The ssj500k Training Corpus for Slovene Language Processing

Simon Krek[1,2], Tomaž Erjavec[1], Kaja Dobrovoljc[1,2], Polona Gantar[2], Špela Arhar Holdt[2], Jaka Čibej[1,2], Janez Brank[1]

[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2]Faculty of Arts, University of Ljubljana, Slovenia
{janez.brank, kaja.dobrovoljc, simon.krek, tomaz.erjavec}@ijs.si
{spela.arharholdt, jaka.cibej, apolonija.gantar}@ff.uni-lj.si

Abstract

This paper presents recent developments and the content of the ssj500k training corpus, the largest and most widely used open-source collection of training data for Slovene language processing, which has been manually annotated with respect to segmentation, tokenisation, lemmatisation, JOS morphosyntax and dependency syntax, Universal Dependencies, semantic role labelling, named entities and verbal multi-word expressions. After a short history of the development of the corpus, we give an overview of the dataset as a whole, and the details of each annotation layer, including a survey of existing natural language processing tools that used it for training. Most ssj500k annotations were carried out using the dedicated Q-CAT querying-supported corpus annotation tool, which is also presented, and the directions for future development of the corpus are discussed.

1. Introduction

Training corpora, i.e. manually annotated language datasets, are essential to the development of natural language processing tools based on supervised machine learning. For Slovene, a South Slavic language with approximately two million speakers, the ssj500k training corpus (Krek et al., 2019) represents the largest collection of such manually annotated training data to date, a result of more than two decades of continuous development.

This first began within the multilingual MULTEXT-East project and its spin-offs (Dimitrova et al., 1998; Erjavec, 2004), providing the pivotal language resources for Slovene morphosyntactic annotation, including the specifications, an inflectional lexicon, and the "1984" morphosyntactically annotated corpus, which was later also used as the starting point for the first Slovene dependency treebank (Džeroski et al., 2006). However, given the limitations in size and diversity of the "1984" corpus, the "Linguistic annotation of Slovene" (Jezikoslovno označevanje slovenščine: JOS)[1] project (2007-2009) aimed to fill this gap by producing a balanced and representative set of text samples, the jos100k corpus (Erjavec and Krek, 2008), manually annotated with lemmas, morphosyntax and dependency syntax (Erjavec et al., 2010), all based on a revised and updated set of MULTEXT-East/JOS annotation guidelines. In the consecutive "Communication in Slovene" (SSJ)[2] project (2008-2013), additional 400,000 words were added to the jos100k corpus to form the ssj500k corpus with JOS/MULTEXT-East morphosyntactic annotation, as well as partial annotation for JOS dependency syntax and named entities (Holozan et al., 2008).

In subsequent years, the annotation of the ssj500k corpus has been incrementally developed within various other projects, such as the "Semantic Role Labelling in Slovene and Croatian" bi-lateral project (Gantar et al., 2018b), the PARSEME verbal multi-word expression annotation (Gantar et al., 2018a), Universal Dependencies for Slovene (Dobrovoljc et al., 2017) and Janes named entity annotation (Fišer et al., 2018), producing several new layers of grammatical annotation. The ssj500k corpus thus represents the largest, richest and most widely used training and testing dataset for Slovene language engineering research and development.

However, despite this pivotal role, the dataset as a whole lacks systematic and exhaustive documentation, not only in terms of its content and recent developments, but also in terms of the infrastructure that surrounds it, such as its formats and the querying and annotation tools. The aim of this paper is to remedy this and provide an overview of the latest ssj500k release, version 2.2 (Krek et al., 2019), by presenting its annotation layers and the state-of-the-art tools that employ them (Sections 2. and 3.), describing the surrounding infrastructure (Sections 4. and 5.) and discussing immediate and long-term future work (Section 6.).

2. Overview of the corpus

The ssj500k corpus (Krek et al., 2019) is a balanced collection of text samples from the FidaPLUS corpus of written modern Slovene (Arhar and Gorjanc, 2007), which includes fiction, non-fiction and periodical documents dating from 1990–2000. In total, the corpus contains 586,248 tokens belonging to 27,829 sentences inside 8,137 paragraphs sampled form 1,655 documents. However, due to temporal and financial constraints, not all annotation layers have been applied to the corpus in its entirety.

As summarised in Table 1, the entire corpus has been manually segmented into sentences and tokens and annotated for lemmatisation and JOS/UD morphosyntax. In contrast, other annotation layers cover only parts of the corpus, ranging from approximately half of the corpus (e.g. multi-word expressions) to approximately one fifth of the corpus (e.g. semantic roles). With the exception of the intermittent UD syntax (Section 3.3.), all layers cover a con-

[1]http://nl.ijs.si/jos/
[2]http://www.slovenscina.eu/

Level	Tokens	Senten.	Docs
Segmentation	586,248	27,829	1,655
lemmatisation	586,248	27,829	1,655
JOS morphosyntax	586,248	27,829	1,655
UD morphology	586,248	27,829	1,655
JOS syntax	235,864	11,411	617
UD syntax	140,670	8,000	581
Semantic roles	112,048	5,501	228
Named entities	194,637	9,488	498
Multi-Word Expressions	280,522	13,511	754

Table 1: Size of ssj500k annotation layers.

