The Universal Dependencies Treebank for Slovenian

Kaja Dobrovoljc[1], **Tomaž Erjavec**[2] **and Simon Krek**[3]

[1]Trojina, Institute for Applied Slovene Studies, Trg republike 3, 1000 Ljubljana, Slovenia
[2]Dept. of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[3]AI Laboratory, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
kaja.dobrovoljc@trojina.si
tomaz.erjavec@ijs.si
simon.krek@ijs.si

Abstract

This paper introduces the Universal Dependencies Treebank for Slovenian. We overview the existing dependency treebanks for Slovenian and then detail the conversion of the ssj200k treebank to the framework of Universal Dependencies version 2. We explain the mapping of part-of-speech categories, morphosyntactic features, and the dependency relations, focusing on the more problematic language-specific issues. We conclude with a quantitative overview of the treebank and directions for further work.

1 Introduction

In syntactic parsing and the field of data-driven natural language processing in general, there has been a growing tendency to harmonize the numerous annotations schemes, developed for linguistic annotation of individual languages or specific language resources, that have prevented direct comparisons of annotated data and the performance of the resultant NLP tools. To overcome this heterogeneity inhibiting both theoretical and engineering advancements in the field, the Universal Dependencies[1] annotation scheme provides a universal inventory of morphological and syntactic categories and guidelines for their application, while also allowing for language-specific extensions, when necessary (Nivre, 2015).

The scheme is based on previous similar standardization projects (Marneffe et al., 2014; Petrov et al., 2012; Zeman, 2008), and has recently been substantially modified to its second version (UD v2), following five successive releases of treebanks pertaining to UD v1 (Nivre et al., 2016). In the v2.0 release[2], 72 treebanks for 47 different languages have been released, including the reference (written) Slovenian UD Treebank, set forward in the remainder of this paper.

2 Dependency Treebanks for Slovenian

The Slovenian UD Treebank represents the third generation of syntactically annotated corpora in Slovenian. The first was the Slovene Dependency Treebank (Džeroski et al., 2006), based on the Prague Dependency Treebank (PDT) annotation scheme (Hajičová et al., 1999) and consisting of approximately 30,000 tokens taken from the Slovenian component of the parallel MULTEXT-East corpus (Erjavec, 2012), i.e., the Slovenian translation of the novel “1984” by George Orwell.

As the PDT’s scheme for analytical layer proved to be too complex given the financial and temporal constraints of subsequent projects, a new, simplified syntactic annotation scheme was developed within the JOS project (Erjavec et al., 2010). Within this scheme, the syntactic annotation layer consists of only 10 dependency relations, following the general assumption that specific syntactic constructions can be retrieved by combining these labels with the underlying word-level morphosyntactic descriptions (MSDs), wherein the JOS MSD tagset[3] is identical to the tagset defined in the MULTEXT-East Version 4 morphosyntactic specifications for Slovene (Erjavec, 2012).

The JOS annotation scheme was first applied to the jos100k corpus (Erjavec et al., 2010) consisting of approximately 100,000 tokens, sampled from the FidaPLUS reference corpus of written Slovene (Arhar and Gorjanc, 2007), and later extended to a larger sample of additional 400,000

[1] http://universaldependencies.org/

[2] While work on the individual treebanks for UD v2.0 has been finished, this version has, at the time of the writing of this paper, not yet been officially released.

[3] http://nl.ijs.si/jos/msd/

Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 33–38,
Valencia, Spain, 4 April 2017.

tokens in the Communication in Slovene (SSJ) project,[4] released as the ssj500k training corpus, with the latest version being v1.4 (Krek et al., 2015). The corpus is manually annotated with MSDs and lemmas but, due to financial constrains, only approximately one half (235,000) of the tokens were annotated on the syntactic layer. This subcorpus, known as the ssj200k treebank, currently represents the largest and the most representative collection of manually syntactically annotated data in Slovenian. It has been used in the development of several data-driven annotation tools (Grčar et al., 2012; Dobrovoljc et al., 2012; Ljubešić and Erjavec, 2016) and was chosen as the basis[5] for the construction of the Slovenian UD Treebank, using the conversion process described below.

3 Conversion from JOS to UD

To maintain a long-term compatibility between the two resources and maximize the level of consistency, the ssj200k conversion from JOS to UD annotation scheme was designed as a completely automatic procedure. Due to several discrepancies between the two annotation schemes, however, numerous conversion rules have been compiled on both morphological and syntactic level, whereas the tokenization, sentence segmentation and lemmatization principles of the original ssj200k treebank (currently) remain unchanged. In particular, we haven't used the option where tokens containing several (syntactic) words can be decomposed; this remains as future work.

