news," which are automatically categorized and balanced by sentiment. Two sets were used for testing: a set of news (TsN, Table 4) collected from a source other than TrN (this set is not balanced), and a set of movie reviews (TsMR, Table 4) collected from a website and tagged with the sentiment, based on the grades that were attached to them (this set is not balanced). These sources were used to develop the Serbian document-level sentiment analysis framework (SAFOS), which applies the maximum entropy approach with the features: of unigrams, bigrams, and trigrams. They used hold-out test sets and 10-fold cross-validation (CV) to evaluate the SAFOS system. The combination of unigram and bigram features reduced by "sentiment feature" mapping produced the best classification accuracy scores for both hold-out tests (accuracy 78.3% for TsMR set and 79.2% for the TsN set). Because it was trained and tested on data from the same domain, it performed better in a 10-fold CV with a 95.6% accuracy rate.

Grljević (2016) presented a sentiment analysis of content from social networks to improve the business of higher education institutions. Sentiment analysis is performed at two levels of granularity: at the document level and the sentence level. On the set of online reviews of professors and lectures (ORPL, Table 4) both a rule-based strategy (based on a vocabulary that was manually built for the requirements of this domain and is available), as well as the approach based on machine learning algorithms (NB, SVM, and k-Nearest Neighbor KNN) were applied. In sentiment classification using machine learning algorithms, the SVM algorithm gives the best performance, with 84.94% accuracy at the review level, and 80.13% accuracy at the sentence level. The classification of the sentiment was done using the sentiment lexicon, by introducing separate dictionaries for 1266 positive and 1521 negative sentiment words, intensifiers (95), neutralizers, negation (31), domain- specific phrases (41), stop words (179), and other words that change the sentiment of the next word in the sentence. The classification accuracy at the level of reviews is 80.71% and at the level of sentences, it is 73.70%.

The first balanced and topically uniform sentiment analysis dataset in Serbian (SerbMR, Table 4) was generated by Batanović et al. (2016) and is available online in versions with two sentiment polarity classes (positive and negative; 1682 documents) and three polarity classes (positive, neutral, and negative; 2523 documents). The sentiment labels, in this dataset, were obtained automatically by converting the numerical ratings attached to each review by its author. This dataset was examined to identify the best machine-learning features and simple text-processing options for sentiment classification. By combining the obtained optimal attributes with NBSVM (combination of polynomial Naive Bayes classifier and support vector method classifier), they achieved an accuracy of up to 85.55% for two and up to 62.69% for three classes. By comparing different methods for morphological normalization, it was concluded that the use of stemmer is better than lemmatization in the case of sentiment analysis. The stemmer of Ljubešić and Pandžić gave the best accuracy results on the dataset SerbMR, 86.11% for two and up to 63.02% for three classes (Batanović and Nikolić, 2017).

According to the studies cited above, identifying the presence of negation is insufficient to ascertain sentiment. The collection of film reviews in (Batanović et al., 2016) is subjected to the traditional method of processing negation, which involves changing the polarity of words that follow a negative signal. For three classes, marking two words after the negation led to the most significant improvement in sentiment analysis accuracy (0.94%), while for two classes, marking only the first word after the negation gave the best improvement in accuracy (0.66%). The processing rules of semantic negation, which improved the classification of short informal texts by sentiment, are described in Ljajić and Marovac (2019). These rules were tested on a set of tweets with topic public personalities that were manually marked with the sentiment (TWPP, Table 4). The machine learning method that uses additional attributes based on the proposed negation processing rules improves sentiment analysis accuracy on a set of tweets for three classes by up to 1.45% and for two classes by up to 0.82%. When this method is applied to a set of tweets containing negation, the improvement in sentiment analysis accuracy increases by up to 2.65% for three classes and up to 1.65% for two classes. For this study's aims, dictionaries of negation signals (25), negative quantifiers (56), and intensifiers, as well as a sentiment dictionary of 5632 sentiment words (reduced to the morphological foundation of 4058 negative and 1574 positive words), were constructed. The impact of various morphological normalizations on sentiment analysis was examined on this set of tweets, and it was discovered that the use of stemmer (Milošević, 2012b) takes precedence over normalization using the morphological dictionary SrbMD (accuracy 85.27%) and that reducing words to 4-grams produces good results with little resource usage (Ljajić et al., 2019).

Aspect-based sentiment analysis deals with the identification of sentiments (negative, neutral, positive) and the determination of aspects (target sentiments) in a sentence. Nikolić et al. (2020) proposed an aspect-based sentiment analysis of student opinion surveys in the Serbian language. Two sets of

data were used for sentiment analysis, which was done at the finest level of granularity of the text - the level of the sentence segment (phrase and sentence).

A collection of official student surveys (OSS, Table 4) makes up the first dataset, while the second dataset set of online reviews of professors and lecturers (OSPL, Table 4) previously created for the paper (Grljević, 2016). The OSS and OSPL corpora were automatically annotated for the sentiment (negative, neutral, positive), then manually annotated for aspects (ranging from lower-level features, such as lectures, helpfulness, materials, and organization, to higher-level aspects, such as professor, course, and other). For aspect classification, a cascade classifier (a collection of SVM binary classifiers trained to distinguish between two distinct aspects) was employed. The quality of the aspect analysis was influenced by the corpus, as seen by the F-measures of 0.89 for the OSS corpus and 0.78 for the OSPL corpus, respectively.

Table 4 Corpora for sentiment analysis

Corpus label	Text type	Number of items	Annotation[1]	Reference
TrN	News	2000	S; AA	(Mladenović et al., 2016)
TsN	News	779	S; AA	(Mladenović et al., 2016)
TsMR	Movie reviews	2237	S; AA	(Mladenović et al., 2016)
ORPL	Education Reviews	3863	S, A; AA + MA	(Grljević, 2016)
OSS	Education Reviews	2472	S,A; AA +MA	(Nikolić et al., 2020)
SerbMR	News	2523	S;AA	(Batanović et al., 2016)
SentiComments.SR	Short texts	3490	S; MA	(Batanović, 2020)
ParlaSent-BCS v1.0	Sentences of parliamentary debates	2600	S; MA	(Mochtak et al., 2022)
TWPP	Tweets	7664	S; MA	(Ljajić and Marovac, 2019)
TWVA	Tweets	8817	S,R; MA	(Ljajić et al., 2022)
MRSA	Music Reviews	1830	S; AA	(Drašković et al., 2022)
SMSSA	SMS messages	6171	S; MA	(Šandrih, 2019)
TW15	Tweets	1643735	S; MA	(Mozetič et al., 2016)

[1]Annotation target: S - sentiment, A - aspect, R - relevance; Annotation type: AA - automated annotation MA - manually annotated

Sentiment analysis includes specific subtasks such as polarity detection, subjectivity detection, sarcasm detection, etc. An annotation approach with six sentiment labels was created to satisfy the requirements of processing particular tasks and enabling multiple interpretations of sentiment (Batanović et al., 2020). SentiComments.SR (Table 4), a corpus of short texts in the Serbian language, has been manually annotated using this multi-level annotation scheme. It contains 3490 short movie comments (length up to 50 tokens) (Batanović, 2020). On this corpus, the outcomes of applying linear classifiers using bag-of-words and/or bag-of-embedding features were evaluated under the influence of different morphological normalizations and negation processing techniques. The combination of bag-of-words and bag-of-embeddings attributes resulted in significant improvements in classification for all sentiment analysis subtasks (F - measure: polarity 0.783, subjectivity 0.885 four-class sentiment analysis 0.655, six-class sentiment analysis 0.586). Due to the insufficient number of sarcastic texts in the corpus, the results of sarcasm detection are not representative.