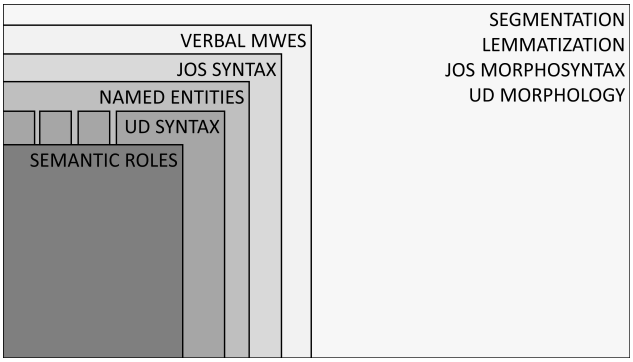

Figure 1: Illustration of nested annotation layers in ssj500k.

tinuous sequence of sentences starting at the beginning of the corpus, and can thus be considered as nested subsets, as illustrated in Figure 1.

3. Annotation Layers

3.1. Segmentation and tokenisation

Manual segmentation and tokenisation in ssj500k corpus follows rules implemented in the Obeliks4J tokeniser for standard Slovene.[3] The rule-based tokenisation and segmentation system was devised with the view that anyone can unambiguously predict how sequences of characters would be split into sentences and tokens in processed texts.

Tokenisation: Whitespace is the principal separator for tokens. Sequences of words that can be written both with or without space (e.g. *kdorkoli*, *kdor koli* "anybody, any body") follow the same principle. During tokenisation, the text is divided into sequences belonging to two categories: words (W) are alphanumeric strings, while the other category includes punctuation and symbol characters (C). C tokens are recognised on the basis of a predefined list of punctuation and symbol characters included in the tokeniser. C tokens may include only one punctuation or symbol character. Sequences of two or more C tokens (e.g. *?!*) are treated as sequences of separate tokens. If a string of alphanumeric characters between two spaces includes C characters, it is split into separate tokens (e.g. *AC/DC* will be split into three tokens *AC, /, DC*), with some exceptions, such as:

- Apostrophe becomes part of a W token if it is used without space on both sides (e.g. *O'Brian* or *mor'va* "we have to").
- Comma and colon become part of a W token if used without space on both sides and if the string contains only digits (e.g. *30:00, 200,000,000*).

Similar rules are applied for other C tokens, including those used in URLs and e-mail addresses where all C characters become part of a single W token (using a regular expression).

Segmentation: sentence segmentation is performed after tokenisation. The basic rule of "sentence terminal character + space + capitalised word" is complemented with several combinations of different punctuation characters.

Markup Type	No. of occurrences
Texts	1,655
Paragraphs	8,137
Sentences	27,829
Punctuation tokens	85,953
Word tokens	500,295

Table 2: Overview of segmentation and tokenisation annotations in ssj500k.

3.2. JOS Annotation

3.2.1. JOS Morphosyntactic Annotation

The morphosyntactic annotation in ssj500k follows the so called JOS morphosyntactic specifications, which are identical to the MULTEXT-East (Erjavec, 2012) morphosyntactic specifications for Slovene. In particular, Version 2.2 of the ssj500k corpus follows MULTEXT-East Version 6.[4] which, for Slovene, lists 1,900 morphosyntactic descriptions (MSDs) and their decomposition into features (e.g. part-of-speech, gender, number, case etc.). Furthermore, the specifications give the MSDs, the attributes, and their values both in English and in Slovene, so it is possible to switch between the morphosyntactic description language. As Table 3 shows, the ssj500k corpus contains approximately 80% of the 1900 theoretically possible MSDs.

Annotation	No. of types
JOS lemmas	38,818
JOS morphosyntactic tags	1,304

Table 3: Overview of JOS morphosyntactic annotations in ssj500k.

As part of the "Communication in Slovene" project, the lemmatisation guidelines for Slovene were also prepared. The guidelines rely on the MULTEXT-East part-of-speech categorisation and define capitalisation rules for the lemmatisation of (potentially) proper-noun-related words, and

[3] https://github.com/clarinsi/Obeliks4J

[4] MULTEXT-East morphosyntactic specifications Version 6 are maintained on https://github.com/clarinsi/mte-msd, and also available at http://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

other problematic issues in Slovene capitalisation (Holozan et al., 2008).

