Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian

Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, Mihailo Škorić

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3954–3962
Marseille, 11–16 May 2020

Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian

Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, Mihailo Škorić
{Faculty of Mining and Geology, Faculty of Philology} University of Belgrade
{Djušina 7, Studentski trg 3} Belgrade, Serbia
{ranka.stankovic, mihailo.skoric}@rgf.bg.ac.rs, cvetana@matf.bg.ac.rs, {branislava.sandrih, misko}@fil.bg.ac.rs

Abstract

The training of new tagger models for Serbian is primarily motivated by the enhancement of the existing tagset with the grammatical category of a gender. The harmonization of resources that were manually annotated within different projects over a long period of time was an important task, enabled by the development of tools that support partial automation. The supporting tools take into account different taggers and tagsets. This paper focuses on TreeTagger and spaCy taggers, and the annotation schema alignment between Serbian morphological dictionaries, MULTEXT-East and Universal Part-of-Speech tagset. The trained models will be used to publish the new version of the Corpus of Contemporary Serbian as well as the Serbian literary corpus. The performance of developed taggers were compared and the impact of training set size was investigated, which resulted in around 98% PoS-tagging precision per token for both new models. The **SR_BASIC** annotated dataset will also be published.

Keywords: Part-of-Speech tagging, lemmatization, corpus, evaluation, Serbian, morphological dictionary

1. Introduction

The task of assigning to each token its Part-of-Speech category (noun, verb, adjective, etc.) is a common Natural Language Processing (NLP) task, known as Part-of-Speech tagging (PoS-tagging). PoS-tagging precedes many other Natural Language Processing tasks, such as Text Classification, Named Entity Recognition, Sentiment Analysis, Question Answering, etc.

Computer programs that perform this task, the so-called 'taggers', can be based on lookup-tables, regular expressions, linguistic (morphological, semantic and syntactic) rules, machine learning methods (Giménez and Màrquez, 2004; Denis and Sagot, 2009; Manning et al., 2014) or state-of-the-art Deep Neural Networks (DNNs) (Huang et al., 2015; Choi, 2016; Akbik et al., 2018).

The first operational tagger and lemmatization model for Serbian were produced as a parameter file (TT11) for the TreeTagger model (Schmid, 1999) after a thorough analysis of the state-of-the-art solutions (Popović, 2010), and subsequently used for various purposes. The research presented in this paper was motivated primarily by the need to enrich the tagset with new tags, such as the grammatical category of gender, harmonize resources manually annotated within different projects over a long period of time, as well as to develop tools for supporting preparation of training sets to be used for different taggers and tagsets in the future. The research was focused on annotation schemata alignment between Serbian morphological dictionaries tagset (presented briefly in Subsection 2.1.), MULTEXT-East tagset (Erjavec, 2012), and the Universal Part-of-Speech tagset (Petrov et al., 2012), for the purpose of preparing datasets for training the TreeTagger and spaCy taggers. The main goal was to automate the process of annotation schema harmonization and preparation of training datasets as much as possible.

The corpus of training set texts needed checking and correction. Corpus correction is a time consuming process, which requires a lot of manual intervention and help of linguistic specialists. Our focus was not only on removing inconsistencies and reducing the ambiguity level, but also on introducing simultaneously new grammatical categories in the annotation schema. In this paper we present a general strategy for corpus correction and its results. The strategy can be used in different tagging environments: in a standalone tool, used strictly for text annotation, such as TreeTagger, but also in a Python module using spaCy library, which can then be used for various NLP applications.

The paper is organized as follows. The resources used in this research are described in Section 2. Section 3. contains a detailed explanation of the pipeline used for training taggers for Serbian: the annotation schemata are presented in Subsection 3.1. and harmonization and transformation to the final schema in Subsection 3.2. The main features of the tools used for tagging are described in Subsection 3.3. Evaluation results of spaCy model are compared with the results obtained by the TreeTagger trained on the same dataset, as well as with a previous version of TreeTagger for Serbian (Utvić, 2011), and discussed in Section 4. The paper ends with concluding remarks and an outline of future work in Section 5.

2. Resources

The main resources used for the production of the new tagger model for Serbian are: (a) Serbian morphological dictionaries (Cvetana Krstev, Duško Vitas, 2015) (SMD); (b) pre-annotated texts (Duško Vitas, Cvetana Krstev, Ranka Stanković, Miloš Utvić, 2019).

2.1. Serbian morphological dictionaries

Serbian morphological dictionaries represent a rich lexical resource, which can be used in various NLP tasks (Krstev, 2008). It is being continually developed and maintained in the lexical database LeXimirka (Stanković et al., 2018), which supports different export functions, including recently added formats particularly designed for tagger training tasks. It comprises of more than 210,000 lemmas,

including simple- and multi-word units (MWUs), proper names, general- and domain-oriented lexica. Its basic tagset is similar to the one used by the Serbian TreeTagger models built in 2011 (TT11) and 2019 (TT19) and it generally corresponds to the traditional notion of Part-of-Speech in Serbian. These basic tags are refined by adding different markers to lexical entries. For instance, the marker +Aux differentiates auxiliary from other verbs, the marker +NProp differentiates proper nouns from other nouns, while markers +ProN and +ProA differentiate nominal from adjectival pronouns. An example of two different noun types in SMD is:

```
vlada,N+HumColl
   // government, common name
Vlada,N+NProp+Hum+First
    // first name, proper name
```

The additional markers enable mapping of SMD tagset to other tagsets without loss of information (e.g. such a mapping was performed for the production of Serbian MULTEXT-East resources (Krstev et al., 2004)).