Sentiment lexicon Senti-Pol-sr (Stanković et al., 2022) was created based on three existing lexicons (NRC, AFFIN, and Bing) and was manually corrected. The dictionary contains 6454 different tokens. Its initial version is available.

The lexicon was utilized to conduct sentiment analysis on a well-balanced dataset extracted from SrpELTeC, which consisted of 1089 sentences that were manually labeled, with each sentiment category containing 363 instances of positive, neutral, and negative sentiments. This approach achieved the best accuracy of 87.8% on SrpELTeC 2 classes and 71.9% on SrpELTeC 3 classes using MNB with the Bag-of-Words approach combined with our sentiment lexicon features. The results of trained models using LR, NB, decision tree, random forest, SVN, and k-NN methods gave the best accuracy of 87.8% for LR. It has also been shown that training on a dataset of labeled movie reviews (SerbMR) indicates that it cannot be successfully used for sentence sentiment analysis in old novels. Drašković et al. (2022) developed a machine-learning model for sentiment analysis using three different data sets. The first set (MRSA, Table 4) was created for this research by collecting music reviews from 13 portals, which made sure that the set was balanced. The second data set is the already mentioned set of movie reviews, while the third set is music album reviews—MARD. MARD was originally composed in English and then translated into Serbian using the Google Translate API. Standard classification models (NB, LR, and SVM) and hybrid models (combining a linear model with NB) were applied to these datasets. The hybrid model NB-LR gave average good results (58% for three classes and 79% for two classes). It is shown that a set of film and music reviews can be used together to improve the quality model. Extending the model with reviews translated from English does not improve performance, due to the different vocabulary and review writing styles, as well as the quality of the translated text. Emoticon influence, informal speech, lexical, and other language features about the mood in the set of SMS messages (SMSSA, Table 4) are presented by Šandrih (2019). They selected 621 features and divided them into three main categories: lexical (based on signs and words), syntactic (emoticons, abbreviations), and stylistic. Using linear SVM classification, an accuracy of 92.3% was obtained. Sentence-based sentiment classification as well as emotion recognition is suggested to improve the classification of SMS messages. Mozetič and Grčar (Mozetič and Grčar) found that the quality of the classification model depends much more on the quality and size of the training data than on the type of trained model by analyzing 1.6 million manually tagged tweets in 15 different European languages, of which 73,783 tweets are in Serbian (TW15, Table 4). Based on the performed experiments, it was shown that there is no statistically significant difference between the performance of the top classification models (five of these models are based on SVM, and for reference, the NB classifier was applied).

Transfer learning is one of the advanced techniques in AI, which allows a pre-trained model to transfer its knowledge to a new model. Transfer learning is frequently used in sentiment analysis to classify sentiments, and it can produce successful results, particularly in the absence of large labeled data sets.

Batanovic presented the results of applying neural language models based on transformer architectures to sentiment analysis subtasks of short texts from the SentiComments.SR corpus (Batanović et al., 2020).Three transformer-based models were used: Multilingual BERT (Devlin et al., 2018), Multilingual distilBERT (Sanh et al., 2019), and XLM MLM (Conneau and Lample, 2019). Fine-tuning multilingual transformer-based models gain the same or better performance than linear models for all sentiment analysis subtasks. For each subtask, XLM MLM produced the best F-measure results: 0.793 for polarity, 0.887 for subjectivity, 0.686 for four-class sentiment analysis, and 0.627 for six-class sentiment analysis.

Based on a sample of parliamentary discussions, Mochtak et al. (2022) demonstrated that using transformer models produces outcomes that are noticeably superior to those obtained using a simpler architecture. The dataset consists of sentences of average length from the corpus of parliamentary proceedings in the region of the former Yugoslavia - Bosnia and Herzegovina, Croatia, and Serbia. A set of 2600 sentences (ParlaSent-BCS v1.0, Table 4), including 876 with only positive, 876 with only negative, and 866 without sentiment words, were chosen for the dataset using the Croatian gold standard sentiment lexicon (Glavaš et al., 2012) (translated to Serbian with a rule-based Croatian-Serbian translator (Klubička et al., 2016)). This dataset contains 1059 sentences from the Serbian parliament. The dataset is manually annotated using the multiple-level annotation schema described by Batanović et al. (2020), and it is available online. A sentiment analysis approach was applied at the sentence level. The results of classification four of the transformer models were compared: FastText (Bojanowski et al., 2017) with pre-trained CLARIN.SI word embeddings (Ljubešić and Erjavec, 2018), XLM-R (Conneau et al., 2019), CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020), and BERTić (Ljubešić and Lauc, 2021). BERTić gave the best results compared to the others (model macro F1 0.7941 ± 0.0101). Compared to Bosnian and Croatian, the Serbian language proved to be the most difficult to predict. Using BERTić for sentiment analysis, Ljajić et al. (2022) expanded the annotated dataset that was used for the topic analysis of tweets containing negative sentiment towards the COVID-19 vaccination. A collection of 8817 vaccination-related tweets in the Serbian language (TWVA, Table 4) were manually labeled as relevant or irrelevant regarding the COVID-19 vaccination sentiment. Relevant tweets were manually marked with sentiment labels: positive, negative, or neutral. On this data set, BERTić correctly categorized tweets as relevant or irrelevant with a 94.7% accuracy rate and correctly classified relevant tweets as negative, positive, or neutral with an 85.7% accuracy rate. The annotated set was expanded by this classifier, and from the original manually annotated 1770 tweets with negative sentiment, another 1516 tweets with negative sentiment were automatically marked, forming the data set used for the topic analysis. The topic analysis was carried out using the latent Dirichlet allocation (LDA) and nonnegative matrix factorization (NMF)

methods. Topics that are potential reasons for vaccine skepticism are highlighted by topic analysis: worries about adverse reactions, efficacy, inadequate testing, mistrust of authorities, and conspiracy theories.