3.1 Mapping of Morphosyntax

In terms of POS categorization, UD introduces a more fine-grained tagset of 17 POS categories in comparison with 12 POS categories in JOS, as it distinguishes between different types of (JOS-defined) verbs (`AUX` vs. `VERB`), conjunctions (`CCONJ` vs. `SCONJ`), characters (`SYM` vs. `PUNCT`), on the one hand, and subsumes the JOS Abbreviation POS as part of the `X` UD POS, on the other. A particularly challenging new category is the determiner (`DET`), reserved for nominal modifiers expressing the reference of the noun phrase in context, not traditionally used in Slavic grammars. For its conversion, a lexicon-oriented approach was adopted, in which pronominal subcategories in JOS were classified as either `DET` or `PRON` based on their typical syntactic behavior and their inflectional features, regardless of their context-specific syntactic role (Figure 1). Thus, predominantly pro-adjectival sub-categories (e.g. possessive or demonstrative pronouns) were converted to `DET`, while pro-nominal (e.g., personal pronouns) remained annotated as `PRON`, with lemmas in some sub-categories distributed between both POS categories (e.g., the JOS indefinite pronouns *nekdo*.`PRON` "somebody" vs. *mnog*.`DET` "many"). Similarly, a pre-determined list of indefinite quantifiers (e.g., *nekaj* "some", *več* "more", *veliko* "a-lot"), annotated as adverbs in JOS, has also been converted to `DET`.

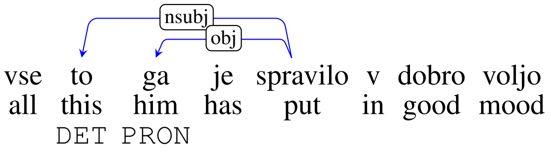
nsubj
obj
vse to ga je spravilo v dobro voljo
all this him has put in good mood
DET PRON

Figure 1: The annotation of JOS demonstrative (*to*) and personal (*ga*) pronouns in UD.

For the Slovenian UD Treebank 22 morphological features have been adopted, among which four are language- (`Gender[psor]`, `Number[psor]`, i.e., gender and number of the possessor with possessive adjectives) or treebank-specific (`NumForm`, `Variant`). In addition to the features not expressed morphologically in Slovenian (`Evident`), or not identifiable using automatic procedures (`Polite`), the Slovenian Treebank currently also lacks the universal `Voice` feature, as no morphological distinction has been made between predicative and attributive uses of participles in the JOS annotation scheme (e.g., *ukradena denarnica* "a stolen wallet" vs. *denarnica je bila ukradena* "the wallet was stolen").

The morphological layer conversion from JOS to UD is performed by a script which uses two semi-ordered tables (one for mapping the POS and the other for features). In total, the POS mapping contains 107 rules, of which 22 simply map a combination of the JOS POS and features to an UD POS, while 85 also specify the lemma of the token. There is only one rule that also takes into account the syntactic relation of the token, namely

[4] http://www.slovenscina.eu/

[5] It should be noted that several errata were discovered in ssj500k v1.4 in the process of conversion to UD v2.0. These were corrected and a new version of ssj500k will be released shortly. It is the new version that was used as the basis for the conversion to UD v2.0.

that for mapping an JOS auxiliary verb to the UD AUX or VERB. The feature mapping table contains 106 rules, of which 85 map a combination of the JOS POS and features, and possibly the already mapped UD POS to a UD feature, and 21 which are lemma-dependent.

3.2 Mapping of Syntax

Although both the JOS and the UD annotation scheme are based on the dependency grammar theory and adopt similar principles regarding the primacy of content words over function words, there are several significant differences between the two frameworks. Most notably, the UD annotation scheme introduces a much broader scope of syntactic analysis in comparison with JOS, where priority was given to parsing of predicates and their valency arguments, whereas semantically 'peripheral' sentence elements, such as sentence adverbs, discourse particles, interjections, vocatives, apposition, punctuation, clausal coordination, juxtaposition, etc. did not receive any syntactic analysis in JOS (as exemplified in Figure 2).

Secondly, the UD scheme also incorporates a much more detailed set of dependency relations (37 universal labels) than JOS (10 labels), as illustrated by the example given in Figure 3, in which the JOS *Atr* relation, intended for annotation of any head-modifier relation in a nominal phrase, converts to various types of nominal dependents in UD, such as different types of modifiers (amod, nmod, nummord, advmod, det, acl). In the same way, no distinction is made in JOS regarding the different syntactic structures of the dependents, whereas UD differentiates between nominal (nsubj, obj/iobj, obl) and clausal (csubj, ccomp, advcl) dependents performing the same syntactic role (see, for example, the two annotations of JOS *Obj* in Figure 2).