2.2. Pre-annotated texts

Various pre-annotated texts were used in this research for training and testing. These texts were tagged mainly using SMD (and its tagset) and the Unitex system,[1] with manually performed disambiguation. Besides the basic SMD PoS classes depicted in Table 3 an additional tagset (nPoS) was introduced for subclasses of adjectives, nouns and verbs, providing the information on grammatical gender and comparative degree for adjectives (Table 1).[2]

PoS		masculine	feminine	neuter
Adjective	**Positive**	A:am	A:af	A:an
	Comparative	A:bm	A:bf	A:bn
	Superlative	A:cm	A:cf	A:cn
Noun		N:m	N:f	N:n
Verb		V:m	V:f	V:n

Table 1: nPoS tagset that includes grammatical gender and adjective comparative degree.

The texts used in this research are shown in Table 2. The text 1984, Serbian translation of Orwell's novel, was annotated according to the MULTEXT-East specification and included in MULTEXT-East resources (version 3) (Krstev et al., 2004). The text Verne, Serbian translation of the novel *Around the world in 80 days*, was prepared using the same specification in the scope of SEE-ERA.net project (Tufiş et al., 2009). Intera is the Serbian part of the multilingual corpus prepared in the scope of the project "Integrated European language data Repository Area" (Gavrilidou et al., 2006). It contains texts from law, health and education domains. Švejk, Floods, History are three short texts selected, respectively, from a novel (*The Good Soldier Schweik*), newspaper articles (reporting on floods in Serbia in 2014) and a history textbook. Novels was composed of excerpts from the Serbian part of ELTeC corpus containing novels published between 1840 and 1920.[3]

Text	Tokens	T-Types	Words	W-Types
1984	108,133	18,117	69,706	18,050
VERNE	73,826	12,298	59,706	12,266
INTERA	1,071,200	56,743	907,643	55,470
ŠVEJK	4,122	1,484	3,347	1,475
FLOODS	4,671	1,798	3,813	1,741
HISTORY	6,596	2,726	5,287	2,622
NOVELS	5,118	2,117	4,236	2,093

Table 2: Pre-annotated texts used for training and testing of the spaCy tagger and the TreeTagger TT19

3. Tagging

3.1. Tagsets

The basic set of PoS-categories/tags that should be assigned to tokens is not generally accepted, even for a specific language. The choice of a tagset usually depends on the foreseen task or project. A tagset tailored to be applicable for PoS-tagging in general is the Universal Part-of-Speech (UPoS) tagset (Petrov et al., 2012) (used by spaCy), and it lists the following 17 categories: adjective (ADJ), adposition (ADP), adverb (ADV), auxiliary (AUX), coordinating conjunction (CCONJ), determiner (DET), interjection (INTJ), noun (N), numerical (NUM), particle (PART), pronoun (PRON), proper noun (PROPN), punctuation (PUNCT), subordinating conjunction (SCONJ), symbol (SYM), verb (VERB) and other (X). It should be noted that the MULTEXT-East tagset (Erjavec, 2012) was also tailored to be universal. SMD uses its own tagset that corresponds closely to Serbian traditional grammars. The Serbian TreeTagger models TT11 and TT19 (see Subsection 3.3.) use modifications of the SMD tagset. A general overview of the tagsets used is presented in Table 3. It should be noted that tags for some PoS differ between tagsets (e.g. ADJ in UPoS vs. A in SMD for adjective).

3.2. Harmonization of annotation schema

Texts used for training and testing (presented in Section 2.) were produced over the last 15 years for different purposes. This resulted in many differences and inconsistencies in tagging, that had to be resolved. Besides the use of different annotating schemata, these issues were:

- Some texts were fully annotated, with lemmas and all grammatical categories (1984, Verne), some were only lemmatized with assigned PoS (Intera, Švejk, Floods, History), while in one text (Novels) values of the grammatical category gender were added.

[1] Unitex/GramLab — Cross Plaform Corpus Processing Suite, https://unitexgramlab.org/

[2] The category of gender is relevant only for some verbal forms.

[3] ELTeC is a corpus prepared in the scope of COST Action CA16204 *Distant Reading for European Literary History*, https://www.distant-reading.net/.

PoS-tag	UPoS	SMD	TT11	MTE
ADJ	✓	✓ A	✓ A	✓ A
ADP	✓	✓ PREP	✓ PREP	✓ S
ADV	✓	✓	✓	✓ R
AUX	✓	×* V+Aux	× V	×* V+a
CCONJ	✓	×* CONJ	× CONJ	×* C+c
DET	✓	×*	×	✓ D
INTJ	✓	✓ INT	✓ INT	✓ I
N	✓	✓+	✓+	✓* N-p
NUM	✓	✓	✓-	✓ M
PART	✓	✓ PAR	✓ PAR	✓ Q
PRON	✓	✓+ PRO	✓+ PRO	✓ P
PROPN	✓	×* N+NProp	× N	×* N+p
PUNCT	✓	×	✓	×
SCONJ	✓	×* CONJ	× CONJ	×* C+s
SYM	✓	×	×	×
VERB	✓	✓+ V	✓+ V	✓V-a
X	✓	×	✓- ?	✓-

(a) Universal PoS-tagset compared with other tagsets

PoS-tag	UPoS	SMD	TT11	MTE
ABB	× X	✓	✓	✓ Y
PREF	×	✓	✓	× X
RN	× NUM	×*	✓	✓ M
SENT	× PUNCT	×	✓	×
?	✓ X	×	✓	✓ X

(b) T11 specific tags compared to other tagsets

Table 3: Tagsets used in UPoS, SMD, TT11 and MULTEXT-East (MTE); asterix (*) signifies that a basic tag is not in the tagset but can be deduced from additional information (markers); plus (+) signifies that the same tag is used for a proper super set; minus (-) signifies that the same tag is used for a proper subset.

- Some texts were tagged with MWUs and named entities (NEs) (Verne, Švejk, Floods, History). Since the taggers developed within this research tagged only simple words these complex units had to be decomposed into simple words. For instance, *Devetnaesti vek* 'Nineteenth century' which was tagged as temporal named entity had to be separated to two tokens: the adjective *devetnaesti* and the noun *vek*.

