The JOS linguistically tagged corpus of Slovene

Tomaž Erjavec,[1] Darja Fišer,[2] Simon Krek,[1] Nina Ledinek[3]

[1] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si, simon.krek@ijs.si

[2] Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
darja.fiser@guest.arnes.si

[3] Fran Ramovš Institute of the Slovenian Language,
Scientific Research Centre of the Slovenian Academy of Sciences and Arts
Novi trg 4, 1000 Ljubljana, Slovenia
NLedinek@zrc-sazu.si

Abstract

The JOS language resources are meant to facilitate developments of HLT and corpus linguistics for the Slovene language and consist of the morphosyntactic specifications, defining the Slovene morphosyntactic features and tagset; two annotated corpora (jos100k and jos1M); and two web services (a concordancer and text annotation tool). The paper introduces these components, and concentrates on jos100k, a 100,000 word sampled balanced monolingual Slovene corpus, manually annotated for three levels of linguistic description. On the morphosyntactic level, each word is annotated with its morphosyntactic description and lemma; on the syntactic level the sentences are annotated with dependency links; on the semantic level, all the occurrences of 100 top nouns in the corpus are annotated with their wordnet synset from the Slovene semantic lexicon sloWNet. The JOS corpora and specifications have a standardised encoding (Text Encoding Initiative Guidelines TEI P5) and are available for research from *http://nl.ijs.si/jos/* under the Creative Commons licence.

1. Introduction

Linguistically annotated corpora are the basis for human language technology and corpus linguistics but are, for a number of languages, still difficult to obtain, esp. as complete datasets. Essential resources are validated part-of-speech, or, better, morphosyntactically tagged corpora; treebanks; and word-sense annotated corpora.

The JOS[1] project aimed to fill this gap in Slovene language resources by producing two freely available annotated corpora in a standardised encoding, the smaller of which is manually annotated with these three levels of linguistic interpretation. We have previously (Erjavec and Krek, 2008) reported on the first stage of this annotation, where the two base corpora were constructed and the morphosyntactic annotation was performed. In this paper we report on the final result of the morphosyntactic annotation, including two Web services, and concentrate on the next two levels, namely the syntactic and lexico-semantic annotation.

2. The JOS Corpora

The two JOS corpora are the 100,000 word jos100k and the 1 million word jos1M. Both were obtained by sampling the FidaPLUS corpus[2] (Arhar and Gorjanc, 2007), a 600 million word reference corpus of Slovene annotated with automatically assigned context disambiguated morphosyntactic descriptions (MSDs) and lemmas. The first step to arrive at the JOS corpora was to convert FidaPLUS to XML in order to maintain a standard format and to enable processing with XML tools, in particular XSLT.

The content of JOS corpora was obtained from FidaPLUS by a two-stage filter and sampling procedure (first over documents, then over paragraphs) meant to ensure that jos100k and jos1M are representative and balanced, consist of clean texts, and do not infringe copyright (Erjavec and Krek, 2008).

The JOS corpora are encoded in XML, with the schema being a parameterisation of the Text Encoding Initiative Guidelines TEI P5 (TEI Consortium, 2007). The XML schema uses the TEI modules for corpora, simple linguistic analysis, linking, and ISO feature-structures. Furthermore, it also introduces some extensions, in particular validation of corpus MSDs directly from the schema. The two corpora have extensive meta-data (TEI headers) giving e.g., the bibliographical information about each texts, the text-type taxonomy, the morphosyntactic tagset and feature definitions, etc. While jos1M is annotated only at the word-level, the jos100k corpus Version 2.0 is annotated for three levels of linguistic description.

Figure 1 shows the corpus mark-up of these tree levels. Words are annotated by their MSD and lemma. Syntactic annotation is stored in stand-off mark-up, with dependency labels marking pointers to the two connected tokens; the sentence id serves as the root. The semantic label from the Slovene wordnet lexicon (identical to the Princeton WordNet synset id) is attached to the term element. Each term element is also marked for its head noun and possibly by a subtype indicating missing synsets (or specific enough hy-

[1] The JOS acronym stands for "Jezikoslovno označevanje slovenščine", i.e. "Linguistic Annotation of Slovene".

