SETimes.SR – A Reference Training Corpus of Serbian

Vuk Batanović,[*] Nikola Ljubešić,[†] Tanja Samardžić[‡]

[*]School of Electrical Engineering, University of Belgrade
Innovation Center, School of Electrical Engineering, University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
vuk.batanovic@ic.etf.bg.ac.rs

[†]Department of Knowledge Technologies
"Jožef Stefan" Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ijs.si

[‡]Language and Space Lab
University of Zürich
Freiestrasse 16, 8032 Zürich, Switzerland
tanja.samardzic@uzh.ch

Abstract

In this paper we present SETimes.SR – a gold standard dataset for Serbian, annotated with regard to document, sentence, and token segmentation, morphosyntax, lemmas, dependency syntax, and named entities. We describe the annotation layers and provide a basic statistical overview of them, and we discuss the method of encoding them in the CoNLL and the TEI format. In addition, we compare the SETimes.SR corpus with the older SETimes.HR dataset in Croatian.

1. Introduction

Annotated corpora of Serbian are still extremely scarce, despite the fact that various linguistic resources for Serbian have been under development since the early nineties. On the other hand, considerable advancements have recently been made in NLP technologies for Croatian, a language closely related to Serbian, thanks to a series of projects that resulted in a number of richly annotated data sets. The availability of the parallel Croatian-Serbian SETimes corpus, initially compiled by Tyers and Alperen (2010) and distributed through the OPUS platform (Tiedemann, 2009), and later improved by Agić and Ljubešić (2014), presents a good opportunity for cross-linguistic annotation transfer from Croatian to Serbian.

In this paper, we present SETimes.SR, a richly annotated gold standard dataset for Serbian, developed via an extensive use of the existing Croatian data and models. The SETimes.SR corpus is annotated on the following levels: document, sentence, and token segmentation, morphosyntax, lemmas, dependency syntax, and named entities. To the best of our knowledge, this is the first publicly available corpus in Serbian that contains all the annotation layers required for a full natural language processing pipeline.

The remainder of this paper is structured as follows: in Section 2, we describe the layers of annotation included in the SETimes.SR corpus and we present a statistical overview of label distributions in each layer. In Section 3, we present the method used to encode the corpus data, and in Section 4 we compare the new SETimes.SR dataset with the older SETimes.HR corpus in Croatian (Agić and Ljubešić, 2014). Finally, in Section 5, we present our conclusions and discuss some directions of future work.

2. Corpus Description

The SETimes.SR corpus contains news stories collected from the now defunct Southeast European Times news portal and written in Serbian using the Ekavian pronunciation and the Serbian Latin script. The SETimes portal provided news in English and languages spoken in southeast Europe, and was also the source for the SETimes.HR annotated corpus in Croatian, whose content is parallel to SETimes.SR on the document level and, for the most part, on the sentence level as well.

2.1. Segmentation

SETimes.SR is segmented into 163 documents, close to four thousand sentences, and almost 87 thousand tokens. Hence, the average document length is around 24 sentences or 532 tokens, while the average sentence length is around 22 tokens. A statistical overview of the corpus is given in Table 1.

All documents are preceded by a tag indicating their name. Tokenized sentences are preceded by a tag stating their original, untokenized text, as well as a tag contain-

Item	Count
Documents	163
Sentences	3 891
Tokens	86 726
Types	17 586
Lemmas	8 619
MSDs	557

Table 1: A statistical overview of the SETimes.SR corpus

ing their numerical ID. The form of all these tags is compliant with the Universal Dependencies v2 specification[1]. Each token is also annotated with a tag indicating its spacing from the following token in the original text, making it possible to reconstruct the original texts from their tokenized forms.