The lemmatisation and morphosyntactic annotations have been used in the development of several locally developed tagging tools, such as the Obeliks (Grčar et al., 2012) and ReLDI (Ljubešić and Erjavec, 2016) taggers, and the ensemble neural network tagger for Slovene (Belej, 2018).

Given that JOS/MULTEXT-East morphosyntactic annotations are also contained (in the `xpos` column) in Universal Dependencies SSJ Slovene treebank (cf. Section 3.3.), which is a subset of the ssj500k corpus, many other lemmatisation and tagging tools competing in the CONLL 2017 and 2018 shared tasks (Zeman and others, 2017; Zeman et al., 2018) have also been trained on this data. The StanfordNLP tool (Qi et al., 2018), in particular, has been trained on full ssj500k morphosyntactic annotations (Ljubešić and Dobrovoljc, 2019)[5] and gives currently the best tagging results for Slovene.

3.2.2. JOS Syntactic Annotation

In both JOS and SSJ projects, the JOS dependency annotation scheme was also developed. The system was designed specifically for Slovene and is based on syntactic dependencies.

The main idea is that the information attributed on lower levels of annotation (MSD, lemmas) need not to be repeated at the syntactic level. The result is an extremely robust system comprising of only 10 labels (Table 4): 5 labels link elements at the word-phrase level, 4 labels pertain to sentence elements (predicate arguments), and the last label links the heads of sentences to the annotation root. The main features of the system are described in (Erjavec et al., 2010), and in more detail in the annotation guidelines (Holozan et al., 2008).

In ssj500k, 11,411 sentences have been annotated using the JOS dependency annotation scheme. This data was primarily used for training a MST-based dependency parser for Slovene (Dobrovoljc et al., 2012), but it has recently also been applied to a neural-network parsing architecture (Qi et al., 2018) to be used in JOS dependency annotation of the Gigafida 2.0 reference corpus of Slovene (Arhar Holdt et al., forthcoming).

Level	Label	Relation	Tokens
Phrases	Atr	Attribute	79,556
	PPart	Predicate part	16,138
	Coord	Coordination	6,530
	Conj	Conjunction	19,278
	MWU	Multi-word unit	828
Arguments	Sb	Subject	11,690
	Obj	Object	15,637
	AdvM	Adverbial of manner	5,762
	AdvO	Other adverbials	14,276
Root	Root	All other relations	66,169

Table 4: JOS dependency relation annotations in ssj500k.

[5] `https://github.com/clarinsi/classla-stanfordnlp`

3.3. Universal Dependencies

Universal Dependencies (UD) is a cross-linguistically consistent grammatical annotation scheme, providing a universal inventory of morphological and syntactic categories and guidelines for their application (Nivre et al., 2016). In its latest release (Nivre et al., 2019), it has been applied to more than 100 treebanks in over 70 world languages, including the written SSJ (Dobrovoljc et al., 2017) and spoken SST (Dobrovoljc and Nivre, 2016) treebanks for Slovene. The SSJ UD treebank is the result of an automatic rule-based conversion of the ssj500k subset annotated for JOS syntax (Section 3.2.), both on the level of morphology and dependency syntax.[6] The UD annotations are thus the only ssj500k layers that have been produced automatically, however, given the carefully designed conversion rules its quality resembles that of manual annotation.

3.3.1. UD morphology

The conversion of JOS morphosyntactic tags (Section 3.2.) to UD part-of-speech categories and morphological features has been performed by a script which uses two semi-ordered tables (one for mapping the POS categories and the other for features). The large majority of rules are direct mappings of specific categories (e.g. conversion of JOS numerals (`M`) with `Form=letter` and `Type=ordinal` to UD adjectives (`ADJ`) with feature `NumType=Ord`) with the exception of the mappings to the UD determiner (DET) and auxiliary (AUX) POS categories, which also require a predefined lexicon and context information, respectively. This mapping has recently been applied to the entire ssj500k corpus (Dobrovoljc et al., 2019), providing much larger training data for UD morphology than the officially released UD SSJ treebank, which also requires full syntactic annotations. The benefits of such larger training set have already been attested in several UD tagging and lemmatisation experiments for Slovene and other South Slavic languages (Dobrovoljc et al., 2019; Ljubešić and Dobrovoljc, 2019).

Annotation	No. of types
UPOS	16
FEATS	1,008
UPOS+FEATS	1,141

Table 5: The number of UD POS categories, UD feature combinations and full UD morphology annotations in ssj500k.

3.3.2. UD syntax

Due to several significant differences between the JOS and UD dependency annotation schemes (Dobrovoljc et al., 2017), most notably a much more fine-grained set of dependency labels in UD (i.e. 32 dependency relations), which also cover phenomena outside the predicate-argument structure (cf. Section 3.2.), approximately 250 rules have been designed for dependency relation and head

[6] The JOS to UD conversion scripts are available at `https://github.com/clarinsi/jos2ud`.

identification, taking into account various lexical, grammatical and contextual features of the tokens. Given the limited scope of such detailed conversion rules, only 8,000 out of 11,411 JOS dependency trees in ssj500k have been successfully converted to UD and released as the official UD SSJ treebank. Nevertheless, this dataset has had important implications for Slovene language processing infrastructure, as it has been featured in the development of many state-of-the-art NLP tools for parsing raw text to Universal Dependencies (Zeman et al., 2018).