[2] http://www.fidaplus.net/

```
<s xml:id="F0020003.557.2">
  <w xml:id="F0020003.557.2.1" lemma="ta" msd="Zk-sei">To</w><S/>
  <w xml:id="F0020003.557.2.2" lemma="biti" msd="Gp-ste-n">je</w><S/>
  <term type="sloWNet" sortKey="kraj" subtype="missing_hyponym" key="ENG20-08114200-n">
    <w xml:id="F0020003.557.2.3" lemma="turističen" msd="Ppnmein">turističen</w><S/>
    <w xml:id="F0020003.557.2.4" lemma="kraj" msd="Somei">kraj</w>
  </term>
  <c xml:id="F0020003.557.2.5">.</c><S/>
</s>
<linkGrp type="syntax" targFunc="head argument" corresp="#F0020003.557.2">
  <link type="ena" targets="#F0020003.557.2.2 #F0020003.557.2.1"/>
  <link type="modra" targets="#F0020003.557.2 #F0020003.557.2.2"/>
  <link type="dol" targets="#F0020003.557.2.4 #F0020003.557.2.3"/>
  <link type="dol" targets="#F0020003.557.2.2 #F0020003.557.2.4"/>
  <link type="modra" targets="#F0020003.557.2 #F0020003.557.2.5"/>
</linkGrp>
```

Figure 1: Example sentence from jos100k: "To je turističen kraj." , lit. "It is a tourist place."

ponyms) in PWN. The MSDs and dependency relations are given their Slovene label in the XML source – however, these can be interchanged with their English equivalents.

Element	n	Gloss
div	248	Sampled text from FidaPLUS
p	1,599	Complete paragraph
s	6,151	Sentence
term	5,430	Wordnet literal
w	100,003	Word token, annotated
c	18,391	Punctuation token
S	98,890	Whitespace
linkGrp	5,961	Syntactic analysis of a sentence
link	112,442	Syntactic dependency relation

Table 1: XML tag counts in jos100k

Table 1 gives the counts over the TEI elements used in jos100k: there are almost 250 texts represented in the corpus, with 1,600 sampled paragraphs containing 6,000 sentences. To corpus contains over 5,000 semantic annotations, with just under 6,000 sentences having syntactic annotations; note that currently about 5% are missing a syntactic analysis.

3. Morphosyntactic annotation

The JOS morphosyntactic specifications (Version 1.1) are the basis for word-level corpus annotation, as they define the Slovene tagset of 1,902 morphosyntactic descriptions (MSDs) and give the decomposition of these MSDs into features. The specifications are, just as the corpora, encoded in TEI P5 and are compatible with the Slovene part of the multilingual MULTEXT-East morphosyntactic specifications Version 4 (Erjavec, 2010). The JOS specifications are written in both Slovene and English, with the MSDs and feature names also being available in both languages. The specifications are provided in source XML and derived HTML, with the MSDs tagset also available in tabular files giving the mappings between various formats.

The manual annotation, performed by a supervised team of students, consisted of correcting the MSDs and lemmas in the two JOS corpora, where the base-line annotations were mapped to the JOS specifications from the FidaPLUS MSDs and lemmas. The annotation of the jos100k corpus was validated twice by different annotators, and the words where the two manual annotations differed were validated for a third time; jos100k can thus serve as a gold standard annotated corpus of Slovene.

The jos1M corpus is also morphosyntactically annotated but project resources did not allow for manual verification of the complete corpus, so only "suspicious" MSDs were validated, about 190,000 words. Experiments have shown that, due to its greater size, the jo1M corpus used as a training set produces better tagging models than jos100k in spite of its mistakes. The main purpose of jos1M is thus to provide a training set for part-of-speech taggers and lemmatisers for Slovene.

The JOS homepage also provides two Web services, a concordancer and an automatic annotation service. The two corpora are available for searching via a Web interface with CQP (Christ, 1994) as the back-end. The web interface allows displaying and querying over words or word-level annotations, i.e. lemmas, MSDs and even morphosyntactic features (e.g. supports queries such as `[clitic="yes" & number="dual"]`) and thus enables detailed grammatical explorations of the corpora.

A tagger and lemmatiser for Slovene (Erjavec and Džeroski, 2004) were trained on jos1M and are offered as a Web service under JOS. Users can submit texts, and receive them tokenised, tagged and lemmatised. The output format is compatible with SketchEngine[3] (Kilgarriff et al., 2004), so users having an account there can upload their processed texts to SketchEngine and use its powerful corpus analysis features over their corpora.

4. Shallow syntactic annotation

Treebanks are a basic language resource, used to study syntactic phenomena and as training and testing datasets for inductive parsers. For Slovene, the first attempt to produce a treebank was SDT (Džeroski et al., 2006) where a part of the MULTEXT-East corpus was annotated with analytic dependency structures according to the Prague Dependency Treebank model (Hajič et al., 2006). However, in the process of manually annotating this corpus it turned out it was

[3] http://www.sketchengine.co.uk/

difficult for students to consistently follow the annotation guidelines.