2.2. Morphosyntax and Lemmas

Morphosyntax in the SETimes.SR corpus is encoded using 13 part-of-speech categories and numerous morphosyntactic attributes particular to each category. This annotation scheme was proposed in the MULTEXT East (MTE) v5 guideline draft for Bosnian[2]. The choice of this scheme set was motivated by our goal to keep the tagset as close as possible to the one applied in SETimes.HR. At the time of our morphosyntactic annotation, the most up-to-date version of the Croatian tagset was the one used for Bosnian, another of the closely related languages originating, together with Croatian and Serbian, from the former Serbo-Croatian. The only major difference between the tags used in our corpus and the Croatian specification is a tag for the synthetic future tense[3]. A list of MTE POS categories and their frequencies in the SETimes.SR corpus is given in Table 2.

The process of morphosyntactic annotation of SETimes.SR is already briefly described in (Samardžić et al., 2017) in relation to syntactic annotation. Following previous findings that models trained on Croatian data achieve very similar tagging accuracies on both Croatian and Serbian texts (Agić and Ljubešić, 2014; Ljubešić et al., 2016), we first processed the Serbian corpus with the best performing model for Croatian (Ljubešić et al., 2016). The output was then manually corrected by two expert annotators. The training set for the Croatian model included, among others, the parallel SETimes.HR data, which made the automatic annotations already very accurate.

With the rising popularity of cross-linguistic Universal Dependencies annotations, we decided to also generate POS tags in accordance with the Universal Dependencies version 2 encoding system, which consists of 17 part-of-speech categories. The UD POS tags were, for the most part, created via automatic mapping from the MTE morphosyntactic descriptors. A notable exception that had to be manually converted were the abbreviations (MTE tag Y), since the UD standard does not provide a separate POS tag for this category. The MTE-UD mapping table and code are available on the SETimes.SR GitHub repository[4]. The frequency distribution of UD POS tags in the SETimes.SR corpus is shown in Table 3.

2.3. Dependency Syntax

Syntactic dependencies are annotated according to the Universal Dependencies version 2 standard, which describes 37 syntactic relations. Among them, 33 are present in the SETimes.SR corpus, and their distribution in it is given in Table 4. As in the case of morphosyntax, the first step in syntactic annotation was processing the corpus with the most up-to-date Croatian model (Agić and Ljubešić, 2015). Again, the training data for the Croatian model consisted of the parallel SETimes.HR data, which made these initial annotations rather accurate. Manual correction was made in 14% of all syntactic edges.

The process of annotation transfer and correction was described in more detail in (Samardžić et al., 2017). Since the time of that publication, the annotation was completed, validated and shared through the Universal Dependencies infrastructure.[5] The same annotation that can be downloaded as a UD treebank is included in the corpus described here.

To assess the reliability of the annotation, we have measured the inter-annotator agreement on a sample of 300 sen-

MTE POS gloss	POS tag	Count	Percentage
Nouns	N	28 322	32.66%
Verbs	V	12 990	14.98%
Punctuation	Z	10 790	12.44%
Adjectives	A	9 372	10.81%
Adpositions	S	8 460	9.75%
Conjunctions	C	6 032	6.96%
Pronouns	P	4 921	5.67%
Adverbs	R	2 847	3.28%
Numerals	M	2 217	2.56%
Particles	Q	410	0.47%
Residuals	X	350	0.40%
Abbreviations	Y	15	0.02%
Interjections	I	0	0%

Table 2: MTEv5 part-of-speech tag distribution in the SETimes.SR corpus

UD POS gloss	UD POS tag	Count	Percentage
Nouns	NOUN	21 144	24.38%
Punctuation	PUNCT	10 787	12.44%
Adjectives	ADJ	10 392	11.98%
Adpositions	ADP	8 460	9.75%
Verbs	VERB	7 439	8.58%
Proper nouns	PROPN	7 188	8.29%
Auxiliary	AUX	5 551	6.40%
Subord. conj.	SCONJ	3 179	3.67%
Determiners	DET	2 901	3.34%
Coord. conj.	CCONJ	2 853	3.29%
Adverbs	ADV	2 847	3.28%
Pronouns	PRON	2 020	2.33%
Numerals	NUM	1 202	1.39%
Particles	PART	410	0.47%
Other	X	350	0.40%
Symbols	SYM	3	0.01% <
Interjections	INTJ	0	0%

Table 3: UD part-of-speech tag distribution in the SETimes.SR corpus

[1] http://universaldependencies.org/format.html

[2] http://nl.ijs.si/ME/V5/msd/html/

[3] Our plan is to define a single tagset for the Serbo-Croatian macro language (ISO 639-3 code *hbs*).