UD Dependency relation	Label	Tokens
Adjectival clause	acl	2394
Adverbial clause modifier	advcl	1302
Adverbial modifier	advmod	11702
Adjectival modifier	amod	11800
Appositional modifier	appos	393
Auxiliary	aux	7161
Case marking	case	13495
Coordinating conjunction	cc	4668
Clausal complement	cc:preconj	62
Preconjunct	ccomp	1216
Conjunct	conj	5297
Copula	cop	2826
Clausal subject	csubj	558
Unspecified dependency	dep	8
Determiner	det	3311
Discourse element	discourse	68
Expletive	expl	2298
Fixed MWE	fixed	455
Flat MWE	flat	43
Flat MWE: foreign	flat:foreign	55
Flat name	flat:name	688
Indirect object	iobj	606
Marker	mark	4701
Nominal modifier	nmod	10352
Nominal subject	nsubj	7242
Numeric modifier	nummod	1610
Object	obj	7036
Oblique nominal	obl	9900
Parataxis	parataxis	1625
Punctuation	punct	18596
Root	root	8000
Open clausal complement	xcomp	1202

Table 6: Overview of UD dependency relations in ssj500k.

3.4. Named Entities

The first version of the corpus was already annotated with named entities (Štajner et al., 2013), but only about a quarter of the text had these annotations. Furthermore, there had been no annotation guidelines for the named entities, with the result that many annotations were inconsistently applied. In the scope of the Janes project (Fišer et al., 2018) we developed guidelines for named entity annotation in Slovene[7] and we used these guidelines to first correct the already existing NE annotations and then to extend the portion of the text that is annotated, so that around half of the ssj500k v2.0 corpus is annotated for named entities.

As summarised in Table 7, these include five categories: persons (i.e. names of people and pets, pseudonyms, fictional characters and named groups of people), person derivatives (i.e. possessive adjectives derived from personal names), locations (i.e. geographical and celestial entities, public places, shopping centres, etc.), organisations (i.e. organisation, company and institution names, media, art-related groups, etc.) and the miscellaneous category that includes various types of entities ranging from names of works of arts to registered names of products or stock market indices. With the exception of possessive adjectives (essential for tools dealing with data anonymisation), most named entity annotations were thus attributed to nouns or noun phrases, usually in a capitalised or acronym form. This subset of the ssj500k corpus has already been used in the development of the Janes-NER (Fišer et al., 2018),[8] a named entity recognition system for South Slavic languages.

NE Type	Tag	No. of tokens
person	PER	2,928
person derivative	DERIV-PER	163
location	LOC	1,967
organisation	ORG	1,356
miscellaneous	MISC	602
Total		7,016

Table 7: Overview of named entity annotations in ssj500k.

3.5. Semantic Roles

Semantic Role Labelling (SRL) refers to the process of detecting and assigning semantic roles to semantic arguments determined by the predicate or a verb in a sentence. In comparison with syntactic trees, semantic role labelling requires a higher level of abstraction: namely, the same semantic content can be expressed with different syntactic relations. A semantic role labelling annotation scheme for Slovene was developed within the project Semantic Role Labelling in Slovene and Croatian (Gantar et al., 2018b) and follows previous SRL efforts for other languages, especially for Czech (the Prague Dependency Treebank tagset; Mikulová et al. (2006)), but also for English (FrameNet; Baker et al. (1998)); PropBank; Palmer et al. (2005)), Croatian (Crovallex; Preradović et al. (2009)) and other languages (AnCora, Taulé et al. (2011); SoNaR, Schuurman et al. (2010)). It was essential to take into account the specifics of the Slovene language, such as rich morphology and free word order. The final version of the tagset consists of 25 semantic labels: 5 for arguments, 17 for adjuncts, and

[7]The NER Guidelines are available from `http://nl.ijs.si/janes/wp-content/uploads/2017/09/SloveneNER-eng-v1.1.pdf`.

[8]`https://github.com/clarinsi/janes-ner`

3 labels for multi-word predicates and other multi-word expressions, as seen in Table 8. Each semantic label was defined and described in detail, which served as the guidelines for the annotation task.

In ssj500k, 5,491 sentences were annotated with semantic roles, with the first 500 sentences used for test annotation. The second phase included automatic annotation of the remaining 4,991 sentences and their manual revision. The resulting annotations in the corpus therefore represent the final role labelling system, tested both on Slovene and Croatian, which was already used for preliminary quantitative analyses (Gantar et al., 2018b). The ssj500k SRL data has also been used for modelling the mate-tools semantic role labeller (Björkelund et al., 2009; Gantar et al., 2018b), which has been applied to the Gigafida 2.0 reference corpus of Slovene.

