sciendo
DOI 10.2478/jazcas-2019-0068

LEVELS OF ANNOTATION IN THE SLOVENE TRAINING CORPUS ssj500k 2.2

MIJA BON – POLONA GANTAR
Faculty of Arts, University of Ljubljana, Slovenia

BON, Mija – GANTAR, Polona: Levels of annotation in the Slovene training corpus ssj 500k 2.2. Journal of Linguistics, 2019, Vol. 70, No 2, pp. 390 – 399.

Abstract: This paper presents the Slovene Training Corpus ssj500k 2.2, which has been annotated on the levels of tokenization, sentence segmentation, part-of-speech tagging, lemmatization, syntactic dependencies, named entities, verbal multi-word expressions, and semantic role labeling. It describes the individual layers of annotation and shows the scope of using the training corpus in the production of various lexicons, such as the lexicon of multi-word units and the valency lexicon of modern Slovene. It concludes by presenting our future work, i.e. the annotation of multi-word expressions based on the Slovene Lexical Database.

Keywords: corpus linguistics, training corpus, corpus annotation, Slovene language

1 INTRODUCTION

A training corpus is a linguistic source, which is generally manually annotated or corrected and used mainly for training statistical models for different purposes, such as part-of-speech tagging or parsing [1]. Training data can be used in supervised machine learning systems, which enable efficient automatic annotation of even very large corpora. The latest version of the Slovene Training Corpus 2.2 [17] consists of two training corpora, the whole of jos100k corpus V2.0 [7] and 400,000 words from training corpus jos1M 1.2. [8]. The training corpus ssj500k 2.2 [17] is freely available at Clarin.si repository under the Creative Commons (CC) license, Attribution-NonCommercial 3.0.[1] Compared to the previous, 2.1 version [16], this version corrects various errata in spacing and text metadata and, in cases where it was possible to do so automatically, adds UD morphological and dependency annotations to the corpus [17].

2 CORPUS DESCRIPTION

The ssj500k 2.2 training corpus contains 586,248 tokens, 27,829 sentences, and 500,295 words/lemmas, manually annotated on six levels: the whole corpus is

[1] http://eng.slovenscina.eu/tehnologije/ucni-korpus, https://www.clarin.si/repository/xmlui/handle/11356/1210.

lemmatized and morphosyntactically (POS) annotated, about half of the corpus is annotated with syntactic dependencies and verbal multi-word expressions, a third of it is annotated with named entities, and approximately a quarter of it with semantic role labels. The whole of the corpus is thus morphosyntactically annotated, with specific parts of the corpus also containing other layers of annotation. Namely, the part of corpus labeled with SRL also includes syntactic annotation, MWEs, and named entities. This is particularly useful in linguistic analysis, where different levels of annotation can be combined to form a more comprehensive overview of a particular linguistic phenomenon.

Level of annotation	*Annotated sentences*
Part-of-speech	27,829
Lemmatization *Verbal multi-word expressions*	27,829 13,511
Syntactic dependencies	11,411
Named entities	9,478
Semantic role labeling	5,491

Tab. 1. Number of annotated sentences on each level of ssj500k 2.2

2.1 Sentence segmentation

The ssj500k 2.2 is segmented into 27,829 sentences with 586,248 tokens. A statistical overview is given in Table 2. The data had already been annotated on the levels of segmentation and tokenization in preliminary corpora. The ssj500k 2.2 corpus was further manually validated and corrected.

