Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages

Tanja Samardžić
URPP Language and Space
University of Zürich
tanja.samardzic@uzh.ch

Mirjana Starović
Leksikom, Belgrade
djelmas@eunet.rs

Željko Agić
IT University of Copenhagen
zeag@itu.dk

Nikola Ljubešić
University of Zagreb
nljubesi@ffzg.hr

Abstract

The paper documents the procedure of building a new Universal Dependencies (UDv2) treebank for Serbian starting from an existing Croatian UDv1 treebank and taking into account the other Slavic UD annotation guidelines. We describe the automatic and manual annotation procedures, discuss the annotation of Slavic-specific categories (case governing quantifiers, reflexive pronouns, question particles) and propose an approach to handling deverbal nouns in Slavic languages.

1 Introduction

The notion Universal Dependencies (UD) refers to an international movement started with the goal to reduce to a minimum cross-linguistic variation in the formalisms used to label syntactic structure (McDonald et al., 2013; Nivre et al., 2016). This goal was defined following multilingual parsing campaigns (Buchholz and Marsi, 2006; Hajič et al., 2009) that revealed substantial cross-linguistic differences in the sets of labels and relations used in different treebanks, making it hard to compare parsers' performances across languages (McDonald and Nivre, 2007).

In this paper, we document the process of building a UD treebank for Serbian underlining the advantages of using the existing general framework, but also data and tools already available for other languages. The availability of shared resources is especially important for languages such as Serbian, which, more than 20 years after the publication of Penn Treebank (Marcus et al., 1994), still has no resource with annotated syntactic structure, lagging behind its close relatives for which UD annotation is available.

Labeled as *automatic conversion with manual corrections* in the UD documentation,[1] our approach consists of four steps: 1) automatic porting of Croatian annotation to Serbian, 2) comparison and adaptation, 3) automatic conversion and correction, and 4) manual correction.

Despite the fact that Serbian can be parsed with the model already available for Croatian, as argued by Agić and Ljubešić (2015), building a Serbian treebank is useful for two reasons. First, it allows learning a more precise model for Serbian, taking into account important syntactic differences such as, for instance, the use of infinitive (Tiedemann and Ljubešić, 2012). Second, improvements and corrections in the Serbian treebank can be ported back and used for updating Croatian treebank. This does not only concern improvements in consistency resulting from detailed manual inspection, but also version updating. In particular, the currently available Croatian treebank follows the UD guidelines version 1 (UDv1), while Serbian follows the current version 2 (UDv2).

2 Applying Croatian Model to Serbian

To port the existing Croatian annotation to Serbian, we use the Croatian data and tools described by Agić and Ljubešić (2015).

The Serbian treebank consists of sentences that are aligned with Croatian sentences in the SETimes.HR corpus (Agić and Ljubešić, 2014) used to produce the first version of the Croatian UD treebank. As morphosyntactic annotation is needed as input for syntactic parsing, we

[1] http://universaldependencies.org/

Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 39–44,
Valencia, Spain, 4 April 2017.

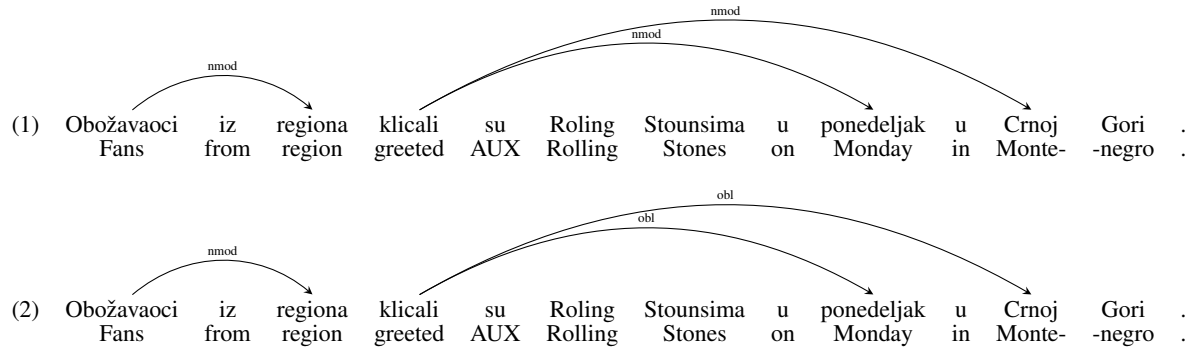

Figure 1: The difference between UDv1 (1) and UDv2 (2) in applying the label `nmod`.

In	Out	Context
auxpass	aux	ALL
csubjpass	csubj	ALL
dobj	obj	ALL
iobj	obl	ALL
nsubjpass	nsubj	ALL
mwe	fixed	ALL
remnant	orphan	ALL
dislocated	NA	ALL
name	flat	ALL
foreign	flat	ALL
nmod	obl	if the PoS of the head is V or A, or N if the lemma ends in *-nje*

Table 1: Automatic conversion from UD v1 to UD v2.