Semantic role	Tag	No. of tokens
actor	ACT	5,269
patient	PAT	6,204
recipient	REC	1,064
origin	ORIG	290
result	RESLT	2,756
time	TIME	1,631
duration	DUR	414
frequency	FREQ	278
location	LOC	1,155
source	SOURCE	136
goal	GOAL	571
event	EVENT	104
aim	AIM	218
cause	CAUSE	355
contradiction	CONTR	128
condition	COND	315
regard	REG	452
accompaniment	ACMP	111
restriction	RESTR	6
manner	MANN	1500
means	MEANS	250
quantity	QUANT	250
multiword predicate	MWPRED	365
modal	MODAL	611
phraseme	PHRAS	211

Table 8: Overview of semantic role annotations in ssj500k.

3.6. Verbal Multi-Word Expressions

ssj500k was also annotated with verbal multi-word expressions (VMWEs) as part of the PARSEME COST Action Shared Task 1.1 (Ramisch et al., 2018), which focused on the automatic identification of VMWEs in running text and developed universal annotation guidelines (Candito et al., 2016) that were tested on 13 languages.

The corpus was annotated with four VMWE categories: a) *inherently reflexive verbs* (IRV), which occur with the independent morpheme *se/si* (*zdeti se* 'to seem'); b) *light-verb constructions* (LVC), which consist of a verb and nominal/prepositional phrase (e.g. *imeti mnenje* 'to have an opinion', *biti v dvomih* 'to be in doubt') and are divided into two subtypes: *full light-verb constructions* (LVC.full), where the verb contributes to the meaning on a predominantly categorical level, and *causal light-verb constructions* (LVC.cause), where the subject can be interpreted as the cause or source of the denoted action (e.g. *spraviti v smeh* 'to make someone laugh'); c) *inherently adpositional verbs* (IAV), which only occur with a prepositional morpheme (*simpatizirati z* 'to sympathise with') or change meaning when occurring with a prepositional morpheme (*biti za* 'be for, to support' vs. *biti* 'to be'); and d) *verbal idioms* (VID), which, even with certain syntactic conversions, must retain a meaning that is independent of the meanings of their elements (e.g. *plačati ceno* 'to pay the price').

A total of 13,511 sentences (approx. 48% of the corpus) were annotated using the FoLiA Linguistic Annotation Tool (FLAT). 2,290 of sentences contain at least one VMWE. On average, one VMWE is present in every fourth sentence. The total number of annotated VMWEs in the training corpus is 3,364, with 1,100 different VMWEs. The most frequent category (see Table 9 is IRV (48%) and the least frequent category is LVC.cause (2%).

MWE Type	Tag	Tokens
verbal idiom	VID	724
inherently reflexive verb	IRV	1627
light verb construction (full)	LVC.full	239
light verb construction (cause)	LVC.cause	64
inherently adpositional verb	IAV	710
Total		3,364

Table 9: Overview of verbal MWE annotations in ssj500k.

4. Format and Distribution

The ssj500k corpus is encoded according to the Text Encoding Initiative Guidelines (TEI Consortium, 2018), in particular, to the TEI parametrisation that is maintained by the CLARIN.SI research infrastructure.[9] The TEI document contains an extensive header, where the metadata of the corpus is given. This includes the usual bibliographic information but also the description of the sources, of the editorial policies, a listing of the tags used in the corpus together with their quantities and explanations, and a list of revisions of the corpus.

Two further elements of the TEI header deserve a mention. First, the header includes taxonomies for the text types included in the corpus and for the various labels of linguistic annotations included in the corpus: the named entity types, the JOS and UD syntactic labels, the types of verbal multiword expressions and of the semantic roles. These taxonomies enable us to give an identifier to each category, which is then referred to in the corpus annotations, while also encoding the gloss of the label both in English and Slovene. The second element is the TEI list of prefix definitions, which contains the definition of prefixes used in

[9]This TEI parametrisation is available from `https://github.com/clarinsi/TEI-schema`.

corpus annotation, e.g. `mte:` for MULTEXT-East morphosyntactic annotations. Each prefix definition contains the prefix, a description of the prefix use, and two regular expressions: a match pattern and a replacement pattern. The two implement the conversion of the string following the prefix into a URI, e.g. that the annotation `mte:Npmsl` corresponds to the URL `http://nl.ijs.si/ME/V6/msd/tables/msd-fslib-sl.xml#Npmsl`.

The body of the corpus is composed of divisions, each containing the samples of one source text of the FidaPLUS corpus, with each division giving also the bibliographic and taxonomic data on the text. The divisions are then composed of paragraphs, and these of sentences. In Figure 2 we give an (abbreviated) example of a sentence together with its annotations.