Element	*n*
Text	1,677
Paragraph	8,137
Sentence	27,829
Token	586,248

Tab. 2. Statistical data on elements of ssj500k 2.2[2]

2.2 Part-of-speech tagging and lemmatization

The entirety of the ssj500k 2.2 was lemmatized and tagged on morphosyntactic (POS) and syntactic levels; it consists of 500,295 words. Part-of-speech tagging was done by using the tagset JOS system [6], which includes 12 POS categories with 1,903 possible attributes [14]. The part-of-speech tagging and lemmatization of

[2] http://eng.slovenscina.eu/tehnologije/ucni-korpus

preliminary corpora was performed by using the Obeliks tool, described in [14], with 91.34% accuracy for all tags and 98.30% for POS only. The lemmatizer had an approximately 98% accuracy [14, p. 4]. The whole corpus was manually corrected.[3]

Because of an increasing interest in the field of NLP in creating consistent annotation schemes that would enable the comparison of annotated data of individual languages, the Slovene JOS annotation scheme was adjusted to conform to the Universal Dependencies framework. The Universal Dependencies v2 standard includes 17 POS categories, and the Slovene corpus uses 16 of those UPOS tags suitable for Slovene: ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, VERB, X. The conversion of the ssj500k into a UD treebank was initially envisioned as an automated process. However, due to numerous differences between the two systems of annotation, especially on the level of syntactic description, a complex system of conversion rules was additionally created.[4]

2.3 Syntactic dependencies

The JOS dependency treebank model, used for surface-syntactic dependency annotation, was designed within the framework of the project Linguistic Annotation of Slovene [19]. The model which is based on syntactic dependencies but which also takes into account the syntactic characteristics of the Slovene language, consists of a robust three-level system. The labels are described in Table 3. A tag is attributed to each token and punctuation. The highest level in the system is occupied by the so-called meta element, which demonstrates either the interrelation of syntactically highest elements in sentences or syntactically less predictable structures, e.g. an ellipsis. [19, p. 51–52].

Groups of labels	Labels	Description
First level labels link elements in different types of phrases.	*dol*	Links heads and modifiers in phrases.
	del	Links parts of verbal phrases.
	prir	Links heads in coordinate structures within clauses.
	vez	Links words or commas in conjunctive function.
	skup	Links (function) words in frozen multi-word structures.
Second level labels link sentence elements.	*ena*	Clause subject.
	dve	Clause object.
	tri	Adverbial of manner.
	štiri	Other adverbials.

[3] During the making of the last version of reference corpus Gigafida 2.0. (`https://viri.cjvt.si/gigafida/`), a new meta-tagger was created, with the accuracy of 94.34% for MSD and 98.66% for lemmatization.

[4] For more about the process see [3].

Third level label links all other structures.	*modra*	Links to root, punctuation, syntactically less predictable structures, parentheses etc.

Tab. 3. Labels in the JOS dependency model (`http://eng.slovenscina.eu/tehnologije/razclenjevalnik`)

At the same time, the Sentence Markup (SMU) tool [2] was additionally developed for manual annotation, visualization, and data search. Surface-syntactic dependency annotation was performed in 11,411 sentences, with approximately half of them re-annotated following the evaluation of the annotation system, POS errors, and quality analysis of the annotation. At least two annotators manually annotated syntactic dependencies. All cases of discrepancy were further examined by a third annotator.

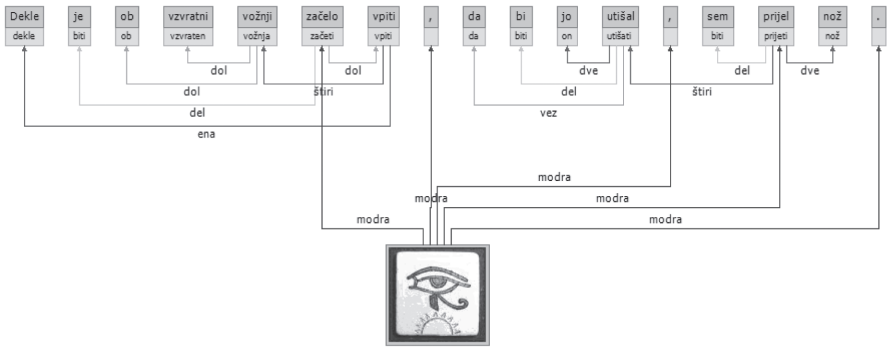

Fig. 1. Syntactic level in the SMU annotation tool

2.4 Named entities

Approximately a third of ssj500k 2.2 (9,478 sentences) was manually annotated with named entity annotations in the WebAnno tool, with the aim of developing a named entity extractor for the Slovene language based on machine learning [20]. The annotation distinguished five types of NE: Person (per), Person Derivative (deriv-per), Location (loc), Organization (org), and Miscellaneous (misc). Apart from standard categories for named entities – i.e. names for people, pets, and groups of people; locations, including named buildings; organizations and institutions; and other proper nouns for things, for example book titles etc. – the category deriv-per was introduced, which annotates personal possessive adjectives, in order to improve anonymization of personal data [21].