In	Out	Context
expl	NA	ALL
reparandum	NA	ALL
det	det:numgov	if the lemma is "koliko"
nummod	nummod:gov	if the word is a cardinal number and the head is in the genitive case
compound	amod nmod flat	if the PoS is A if the PoS is N otherwise if the lemma is not "sebe"
ALL	compound	if the lemma is "sebe"
ALL	det	if the word is a "possessive pronoun"
ALL	xcomp	if the head word is the modal "moći"

Table 2: Automatic version-independent updates.

add morphosyntactic definitions (MSD) following the modified Multext-East version 4 format (Erjavec, 2012) documented in the draft of version 5.[2] MSD annotation is first added automatically using the state-of-the-art Croatian tagger described by Ljubešić et al. (2016), and then corrected manually by two experts native in Serbian, resulting in gold MSD labels.

Once morphologically annotated, the Serbian side of SETimes.HR, coined SETimes.SR, was then parsed using the `mate-tools`, a graph-based dependency parser (Bohnet, 2010) trained on the Croatian UD v1.2 treebank data. The parser was trained with default parameters.

3 Category Comparison and Adaptation

In this step, we perform manual inspection of a sample of parsed sentences in order to decide what categories and relations to use for Serbian. We extract and evaluate a handful of examples of all annotated relations, comparing the annotation to the general guidelines and to the language-specific entries for Croatian and other contemporary Slavic languages available in the current UD set: Bulgarian, Croatian, Czech, Polish, Russian, Slovak, Slovenian and Ukrainian.

We introduce two kinds of changes with respect to the initial set of categories implemented by the Croatian model. With the first set of changes, we convert general relations UDv1 to UDv2. With the second set of changes, we correct the existing annotation in order to resolve some of the issues raised on the UD web site and improve the descriptive adequacy of the annotation.

3.1 Version Updating

The most important conceptual novelty in the UDv2 guidelines, at least when it comes to Slavic syntax, is the treatment of core vs. oblique arguments of predicates. Based on well-established typological distinctions (Thompson, 1997; Andrews, 2007), UDv1 guidelines stated that a distinction should be made between core and oblique arguments, rather than between complements and adjuncts. Both `obj` and `iobj` were intended to

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

be used for core arguments only, while other labels were intended for oblique arguments.

However, the Slavic treebanks that we consulted systematically use `iobj` to annotate oblique dependents. We believe that this is partly due to sometimes underspecified general guidelines and partly to the strong tradition of making the complement vs. adjunct distinction, which creates the need to distinguish between two kinds of oblique dependents (complements obligatory, adjuncts optional).

We adopt the distinction between core and oblique arguments by implementing the rows 3 and 4 in Table 1. We use `obj` only for direct objects (bare nominal dependents with accusative case) and the new label `obl` for all the other verb dependents, most of which are currently annotated with `iobj` in Croatian and all the other Slavic treebanks. Our new label `obl` includes Serbian counterparts of "dative subjects" indicated as a special construction in Russian documentation.

Another important change is narrowing the use of the relation `nmod` to the nominal domain, as illustrated in Figure 1. We implement this as shown in Table 1, row 11.

Three changes, (rows 1, 2, 5 in Table 1) are made following the UDv2 treatment of passive. We note that the change in the new version of the guidelines is convenient for describing Serbian, as well as other Slavic languages, because the distinction between passive and other intransitive constructions is considerably blurred in these languages.

Finally, we update the relations used for different kinds of conventionalised expressions (rows 6-10 in Table 1, NA as output means that the relation is removed from the list).

3.2 Version-independent Updates

A number of changes are made after inspecting Croatian counterparts of the constructions listed under "special constructions" in the UD language-specific documentations for Slavic languages (available only for Czech, Russian, and Bulgarian) with the goal to improve cross-linguistic parallelism. We make decisions on several issues discussed in this section.

The most prominent specific constructions, discussed in Czech and Russian documentations, are those involving **case governing quantifiers**, such as *koliko*, 'how much, how many', *nekoliko* 'some, several', *mnogo* 'much, many', *malo* 'little, few'. What is special in these constructions is that the case of the head nominal does not depend on the function of the nominal in a clause, but is determined by the quantifier (genitive case is required). To capture this phenomenon, general labels `nummod` and `det` are extended to `nummod:gov` and `det:numgov`, respectively. This specification is applied only in Czech and Russian, although it is relevant to the other Slavic languages too. In this case, we decide to follow Czech and Russian, as shown in Table 2, rows 3–4. We do not follow Czech in using `det:nummod` for those quantifiers that do not govern the case. Since this relation is syntactically equivalent to the simple `det` relation (quantifier agrees with the quantified noun in case), we leave the simple label.