- Same word tokens were assigned different lemmas in different texts. The reason for this was that texts were tagged with SMD, which evolved in time and many entries were enhanced and corrected. For instance, numerous lemmas of adjectives were represented in SMD using their definite form which is predominantly used (e.g. *počasni* 'honorable'); the use of SMD to process various texts revealed that many of these adjectives are used in indefinite form as well, and the representation of adjective lemmas in SMD was accordingly corrected (*počasan*), since adjective lemmas are in indefinite form (if it exists).

- Certain words were not consistently PoS-tagged in all texts as a result of the evolution of SMD and/or the view of the annotator. For instance, some annotators have assigned a tag ADV (adverb) to the word *danima* 'lasting several days', whereas others have regarded it as the instrumental case in plural of *dan* 'day' and used the tag N.

- There are numerous lemma variants in SMD which decline differently; however, some inflected forms in their respective paradigms coincide. For instance, lemmas *komunista* and *komunist* 'communist' share one singular form (instrumental case *komunistom*) and almost all plural forms (*komunisti*, *komunista*, *komuniste*, *komunistima*) and the annotator could choose either of the two lemmas, thus introducing the unwelcome variability.

Some of these issues were resolved manually (for shorter texts), while others were resolved automatically, when possible, namely where there was no ambiguity or some rule could be formulated and used. Our aim to include grammatical categories of comparative degree (for adjectives) and gender (for nouns, adjectives, some forms of verbs and some types of pronouns and numbers) into tagger models required an update of the training corpus and addition of respective category values to texts where they were missing. The biggest challenge was to add new grammatical categories to the Intera corpus. Having in mind its size, it had to be done automatically using morphological dictionaries introduced in Section 2.1. In the first step unambiguous information was added, while in the second step we tried to resolve ambiguities through several heuristic rules, including frequencies from texts that were fully annotated manually. Some discrepancies were detected and manually corrected, like adjacent tokens ADJ NOUN that did not agree in gender. However, it cannot be said that all problems of these kind were detected and solved.

An additional problem was that training corpora used different tagsets (described in Subsection 3.1.). All texts had to be mapped to tagsets used by the existing tagger model TT11 and the two new tagger models TT19 and SerSpaCy (see Subsection 3.3.). Although most of the texts were tagged with SMD before mapping to some other tagset, the initial SMD version was not available for all texts (e.g. Intera) and various mapping procedures had to be developed. Mapping from SMD to UPoS tagset was rather straightforward. For some tags it was direct – A → ADJ, PREP → ADP, ADV → ADV, INT → INTJ, PAR → PART. For others, markers assigned to lemmas were used as an additional indicator: V-Aux → V, V+Aux → AUX, PRO+ProA → DET, PRO-ProA → PRON, N+NProp → PROPN, N-NProp → NOUN.

For distinguishing between coordinating (CCONJ) and subordinating (SCONJ) conjunctions a special list was prepared, since this information was only recently introduced in SMD. Roman numerals (RN) and other numerals (NUM) were annotated as NUM, while ABB (abbreviations and acronyms), PREF (prefixes), ? (other) and X (residual) from all tagsets were mapped to X (other). Tag SYM was not used. Sentence delimiters for spaCy were controlled by numbering, so both SENT (end sentence tag) and PUNCT

tags were encoded as PUNCT.

In order to perform this complex harmonization task we developed a procedure to automatically add additional layers to annotated texts for each annotation schema. A MS SQLServer database that contains all annotated text, which is equipped with auxiliary tables and functions from LeXimirka (Stanković et al., 2018), supported the whole endeavour. All sentences and tokens of annotated texts in the database were enumerated. Duplicate sentences were detected and labeled in order to exclude them from the training set.

3.3. Systems for Morpho-Syntactic Tagging

The first operational tagger and a lemmatization model for Serbian was produced as the parameter file TT11 for TreeTagger, a supervised ML-tagger based on Hidden Markov Models (HMMs) that uses decision trees for smoothing (Schmid, 1999). A manually annotated training corpus, a full lexicon containing all allowed pairs (PoS, lemma) assigned to particular token and a list of PoS-tags related to open class words are required to automatically produce a parameter file for TreeTagger. TT11 tagger model was produced from the training corpus Intera (see Subsection 2.2.) and the full lexicon of 1.3+ million tokens (including punctuation) previously derived from then latest version of SMD (Utvić, 2011). TT11 tagset consists of 16 tags, most of them acquired from SMD as labels for major Parts-of-Speech (Table 3 column T11). TT11 was used to annotate SrpKor2013, current version of SrpKor – Corpus of Contemporary Serbian.[4]

As pointed out in (Utvić, 2011) "TreeTagger isn't a 'true' lemmatizer", it assigns "the most likely Part-of-Speech tag" and "simply concatenates lemma from a full lexicon, which corresponds to the chosen Part-of-Speech. Hence, word forms with the same Part-of-Speech, but different lemma cannot coexist in the full lexicon."

A new TreeTagger was produced for this research – TT19, based on the same technology as TT11, the only difference being the set of resources used for training. Both the training corpus and the lexicon were expanded. Several smaller annotated corpora were added to Intera: 1984, Švejk and Floods, and the lexicon was expanded to over 2.1+ million tokens (including punctuation and other non-alphanumeric symbols which occur in the training set required for training using TreeTagger). As this tagger was meant to tag not only PoS, but also grammatical categories (nPoS), the annotation set in both the lexicon and the training corpora was expanded accordingly.