On the other hand, some semantic information is lost when converting data from JOS to UD, as JOS distinguishes between different types of arguments given their semantic role, such as between different types of adverbials or between semantically (non-)obligatory prepositional phrases, whereas UD only adopts the distinction between core arguments (i.e., subjects, objects, clausal complements) on the one hand, and oblique modifiers on the other, regardless of the degree of their obligatoriness in terms of valency and semantics.

In addition to categorization differences, the principles for determining the head-dependant direction mostly remain the same, with the exception of some specific constructions and the copula relation, in which the copula is dependent on the nonverbal predicate (see the cop relation in Figures 2 and 3).

In total, 32 different dependency relations have been used in the Slovenian UD treebank, including three extensions, i.e., cc:preconj for annotation of preconjuncts, flat:name for relations within personal names, and flat:foreign for relations within strings of foreign tokens. The eight missing universal relations in the treebank relate either to phenomena that do not occur in Slovenian (clf, compound), have not been found in the ssj200k treebank (dislocated, goeswith, reparandum) or do not enable reliable automatic identification (list, orphan, vocative).[6]

Among many syntactic particularities that have also be identified in other Slavic languages (Zeman, 2015), language-specific issues requiring additional consideration in the future include the treatment of (in)direct objects (with the iobj label currently only assigned in case of two competing objects), the inventory of TAMVE particles that could have been annotated as AUX/aux (such as *ne* ”not”, *lahko* “may” or *naj* “should”), and the treatment of the *se* reflexive pronoun (currently annotated as expl in Slovenian, regardless of its specific semantic role).

In total, the script for conversion of syntactic layer includes approximately 250 rules for dependency relation identification and/or head attachment, taking into account the lexical, morphological and syntactic features of individual tokens, their dependants or parents, as well as the features of tokens in the surrounding context. The conversion is performed in several iterations over tokens of a sentence, starting with the conversion of existing JOS-annotated constructions, and followed by different heuristics for annotation of previously un-annotated phenomena, including rules for root identification and punctuation attachment. In the last stage of the conversion, some mistakes and inconsistencies identified in the original ssj200k corpus are also corrected.

[6] Some of these relations, however, do occur in the manually annotated Spoken Slovenian UD Treebank (Dobrovoljc and Nivre, 2016).

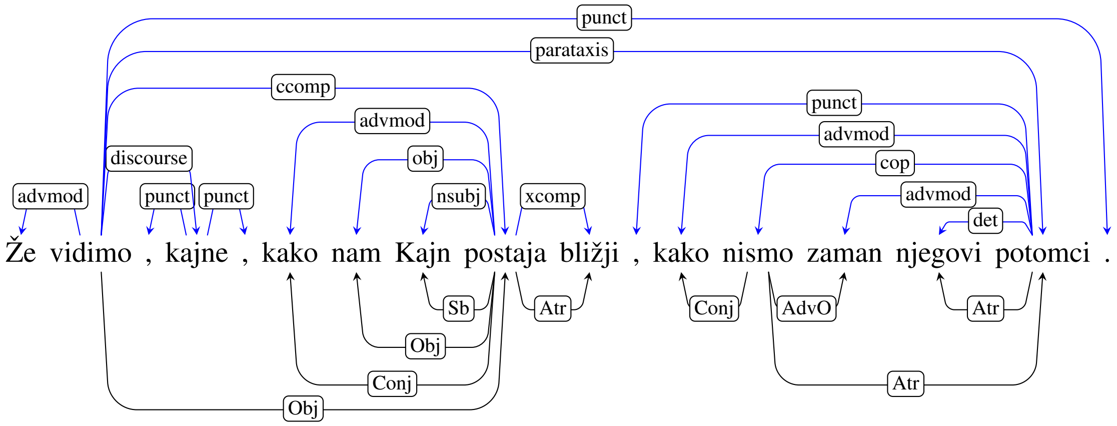

Figure 2: The comparison of UD (above) and JOS (below) annotation schemes in terms of complexity of dependency trees. All unanalysed tokens in JOS have been annotated as direct dependents of the root element.

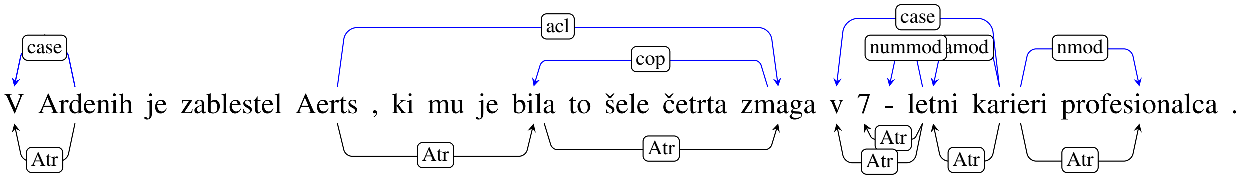
case
acl
cop
case
nummod
amod
nmod
V Ardenih je zablestel Aerts , ki mu je bila to šele četrta zmaga v 7 - letni karieri profesionalca .
Atr
Atr
Atr
Atr
Atr
Atr
Atr
Atr

Figure 3: The comparison of UD (above) and JOS (below) annotation schemes in terms of complexity of dependency relation taxonomy.