The other constructions addressed in Czech documentation is "**reflexive pronoun**", whose short form can be assigned a whole range of functions. Czech documentation lists the following relations: `dobj`, `iobj`, `nmod`, `auxpass:reflex`, `expl`, and `discourse`. While annotation of this form is not explicitly addressed in the documentation of the other Slavic languages, it can have similar functions, which are likely to be annotated using different subsets of the relations listed above (for instance, the label `auxpass:reflex` is not used in any other Slavic language).

Croatian departs from all the other Slavic languages by using the relation `compound` for most of the instances of this form, rather than annotating fine-grained distinctions. This decision is based on the view of this form as a detachable morpheme belonging to the verb to which it is attached both in lexical and morphological sense. In this view, the "reflexive pronoun" becomes parallel with English or German verb particles, and the relation used for these particles can be applied to it. We note that this view is supported by substantial theoretical findings showing that the short reflexive form is not just a prosodic variant of the full reflexive pronoun and that, in fact, it is not a pronoun at all (Sells et al., 1987; Moskovljević, 1997). Furthermore, Reinhart and Siloni (2004) and Marelj (2004) argue that this form should be analysed in the same way in all its uses: as a free morpheme marking absence of one of the verb's core dependents. The functions listed above, and a whole range of other functions usually not mentioned in

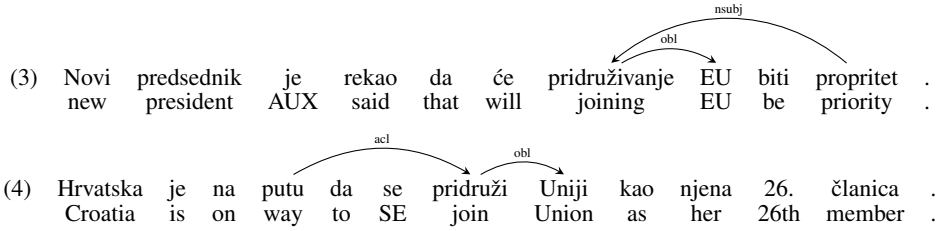

Figure 2: Parallelism between deverbal nouns (*pridruživanje*) and their source verbs (*pridružiti*).

grammars, are higher-level interpretations of the same syntactic form. Annotating these functions, in our opinion, should not be part of UD.

Based on these arguments, we follow Croatian in using the label `compound`, despite the fact that this is not in accordance with the other Slavic treebanks. We extend this relation to all instance of the short reflexive form and eliminate all the other labels (e.g., `dobj`), that are occasionally found in the initial annotation, as shown in the row 6 in Table 2. We also eliminate all the other uses of the relation `compound` (row 5 in Table 2).

The last specific construction, addressed in Bulgarian documentation, is the particle used to form **YES/NO questions**. This particle is assigned the relation `discourse` in Bulgarian, while the relation `mark` is used in Croatian. In this case too, we follow Croatian annotation as this particle does not link the sentence to a broader context, but rather marks the function of the sentence itself.

The revision of the relations resulted in removing two labels found not to be used in the annotation (rows 1-2 in Table 2).

In addition to the constructions listed in language-specific documentations, we note one more form whose annotation needs to be specifically documented: **deverbal nouns**. This category is not specific to Slavic languages, but its annotation might be due to a specific realisation of the distinction between result and process deverbal nominals (Grimshaw, 1990).

Deverbal nouns can have a different degree of nominal and verbal properties across languages and within a language. Those whose meaning is a result are closer to the nominal side of the scale, while those that describe a process are closer to the verbal side. While result nouns can be annotated as other abstract nouns, process deverbal nouns keep the initial verbal (non-finite) dependencies, which means that their dependents should be annotated in the same way as the dependents of the verbs from which they are derived (like infinitives and some participles). Some examples in general UD guidelines suggest that English *-ing* forms with nominal functions are treated as verbs in this respect.

Size in Tokens	Automatic N	Automatic %	Manual N	Manual %	Start–End N	Start–End %
26708	4499	17	3785	14	7423	28

Table 3: The amount of changed annotations in automatic conversion, manual correction, and in the resulting treebank compared with the initial annotation ported from Croatian (Start–End).

Serbian (and Croatian) morphology allows drawing a relatively clear difference between result and process deverbal nouns: the suffix *-nje* is used to derive process nouns in a rather regular way, while a number of idiosyncratic suffixes are used to derive result nouns. We mark this distinction by annotating the dependents of deverbal nouns ending in *-nje* ((3) in Figure 2) in the same way as the dependents of the non-finite forms of their source verbs ((3) in Figure 2), while keeping their nominal function. We treat the other deverbal nouns (derived with other suffixes) as regular nominals.