[4] http://github.com/vukbatanovic/SETimes.SR/

[5] http://hdl.handle.net/11234/1-2837

UD syntactic tag	Count	Percentage
punct	10 783	12.43%
case	8 537	9.84%
nmod	7 914	9.13%
amod	7 542	8.70%
obl	6 894	7.95%
nsubj	6 730	7.76%
aux	4 101	4.73%
root	3 891	4.49%
conj	3 250	3.75%
obj	3 064	3.53%
mark	2 850	3.29%
flat	2 841	3.28%
cc	2 666	3.07%
advmod	2 391	2.76%
nummod	1 798	2.07%
acl	1 692	1.95%
det	1 570	1.81%
cop	1 329	1.53%
ccomp	1 210	1.40%
compound	1 197	1.38%
parataxis	1 187	1.37%
xcomp	973	1.12%
appos	672	0.77%
advcl	636	0.73%
fixed	414	0.48%
discourse	315	0.36%
csubj	164	0.19%
orphan	79	0.09%
goeswith	19	0.02%
list	11	0.01%
dep	3	0.01% <
iobj	2	0.01% <
vocative	1	0.01% <

Table 4: UD syntactic relation distribution in the SETimes.SR corpus

s1-s100		HR1	HR2	HR3	SR
N=2275	HR1	-	93%	93%	91%
	HR2	156	-	94%	92%
Agr=92%	HR3	159	126	-	92%
	SR	194	174	179	-
s101-s200		HR1	HR2	HR3	SR
N=2194	HR1	-	94%	94%	92%
	HR2	132	-	94%	92%
Agr=93%	HR3	114	140	-	91%
	SR	168	169	187	-
s201-s300		HR1	HR2	HR3	SR
N=2246	HR1	-	94%	94%	92%
	HR2	128	-	93%	92%
Agr=93%	HR3	142	153	-	91%
	SR	178	190	197	-

Table 5: UD annotation agreement between three Croatian native speakers (HR) and one Serbian (SR). The lower sides show the number of disagreements, the upper sides the agreement scores; N=number of tokens; Agr=average agreement scores.

tences, split into three groups of 100 sentences annotated at different time points. Each of these groups was annotated by four annotators: three Croatian native speakers and one Serbian.

The agreement scores between each pair of annotators are shown in Table 5. The agreement measure we use is the proportion of identically annotated tokens (same morphosyntactic label, dependency link, and dependency label) out of all annotated tokens (the upper sides in Table 5). The overall average agreement is slightly below 93%. It is a bit higher within the group of Croatian annotators, and a bit lower between the Serbian annotator and the Croatian group. This distinction, however, is not necessarily due to linguistic differences, but rather due to the fact that the Croatian team was trained together and separately from the Serbian annotator.

2.4. Named Entities

Named entity annotations are encoded in the IOB2 format and include the following five types of entities:

- Person (PER)
- Person derivative (DERIV-PER)
- Location (LOC)
- Organization (ORG)
- Miscellaneous (MISC)

The PER, LOC, ORG, and MISC categories are standard, while the DERIV-PER tag was introduced in order to mark personal possessive adjectives, e.g. ***Darvinova*** *teorija* 'Darwin's theory'. This addition is intended to potentially improve personal data anonymization methods in Serbian. This annotation scheme was originally developed during the annotation of the Slovene ssj500k and Janes-Tag datasets[6].