As can be seen, each sentence contains words, punctuation symbols and white-space elements, and possibly segments marking-up the named entities. Each token is annotated for its MULTEXT-East morphosyntactic description, its UD morphological features, its lemma and is also given an identifier, so that linking-type annotations can refer to it. The linking annotations for the two types of syntactic dependencies (UD and JOS), verbal multiword units (not present in the example sentence) and semantic roles are then given at the end of the sentence as link groups. These contain links, which give the label of the relation and the links to the head and argument of the relation.

The ssj500k v2.2 corpus is available for download under the CC BY-NC-SA licence from the CLARIN.SI repository (Krek et al., 2019), where it is archived in several formats:

- in TEI, with all annotations in English (e.g. `Ncmsn` for Noun, Type=common, Gender=masculine, Number=singular, Case=nominative);
- in TEI, with all annotations in Slovene (e.g. `Somet` for Samostalnik, vrsta=obča, spol=moški, število=ednina, sklon=imenovalnik);
- in CONLL-U format, automatically derived from the TEI version;
- in the so-called vertical format used by well-known concordancers such as CQP and Sketch Engine, automatically derived from the TEI version; this pack also includes the registry file needed to mount the corpus under one of the mentioned concordancers.

The corpus is also available in the noSketch Engine and KonText[10] concordancers installed at CLARIN.SI, and its repository entry is linked to the concordancers. It is thus possible to search and display the corpus, including most of the provided annotations.

5. Q-CAT Annotation and Querying Tool

Most annotations presented in sections above were carried out using the Q-CAT Querying-Supported Corpus Annotation Tool (TBA), which has been specifically designed for the annotation of the ssj500k corpus and has evolved accordingly with the new annotation layers being added through the years. Recently, however, the tool has been substantially improved so as to allow user modifications of annotation settings, including the addition of new annotation layers of various types, and complex queries employing one or more annotation layers. Several functions improving the overall user experience have also been added, including dynamic user-defined visualisations.

In short, the Q-CAT GUI, a .NET application, which runs on the Windows operating system, allows the user to annotate sentences with three types of annotation: tags, links and chunks. A tag is the annotation attributed to individual tokens, which typically include information on its surface form, base form (lemma) and morphosyntactic tag. A link is a directed connection from one token to another (including an optional root element), while a chunk is annotation attributed to a set of one or more tokens, which are not necessarily adjacent. The sentence is displayed graphically with tags being visualised as blue tiles attaching to the surface form, links as arrows and chunks as coloured rectangles (cf. Figure 3).

The user may define and customise one or more levels of annotation (for example: named entities, multi-word expressions, etc.), with each level consisting of one or more annotation types (for example: named entities might be persons, organisations, or locations). Each annotation type can be associated with a shortcut key for easier editing and with a colour for the purposes of visualisation. A sentence may be annotated with annotations belonging to any number of levels, but only two levels of the type link can be shown simultaneously (one above the sentence and one below it).

Q-CAT also supports queries that search over all the sentences in the input TEI file, looking for groups of words that satisfy the search constraints. These can be specified on the level of tags, chunks or links, as shown in Figure 4, which illustrates the query or actants (SRL link of type `ACT`) of main verbs (JOS morphosyntactic tag `Vm`), which are not annotated as subjects (JOS dependency label `Subj`). All sentences and tokens matching the constraints are displayed, counted and can be saved either as a subcorpus (`.xml`) or a tab-separated list (`.txt`).

6. Conclusions and Future Work

We have presented the content and the recent developments of the ssj500k training corpus for Slovene, which has been manually annotated with respect to segmentation, tokenisation, lemmatisation, JOS morphosyntax and dependency syntax, Universal Dependencies, semantic role labelling, named entities and verbal multi-word expressions. As such, the ssj500k represents the largest open source collection of grammatically annotated training data for Slovene language processing to date, which has been used in the development of many supervised machine learning systems for Slovene. In addition to that, the ssj500k corpus represents one of the fundamental language resources for Slovene language research, with a well-maintained infrastructure enabling its downloading, browsing, querying and annotation.