7 Named entity recognition

Named-entity recognition (NER) is a task that seeks to locate and classify named entities mentioned in unstructured text into categories such as personal names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. Named entity recognition (NER) as an NLP task is fairly old, gaining popularity with Message Understanding Conferences in the mid-1990s (Sekine, 2004), However, NER for Serbian has not been addressed substantially until the 2010s.

Vitas and Pavlović-Lažetić (2008) developed a system that uses morphological and lexical analysis in combination with dictionaries (Serbian and transcribed English first names, and geographical locations) for recognition of people's names and geographical entities. The system is using e-dictionaries and transducer-based rules or grammars for disambiguation of proper names and geopolitical entities (Krstev et al., 2007; Vitas and Pavlović-Lažetić, 2008).

Ljubešić et al. (2013) proposed a first system based on machine learning and conditional random fields for the recognition of names, organizations, and locations for Croatian and Slovene, which are closely related to Serbian. They have used a set of annotated web and news corpora (SETimes, Vjesnik, and corpora for both Slovene and Croatian developed as a student project (Filipić et al., 2012)) to train their method. For features, they used linguistic features and distributional similarity features calculated from large unannotated monolingual corpora. Their experiments showed that distributional features improve the F1 score by 7-8 points, while morphological features can improve by additional 3-4 points. However, as the size of the dataset increases, the morphosyntactic and distributional features lose their importance for NER. They have made resources used for building this NER system publicly available.

Another approach, based on the previous application of rules encoded in transducers and thesauri (Krstev, 1997) was enhanced for recognition of personal names and geopolitical names (Krstev et al., 2014). Dictionaries are used for matching tokens and phrases, while recursive transition networks (grammar graphs) from Unitex (Paumier et al., 2002) are used to resolve ambiguities (e.g. taking into account grammatical rules such as case-number-gender agreement). They reported that the system prefers precision over recall, with a precision of 0.96 and a recall of 0.88.

For the purpose of comparing NER approaches in multi-lingual aligned texts (bitexts), a system called NERosetta was developed (Krstev et al., 2013; Krstev C et al., 2013). To illustrate the system, 7 bitexts involving 5 languages (French, English, Greek, Serbian, Croatian) and 5 different NER systems were used (1 for Serbian (Krstev et al., 2014), 1 for Croatian (Ljubešić et al., 2013),

1 for English (Stanford NER) and 2 for French). The entities that were evaluated were Person, Organization, and Location, with some of the NER systems providing annotations for time, date, money, percent, and others. The demo application is available on the web [22].

A dictionary approach with the addition of transducer-based grammars (Krstev et al., 2014) was used to create a gold standard data set based on news articles annotated with personal names. This data set was then used to train machine learning-based approaches, namely Stanford NER and SpaCy (Šandrih et al., 2019). Their evaluation indicated that the rule-based approach performed the best (based on the F1-score), while Stanford NER had the best recall.

Tanasijević (2019) developed a system for labeling cultural heritage documents with metadata. In order to do this, she developed a system that recognizes entities, such as years and person names, as well as topics of the tagged documents.

The transformer-based model was also introduced for several tasks in Serbian, Croatian, and Slovene, including NER (Ljubešić and Lauc, 2021). The model was pre-trained on web-crawled texts in Serbian, Bosnian, Croatian, and Slovene, consisting of 8 billion tokens, and then fine-tuned for NER on several openly available datasets, such as SETimes.SR (Batanović et al., 2018), corpora of news articles, or ReLDI-sr (Ljubešić et al., 2017), corpora of annotated tweets. For reference, authors compared this model with CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020) and multilingual BERT (Devlin et al., 2018), where language-specific BERT-based models significantly outperformed multi-lingual BERT.

Apart from a general domain for Serbian, a decent amount of work has been done for medical named entity recognition. One of the previously described systems for a general domain was adapted for medical de-identification of clinical texts (Jaćimović et al., 2014, 2015). The system recognized persons, dates, geographic locations, organizations, and numbers using vocabularies and transducer grammar rules. The authors reported an overall F1 score of 0.94. On the other hand, (Puflović et al., 2016) created a model based on character and word n-grams. The dataset they used was obtained from a neurological clinic and their system was designed to recognize names of diseases, names of medications, abbreviations, and numbers that represent dosage, dates or times, and medical treatment success. They have manually checked 100 documents, reporting accuracy ranging from 64% to 90%. A mathematical model for medical term recognition was proposed by (Avdić et al., 2020). They have proposed three methods. The first method was based on the dictionary matching of terms. The second method uses a formula for labeling words contained in the training set, where confidence is calculated by the number of instances in which a word is labeled with a certain label in the training set divided by the total count of that word in the training set. The third method is an extension of the second method, where several rules are added to terms with errors

[22] http://www.korpus.matf.bg.ac.rs/nerosetta/

and abbreviations, so they can be tagged well. Labeled entities included medical terms (symptoms, symptom descriptions, diagnoses, biochemical analyses, Latin words, anatomic names of organs, therapies, and other medical terms) and non-medical terms (numbers, negation symbols, and other words). The best-performing model was the third one, with an F1 score of 0.937, while the highest F1 score for medical terms was 0.896. The methods based on deep neural networks and multilingual language models were proposed by Kaplar et al. (2022). They used a manually annotated corpora from the Clinic for Nephrology at the University Clinical Center of Serbia (203 discharge summaries annotated by 2 computer science Ph.D. students adapting the 2012 i2b2 temporal relation challenge annotation schema). They have created models based on conditional random fields (CRF), multilingual transformers (BERT Multilingual and XLM RoBERTa), long short-term memory (LSTM) recurrent neural networks, and their ensembles. CRF method had hand-crafted features that are commonly used in literature (word, word stem, shape of the word, previous 3 words, next 3 words, etc.). For the LSTM model, the authors used gensim's word2vec model before feeding the embeddings to the LSTM network, followed by the CRF token classifier. The study showed that the highest precision was achieved with the CRF-based model, while the highest recall had a multi-lingual transformer model. The best F1 score had an LSTM-CRF-based model. The best performance was achieved by creating an ensemble of the models with majority voting (F1 score of 0.892).

8 Language models

A language model determines word probability in a sequence. In order to create a language model, many approaches were proposed, ranging from simply calculating word appearance in a larger text corpus to adding more lexical and syntactic features to learning word probabilities. Early language models were purely statistical, while since 2014, we have seen a proliferation of neural language models - language models based on neural networks.

Language models are prerequisites for many natural language tasks. Therefore, many works in classification (Graovac, 2014), sentiment analysis (Milošević, 2012a; Jolić, 2015; Grljević, 2016; Batanović and Nikolić, 2017; Ljajić and Marovac, 2019) or named entity recognition (Šandrih et al., 2019), used traditional n-gram language models, at times enriched with lexical, morphological or syntactic features. These systems were previously described in this review.