4 The Slovenian UD Treebank

Many constructions in the ssj200k corpus could not be converted automatically, among which different types of clausal coordination, juxtaposition and predicate ellipsis prevail. Sentences with such constructions were therefore omitted from the conversion and the resulting Slovenian UD Treebank has about 40% less tokens than the original ssj200k treebank. Nevertheless, it remains comparable to UD treebanks available for other languages (Nivre and et al., 2016), both in terms of size and average sentence length (Table 1).

	sl-ud	**ud-avg**	**ssj200k**
	(UD 2.0)	(UD 1.4)	(v1.4)
tokens	140,670	191,697	235,865
sentences	8,000	10,560	11,411
tok./sent.	17.6	18.2	20.7

Table 1: The size of Slovenian UD Treebank (sl-ud) in comparison with the average UD Treebank (ud-avg) and the original ssj200k treebank.

This latest version of the Slovenian UD Treebank is planned to be released as part of UD version 2.0, scheduled for March 2017, under the CC BY-NC-SA 4.0 license. The treebank maintains full compatibility with the original ssj200k treebank, encoded according to the XML-based Text Encoding Initiative (TEI) Guidelines (TEI Consortium, 2012), by listing the original JOS morphosyntactic and syntactic annotations as part of the XPOSTAG and MISC CONLL-U[7] columns, respectively, and by keeping the original ssj200k/FidaPLUS sentence identifiers as part of the CONLL-U comment line.

5 Conclusions

This paper presented the latest Slovenian UD Treebank, obtained with automatic conversion from the ssj500k Treebank, which uses the JOS annotation scheme. This new language resource represents a valuable contribution to the Slovenian NLP landscape, where research on dependency parsing and syntactically annotated data is still scarce (Krek, 2012). In addition to further improvements of the treebank, both in terms of size and annotation quality, priority in future work

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

should be given to evaluation of impact of the new annotation scheme on tagging/parsing accuracy, and its potential transfer to other reference corpora for Slovenian.

Acknowledgments

The first author would like to than Joachim Nivre and Dan Zeman for their invaluable inspiration and help. The work presented here was supported by the IC1207 COST Action PARSEME (PARSing and Multi-word Expressions) and Slovenian research programme P2-0103 “Knowledge Technologies”.

References

Špela Arhar and Vojko Gorjanc. 2007. Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa (The FidaPLUS Corpus: A New Generation of the Slovene Reference Corpus). *Jezik in slovstvo*, 52(2):95–110.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Kaja Dobrovoljc, Simon Krek, and Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino (Dependency Parser for Slovene). In *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana, Slovenia.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtsky, and Andreja Žele. 2006. Towards a Slovene Dependency Treebank. In *Fifth International Conference on Language Resources and Evaluation, LREC’06*, Paris. ELRA.

Tomaž Erjavec, Darja Fišer, Simon Krek, and Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (Obeliks: a statistical morphosyntactic tagger and lemmatiser for Slovene). In *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana, Slovenia.

Eva Hajičová, Zdeněk Kirschner, and Petr Sgall. 1999. A Manual for Analytic Layer Annotation of the Prague Dependency Treebank (English translation). Technical report, ÚFAL MFF UK, Prague, Czech Republic.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, and Nanika Holz. 2015. *Training corpus ssj500k 1.4*. Slovenian language resource repository CLARIN.SI.

Simon Krek. 2012. *Slovenski jezik v digitalni dobi – The Slovene Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at http://www.meta-net.eu/whitepapers.

Nikola Ljubešić and Tomaž Erjavec. 2016. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: a Cross-Linguistic Typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Joakim Nivre and et al. 2016. *Universal Dependencies 1.4*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague. http://hdl.handle.net/11234/1-1827.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, 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 *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 9041 of *Lecture Notes in Computer Science*, pages 3–16. Springer International Publishing.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

TEI Consortium, editor. 2012. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Daniel Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 213–218, Marrakech, Morocco. European Language Resources Association.

Daniel Zeman. 2015. Slavic Languages in Universal Dependencies. In *Proceedings of the conference "Natural Language Processing, Corpus Linguistics, E-learning"*, pages 151–163, Bratislava, Slovakia. RAM-Verlag.