An independent tagger that relied on SMD (simple- and multi-word units) and a grammar for named entity tagging (Krstev et al., 2014) was produced, and it used Conditional Random Fields (Constant et al., 2018). Its major novelty was tagging of not only simple words, but also MWUs and named entities. However, it did not perform lemmatization.

spaCy (Explosion, 2019) is a commercial open-source library for advanced NLP in Python. Its model for tagging is based on a DNN that is designed to predict a "super tag" with a PoS, morphology and a dependency label (Honnibal and Montani, 2017). The developers reveal that this is due to the underlying architecture that contains a hidden convolutional layer shared between a tagger, parser and a named entity recognizer. As a consequence, it is convenient to train all three models at once. spaCy features pretrained models for many languages,[5] which can be easily downloaded and applied.

The potential of this Python module has been already recognized in many applications. Authors in (Ribeiro et al., 2018) applied spaCy for parsing user-stories (i.e. short textual narratives that contain certain information about the subject, object and motive) into tokens. Since the resulting tokens contain the term itself, its PoS-tag and relationships with other tokens, they are subsequently used to infer concepts and relationships contained in user-stories for automated extraction of conceptual models. In (Ribeiro et al., 2018) authors consider the phenomenon of ML-based models for NLP tasks in general, which make different predictions for input instances that are semantically extremely similar. They propose a system for detecting these semantic-preserving perturbations in input data that induce changes in the model's predictions, and apply spaCy in the PoS-tagging step.

4. Results and Evaluation

Beside the featured language models, spaCy also allows training of new language models. The easiest way to do this is by using the Command Line Interface (CLI).[6] We used first the *convert* command, which enables conversion of input files in CoNLL-U format to spaCy's json format. Sentences in CoNLL-U format[7] are split into multiple token lines, where the number of lines equals the number of tokens in a sentence. Each line contains the following ten fields, separated with tabulator character: 1) word index (integer starting from 1 for each new sentence), 2) word form or punctuation symbol, 3) lemma or stem of word form, 4) Universal Part-of-Speech tag, 5) language-specific tag with nPOS introduced in Table 1 (or underscore if not available), 6) list of morphological features from the universal feature inventory or from a defined language-specific extension (or underscore if not available), 7) head of the current word, which is either a value of ID or zero, 8) Universal dependency relation to the HEAD, 9) enhanced dependency graph in the form of a list of head-deprel pairs, and 10) any other annotation. In our dataset, there were values only at positions 1)–5).

spaCy models are trained to predict class labels from any custom tagset that is available in training data. Yet, before the training procedure, it is obligatory to define a mapping from this custom tagset to the Universal Part-of-Speech tagset. This is done by listing key-value pairs for each of the custom tags. For example, the tags N:m, N:f and N:n

[4] SrpKor, http://www.korpus.matf.bg.ac.rs/

[5] Existing spaCy models, https://spacy.io/usage/models

[6] CLI for spaCy, https://spacy.io/api/cli

[7] CoNLL-U format, https://universaldependencies.org/format.html

from our tagset were all mapped to the NOUN tag. Afterwards, we validated our training and development sets using the *debug-data* command. This option checks if there is an overlap between training and evaluation data, and if all fine-grained PoS-categories are mapped to the corresponding Universal PoS-categories. It also reports if there is a class imbalance. After all the checks were passed, the data could be used for further steps.

The texts were gradually expanded, which resulted in five different input subsets (dubbed as SR_BASIC and SR-{25,50,75,100}). The subset SR_BASIC consisted of: 1984, Švejk and Floods and 5% of Intera given in Table 2). The subsets SR-{25,50,75,100} were supersets of SR_BASIC, containing {25%, 50%, 75%, 100%} of Intera respectively. Each of the subsets was split into training, development and test sets (80%:10%:10%). In Table 4 the distributions of number of sentences, tokens and words per training, development and test sets are given.

Using the CLI *train* command, we trained five different models. Each model was trained in 30 iterations, using the default parameters setting. The performance for each of these models, followed by a comparison with other PoS-taggers, is reported in the Subsections 4.1. and 4.2.

4.1. Mutual tagger comparisons

First we compared two taggers trained on the largest training set (SR-{100}). We used it to train the following taggers: 1) SerSpaCy, a spaCy model, and 2) TT19, a TreeTagger model. We used T11 as the baseline for evaluation, since it was already tested and evaluated as part of a previous research (Popović, 2010; Utvić, 2011). Besides the standard PoS tags, SerSpaCy and TT19 models included nPoS tags with grammatical categories. In this section, we present a comparison of SerSpaCy and TT19 results on nPOS. TreeTagger models perform lemmatization, which was also taken into consideration.

Evaluation was performed on four different manually annotated set of texts. Test set was compiled of 10% of each text used for training, and it can give a rough idea on how models perform when tagging similar, already familiar text. Verne, History and Novels represent texts previously unknown to the taggers and show their performance in real world scenario.

Figure 1 compares performance of taggers for PoS task against different test sets. While the spaCy tagger shows overall top performance on the Test set, TreeTagger shows better results when tagging unfamiliar texts. Unlike spaCy, both versions of TreeTagger were trained using as additional input annotated dictionaries with over two million word forms with their possible Part-of-Speech and lemma. Perhaps a similar input into spaCy's training could improve its performance on previously unknown texts, which will be a part of the follow up research on this subject. In addition, TT19, which had used a bigger lexicon also performed better than TT11 on unfamiliar texts.

Figure 2 compares the nPoS tagging between spaCy and TreeTagger. As in the case of PoS, spaCy shows better results on familiar, while treetagger shows better result when tagging unfamiliar text. Although TreeTagger TT19 seems to have better overall results, the performance of both taggers drops significantly when tagging unknown text.