[10] KonText (Machálek, 2020) is an alternative front-end to Manatee (Rychlý, 2007), the back-end of the Sketch Engine concordancer. It is available from `https://github.com/ufal/lindat-kontext`.

```
<s xml:id="ssj1.1.2">
  <w ana="mte:Ncmsn" msd="UposTag=NOUN|Case=Nom|Gender=Masc|Number=Sing"
    lemma="dogodek" xml:id="ssj1.1.2.t1">Dogodek</w><c> </c>
  <w ana="mte:Sl" msd="UposTag=ADP|Case=Loc"
    lemma="v" xml:id="ssj1.1.2.t2">v</w><c> </c>
  <seg type="name" subtype="loc">
    <w ana="mte:Npmsl" msd="UposTag=PROPN|Case=Loc|Gender=Masc|Number=Sing"
       lemma="Ankaran" xml:id="ssj1.1.2.t3">Ankaranu</w>
  </seg><c> </c>
  ...
  <w ana="mte:Ncfsn" msd="UposTag=NOUN|Case=Nom|Gender=Fem|Number=Sing"
     lemma="nesreča" xml:id="ssj1.1.2.t7">nesreča</w>
  <pc ana="mte:Z" msd="UposTag=PUNCT" xml:id="ssj1.1.2.t8">.</pc>
  <linkGrp corresp="#ssj1.1.2" targFunc="head argument" type="UD-SYN">
    <link ana="ud-syn:root" target="#ssj1.1.2 #ssj1.1.2.t1"/>
    <link ana="ud-syn:case" target="#ssj1.1.2.t3 #ssj1.1.2.t2"/>
    ...
  </linkGrp>
  <linkGrp corresp="#ssj1.1.2" targFunc="head argument" type="JOS-SYN">
    <link ana="jos-syn:Atr" target="#ssj1.1.2.t5 #ssj1.1.2.t1"/>
    <link ana="jos-syn:Atr" target="#ssj1.1.2.t3 #ssj1.1.2.t2"/>
    ...
  </linkGrp>
  <linkGrp corresp="#ssj1.1.2" targFunc="head argument" type="SRL">
    <link ana="srl:ACT" target="#ssj1.1.2.t5 #ssj1.1.2.t1"/>
    <link ana="srl:PAT" target="#ssj1.1.2.t5 #ssj1.1.2.t7"/>
  </linkGrp>
</s>
```

Figure 2: An example of an TEI annotated sentence *The event in Ankaran was a dramatic accident.* The content is abbreviated for display.

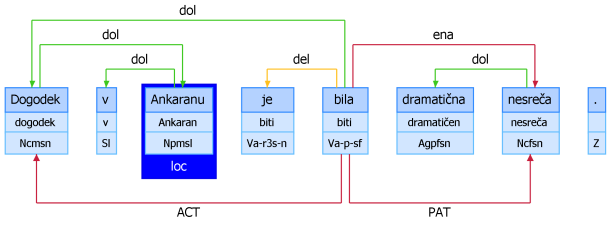

Figure 3: An example of an annotated sentence *The event in Ankaran was a dramatic accident.* in the Q-CAT tool.

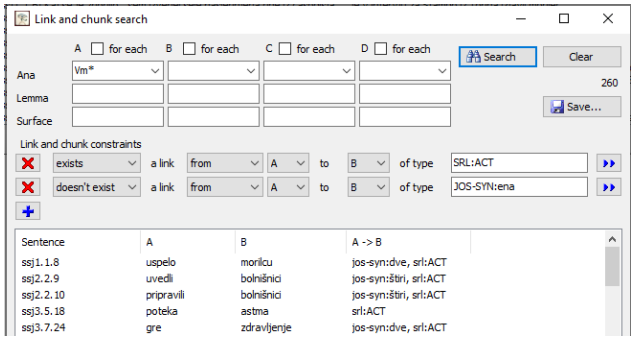

Figure 4: An example of a querying constraint in Q-CAT with results.

To support future research in both natural language engineering and linguistics, we will strive to ensure a continuous development of this dataset in the future as well. In addition to enlarging the coverage of existing annotation layers, especially those covering only part of the corpus, our immediate future work concerns adding new annotation layers, emerging from ongoing annotation campaigns, such as the annotation of non-verbal multi-word expressions (Gantar et al., 2019) and coreference resolution (Žitnik and Bajec, 2018), with possible modifications of the TEI and vertical formats as well.

Since the beginnings of ssj500k, several manually annotated training datasets for non-standard Slovene have also emerged, such as the Janes-Tag corpus of computer-mediated communication (Fišer et al., 2018), the goo300k corpus of historical texts (Erjavec, 2015) and the SST treebank of spoken Slovene (Dobrovoljc and Nivre, 2016), which have been shown to significantly improve non-standard language processing. In line with this, there is an ongoing project investigating the potential removal of the (relatively scarce) non-standard phenomena from the existing version of the ssj500k corpus, with the aim of providing a more reliable dataset for mainstream natural language processing tools, primarily developed for processing of standard Slovene.

7. Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. The work

described in this paper was funded by the Ministry of Education, Science and Sport within the "Communication in Slovene" project (3311-08-986003), and the Slovene Research Agency within the national research programmes "Language Resources and Technologies for Slovene" (P6-0411) and "Knowledge Technologies" (P2-0103), the national basic research projects "Linguistic annotation of Slovene language: methods and resources" (J2-9180) and "New grammar of contemporary standard Slovene: sources and methods" (J6-8256), and the Young Researcher Programme (MR-37487).