Ostrogonac (2018) in his PhD thesis does a review and comparison of language models for Serbian up to 2018. In this work, he proposes the first neural language model for Serbian, based on recurrent neural networks trained on a corpus of morphologically annotated text in Serbian. Also, he creates a hybrid model that uses parts-of-speech and lemmas, and matches sequences of words to either n-grams in corpus, or to partially lemmatized sequences. These models are compared with more traditional n-gram models for correcting semantic

and grammatical errors in the text. The error is detected by setting a threshold for a difference in log likelihood between a language model with morphological features and one without it. While setting thresholds may be challenging, it showed the potential use cases for specifically trained neural language models for Serbian.

There has been a significant effort done by international researchers to create multilingual neural language models. Some of these models included also Serbian, such as FastText (Bojanowski et al., 2017), multilingual BERT (Devlin et al., 2018), XLM-R (Conneau et al., 2019), and XLM MLM (Conneau and Lample, 2019). Batanović in his Ph.D. thesis (Batanović, 2020), compared a number of n-gram language models for the tasks of sentiment analysis and text similarity. He further compared these language models and methods with fine-tuned multi-lingual transformer-based models (multilingual BERT base (Devlin et al., 2018), DistilBERT Multilingual (Sanh et al., 2019), and XLM MLM (Conneau and Lample, 2019)), showing transformer models in all cases outperforming all n-gram based models (including ones containing a large amount of morphological, lexical, and syntactic features).

The first, and, at the time of writing of this paper, the only transformer-based language model specifically trained for Serbian, Croatian, Bosnian, and Montenegrin is BERTić (Ljubešić and Lauc, 2021). BERTić is trained using the ELECTRA approach (Clark et al., 2020) for training transformer models. This approach involves training a smaller generator model and the main discriminator model with the task to discriminate whether a specific word is an original word from the text or a word generated by the generator model. The model is trained on a corpus of 8 billion tokens crawled from the web in Serbian, Croatian, Bosnian, and Montenegrin. While there was previously a BERT-based model for Croatian and Slovenian, called CroSloBERT (Ulčar and Robnik-Šikonja, 2020), BERTić outperformed it on almost all tasks (morphological annotation, NER, social media geolocation prediction, commonsense causal reasoning task). This is mainly because of the bigger corpus, and computational efficiency of the ELECTRA approach that was used.

9 Conculsion and future directions

Research on natural language processing for the Serbian language has a long tradition, going back to the second half of the 1990s. During this time, many approaches for lexical, morphological, syntactic, and semantic processing of text were explored. In the past decade, the number of researchers and research published on natural language processing for Serbian significantly increased. Several Universities and research institutes in Serbia established natural language research groups.

The Serbian language is a highly inflected language and therefore many challenges in natural language processing are specific to Serbian, such as the most efficient way for tokenization, handling the inflections in various tasks, handling negations, etc. While some work has been done on these challenges,

they are still open research questions. Basic lexical and morphological tasks, such as transliteration, diacritic restoration, tokenization, stemming, lemmatization, and part-of-speech tagging are quite well-researched, with many approaches presented, evaluated, and compared. Some of the classification tasks, such as sentiment analysis were, as well, extensively researched. Sentiment analysis seems to gain a lot of interest after 2012. Named entity recognition has also been researched for several named entities, such as proper and personal names, and locations. Also, few approaches have been proposed for biomedical NER.

On the other hand, some methods and tasks were still not adequately addressed for Serbian. Many classification tasks, except sentiment analysis, have not been explored and language resources for them are missing. As it was previously said, methods for only a basic set of named entities have been proposed. Domain-specific classification and named entity recognition methods are still missing.

Methods in the semantic web, ontology, and semantic networks have not much proliferated in the area of Serbian NLP, as only a few papers are touching on this subject. Most significant research in network space has been done in developing Serbian WordNet, but this is a rather morphological and lexical network, then something that can be considered a semantic network. Language resources for many of the tasks are still missing.

While the language-specific BERT-based model has been trained, there is only a single initiative to create this kind of language model. Also, resources such as sentence embedding or document embedding methods have not been yet developed. These methods would also contribute significantly to the creation of methods for summarization, question answering, language-specific semantic search, or machine translation.

At the moment, there is a proliferation of large language models, such as GPT-3 (Brown et al., 2020), Lambda (Thoppilan et al., 2022) or ChatGPT (Ouyang et al., 2022). While these models are multilingual and can generate text in Serbian, there has not yet been much research on prompt engineering or fine-tuning these language models for Serbian.

10 Acknowledgements

This paper is partially supported by the Ministry of Education, Science, and Technological Development of the Republic of Serbia, Projects No. III44007.

References

Agić, Ž., N. Ljubešić, and D. Merkler 2013. Lemmatization and morphosyntactic tagging of croatian and serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pp. 48–57.

Akhmetov, I., A. Pak, I. Ualiyeva, and A. Gelbukh. 2020. Highly language-independent word lemmatization using a machine-learning classifier. *Computación y Sistemas 24*(3): 1353–1364 .

Andonovski, J. 2020. *Mreža otvorenih podataka i jezički resursi u procesu izgradnje srpsko-nemačkog literarnog korpusa*. Ph. D. thesis.

Andonovski, J., B. Šandrih, and O. Kitanović. 2019. Bilingual lexical extraction based on word alignment for improving corpus search. *The Electronic Library* .

Aurousseau, M. 1953. Transliteration of cyrillic script. *Nature* 171: 940–940 .

Avdić, A. 2021. *Realizacija servisa pametnog zdravstva i njihova integracija u koncept pametnih gradova*. Ph. D. thesis.

Avdić, A., U. Marovac, and D. Janković. 2020. Automated labeling of terms in medical reports in serbian. *Turkish Journal of Electrical Engineering and Computer Sciences 28*(6): 3285–3303 .

Balvet, A., D. Stosic, and A. Miletic 2014. Talc-sef a manually-revised pos-tagged literary corpus in serbian, english and french. In *LREC 2014*.

Batanović, V., M. Cvetanović, and B. Nikolić 2018. Fine-grained semantic textual similarity for serbian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Batanović, V., M. Cvetanović, and B. Nikolić. 2020. A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. *PLoS One 15*(11): e0242050 .

Batanović, V., B. Furlan, and B. Nikolić. 2011. Softverski sistem za odredjivanje semantičke sličnosti kratkih tekstova na srpskom jeziku. *Zbornik radova sa 19. telekomunikacionog foruma (TELFOR 2011)*: 1249–1252 .