7,015 named entities were marked in 9,478 sentences; this amounts to 1.35 named entities per sentence on average. The distribution of named entities by categories is given in Table 4.

Named entity	*n*	%
Loc	1,968	28%
Org	1,338	19%
Per	2,927	41.5%
Deriv-per	180	2.5%
Misc	602	9%
Total	7,015	100%

Tab. 4. Statistical data on named entities in the annotated part of ssj500k 2.2

2.5 Verbal multi-word expressions

The annotation scheme for verbal multi-word expressions (VMWEs) was based on categories developed within the international PARSEME COST Action Shared Task 1.1, adapted to the Slovene language [9, 10]. VMWE annotation includes the following four categories:

- inherently reflexive verbs (IRV),
- light-verb constructions, divided into full (LVC.full) and cause (LVC.cause),
- inherently adpositional verbs (IAV),
- verbal idioms (VID).

13,511 sentences were manually annotated following the Guidelines developed within the PARSEME shared task 1.1. In the first phase, 11,411 sentences were annotated by two annotators in accordance with the first version of the Guidelines. Discrepancies in annotations were discussed and adjusted accordingly. During the second phase, categories were automatically modified to comply with the second version of the Guidelines and manually checked. Additionally, 2,100 sentences were manually annotated by individual annotators in accordance with the modified Guidelines [12]. The first phase of annotation was performed in the SMU tool adjusted for VMWE labeling; the second phase employed the FLAT annotation platform, which enables labeling strings of text using previously defined categories ([10, p. 86], [12]).

The 13,511 annotated sentences contain 3,364 VMWEs in all forms (as they appear in sentences), with slightly fewer than 1,100 different expressions. 2,920 sentences (approximately 22%) contain at least one VMWE. Overall, the distribution of VMWEs in the annotated part of the ssj500k 2.2 is 0.25 VMWE per sentence; in other words, there is one VMWE present, on average, in every fourth sentence ([10, p. 86-87], [12]).

VMWE category	*n*	*%*
IRV	1,627	48%
IAV	710	21%
VID	724	22%
LVC-cause	64	2%
LVC.full	239	7%
together	3,364	100%

Tab. 5. Distribution of VMWEs in ssj500k 2.2

2.6 Semantic Role Labelling

Semantic Role Labeling (SRL) refers to the process of detecting and assigning semantic roles to semantic arguments determined by the predicate or verb of a sentence. The framework for semantic role labeling was developed within the bilateral project Semantic Role Labeling in Slovene and Croatian ([15], [11]). It follows the path of previous SRL efforts (PDT, Vallex, FrameNet, Propbank etc.) while also considering the specifics of both target languages. The SRL tagset, based on the Prague Dependency Treebank, consists of 25 semantic labels: 5 arguments, 17 adjuncts, and 3 labels for multi-word predicates [11, p. 93–94], as seen in Table 6 and described in detail in [15].

agent	*ACT*
patient	PAT
recipient	REC
origin	ORIG
result	RESLT
location	LOC
source (location)	SOURCE
goal (location)	GOAL
event	EVENT
time	TIME
duration	DUR
frequency	FREQ
aim	AIM
cause	CAUSE
contradiction	CONTR
condition	COND
regard	REG

accompaniment	ACMP
restriction	RESTR
manner	MANN
means	MEANS
quantification	QUANT
multi-word predicate	MWPRED
modal	MODAL
phraseological unit	PHRAS

Tab. 6. SRL tagset for Slovene

A total of 5,491 sentences were annotated with semantic roles. The first 500 manually annotated sentences were used for automatic labeling, using mate-tools semantic role labeler with the German feature set [11, p. 94]. The second phase included automatic annotation of the remaining 4,991 sentences and their manual verification by five annotators [11, p. 94–95]. The annotation was performed in the SMU tool.