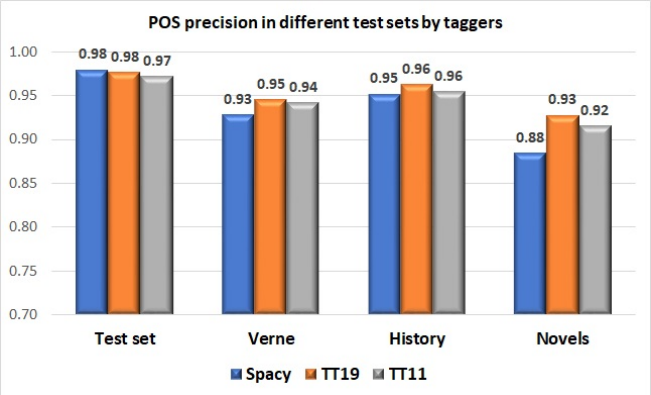

Figure 1: Part-of-Speech tagging accuracy per token on test sets, for each of trained models

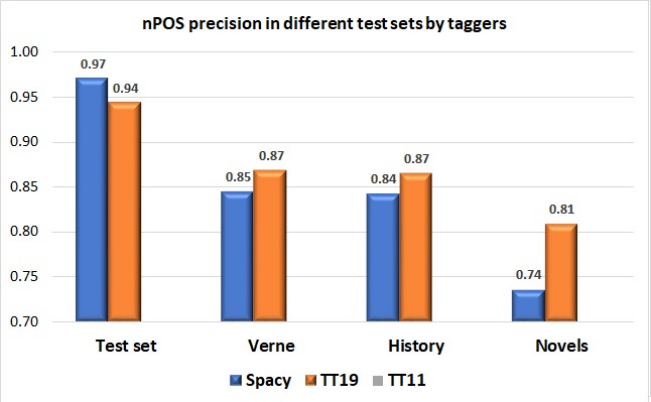

Figure 2: nPoS-tagging accuracy per token on test sets

Comparison of lemmatization precision of the old and new TreeTagger models is shown in Figure 3. It can be observed that there is little difference in results for the test sets, both familiar and unfamiliar, with a descent precision for both, where the new tagger performed slightly better, especially for the Novels test set. This comes as no surprise, due to the fact that it is a very specific text, which is fully covered by the new dictionary used for the TT19 model.

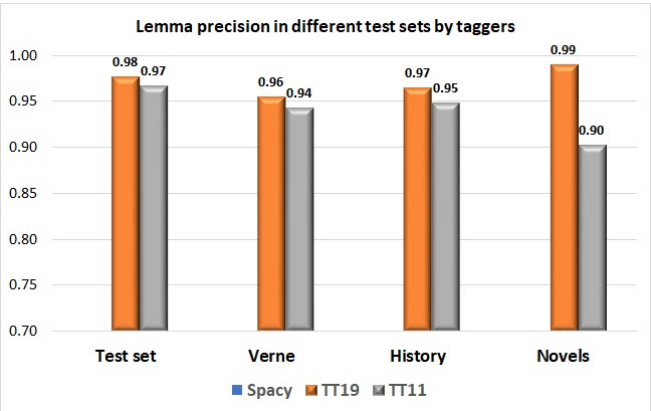

Figure 3: Precision of lemmatization per token, obtained by two TreeTagger based taggers

	sentences			tokens			words		
	train	**dev**	**test**	**train**	**dev**	**test**	**train**	**dev**	**test**
sr_basic	7,677	965	965	141,438	18,055	17,985	119,378	15,194	15,116
sr_25	10,661	1,336	1,348	233,027	29,486	29,801	201,489	25,432	25,762
sr_50	20,432	2,555	2,563	450,428	56,280	56,052	387,242	48,337	48,137
sr_75	30,682	3,822	3,839	712,756	88,447	87,965	611,505	75,858	75,477
sr_100	40449	5053	5,056	952,291	118,238	117,529	819,360	101,830	101,177

Table 4: Training, development and test sets size distribution

We were interested not only in the overall performance of developed taggers and models, but also in their performance for specific PoS-tags. Figures 4, 5 and 6 show the tagging precision per test sets for selected PoS. While tagging results differ for different PoS, it must be noted that the most frequent classes, such as adjectives, nouns and verbs, have the greatest impact on the overall average result and in the feedback during tagger training. Further research could lead to the improvement for classes where the results were not satisfactory, e.g. adverbs and particles. The data set harmonization could improve results for adverbs, which were shown to be the most ambiguous part of speech.

4.2. Training set size impact

For each of the five spaCy trained models, F_1, precision and recall per each tag are represented visually in Figure 7. For the SR_BASIC model, F_1, precision and recall per each tag, as well as the number of tokens in the test set, are given in Table 5. For the SR_{25,50,75,50} models, F_1, precision and recall per each tag, as well as the number of tokens in the test set, are given in Table 6. The value "MISS" represents the number of tokens that were not aligned with the corresponding token in the test set by the spaCy tokenizer.

SR_BASIC				
	F1	**P**	**R**	**#**
ADJ	.91	.91	.92	1713
ADP	.99	1	.99	1458
ADV	.89	.89	.89	787
AUX	.99	.99	.99	1235
CCONJ	.94	.95	.94	838
DET	.97	.98	.95	736
INTJ	.5	.5	.5	4
NOUN	.97	.97	.96	3984
NUM	.98	.97	.99	305
PART	.91	.93	.89	673
PRON	.94	.92	.96	527
PROPN	.91	.92	.9	407
PUNCT	.99	.98	1	2817
SCONJ	.97	.97	.97	554
MISS	0	0	0	55
VERB	.95	.94	.95	1807
X	.82	.74	.93	85

Table 5: F_1, precision, recall and number of samples per each tag

As already mentioned in Section 3.3. spaCy offers pretrained PoS-taggers for many languages. The best performing available models, along with their PoS per-token accuracy scores on the corresponding test set, are: en_core_web_lg for English (97.21%), es_core_news_md for Spanish (97.03%), el_core_news_md for Greek (96.51%), de_core_news_md for German (96.44%), it_core_news_sm for Italian (96.06%), fr_core_news_md for French (95.15%), nl_core_news_sm for Dutch (91.12%), etc.[8] The performance of our models, especially the full model SR-100, is thus comparable to the performance of the available pretrained models.