Almost seven thousand named entities were encountered in SETimes.SR, or around 42 per document, which is high, but not surprising given the journalistic nature of the texts within the corpus. The distribution of named entity types in SETimes.SR is shown in Table 6, while Table 7 contains the distribution of tokens belonging to a named entity.

Named entity type	Count	Percentage
Person	1 884	27.35%
Person derivative	75	1.09%
Location	2 678	38.88%
Organization	1 953	28.35%
Miscellaneous	298	4.33%
Total	6 888	100%

Table 6: Distribution of named entities in the SETimes.SR corpus

[6] `http://nl.ijs.si/janes/wp-content/uploads/2017/09/SlovenianNER-eng-v1.1.pdf`

Named entity type	Token count	Percentage
Person	3 045	3.51%
Person derivative	75	0.09%
Location	3 137	3.62%
Organization	3 369	3.88%
Miscellaneous	788	0.91%
Total	10 414	12.01%

Table 7: Distribution of named entity tokens with regard to the entire SETimes.SR corpus

The annotation of named entities was performed in the online tool WebAnno. Two annotators performed independent annotations, while a third annotator curated the collisions between them.

3. Corpus Encoding and Publishing

The working version of the SETimes.SR corpus was encoded in a modified variant of the tabular CoNLL-X format (Buchholz and Marsi, 2006), which consists of the following columns:

1. ID, token index in a sentence
2. FORM, token surface form
3. LEMMA, token lemma
4. POS, part of speech according to the MULTEXT East v5 standard
5. MSD, morphosyntactic description according to the MULTEXT East v5 standard
6. MSDFEAT, morphosyntactic features according to the MULTEXT East v5 standard
7. ___, a column left blank in order to preserve formatting equivalence with the hr500k corpus (Ljubešić et al., 2018), which contains older, non-UD dependency relation tags in this position
8. UDDEPREL, dependency relation (head, label) according to the UDv2 standard
9. UPOS+FEATS, part of speech and morphological features according to the UDv2 standard
10. UDSPEC, UDv2 language specific feature tag, used to encode the spacing between tokens in the original sentence texts
11. NER, named entity annotations encoded through IOB2

The CoNLL-type format was then converted to XML according to the TEI, (Guidelines for Electronic Text Encoding and Interchange (TEI Consortium, 2017)), in order to ensure (meta-)data persistence. Apart from the automatic conversion of the text and its annotations, this also involved writing the `teiHeader` element, which gives the metadata of the corpus, containing its name, authors, license, source description, annotation vocabulary, tag usage, revision history etc.

Each sentence in the TEI encoding (`s`), as well as each token (words (`w`) and punctuation symbols (`pc`)), is assigned a unique ID, as illustrated in Figure 1. White space in the sentence is also marked-up, with `c`. The `@lemma` attribute contains the lemma of the words, while the MULTEXT-East MSD is given in the `@ana` attribute. The UD parts of speech and features are placed within the `@msd` attribute, which is an attribute newly introduced into the TEI. Note that the double pipe symbol is used to separate the universal features from the (Serbian) language specific ones. The reason why the MULTEXT-East MSDs are not given in the `@msd` attribute, as might be expected, is that while `@msd` can contain any string, the `@ana` is defined as a pointer, which MULTEXT-East MSDs can be, but UD features cannot. We explain below in more detail the functioning of TEI pointers for linguistic labels as used in the SETimes.SR corpus. Named entities are encoded in-line, by simply using the standard TEI `name` element. Within it, the `@type` attribute contains the type of the named entity.