Fig. 2. SRL layer of annotation in the SMU tool

All 25 semantic labels were found in the corpus; predictably, however, the most frequent ones were argument roles of PAT and ACT, the former with a significantly higher frequency than the latter, and RESLT. These were followed by adjunct roles of TIME, MANN, and LOC [11, p. 96].

Slovene was found to have relatively stable patterns for its most frequent verbs, such as *biti* 'to be', *imeti* 'to have', *dobiti* 'to get', *morati* 'must', *moči* 'can', *hoteti* 'will', *želeti* 'want', *reči* 'to say', *povedati* 'to tell', e. g. [11, p. 95–97]:

'to have' *imeti*

- WHO (ACT) has WHAT (PAT) [for WHOM (REC), from whom (ORIG), where (LOC), when (TIME) …]

'must' *morati*

- WHO (ACT) must INF (MODAL)

'to go' *iti*

- WHO (ACT) goes WHERE (GOAL) [how (MANN), when (TIME), under what conditions (COND) …]
- to go (PHRAS)
- to go SUPINE (MWPRED).

3 CONCLUSION AND FUTURE WORK

The Slovene Training Corpus ssj500k 2.2 was primarily intended for machine learning and linguistic analysis. So far, the training corpus has been used for automatic tagging of the Slovene reference corpus Gigafida 2.0, creating a lexicon of MWEs, machine learning for automatic annotation of corpora Gigafida and Kres on the level of MWEs and SRL, as well as the analysis of sentence patterns and building of valency lexicons.

In the future we plan to continue working on new layers of annotations based on the Slovene Lexical Database, firstly with a newly developed typology of MWEs. The new typology enables us to also identify non-verbal MWEs, such as noun MWEs (*rdeče številke* lit. red numbers, 'deficit', *kaplja v morje* lit. a drop in the see, 'negligible amount', fixed prepositional phrases (*med drugim* 'among others', *v skladu s/z* 'in accordance with), and multi-word discourse markers (*v tem primeru* 'in this case'). Our aim is to recognize and define MWEs as part of language with individual meaning and/or syntactic function. MWEs are categorized into three types: phraseological units – PU (*kaplja čez rob* 'the last straw', *leta tečejo* 'years go by'), *za vraga* 'heck'), fixed expressions – FE (*varnostni trikotnik* 'warning triangle'), *črna luknja* 'black hole', *d. o. o.* (*družba z omejeno odgovornostjo* 'limited company'), and syntactic combinations – SC (*na prostem* 'in the open', *za zdaj* 'for now', *v skladu s/z* 'in accordance with', *in tako naprej* 'and so on', *po eni strani – po drugi strani* 'on the one hand – on the other hand'). Collocations and extended collocations, which can also be seen as MWEs, will be extracted from the corpus via Sketch Engine and other tools developed within the project New Grammar of Contemporary Standard Slovene. Currently, the first phase of the manual annotation of MWEs is in progress. Following this, IAA will be analyzed to prove or disprove the consistency of categories. In the final stage, a lexicon of MWEs will be completed and made part of the Dictionary of Modern Slovene [13].

ACKNOWLEDGMENTS

Annotation levels were defined within the framework of the national project *Nova slovnica sodobne standardne slovenščine: viri in metode* (New grammar of contemporary standard Slovene: sources and methods, ARRS J6-8256); corpus annotation and analyses were completed as part of the *ARRS P6-0215* research program (Slovene language – basic, contrastive, and applied studies). Corpus development has been a part of the infrastructural program of the Center for Language Resources and Technologies at the University of Ljubljana.