For each of the trained spaCy models, accuracy per token (measured on all tokens including punctuation) on each test set is displayed in Figure 8. It can be observed that the training set size directly influences the model's performance: the larger the training set, the higher is the accuracy.

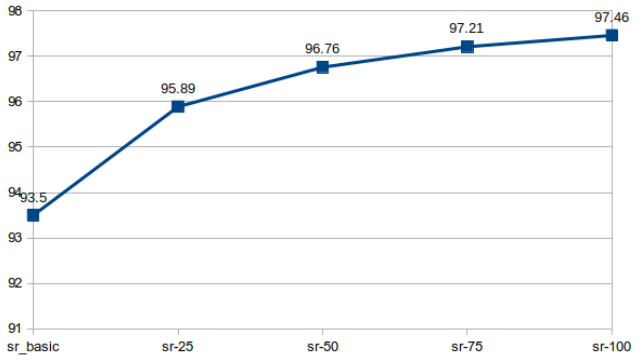

Figure 8: Accuracy per token on test sets, per each spaCy model

5. Conclusion

Publishing of new versions of annotated Serbian corpus requires an accurate automatic tagging system. While first cross-validation results suggested very high accuracy of the selected taggers, their performance on previously unknown texts reveals that there is room for further improvement.

We may conclude that the key to improvement of these tagger lies in the improvement of the training set. Addition of new grammatical categories where they are missing, as well as locating and fixing possible annotating inconsistencies, will be conducted as part of future research on this subject. We believe that evaluation on different text types not used in training phase, will give us solid ground to decide which tagger to use for annotation of new versions of the Corpus of Contemporary Serbian. To that end, other

[8]The best performing spaCy tagger models on December 1, 2020.

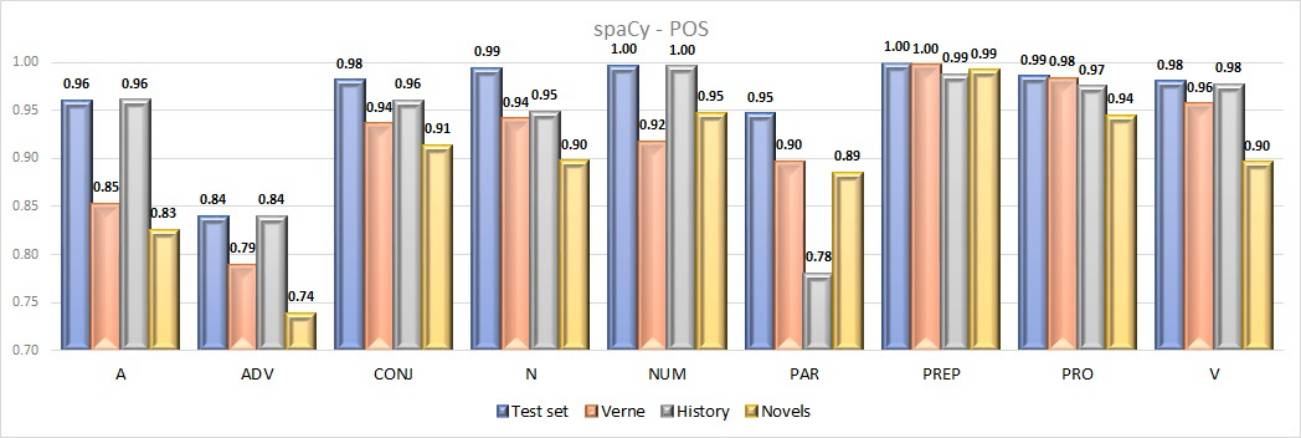

Figure 4: Results of PoS-tagging with SerSpaCy tagger on different test sets and for different PoS-tags

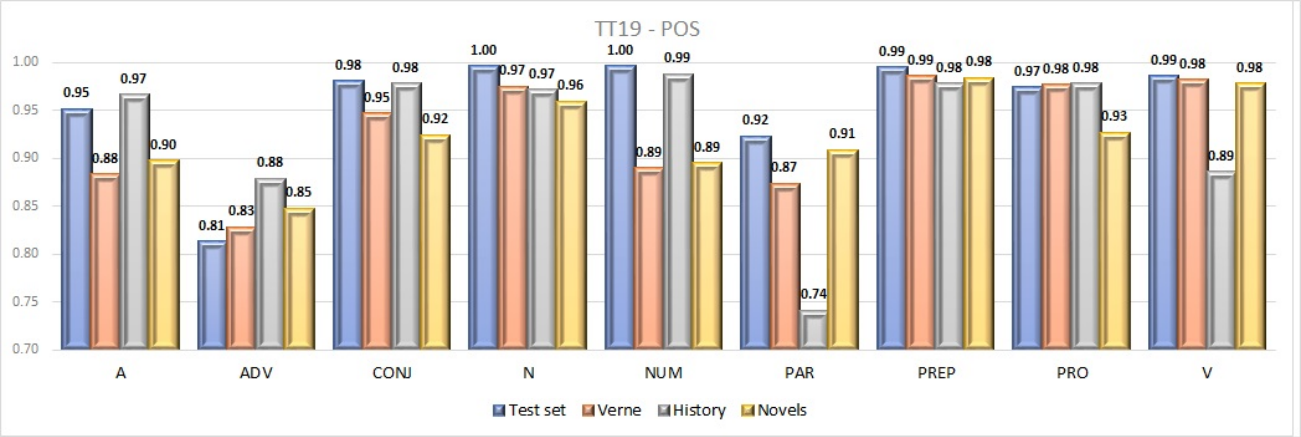

Figure 5: Results of PoS-tagging with TT19 tagger on different test sets and for different PoS-tags