The final layer of annotation are the UD dependencies, which are encoded in a stand-off format, using the link group (`linkGrp`) element. `linkGrp` is an element of `s` and has attributes specifying its type (here used for the

```
<s xml:id="s2">
  <name type="loc">
    <w xml:id="s2.1" lemma="Kosovo" ana="mte:Npnsn"
       msd="UposTag=PROPN|Case=Nom|Gender=Neut|Number=Sing">Kosovo</w>
  </name>
  <c> </c>
  ...
  <w xml:id="s2.10" lemma="pritužba" ana="mte:Ncfpg"
     msd="UposTag=NOUN|Case=Gen|Gender=Fem|Number=Plur||SpaceAfter=No">pritužbi</w>
  <pc xml:id="s2.11" ana="mte:Z" msd="UposTag=PUNCT">.</pc>
  <linkGrp targFunc="head argument" type="UD-SYN">
    <link ana="ud-syn:nsubj" target="#s2.3 #s2.1"/>
    <link ana="ud-syn:advmod" target="#s2.3 #s2.2"/>
    ...
  </linkGrp>
</s>
```

Figure 1: TEI encoding of a corpus sentence

Serbian:	Mogu samo ***da*** *zaključim* ***da*** *nećemo postići napredak* pred Savetom.
Croatian:	Mogu samo *zaključiti kako nećemo ostvariti napredak* u Vijeću.
English:	I can only *conclude that we are not going to progress* in the Council.

Figure 2: Differences between Serbian and Croatian in the usage of the *da* subordinating conjunction

annotation layer label) and the ordering of the arguments of the links. It also contains the links themselves. Each link is comprised of a link label and pointers to the IDs of the link head and argument. In cases where a syntactic dependency has the (virtual) root as its head, the sentence ID is used as the ID of the head (in the example in Figure 1 that would be `#s2`).

As mentioned, the `@ana` attribute is a pointer, which usually contains a local reference to an ID (e.g. `#s2.1`) or a fully qualified URI. TEI has another option for its pointers, namely using a prefix before the ID and separated from it by a colon (e.g. `mte:Npnsn`). Such pointers are then resolved using the `prefixDef` element in the TEI header, which defines the prefixing schema used, showing how abbreviated URIs using the scheme may be expanded into full URIs. In the case of the SETimes.SR corpus all the prefixes are simply expanded to local references, which are given in the TEI header. The only exception are the MULTEXT-East MSDs, which are defined in the `back` element of the TEI document as a feature-structure giving the decomposition of the MSD into its features. It is therefore quite simple, using just the TEI encoded corpus, to move, for example, from `mte:Mdo` to `Category = Numeral, Form = digit, Type = ordinal`.

The TEI encoded corpus, which is to be regarded as the canonical version of SETimes.SR, was then automatically converted to the so-called vertical format, which is used by CQP-based concordancers, in particular by the (no)Sketch Engine (Rychlý, 2007). The vertical format is able to encode hierarchical structures (e.g. sentences and names), and token annotations (e.g. lemmas and MSDs), but not links between tokens (e.g. dependencies). To nevertheless preserve as much of this information as possible, the dependencies are annotated next to tokens, so that the argument token is annotated with the dependency label and head lemma.

Finally, the TEI, vertical and CoNLL encodings of SETimes.SR were deposited to the CLARIN.SI repository[7], where the data is available under a Creative Commons license. The corpus is also available for exploration via the CLARIN.SI noSketch Engine and KonText concordancers, to which the links are included on the CLARIN.SI repository page.

4. Comparison with SETimes.HR

Since the SETimes.HR corpus in Croatian preceded SETimes.SR and was instrumental in its creation, it is interesting to compare the two corpora and identify the similarities and differences between them. Instead of the original SETimes.HR corpus (Agić and Ljubešić, 2014), we consider the SETimes.HR portion of the hr500k corpus, since it contains both the new annotation layers, as well as updates and corrections within the original annotation layers (Ljubešić et al., 2018).

Item	SR	HR
Documents	163	163
Sentences	3 891	3 757
Tokens	86 726	83 630

Table 8: A comparison between the SETimes.SR corpus and the SETimes.HR part of the hr500k corpus

Both corpora consist of the same number of documents gathered from the same source, but the Croatian one contains fewer sentences and tokens, as shown in Table 8. Work is currently under way to insert any missing sentences from the original SETimes parallel corpus (Tyers and Alperen, 2010) into both the Serbian and the Croatian dataset, thereby reducing the sentence and token count differential between them to a minimum, enabling maximal parallelism.