For this reason, the JOS dependency model was developed, which, although based on the Prague model, is considerably simpler. It reduces the number of possible dependencies to 10 and lays down easier to follow guidelines for manual annotation. In the first stage, the shallow dependency model was elaborated, annotator guidelines were written, and a 500 sentence corpus was carefully annotated, to test the model and serve as a base of examples for the annotators. Then jos100k was annotated in full by a team of supervised students. The corpus was first annotated in parallel by two annotators with a dedicated graphical editor, and the differences resolved by a third annotator. The result is the first syntactically annotated corpus of Slovene.

In Table 2 we give the syntactic dependencies with their Slovene names and English equivalents, the number of times they were used in the annotation of jos100k and a short gloss. While a detailed exposition of their meaning is outside the scope of this paper, a brief description of the dependencies follows. Root links the abstract node of the clause or sentence to elements which form further connections in the dependency tree. PPart links elements without a dependency relation in the usual head-dependent sense which are consequently defined merely as parts of a word phrase, typically parts of a predicate. Atr links heads to their dependents in word phrases. Sb/Obj/AdvM/AdvO link subjects, objects and adverbials, although these dependencies do not comply entirely with their definitions in traditional Slovene grammars. Coord is used to link parts of coordinate structures on the phrase level. Conj is used in combination with the Coord relation to link the source of Coord via its target as the source of Conj into a triangle identifying the two heads of a coordinate structure and the corresponding conjunction. MWU links words with a very strong tendency to appear together as a group forming a multiword unit and which do not show characteristics of a head-dependent phrase structure.

Relation	Name	n	Gloss
modra	Root	32,912	Root of the tree
del	PPart	7,879	Predicate part
dol	Atr	36,873	Attribute
ena	Sb	5,641	Subject
dve	Obj	7,445	Object
tri	AdvM	2,762	Adverbial of manner
stiri	AdvO	6,827	Adverbials, other
prir	Coord	2,896	Coordination
vez	Conj	8,858	Conjunction
skup	MWU	349	Multi-Word Unit

Table 2: Syntactic dependencies in jos100k

The purpose of the corpus is to serve as a training and testing set for dependency learners / parsers. Current experiments with MST[4] show that with 10-fold cross validation over jos100k labeled accuracy is 80% and unlabeled accuracy is 84%.

[4]The MSTParser (McDonald et al., 2006) achieved the best results over the SDT Slovene treebank in the 2006 CoNLL-X shared task on multilingual dependency parsing.

5. Lexical sense assignments

Word sense disambiguation is a challenging tasks for human language technologies but in order to develop programs to perform it, a manually annotated sense tagged corpus needs to be available. For Slovene, a semantic lexicon based on Wordnet, called sloWNet, has been developed semi-automatically using various language resources, in particular a bilingual dictionary, parallel corpora and open source lexical resources, such as Wikipedia and the Eurovoc thesaurus (Fišer and Sagot, 2008).

The latest version of sloWNet contains about 20,000 unique literals which are organized into almost 17,000 synsets. It is rich in basic concepts as well as specific ones. The former were mostly obtained from the dictionary and parallel corpus while the latter come from Wikipedia. sloWNet mostly contains nominal synsets, although there are some verbal and adjectival synsets as well. Apart from single word literals, there are also plenty of multi-word expressions. A comparison of nouns in sloWNet and the jos100k corpus showed that sloWNet nouns cover 30% of the nouns present in jos100k, with 90% coverage of the top third of the nouns ranked by frequency.

The main goal of lexical sense assignment was to obtain the first semantically annotated corpus for Slovene. However, because sloWNet had been created automatically and had been based on a foreign-language resource, our secondary goal was to check the coverage of the senses it contains compared to the senses represented in the corpus and thereby evaluate the developed lexicon in a practical semantic task and to improve it.

In this first attempt of semantically annotating Slovene (Fišer and Erjavec, 2010), we limited the task to nouns only because sense assignment for nouns is the easiest and because they are currently best covered in sloWNet. We extracted all the common nouns that exist in sloWNet which have more than one sense and appear in jos100k with a frequency of at least 30. There were 102 such nouns, most of which belong to the Basic Concept Sets in wordnet. This yielded a total of 5,430 tokens, which means that on average there are about 54 annotation examples for each noun included in the annotation process. The most frequent noun is leto/year with almost 350 examples and the next most frequent dan/day with 150.

The annotation procedure consisted of several stages: the annotators started from sloWNet in which they checked all senses of a given word and corrected any errors they found. In the second step, the annotators turned their attention to the concordances and tried to assign a wordnet sense to each occurrence of the given word in the corpus. If they came across a meaning of a word or a phrase they could not find in sloWNet, they added it to the wordnet. In the end, the annotations were consolidated and validated by a referee.