Batanović, V., N. Ljubešić, and T. Samardžić 2018. Setimes. sr–a reference training corpus of serbian. In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)*, pp. 11–17.

Batanović, V., N. Ljubešić, T. Samardžić, and M.M. Petrović. 2020. Otvoreni resursi i tehnologije za obradu srpskog jezika. *Proc. of the Primena slobodnog softvera i otvorenog hardvera* .

Batanović, V. and B. Nikolić. 2017. Sentiment classification of documents in serbian: The effects of morphological normalization and word embeddings. *Telfor Journal 9*(2): 104–109 .

Batanović, V., B. Nikolić, and M. Milosavljević 2016. Reliable baselines for sentiment analysis in resource-limited languages: The serbian movie review dataset. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2688–2696.

Batanović, V.V. 2020. *Metodologija rešavanja semantičkih problema u obradi kratkih tekstova napisanih na prirodnim jezicima sa ograničenim resursima.* Ph. D. thesis, Univerzitet u Beogradu-Elektrotehnički fakultet.

Batanović V, V., N. Ljubešić, T. Samardzić, and T. Erjavec. 2018. Training corpus setimes. sr 1.0 .

Bogdanović, M. and J. Tošić. 2022. Corpus of legislation texts of republic of serbia 1.0. Slovenian language resource repository CLARIN.SI.

Bogetić, K. and V. Batanović. 2020a. Annotated corpus of serbian language-related news articles MetaLangNEWS-sr. Slovenian language resource repository CLARIN.SI.

Bogetić, K. and V. Batanović. 2020b. Annotated corpus of serbian language-related news comments MetaLangNEWS-COMMENTS-sr. Slovenian language resource repository CLARIN.SI.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* 5: 135–146 .

Brown, T., B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33: 1877–1901 .

Clark, K., M.T. Luong, Q.V. Le, and C.D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* .

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* .

Conneau, A. and G. Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems* 32 .

Corpora etc, C. 1992. Serbo-croatian text corpus. Oxford Text Archive.

Dalianis, H. and B. Jongejan 2006. Hand-crafted versus machine-learned inflectional rules: The euroling-siteseeker stemmer and cst's lemmatiser. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Delić, V., M. Sečujski, N. Jakovljević, M. Janev, R. Obradović, and D. Pekar. 2010. Speech technologies for serbian and kindred south slavic languages. *Advances in Speech Recognition*: 141–164 .

Devlin, J., M.W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Dobrić, N. 2012. Savremeni jezički korpusi na zapadnom balkanu–istorijat, trenutno stanje i budučnost. *Slavistična revija 60*(4): 677–692 .

Drašković, D., D. Zečević, and B. Nikolić. 2022. Development of a multilingual model for machine sentiment analysis in the serbian language. *Mathematics 10*(18): 3236 .

Eberhard, D.M., F.S. Gary, and D.F. Charles. 2022. Ethnologue: Languages of the world.

Filipić, L., T. Jurić, and M. Stupar. 2012. Strojno prepoznavanje naziva u tekstovima pisanima hrvatskim jezikom. *Studentski znanstveni rad, Rektorova nagrada, Filozofski fakultet, Sveučilište u Zagrebu* .

Furlan, B., V. Batanović, and B. Nikolić. 2013. Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems 55*(3): 710–719 .

Gerych, I. 1965. *Transliteration of Cyrillic alphabets*. Ph. D. thesis, University of Ottawa (Canada).

Gesmundo, A. and T. Samardzic 2012. Lemmatising serbian as category tagging with bidirectional sequence classification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2103–2106.

Glavaš, G., J. Šnajder, and B. Dalbelo Bašić 2012. Semi-supervised acquisition of croatian sentiment lexicon. In *International Conference on Text, Speech and Dialogue*, pp. 166–173. Springer.

Graën, J., M. Bertamini, M. Volk, M. Cieliebak, D. Tuggener, and F. Benites 2018. Cutter–a universal multilingual tokenizer. In *CEUR Workshop Proceedings*, Number 2226, pp. 75–81. CEUR-WS.

Graovac, J.B. 2014. *Prilog metodama klasifikacije teksta: matematički modeli i primene*. Ph. D. thesis.

Grass, T., D. Maurel, and O. Piton 2002. Description of a multilingual database of proper names. In *Advances in Natural Language Processing: Third International Conference, PorTAL 2002 Faro, Portugal, June 23–26, 2002 Proceedings*, pp. 137–140. Springer.

Grljević, O. 2016. *Sentiment u sadržajima sa društvenih mreža kao instrument unapredjenja poslovanja visokoškolskih institucija*. Ph. D. thesis.

Holmes, G., A. Donkin, and I.H. Witten 1994. Weka: A machine learning workbench. In *Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference*, pp. 357–361. IEEE.

Jaćimović, J., C. Krstev, and D. Jelovac 2014. Automatic de-identification of protected health information. In *Proceedings of the 17th International Multiconference INFORMATION SOCIETY–IS 2014, Language Technologies, October 9th- 10th, 2014, Ljubljana, Slovenia*, pp. 73–78. Jožef Stefan Institute, Ljubljana, Slovenia.

Jaćimović, J., C. Krstev, and D. Jelovac. 2015. A rule-based system for automatic de-identification of medical narrative texts. *Informatica 39*(1) .

Jolić, N. 2015. Klasifikacija sentimenta u twitter postovima korišćenjem udaljenog nadzora .

Juršic, M., I. Mozetic, T. Erjavec, and N. Lavrac. 2010. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science 16*(9): 1190–1214 .

Kaplar, A., M. Stošović, A. Kaplar, V. Brković, R. Naumović, and A. Kovačević. 2022. Evaluation of clinical named entity recognition methods for serbian electronic health records. *International Journal of Medical Informatics*: 104805 .

Kešelj, V. and D. Šipka. 2008. A suffix subsumption-based approach to building stemmers and lemmatizers for highly inflectional languages with sparse resources. *INFOtheca-Journal of Informatics & Librarianship* 9 .

Klajn, I. 2005. *Gramatika srpskog jezika* (I ed.). Belgrade: Zavod za udžbenike i nastavna sredstva.

Klubička, F., G. Ramírez-Sánchez, and N. Ljubešić 2016. Collaborative development of a rule-based machine translator between croatian and serbian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pp. 361–367.

Koeva, S., C. Krstev, and D. Vitas 2008. Morpho-semantic relations in wordnet–a case study for two slavic languages. In *Global wordnet conference*, pp. 239–253. University of Szeged, Department of Informatics.

Kostić, A. 2014. Electronic corpus of serbian language from 12th to 18th century. *Review of the National Center for Digitization* .

Kovačević, L., V. Injac, and D. Begenišić. 2004. *Bibliotekarski terminološki rečnik: englesko-srpski, srpsko-engleski*. Narodna biblioteka Srbije.