References

[1] Arhar, Š. (2009). Učni korpus SSJ in leksikon besednih oblik za slovenščino. Jezik in slovstvo, 54, pages 3–4.

[2] Dobrovoljc, K., Krek, S., and Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino.

[3] Dobrovoljc, K. Erjavec, T., and Krek, S. (2016). Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. Konferenca Jezikovne tehnologije in digitalna humanistika, pages 190–192. Ljubljana.

[4] Dobrovoljc, K., Erjavec, T., and Krek, S. (2017). The Universal Dependencies Treebank for Slovenian.

[5] Dobrovoljc, K., Erjavec, T., and Krek, S. UD Slovenian SSJ. Accessible at: `https://universaldependencies.org/treebanks/sl_ssj/index.html`.

[6] Erjavec, T., Fišer, D., Krek, S., and Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), pages 1806–1809. Paris, ELRA.

[7] Erjavec, T., Krek, S., and Fišer, D. (2010). jos100k corpus V2.0. Accessible at: `http://hdl.handle.net/11356/1213`.

[8] Erjavec, T., Krek, S., and Dobrovoljc, K. (2019). Training corpus jos1M 1.2, Slovenian language resource repository CLARIN.SI. Accessible at: `http://hdl.handle.net/11356/1213`.

[9] Gantar, P., Krek, S., and Kuzman, T. (2017). Verbal multiword expressions in Slovene. Europhras 2017, pages 247–259. Springer.

[10] Gantar, P., Arhar Holdt, Š., Čibej, J., Kuzman, T., and Kavčič, T. (2018). Glagolske večbesedne enote v učnem korpusu ssj500k 2.1. In Proceedings of the conference on Language Technologies & Digital Humanities, pages 85–92.

[11] Gantar, P., Štrkalj Despot, K., Krek, S., and Ljubešić, N. (2018). Towards Semantic Role Labeling in Slovene and Croatian. In Proceedings of the conference on Language Technologies & Digital Humanities, pages 92–98.

[12] Gantar, P., Arhar Holdt, Š., and Čibej, J. (in print). Structural and Semantic Classification of Verbal Multi-Word Expressions in Slovene. Contributions to Contemporary History.

[13] Gorjanc, V., Gantar, P., Kosem, I., and Krek, S. (2017). Dictionary of Modern Slovene: Problems and Solutions. Ljubljana, Založba FF.

[14] Grčar, M., Krek, S., and Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In Zbornik Osme konference Jezikovne tehnologije. Ljubljana, Institut Jožef Stefan.

[15] Krek, S., Gantar, P., Dobrovoljc, K., and Škrjanec, I. (2016). Označevanje udeleženskih vlog v učnem korpusu za slovenščino. In Proceedings of the Conference on Language Technologies & Digital Humanities, pages 106–110. Faculty of Arts. University of Ljubljana.

[16] Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L., and Zajc, A. (2018). Training corpus ssj500k 2.1, Slovenian language resource repository CLARIN.SI. Accessible at: `http://hdl.handle.net/11356/1181`.

[17] Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L., and Zajc, A. (2019). Training corpus ssj500k 2.2, Slovenian language resource repository CLARIN.SI. Accessible at: `http://hdl.handle.net/11356/1210`.

[18] Ledinek, N., and Erjavec, T. (2009). Odvisnostno površinskoskladenjsko označevanje slovenščine: specifikacije in označeni korpusi. Simpozij Obdobja 28, pages 219–224.
[19] Ledinek, N. (2014). Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine. Ljubljana, Založba ZRC, ZRC SAZU.
[20] Štajner, T., Erjavec, T, and Krek, S. (2013). Razpoznavanje imenskih entitet v slovenskem besedilu. Slovenščina 2.0, 2, pages 58–82. Accessible at: http://slovenscina2.0.trojina.si/arhiv/2013/2/Slo2.0_2013_2_04.pdf.
[21] Zupan, K., Ljubešić, N., and Erjavec, T. (2017). Annotation guidelines for Slovenian named entities Janes-NER.