The annotators assigned over 500 different synsets to the 5,430 examples, i.e. about 5 senses per noun on average. Five of the nouns were monosemous in the corpus (e.g., muzej/museum), while the most polysemous noun annotated was čas/time for which a total of 15 senses were used; 27 examples, mostly proper names and culture-specific metaphorical expressions, were left unannotated

because no appropriate wordnet synset could be found.
The annotation process was complicated due to a number of factors, such as the well-known fine-grainedness of senses in Wordnet, problems with single vs. multiword terms, etc. To test the quality of the annotation, 500 previously unseen random occurrences of the target words were given to two of the annotators, and their annotations were compared to those in jos100k. The interannotator agreement was 66% which is slightly worse than usual figures for wordnet sense assignment; however, as mentioned, we chose high-frequency nouns, which also exhibit the most polysemy and are therefore hard to annotate.
The validation of sloWNet with corpus annotations has shown that most core senses that were required to annotate the corpus had already been present in sloWNet whereas the same is not true for peripheral senses and especially for multi-word expressions which had to be added by the annotators in many cases. Multi-word expressions were especially difficult, as in almost half of the cases no exactly appropriate sense could be found in wordnet. This suggests that sloWNet will have to be further extended in order to ensure a thorough coverage of the sense inventory relevant for Slovene.

6. Conclusions

The paper presented the JOS linguistically annotated corpora, focusing on the jos100k corpus, which contains manually assigned morphosyntactic descriptions, lemmas, shallow dependency structures, and wordnet senses for selected nouns. The purpose of the corpus mostly to serve as a testbed for development of Slovene language part-of-speech taggers, lemmatisers, dependency parsers and word-sense disambiguators. An especially interesting topic is using the combination of annotations to produce better quality annotators. Apart from HLT uses, the corpus could also be interesting for linguists even though the annotations often differ in meaning from linguistically expected terms.
The JOS corpora and associated resources are available in the source XML TEI P5 encoding, as well as in several derived formats more suitable for immediate processing and exploitation. In addition to the language resources, the project also offers two services interesting for linguistic uses: a Web concordancer over the corpora, and an annotator that tokenises, tags and lemmatises Slovene texts.
The JOS resources are available from the homepage of the project[5] and the corpora can be directly downloaded under the Creative Commons Attribution-Noncommercial 2.5 Slovenia license. The presented corpora are the first such publicly available resources for Slovene, and should advance HLT research for the language.
Further work on extending the JOS Slovene language resources is being undertaken in the Slovene long term project SSJ,[6] which has as its goals building annotated corpora, a lexical database and language reference materials for Slovene. Here, JOS corpora will be enlarged, resulting in e.g., a 1 million word fully manually validated morphosyntactically tagged corpus, a 400,000 word syntactically tagged corpus, and a 100 million word balanced automatically tagged and parsed corpus of Slovene.

[5] http://nl.ijs.si/jos/

[6] Sporazumevanje v slovenskem jeziku / Communication in Slovene, http://www.slovenscina.eu/

Acknowledgments

The work described in this paper was supported by Slovenian Research Agency project J2-9180 JOS and the EU 6FP-033917 project SMART.

7. 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).

Oliver Christ. 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of COMPLEX '94: 3rd conference on Computational Lexicography and Text Research*, pages 23–32, Budapest, Hungary.

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 and Sašo Džeroski. 2004. Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, 18(1):17–41.

Tomaž Erjavec and Simon Krek. 2008. The JOS morphosyntactically tagged corpus of Slovene. In *Sixth International Conference on Language Resources and Evaluation, LREC'08*, Paris. ELRA.

Tomaž Erjavec. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Seventh International Conference on Language Resources and Evaluation, LREC'10*, Paris. ELRA.

Darja Fišer and Tomaž Erjavec. 2010. sloWNet: Construction and Corpus Annotation. In *Proceedings of Fifth International Conference of the Global WordNet Association (GWC'10)*, Mumbay.

Darja Fišer and Beinot Sagot. 2008. Combining Multiple Resources to Build Reliable Wordnets. In *Proceedings of Text Speech and Dialogue Conference, TSD'08*, Brno.

Jan Hajič, Jarmila Panevová, Eva Hajičova, Petr Pajas, Petr Sgall, Jan Štěpánek, Jǐrí Havelka, and Marie Milkulová. 2006. Prague Dependency Treebank 2.0. Catalog Number LDC2006T01.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress*, pages 105–116, Lorient, France.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. In *Tenth Conference on Computational Natural Language Learning (CoNLL-X)*.

TEI Consortium, editor. 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*.