Krstev, C. 1997. *Jedan prilaz informatiekom modeliranju teksta i algoritmi njegove transformacije*. Ph. D. thesis.

Krstev, C., J. Jaćimović, and D. Vitas 2012. Recognition and normalization of some classes of named entities in serbian. In *Proceedings of the Fifth Balkan Conference in Informatics*, pp. 52–57.

Krstev, C., D. Maurel, et al. 2007. A note on the semantic and morphological properties of proper names in the prolex project. *Lingvisticae Investigationes 30*(1): 115–133 .

Krstev, C., I. Obradović, M. Utvić, and D. Vitas. 2014. A system for named entity recognition based on local grammars. *Journal of Logic and Computation 24*(2): 473–489 .

Krstev, C., G. Pavlović-Lažetić, and I. Obradović. 2004. Using textual and lexical resources in developing serbian wordnet. *Romanian Journal of Information Science and Technology 7*(1-2): 147–161 .

Krstev, C., R. Stanković, I. Obradović, D. Vitas, and M. Utvić 2010. Automatic construction of a morphological dictionary of multi-word units. In *International Conference on Natural Language Processing*, pp. 226–237. Springer.

Krstev, C., S.V. Stanković, and D. Vitas 2014. Approximate measures in the culinary domain: Ontology and lexical resources. In *Proceedings of the 9th Language Technologies Conference IS-LT*, pp. 38–43.

Krstev, C. and D. Vitas 2005. Corpus and lexicon-mutual incompletness. In *Proceedings of the Corpus Linguistics Conference*, Volume 14, pp. 17.

Krstev, C. and D. Vitas. 2011. An aligned english-serbian corpus. *ELL-SIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality)* 1: 495–508 .

Krstev, C., D. Vitas, and G. Pavlović-Lažetić 2008. Resources and methods in the morphosyntactic processing of serbo-croatian. In *Formal Description*

of Slavic Languages: The Fifth Conference, pp. 3–17.

Krstev, C., D. Vitas, and A. Savary 2006. Prerequisites for a comprehensive dictionary of serbian compounds. In *International Conference on Natural Language Processing (in Finland)*, pp. 552–563. Springer.

Krstev, C., A. Zečević, D. Vitas, and T. Kyriakopoulou 2013. Nerosetta–an insight into named entity tagging. In *6th Language and Technology Conference*, pp. 168–172.

Krstev C, C., A. Zečević, D. Vitas, and T. Kyriacopoulou 2013. Nerosetta for the named entity multi-lingual space. In *Language and Technology Conference*, pp. 327–340. Springer.

Laporte, E. 2003. The relex network. *Преузето са http://infolingu. univmlv. fr/Relex/Relex. htm* .

Lemmenmeier-Batinić, D., N. Ljubešić, and T. Samardžić. 2021. Corpus of serbian forms of address 1.0. Slovenian language resource repository CLARIN.SI.

Ljajić, A. and U. Marovac. 2019. Improving sentiment analysis for twitter data by handling negation rules in the serbian language. *Computer Science and Information Systems 16*(1): 289–311 .

Ljajić, A., U. Marovac, and M. Stanković. 2019. Comparison of the influence of different normalization methods on tweet sentiment analysis in the serbian language. *Facta Universitatis, Series: Mathematics and Informatics*: 683–696 .

Ljajić, A., N. Prodanović, D. Medvecki, B. Bašaragin, J. Mitrović, et al. 2022. Uncovering the reasons behind covid-19 vaccine hesitancy in serbia: Sentiment-based topic modeling. *Journal of Medical Internet Research 24*(11): e42261 .

Ljubešić, N., D. Boras, and O. Kubelka. 2007. Retrieving information in croatian: Building a simple and efficient rule-based stemmer .

Ljubešić, N. and K. Dobrovoljc 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of slovenian, croatian and serbian. In *Proceedings of the 7th workshop on balto-slavic natural language processing*, pp. 29–34.

Ljubešić, N., K. Dobrovoljc, and D. Fišer. 2015. Mwelex–mwe lexica of croatian, slovene and serbian extracted from parsed corpora. *Informatica 39*(3) .

Ljubešić, N. and T. Erjavec. 2018. Word embeddings clarin. si-embed. hr 1.0, slovenian language resource repository clarin. si (2018).

Ljubešić, N., T. Erjavec, and D. Fišer 2016. Corpus-based diacritic restoration for south slavic languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3612–3616.

Ljubešić, N., T. Erjavec, M. Miličević, and T. Samardžić. 2017. Serbian twitter training corpus reldi-normtagner-sr 2.0 .

Ljubešić, N., M. Esplà-Gomis, S. Ortiz Rojas, F. Klubička, and A. Toral. 2016. Serbian-english parallel corpus srenWaC 1.0. Slovenian language resource repository CLARIN.SI.

Ljubešić, N. and F. Klubička 2014, April. bs,hr,srWaC - web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, Gothenburg, Sweden, pp. 29–35. Association for Computational Linguistics.

Ljubešić, N. and F. Klubička. 2016. The serbian web corpus srwac. *Ljubljana: Jožef Stefan Institute* .

Ljubešić, N. and F. Klubička. 2016. Serbian web corpus srWaC 1.1. Slovenian language resource repository CLARIN.SI.

Ljubešić, N., F. Klubička, Ž. Agić, and I.P. Jazbec 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)*, pp. 4264–4270.

Ljubešić, N. and D. Lauc. 2021. Bertić–the transformer language model for bosnian, croatian, montenegrin and serbian. *arXiv preprint arXiv:2104.09243* .

Ljubešić, N., F. Markoski, E. Markoska, and T. Erjavec. 2021. Comparable corpora of south-slavic wikipedias CLASSLA-wikipedia 1.0. Slovenian language resource repository CLARIN.SI.

Ljubešić, N. and P. Rupnik. 2022. The twitter user dataset for discriminating between bosnian, croatian, montenegrin and serbian twitter-HBS 1.0. Slovenian language resource repository CLARIN.SI.

Ljubešić, N., M. Starović, T. Kuzman, and T. Samardžić. 2022. Choice of plausible alternatives dataset in serbian COPA-SR. Slovenian language resource repository CLARIN.SI.

Ljubešić, N., M. Stupar, T. Jurić, and Ž. Agić. 2013. Combining available datasets for building named entity recognition models of croatian and slovene. *Slovenščina 2*(1): 35–57 .

Magner, T. 2001. Digraphia in the territories of the croats and serbs. *International Journal of Sociology and Languages*: 11–26. https://doi.org/doi:10.1515/ijsl.2001.028 .