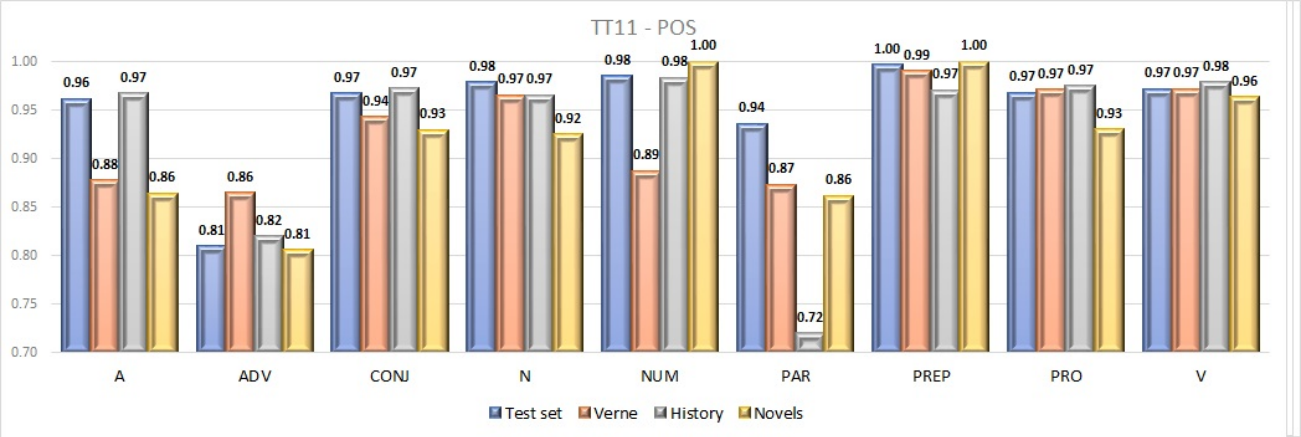

Figure 6: Results of PoS-tagging with TT11 tagger on different test sets and for different PoS-tags

taggers as well as new tagging technologies will be taken into consideration and tested in order to find the best solution for Serbian, a highly-inflected language without fixed word order, for instance RNNTagger.[9] Since CRF tagger for Serbian and Croatian language obtained the accuracy over 98%, as reported in (Ljubešić et al., 2016), we plan to test it on our manually annotated datasets, in order to get a more complete picture of possible solutions for the Serbian language.

Once prepared, the models for tagging will be used to tag the Serbian part of ELTeC (European Literary Text Collection) corpus.[10] It should be noted that within this action the textometry tool TXM (Heiden, 2010) is used for many

[9] RNNTagger, https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/

[10] The current version of ELTeC corpus is available at https://zenodo.org/communities/eltec/.