8. 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.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Primož Belej. 2018. Oblikoskladenjsko označevanje slovenskega jezika z globokimi nevronskimi mrežami. Master's thesis, Faculty of Computer Science and Informatics, University of Ljubljana.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado, June. Association for Computational Linguistics.

Marie Candito, Fabienne Cap Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Carlos Herrero, and et al. 2016. Parseme shared task 1.0 annotation guidelines - version 1.6b (last updated on november 26, 2016). Technical report, COST Action IC1207 – Parsing and multi-word expressions. Towards linguistic precision and computational efficiency in natural language processing (PARSEME).

Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevic, and Dan Tufis. 1998. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern European languages. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 315–319, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

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 (dependeny parser for slovene). In *Proceedings of the 8th Language Technologies Conference*, volume C, pages 42–47, Ljubljana, Slovenia, October. IJS.

Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017. The Universal Dependencies Treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, BSNLP@EACL 2017*, pages 33–38.

Kaja Dobrovoljc, Tomaž Erjavec, and Nikola Ljubešić. 2019. Improving UD processing via satellite resources for morphology. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 24–34, Paris, France, 26 August. Association for Computational Linguistics.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtsky, and Andreja Žele. 2006. Towards a Slovene dependency treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Tomaž Erjavec and Simon Krek. 2008. The JOS morphosyntactically tagged corpus of Slovene. In *LREC 2008*.

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. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 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.

Tomaž Erjavec. 2015. The IMP historical Slovene language resources. *Language Resources and Evaluation*, 49(3):753–775. `https://doi.org/10.1007/s10579-015-9294-7`.

Darja Fišer, Nikola Ljubešić, and Tomaž Erjavec. 2018. The Janes project: language resources and tools for Slovene user generated content. *Language Resources and Evaluation*. `https://rdcu.be/7RX4`.

Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Taja Kuzman, and Teja Kavčič. 2018a. Glagolske večbesedne enote v učnem korpusu ssj500k 2.1. In *Proceedings of the conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Polona Gantar, Kristina Štrkalj Despot, Simon Krek, and Nikola Ljubešić. 2018b. Towards semantic role labeling in slovene and croatian. In *Proceedings of the conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Polona Gantar, Jaka Cibej, and Mija Bon. 2019. Slovene multi-word units: Identification, categorization, and

representation. In *Computational and Corpus-Based Phraseology - Third International Conference, Europhras 2019, Malaga, Spain, September 25-27, 2019, Proceedings*, pages 99–112.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene). In *Proceedings of the 8th Language Technologies Conference*, volume C, pages 89–94, Ljubljana, Slovenia, October. IJS.

Peter Holozan, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman, and Aleš Velušček. 2008. Specifikacije za učni korpus. Projekt "Sporazumevanje v slovenskem jeziku" (Specifications for the Training Corpus. The "Communication in Slovene" project). Technical report, Project "Communication in Slovene".

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2019. *Training corpus ssj500k 2.2*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1210`.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, BSNLP@ACL 2019*.

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).

Tomáš Machálek. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France, May. European Language Resources Association.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2006. Annotation on the tectogrammatical level in the prague dependency treebank. annotation manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep.

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 et al. 2019. *Universal Dependencies 2.4*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, March.

Nives Mikelić Preradović, Damir Boras, and Sanja Kišiček. 2009. Crovallex: Croatian verb valence lexicon. In *Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces*, pages 533–538, June.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency Parsing from Scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium, October. Association for Computational Linguistics.

Carlos Ramisch, Silvo Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, and et al. 2018. Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. In *LAW-MWE-CxG 2018, The 12th Linguistic Annotation Workshop (LAW XII) and the 14th Workshop on Multiword Expressions (MWE 2018)*, pages 222–240, Santa Fe. Ljubljana University Press, Faculty of Arts.

Pavel Rychlý. 2007. Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masarykova univerzita.

Ineke Schuurman, Véronique Hoste, and Paola Monachesi. 2010. Interacting semantic layers of annotation in SoNaR, a reference corpus of contemporary written Dutch. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Mariona Taulé, M. Antònia Martí, and O. Borrega. 2011. Argument Structure Guidelines for Catalan and Spanish. Working paper 4: TEXT-MESS 2.0 (Text-Knowledge 2.0). Technical report, University of Barcelona.

TEI Consortium, editor. 2018. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Daniel Zeman et al. 2017. CoNLL 2017 Shared Task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October. Association for Computational Linguistics.

Tadej Štajner, Tomaž Erjavec, and Simon Krek. 2013.

Named entity recognition in slovene text. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2):58–81, Dec.

Slavko Žitnik and Marko Bajec. 2018. Odkrivanje koreferenčnosti v slovenskem jeziku na označenih besedilih iz coref149. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 6:37–67, 06.