Tables 9 and 10 contain comparisons of part-of-speech tag frequencies, according to the MTEv5 and the UDv2 standard, respectively. The relationship between the Serbian and the Croatian corpus frequencies for each tag is analyzed using the $\tilde{\chi}^2$ test, quantifying the probability that the difference between the observed and the expected frequencies is due to chance. We use the Phi (Φ) coefficient to measure the effect size.

The largest differences exist with regard to the conjunction category or, more specifically, subordinating conjunctions. This difference is chiefly due to the ”*da*” subordinating conjunction, which is used much more frequently in Serbian than in Croatian (SETimes.SR: 2302, SETimes.HR: 507, $\tilde{\chi}^2$ = 1099.97, p = **3.3E-241**, Φ = 0.08035). In Serbian, unlike Croatian, ”*da*” is used in complex predicates involving modal and phase verbs, as well as within a complex form of the future tense. Figure 2 presents an example of these differences between Serbian and Croatian. We also detected a significant stylistic difference regarding the frequency of pronouns, which stems from the fact that the texts in SETimes.HR employ zero anaphora more often than those in SETimes.SR.

We did not compare the frequencies of dependency relation tags between SETimes.SR and SETimes.HR since somewhat different dependency annotation guidelines were used for each corpus (e.g. the UD syntactic tag *expl* appears 971 times in SETimes.HR but is not used at all in the Serbian corpus). On the other hand, we did perform a comparison regarding the named entity annotation frequencies, but they were very similar and no statistically significant differences could be found.

[7] `http://hdl.handle.net/11356/1200`

MTE POS gloss	POS tag	SR	HR	$\tilde{\chi}^2$	p-value	Φ
Nouns	N	28 322	**28 009**	13.36	**0.00026**	0.00886
Verbs	V	12 990	**12 605**	0.29	0.59139	0.00130
Punctuation	Z	10 790	**10 778**	7.63	**0.00575**	0.00669
Adjectives	A	9 372	**9 322**	5.01	**0.02519**	0.00542
Adpositions	S	8 460	**8 282**	1.04	0.30789	0.00247
Conjunctions	C	**6 032**	4 752	116.15	**4.4E-27**	0.02611
Pronouns	P	**4 921**	4 306	22.83	**1.8E-06**	0.01158
Adverbs	R	2 847	**2 785**	0.28	0.59377	0.00129
Numerals	M	**2 217**	2 081	0.77	0.37937	0.00213
Particles	Q	410	**481**	8.38	**0.00378**	0.00702
Residuals	X	**350**	205	32.43	**1.2E-08**	0.01380
Abbreviations	Y	15	**24**	1.95	0.16304	0.00338
Interjections	I	0	0	—	—	—
Total		86 726	83 630	—	—	—

Table 9: MTEv5 part-of-speech frequency comparison between SETimes.SR and the SETimes.HR part of the hr500k corpus. Frequencies that are larger than expected and p-values below the 0.05 level are in bold.