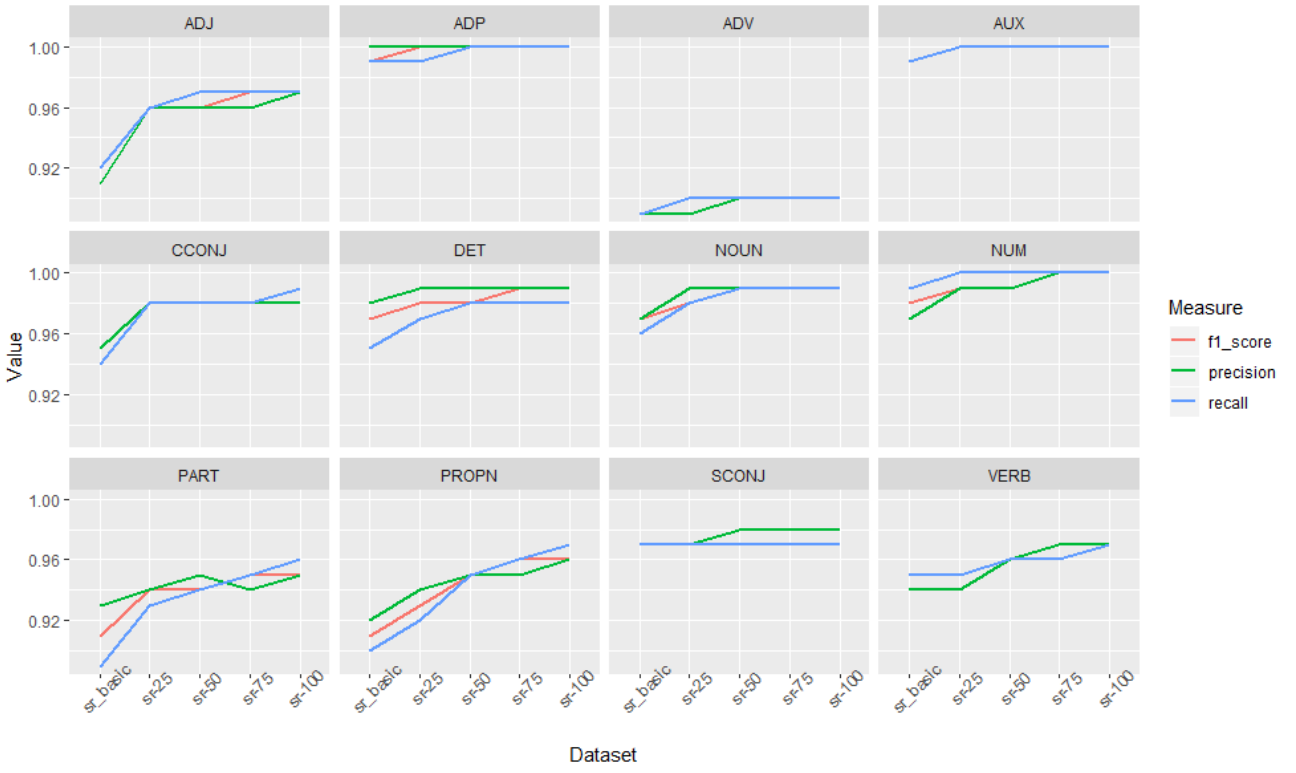

Figure 7: Precision, recall and F1 measure per each tag and spaCy model

	SR_25				SR_50				SR_75				SR_100			
	F_1	P	R	#	F_1	P	R	#	F_1	P	R	#	F_1	P	R	#
ADJ	.96	.96	.96	3856	.96	.96	.97	6769	.97	.96	.97	10326	.97	.97	.97	13692
ADP	1	1	.99	3010	1	1	1	5556	1	1	1	8401	1	1	1	11221
ADV	.9	.89	.9	745	.9	.9	.9	1410	.9	.9	.9	2111	.9	.9	.9	2861
AUX	1	1	1	1097	1	1	1	2290	1	1	1	3475	1	1	1	4781
CCONJ	.98	.98	.98	1633	.98	.98	.98	2854	.98	.98	.98	4562	.98	.98	.99	6016
DET	.98	.99	.97	1031	.98	.99	.98	2007	.99	.99	.98	3343	.99	.99	.98	4731
INTJ	.86	1	.75	4	.75	.75	.75	4	.8	.86	.75	8	.95	1	.91	11
NOUN	.98	.99	.98	8738	.99	.99	.99	15973	.99	.99	.99	24873	.99	.99	.99	33344
NUM	.99	.99	1	888	.99	.99	1	1994	1	1	1	3440	1	1	1	4464
PART	.94	.94	.93	729	.94	.95	.94	1392	.95	.94	.95	2162	.95	.95	.96	2910
PRON	.94	.92	.95	259	.92	.9	.94	526	.93	.92	.94	826	.93	.92	.94	1130
PROPN	.93	.94	.92	687	.95	.95	.95	1179	.96	.95	.96	1636	.96	.96	.97	2006
PUNCT	.98	.96	1	3889	.98	.96	1	7565	.97	.95	1	11839	.98	.95	1	15570
SCONJ	.97	.97	.97	662	.98	.98	.97	1382	.98	.98	.97	2243	.98	.98	.97	3124
MISS	0	0	0	150	0	0	0	353	0	0	0	651	0	0	0	785
VERB	.95	.94	.95	2234	.96	.96	.96	4283	.97	.97	.96	6891	.97	.97	.97	9411
X	.92	.88	.97	189	.97	.96	.97	515	.97	.96	.98	1178	.97	.97	.98	1472

Table 6: F_1, precision, recall and number of samples per each tag

purposes, and it relies for tagging on TreeTagger models. When the baseline annotated dataset (SR_BASIC) is completely cleaned and mended it will be published, and it will be the biggest public dataset for Serbian of this type, with over 200.000 annotated tokens.

6. Acknowledgements

This research was partially supported by Serbian Ministry of Education and Science under the grants #III 47003 and ON 178006, and by COST Action CA16204 Distant Reading for European Literary History.

7. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Choi, J. D. (2016). Dynamic Feature Induction: The Last Gist to the State-of-the-Art. In *Proceedings of the 2016*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 271–281.

Constant, M., Krstev, C., and Vitas, D. (2018). Lexical analysis of serbian with conditional random fields and large-coverage finite-state resources. In Zygmunt Vetulani, et al., editors, *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2015. Lecture Notes in Computer Science, vol 10930.*, pages 277–289. Springer International Publishing, Cham.

Denis, P. and Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, PACLIC 23, Hong Kong, China, December 3-5, 2009*, pages 110–119.

Erjavec, T. (2012). Multext-east: morphosyntactic resources for central and eastern european languages. *Language resources and evaluation*, 46(1):131–142.

Gavrilidou, M., Labropoulou, P., Piperidis, S., Giouli, V., Calzolari, N., Monachini, M., Soria, C., and Choukri, K. (2006). Language resources production models: the case of the intera multilingual corpus and terminology. *Politics*, 39(39):78.

Giménez, J. and Màrquez, L. (2004). Svmtool: A general POS tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*.

Heiden, S. (2010). The txm platform: Building open-source textual analysis software compatible with the tei encoding scheme. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 389–398.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *To appear*.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging.

Krstev, C., Vitas, D., and Erjavec, T. (2004). Morpho-Syntactic Descriptions in MULTEXT-East-the Case of Serbian. *Informatica*, 28(4):431–436.

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

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

Ljubešić, N., Klubička, F., Agić, Ž., and Jazbec, I.-P. (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4264–4270, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Petrov, S., Das, D., and McDonald, R. (2012). A Universal Part-of-Speech Tagset. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Popović, Z. (2010). Taggers Applied on Texts in Serbian. *INFOtheca*, 11(2):21a–38a, December.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Semantically Equivalent Adversarial Rules for Debugging NLP Models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Stanković, R., Krstev, C., Lazić, B., and Škorić, M. (2018). Electronic dictionaries–from file system to lemon based lexical database. In *Proceedings of LREC*, pages 18–W23.

Tufiş, D., Koeva, S., Erjavec, T., Gavrilidou, M., and Krstev, C. (2009). Building language resources and translation models for machine translation focused on south slavic and balkan languages. *Scientific results of the SEE-ERA .NET pilot joint call*, page 5.

Utvić, M. (2011). Annotating the Corpus of Contemporary Serbian. *INFOtheca*, 12(2):36a–47a, December.

8. Language Resource References

Cvetana Krstev, Duško Vitas. (2015). *Serbian Morphological Dictionary - SMD*. University of Belgrade, HLT Group and Jerteh, Lexical resource, 2.0.

Duško Vitas, Cvetana Krstev, Ranka Stanković, Miloš Utvić. (2019). *Sr-Basic: Annotated corpus of Serbian, basic data set*. University of Belgrade, HLT Group and Jerteh, Annotated corpus, 1.0.

Explosion. (2019). *spaCy 2.2*. https://spacy.io/.