Marovac, U., A. Avdić, and A. Ljajić. 2021. Creating a stop word dictionary in serbian. *Scientific Publications of the State University of Novi Pazar Series A: Applied Mathematics, Informatics and mechanics 13*(1): 17–25 .

Marovac, U., A. Ljajić, E. Kajan, and A. Avdić. 2013. Similarity search in text data for the serbian language. *Proceedings of ICEST*: 607–610 .

Marovac, U., A. Pljasković, A. Crnišanin, and E. Kajan 2012. N-gram analysis of text documents in serbian language. In *2012 20th Telecommunications Forum (TELFOR)*, pp. 1385–1388. IEEE.

Matthews, W.K. 1952. The latinisation of cyrillic characters. *The Slavonic and East European Review 30*(75): 531–548 .

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*, pp. 30–34.

Miličević, M. and N. Ljubešić. 2016. Tviterasi, tviteraši or twitteraši? producing and analysing a normalised dataset of croatian and serbian tweets. *Slovenščina 2.0: empirical, applied and interdisciplinary research 4*(2): 156–188 .

Miller, G.A. and C. Fellbaum. 2007. Wordnet then and now. *Language Resources and Evaluation* 41: 209–214 .

Milošević, N. 2012a. Mašinska analiza sentimenta rečenica na srpskom jeziku. *Master's Degree Thesis* .

Milošević, N. 2012b. Stemmer for serbian language. *arXiv preprint arXiv:1209.4471* .

Milosević, N. and G. Nenadić. 2016. As cool as a cucumber: Towards a corpus of contemporary similes in serbian. *arXiv preprint arXiv:1605.06319* .

Milosević, N. and G. Nenadić. 2018. Creating a contemporary corpus of similes in serbian by using natural language processing. *arXiv preprint arXiv:1811.10422* .

Mitrovic, J. 2014. Electronic tools and resources for multi-word unit detection and research in serbian. In *The 2th General Meeting of The IC1207 COST Action, PARSEME, Athens, Greece*, pp. 10–11.

Mitrović, J., M. Mladenović, and C. Krstev 2015. Adding mwes to serbian lexical resources using crowdsourcing. In *poster presented at The 5th PARSEME general meeting. Iași, Romania*, pp. 23–24.

Mitrović, J., C. O'Reilly, M. Mladenović, and S. Handschuh. 2017. Ontological representations of rhetorical figures for argument mining. *Argument & Computation 8*(3): 267–287 .

Mladenović, M. 2016. *Informatički modeli u analizi osećanja zasnovani na jezičkim resursima*. Ph. D. thesis.

Mladenović, M. and J. Mitrović 2013. Ontology of rhetorical figures for serbian. In *International Conference on Text, Speech and Dialogue*, pp. 386–393. Springer.

Mladenović, M., J. Mitrović, and C. Krstev 2014. Developing and maintaining a wordnet: Procedures and tools. In *Proceedings of the Seventh Global Wordnet Conference*, pp. 55–62.

Mladenović, M., J. Mitrović, C. Krstev, and D. Vitas. 2016. Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems 46*(3): 599–620 .

Mladenović, M., S.V. Stanković, and V. Pajić. 2020. Two ways for the automatic generation of application ontologies by using balkanet. *International Journal on Semantic Web and Information Systems (IJSWIS) 16*(2): 18–41 .

Mochtak, M., P. Rupnik, and N. Ljubešič. 2022. The parlasent-bcs dataset of sentiment-annotated parliamentary debates from bosnia-herzegovina, croatia, and serbia. *arXiv preprint arXiv:2206.00929* .

Mozetič, I. and M. Grčar. Smailovi c j. 2016. *Multilingual Twitter sentiment classification: the role of human annotators. PLOS ONE 11*(5): e0155036 .

Mozetič, I., M. Grčar, and J. Smailović. 2016. Twitter sentiment for 15 european languages. Slovenian language resource repository CLARIN.SI.

Nenadić, G. 2004. Creating digital language resources. *Pregled nacionalnog centra za digitalizaciju* (5): 191–30 .

Nikolić, N., O. Grljević, and A. Kovačević. 2020. Aspect-based sentiment analysis of reviews in the domain of higher education. *The Electronic*

Library 38(1): 44–64 .

Nikolić, V. 2016. *Modelovanje i pretraživanje nad nestruktuiranim podacima i dokumentima u e-Upravi Republike Srbije*. Ph. D. thesis.

Odebrecht, C., L. Burnard, and C. Schöch. 2021. European literary text collection (eltec): April 2021 release with 14 collections of at least 50 novels. zenodo.

Ostrogonac, S. 2018. *Modeli srpskog jezika i njihova primena u govornim i jezičkim tehnologijama*. Ph. D. thesis, University of Novi Sad (Serbia).

Ostrogonac, S., B. Rastović, and E. Liliom. 2020. A python package for text processing for serbian–nlpheart. *Scientific Technical Review 70*(3): 41–45 .

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* .

Paumier, S., F. Malchok, C. Marschner, C. Martineau, C. Martínez, D. Maurel, S. Nagel, A. Neme, M. Petit, J. Stiehler, et al. 2002. Unitex 3.2 .

Pavlovic-Lazetic, G. and J. Graovac 2010. Ontology-driven conceptual document classification. In *KDIR*, pp. 383–386.

Pavlović-Lažetić, G., D. Vitas, and C. Krstev 2004. Towards full lexical recognition. In *Text, Speech and Dialogue: 7th International Conference, TSD 2004, Brno, Czech Republic, September 8-11, 2004. Proceedings 7*, pp. 179–186. Springer.

Petrović, D. 2020. *Analiza strukture kolekcije pravnih dokumenata na osnovu njihove povezanosti preko odredjenih jezičkih izraza*. Ph. D. thesis, Универзитет у Нишу, Електронски факултет.

Petrović, D. and S. Janičijević 2019. Domain specific word embedding matrix for training neural networks. In *2019 International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI)*, pp. 71–714. IEEE.

Petrović, D. and M. Stanković. 2019. The influence of text preprocessing methods and tools on calculating text similarity. *Ser. Math. Inform 34*(5): 973–994 .

Plisson, J., N. Lavrač, D. Mladenić, and T. Erjavec. 2008. Ripple down rule learning for automated word lemmatisation. *Ai Communications 21*(1): 15–26 .

Popović, M. and M. Arčan. 2016. Post-edited and error annotated machine translation corpus {PErr} 1.0. Slovenian language resource repository CLARIN.SI.

Popović, Z. 2008. *Evaluacija programa za obeležavanje (etiketiranje) teksta na srpskom jeziku*. Ph. D. thesis.

Popović, Z. 2010. Taggers applied on texts in serbian. *INFOtheca-Journal of Informatics & Librarianship 11*(2) .