As a result of this step, we did not manage to eliminate all the differences with other Slavic treebanks, but we believe that our analysis provides a good basis for future steps in this direction. Relatively frequent versioning planned within the UD work framework makes room for continuous improvements and adaptations. This can be expected to move the current annotation to a more synchronised state through active cross-linguistic exchange enabled by the common framework and based on sound arguments.

4 Automatic Conversion and Manual Correction

Here we describe the implementation of the described updates in 1200 sentences, out of the planned 3900.

Tables 1 and 2 show the full list of changes introduced automatically by means of a custom Python script that takes as input parsed sentences in the CoNLL-X format and outputs the same format with the changes. The tables contain all the changes discussed in the previous section, together with a number of changes performed to address issues concerning the current Croatian annotation that have been raised so far on the UD web site and that have not been addressed through the version updating (rows 5, 7, 8 in Table 2).

The processed files are then imported into DgAnnotator[3] and corrected by three experts, Croatian native speakers, coordinated and supervised by a Serbian expert. Manual correction included idiosyncratic or complex cases that could not be performed automatically. In addition to parser's errors, these corrections addressed shortcomings identified on the UD web site. In particular, we manually correct instances of relative pronouns, such as *što* 'what', *koji* 'which', that were annotated with `mark`. We assign such words a function that they have in the subordinate clause, mostly `nsubj` and `obj`.

Table 3 shows the amount of corrections made in each step. The counts refer to the number of tokens for which either the dependency link or relations are changed. We can see that a total of 28% tokens were changed between the initial ported annotation and the final Serbian treebank. Slightly more changes were made automatically than manually (17% vs. 14%). The fact that the sum of the changes is higher than the difference between initial and final annotation means that the annotators had to change back a number of annotations after the automatic conversion. This number is rather low (3% of tokens) but further inspections might show a way to improve automatic conversion. The percentage of manually corrected annotations is lower than it would be expected based on the parsing accuracy score of 79.6% reported by Agić and Ljubešić (2015). This is due to the fact that the Serbian side of the SETimes corpus is very similar to the Croatian side on which the parser was trained.

[3] `http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/`

5 Conclusion and Future Work

By describing the development of a new UD treebank for Serbian, we have demonstrated how the existing UD infrastructure can be used to improve cross-linguistic parallelism in syntactic annotation, but also to reduce costs of development of new treebanks. Such an infrastructure is especially useful for Slavic languages, whose syntax is similar enough to take advantage of cross-linguistic automatic parsing and common annotation guidelines.

The remaining 2700 sentences will be annotated and made available through the UD infrastructure by the end of April 2017, together with our language-specific guidelines and detailed statistics.

Acknowledgments

The annotation described in this paper is funded by the Swiss National Science Foundation grant No. 160501. We are thankful to our collaborators Daša Farkaš, Danijela Merkler and Matea Srebačić for their valuable contribution.

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'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Željko Agić and Nikola Ljubešić. 2015. Universal dependencies for Croatian (that work for Serbian, too). In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pages 1–8, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Avery D. Andrews. 2007. The major functions of the noun phrase. In Timothy Shopen, editor, *Language Typology and Syntactic Description Clause Structure*, pages 132–223, Cambridge, United Kingdom. Cambridge University Press.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing.

In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, June. Association for Computational Linguistics.

Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic resources for central and eastern European languages. *Lang. Resour. Eval.*, 46(1):131–142, March.

Jane Grimshaw. 1990. *Argument Structure*. MIT Press, Cambridge, Mass.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado, June. 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 Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.

Marijana Marelj. 2004. *Middles and argument structure across languages*. LOT, Utrecht.

Ryan McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 122–131, Prague, Czech Republic, June. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Jasmina Moskovljević. 1997. Leksička detranzitivizacija i analiza pravih povratnih glagola u srpskom jeziku. *Južnoslovenski filolog*, LII:107–114.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Tanya Reinhart and Tal Siloni. 2004. Against the unaccusative analysis of reflexives. In Artemis Alexiadou, Elena Anagnostopoulou, and Martin Everaert, editors, *The Unaccusativity Puzzle: Studies on the syntax-lexicon interface*, pages 159–181. Oxford University Press.

Peter Sells, Annie Zaenen, and Draga Zec. 1987. Reflexivization variation: Relations between syntax, semantics, and lexical structure. In Masayo Iida and Draga Zec Stephen Wechsler, editors, *Working Papers in Grammatical Theory and Discourse Structure*, pages 169–238, Stanford, CA. CSLI.

Sandra A. Thompson. 1997. Discourse motivations for the core-oblique distinction as a language universal. In Akio Kamio, editor, *Directions in Functional Linguistics*, pages 59–82, Amsterdam, the Netherlands. Benjamins.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India, December. The COLING 2012 Organizing Committee.