UD POS gloss	UD POS tag	SR	HR	$\tilde{\chi}^2$	p-value	Φ
Nouns	NOUN	21 144	**20 913**	8.95	**0.00278**	0.00725
Punctuation	PUNCT	10 787	**10 774**	7.58	**0.00589**	0.00667
Adjectives	ADJ	10 392	**10 210**	2.02	0.15485	0.00345
Adpositions	ADP	8 460	**8 282**	1.04	0.30789	0.00247
Verbs	VERB	**7 439**	6 988	2.67	0.10213	0.00396
Proper nouns	PROPN	7 188	**7 119**	2.76	0.09690	0.00402
Auxiliary	AUX	5 551	**5 617**	6.88	**0.00870**	0.00636
Subordinating conjunctions	SCONJ	**3 179**	2 017	225.89	**4.7E-51**	0.03641
Determiners	DET	**2 901**	2 699	1.82	0.17748	0.00327
Coordinating conjunctions	CCONJ	**2 853**	2 735	0.04	0.83357	0.00051
Adverbs	ADV	2 847	**2 785**	0.28	0.59377	0.00129
Pronouns	PRON	**2 020**	1 607	33.75	**6.3E-09**	0.01408
Numerals	NUM	1 202	**1 173**	0.07	0.78559	0.00066
Particles	PART	410	**481**	8.38	**0.00378**	0.00702
Other	X	**350**	205	32.43	**1.2E-08**	0.01380
Symbols	SYM	3	**25**	16.53	**4.8E-05**	0.00985
Interjections	INTJ	0	0	—	—	—
Total		86 726	83 630	—	—	—

Table 10: UD part-of-speech frequency comparison between SETimes.SR and the SETimes.HR part of the hr500k corpus. Frequencies that are larger than expected and p-values below the 0.05 level are in bold.

5. Conclusion

In this paper we have presented SETimes.SR - the first publicly available gold standard corpus of Serbian annotated on the level of document, sentence, and token segmentation, morphosyntax, lemmas, dependency syntax, and named entities. We have described and given a statistical overview of each annotation layer, and presented the way in which the annotations are encoded. We have also compared the new SETimes.SR corpus with the older SETimes.HR dataset in Croatian.

We believe that the creation of SETimes.SR is an important first step in bridging the gap between Serbian and other Slavic languages, such as Czech or Slovene, for which numerous linguistic resources and tools are available. We also hope that the introduction of the SETimes.SR corpus will promote and accelerate the development of other NLP resources and tools for Serbian. In the future, we plan to continue working on the corpus by expanding it with new kinds of annotations, such as a coreference layer. We will also consider enlarging the corpus with additional data.

6. Acknowledgements

This work is supported by the Regional Linguistic Data Initiative (ReLDI) via the Swiss National Science Foundation grant 160501, and the Slovenian Research Infrastructure CLARIN.SI. Vuk Batanović is also partially supported by the III44009 research grant of the Ministry of Education, Science and Technological Development of the Republic of Serbia.

7. References

Željko Agić and Nikola Ljubešić. 2014. The SETIMES.HR Linguistically Annotated Corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC*

2014), pages 1724–1727, Reykjavik, Iceland. European Language Resources Association (ELRA).

Željko Agić and Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that work for Serbian, too). In *Proceedings of the Fifth Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, pages 1–8, Hissar, Bulgaria. Association for Computational Linguistics.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on Multilingual Dependency Parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X 2006)*, pages 149–164, New York City, NY, USA. Association for Computational Linguistics.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Nikola Ljubešić, Željko Agić, Filip Klubička, Vuk Batanović, and Tomaž Erjavec. 2018. hr500k – A Reference Training Corpus of Croatian. In *Proceedings of the 2018 Language Technologies and Digital Humanities Conference (JT-DH 2018)*. Ljubljana, Slovenia.

Pavel Rychlý. 2007. Manatee/Bonito - A Modular Corpus Manager. In *Proceedings of the First Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*, pages 65–70, Brno, Czech Republic. Masaryk University.

Tanja Samardžić, Mirjana Starović, Željko Agić, and Nikola Ljubešić. 2017. Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, pages 39–44, Valencia, Spain. Association for Computational Linguistics.

TEI Consortium. 2017. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Jörg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (RANLP)*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Francis M. Tyers and Murat Serdar Alperen. 2010. South-East European Times: A parallel corpus of Balkan languages. In *Proceedings of the LREC 2010 Workshop on Exploitation of Multilingual Resources and Tools for Central and (South) Eastern European Languages*, pages 49–53. European Language Resources Association (ELRA), Valletta, Malta.