Puflović, D., G. Velinov, T. Stanković, D. Janković, and L. Stoimenov. 2016. A supervised named entity recognition for information extraction from medical records.

Reshamwala, A., D. Mishra, and P. Pawar. 2013. Review on natural language processing. *IRACST Engineering Science and Technology: An International Journal (ESTIJ) 3*(1): 113–116 .

Samardžić, T., N. Ljubešić, and M. Miličević. 2015. Regional linguistic data initiative (reldi) .

Samardžić, T., M. Starović, Ž. Agić, and N. Ljubešić. 2017. Universal dependencies for serbian in comparison with croatian and other slavic languages .

Šandrih, B. 2019. Sms sentiment classification based on lexical features, emoticons and informal abbreviations. *Serdica Journal of Computing 13*(1-2): 081p–096p .

Šandrih, B., C. Krstev, and R. Stanković 2019. Development and evaluation of three named entity recognition systems for serbian-the case of personal names. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 1060–1068.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* .

Sečujski, M. and V. Delić. 2008. A software tool for automatic part of speech tagging in serbian language. *Applied Linguistics 1*(9): 97–103 .

Sečujski, M.S. and A.D. Kupusinac. Automatska morfološka anotacija tekstova na srpskom jeziku korišćenjem hmm .

Sekine, S. 2004. Named entity: History and future. *Project notes, New York University*: 4 .

Stanković, R., M. Košprdić, M.I. Nešić, and T. Radović 2022. Sentiment analysis of serbian old novels. In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, pp. 31–38.

Stanković, R., C. Krstev, I. Obradović, A. Trtovac, and M. 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, May 2012, Istanbul, Turkey*, pp. 1710–1717.

Stanković, R., J. Mitrović, D. Jokić, and C. Krstev 2020. Multi-word expressions for abusive speech detection in serbian. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pp. 74–84.

Stanković, R., I. Obradović, C. Krstev, and D. Vitas 2011. Production of morphological dictionaries of multi-word units using a multipurpose tool. In *Proceedings of the Computational Linguistics-Applications Conference, October 2011, Jachranka, Poland.*

Stanković, R., B. Trivić, O. Kitanović, B. Blagojević, and V. Nikolić. 2011. The development of the geolissterm terminological dictionary. *INFOtheca-Journal of Informatics & Librarianship 12*(1) .

Stanković, R., C. Krstev, B.Š. Todorović, and M. Škoric. 2021. Annotation of the serbian eltec collection. *Infotheca-Journal for Digital Humanities 21*(2): 43–59 .

Stanković, R., B. Šandrih, C. Krstev, M. Utvić, and M. Skoric 2020, May. Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for Serbian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, pp. 3954–3962. European Language Resources Association.

Tanasijević, I. 2019. Toward automatic tagging of cultural heritage documents. *IPSI Transactions on Advanced Research, TAR 15*(1) .

Thoppilan, R., D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* .

Tiedemann, J. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, Volume 2012, pp. 2214–2218.

Todorović, B.Š., R. Stanković, C. Krstev, and M.I. Nešić 2021. Serbian ner&beyond: The archaic and the modern intertwinned. In *Deep Learning Natural Language Processing Methods and Applications–Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2021)*, pp. 1252–1260.

Tomašević, A.D. 2018. *Razvoj modela za upravljanje rudarskom projektnom dokumentacijom.* Ph. D. thesis.

Ulčar, M. and M. Robnik-Šikonja 2020. Finest bert and crosloengual bert: less is more in multilingual models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pp. 104–111. Springer.

Utvić, M. 2011. Annotating the corpus of contemporary serbian. In *Proceedings of the INFOtheca '12 Conference*, pp. 36–47.

Utvić, M. 2014. *Izgradnja referentnog korpusa savremenog srpskog jezika.* Ph. D. thesis.

Utvić, M., R. Stanković, A. Tomašević, M. Škorić, and B. Lazić. 2019. Pretraga korpusa zasnovana na upotrebi eksternih leksičkih resursa putem veb-servisa. *Naučni sastanak slavista u Vukove dane-Vol. 48/3 Srpski jezik i njegovi resursi* .

Vasiljević, N. 2015. *Automatska obrada pravnih tekstova na srpskom jeziku.* Ph. D. thesis.

Vitas, D., S. Koeva, C. Krstev, and I. Obradović 2008. Tour du monde through the dictionaries. In *Actes du 27eme Colloque International sur le Lexique et la Gammaire*, pp. 249–256.

Vitas, D. and C. Krstev. 2006. Literature and aligned texts. *Readings in Multilinguality*: 148–155 .

Vitas, D. and C. Krstev. 2009. Serbian language and sstbi. *SNTPI '09 - Naučno-stručni skup Sistem naučnih, tehnoloških i poslovnih informacija* .

Vitas, D., C. Krstev, and É. Laporte. 2006. Preparation and exploitation of bilingual texts. *Lux Coreana* (1): 110–132 .

Vitas, D., C. Krstev, I. Obradović, L. Popović, and G. Pavlović-Lažetić 2003a. An overview of resources and basic tools for processing of serbian written texts. In *Proc. of the Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics.* Citeseer.

Vitas, D., C. Krstev, I. Obradović, L. Popović, and G. Pavlović-Lažetić 2003b. Processing serbian written texts: An overview of resources and basic tools. In *Workshop on Balkan Language Resources and Tools*, Volume 21, pp. 97–104.

Vitas, D., P. Ljubomir, K. Cvetana, O. Ivan, P.L. Gordana, and S. Mladen. 2012. The serbian language in the digital age. *META-NET White Paper Series, G. Rehm, H. Uszkoreit (eds.)* .

Vitas, D., G. Nenadić, and C. Krstev 1998. Electronic edition of serbian translation of plato's republic aligned with 17 languages. In *East meets West – A compendium of Multilingual Resources*, TELRI Association e.V., Institut fur deutsche Sprache, Mannheim.

Vitas, D. and G. Pavlović-Lažetić. 2008. Resources and methods for named entity recognition in serbian. *INFOtheca-Journal of Informatics & Librarianship* 9 .

Vujanić, M. ed. 2007. *Rečnik srpskoga jezika*. Matica srpska.

Vujičić-Stanković, S. 2016. *Ekstrakcija informacija vodjena ontologijama:(model za srpski jezik)*. Ph. D. thesis.

Vujičić-Stanković, S., C. Krstev, and D. Vitas 2014. Enriching serbianwordnet and electronic dictionaries with terms from the culinary domain. In *Proceedings of the Seventh Global Wordnet Conference*, pp. 127–132.

Vuković, T. 2020. Spoken torlak dialect corpus 1.0 (transcription). Slovenian language resource repository CLARIN.SI.

Wankhade, M., A.C.S. Rao, and C. Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*: 1–50 .