The JOS morphosyntactically tagged corpus of Slovene

Tomaž Erjavec, Simon Krek

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

Abstract

The JOS morphosyntactic resources for Slovene consist of the specifications, lexicon, and two corpora: jos100k, a 100,000 word balanced monolingual sampled corpus annotated with hand validated morphosyntactic descriptions (MSDs) and lemmas, and jos1M, the 1 million word partially hand validated corpus. The two corpora have been sampled from the 600M word Slovene reference corpus FidaPLUS. The JOS resources have a standardised encoding, with the MULTEXT-East-type morphosyntactic specifications and the corpora encoded according to the Text Encoding Initiative Guidelines P5. JOS resources are available as a dataset for research under the Creative Commons licence and are meant to facilitate developments of HLT for Slovene.

1. Introduction

Linguistically annotated corpora are the basis for human language technology research 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, necessary for training taggers, themselves a basic infrastructure for more advanced HLT tasks.

For Slovene, the MULTEXT-East[1] resources (Erjavec, 2004) have so far contained the only available manually validated tagged corpus, i.e., the Slovene translation of the novel ”1984” by G. Orwell. And while the MULTEXT-East tagset and corpus encoding practices have been adopted for a number of Slovene corpora, among them the reference corpora FIDA (Erjavec et al., 1998) and FidaPLUS (Arhar and Gorjanc, 2007), the “1984” corpus itself is nevertheless small (100,000 words) and contains only one translated novel, resulting in very brittle tagging models (Erjavec and Sárossy, 2006). Furthermore, the years of using the Slovene MULTEXT-East tagset have shown that it could do with several modifications.

The JOS project attempts to remedy the lack of Slovene language resources by producing standardised and freely available annotated corpora, in the first instance a revised set of morphosyntactic specifications including a tagset and better morphosyntactically annotated corpora. This paper reports on the first results, the gold-standard jos100k corpus, and the current work on jos1M. The two corpora contain sampled paragraphs from FidaPLUS and are annotated with disambiguated and validated morphosyntactic descriptions and lemmas.

The rest of the paper is structured as follows: Section 2. describes the FidaPLUS corpus and discusses the sampling procedure; Section 3. introduces the JOS morphosyntactic specifications and tagset and their relation to the MULTEXT-East ones; Section 4. explains the manual annotation procedure; Section 5. mentions the encoding of the corpus; Section 6. its availability; and Section 7. gives conclusions and plans for further work.

2. Sampling FidaPLUS

FIDA[2] (Erjavec et al., 1998) is a 100M word monolingual reference corpus of modern day Slovene, and contains varied texts dating from 1990-2000, with over a quarter being from 1999. FIDA is encoded in SGML, and follows the Text Encoding Initiative Guidelines, TEI P3 (Sperberg-McQueen and Burnard, 1999), with words annotated with their lexical (i.e., ambiguous) Slovene MULTEXT-East (Erjavec, 2004) morphosyntactic descriptions (MSDs) and lemmas. FidaPLUS[3] (Arhar and Gorjanc, 2007) significantly extends FIDA (to 600 million words and -2006) and provides, in addition to lexical tags, also automatically assigned context disambiguated MSDs and lemmas. The entire text processing chain, including up-conversion from source formats, tokenisation, lexical processing and disambiguation was performed with the proprietary software by the Slovene HLT company Amebis.[4]

FidaPLUS is freely available for research via a Web concordancer, and is a very useful tool for research into Slovene language. But, outside FIDA/FidaPLUS projects partner institutions, it is not available as a dataset and so cannot serve as the basis for HLT-type research. As a training set for PoS taggers it also suffers from the drawback that it was tagged fully automatically; and while the Amebis tagger gives state-of-the-art performance for Slovene, nevertheless about 15% of the words have annotation errors, including out-of-vocabulary words, about 2%, which are not assigned annotations. FidaPLUS does, however, offer an excellent basis on which to develop a corpus for HLT research.

The first step to arrive at the JOS corpora was to convert FidaPLUS to TEI P4 XML (Sperberg-McQueen and Burnard, 2002) in order to maintain a standard format and to enable processing with XML tools, in particular XSLT. For FIDA (SGML, TEI P3) this step would have been relatively simple, as XML is a subset of SGML, and TEI P4 is backward compatible with TEI P3. Unfortunately, in the production of FidaPLUS neither SGML nor MULTEXT-East conformance was enforced, so the conversion turned out to be a

[1] http://nl.ijs.si/ME/

[2] http://www.fida.net/, now no longer maintained.

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

[4] http://www.amebis.si/

more laborious task than would otherwise be the case. The clean-up procedure was, for the most part, done via a series of Perl scripts, resulting in the XML TEI P4 encoded corpus we call Fida+X. This corpus is slightly smaller than FidaPLUS because the texts for which Perl heuristics were insufficient to make them valid TEI P4 were dropped. Fida+X then served as the source for making the JOS corpora.

2.1. Sampling procedure

The content of jos100k and jos1M corpora was obtained from the 600M word FIDA+X by a two stage filter and sampling procedure meant to help JOS corpora achieve the following characteristics:

- Are representative and balanced.
  The representativeness of JOS corpora follows from this attribute holding for FidaPLUS. As far as balance is concerned, FidaPLUS contains a large percentage of newspaper text, and a relatively small one of fiction and esp. professional writing (technical, academic prose). Simply adopting the balance of FidaPLUS would leave the much smaller JOS corpora with very small amounts of such texts.

- Consist of clean text.
  Given that the corpora will be linguistically annotated, it is worth ensuring that the corpus contains only legitimate text paragraphs and tokens: FidaPLUS contains duplicates and cases where the up-conversion produced very short texts, paragraphs or sentences, or did not remove all formatting information.

- Do not infringe copyright.
  While complete text might be preferable for certain types of analysis, this would be questionable for copyright reasons, while short samples from the texts are not problematic. A sampling procedure has the further advantage that it makes the corpus more varied.

The first sampling step randomly selects complete texts from the Fida+X corpus, but excludes anomalous texts and prefers certain text types to others. First, a filter discards texts that are too short or too long, have too much formatting in them, or are badly formed according to various other heuristics. Second, ponders are given to text types and other metadata, so that, in short, the bias of FidaPLUS towards newspapers is somewhat counteracted, mostly towards technical writing.

Taking these texts as the input, the second step selects random paragraphs, again subject to some constraints: the minimum and maximum size of the paragraph, and that each paragraph is unique in the corpus (duplicates are discarded by using CRC on their text).

These two steps were run twice, with different settings. For jos100k, the first step was set to produce a 10M word corpus, and for jos1M a 100M word one, i.e., for both corpora on the average 1% of paragraphs was selected from the texts.

Table 1 gives an overview of the sizes of the two JOS corpora, in terms of the number of texts, paragraphs, sentences, words and (word and punctuation) tokens.

Corpus	jos100k	jos1M
`<text>`	249	2,565
`<p>`	1,599	15,758
`<s>`	6,151	60,291
`<w>`	100,003	1,000,019
`<w>+<c>`	118,394	1,182,945

Table 1: Tagcount of JOS corpora

In FidaPLUS each text is annotated according to the FIDA typology with categories of text type and text medium. As mentioned, the first step of the sampling procedure gave weights to these categories, so the proportions in JOS are different from those in FidaPLUS. Table 2 and Table 3 give the proportions of the top categories in the two JOS corpora. The weights for some text types were somewhat changed between jos100k and jos1M, which is reflected in the differing proportions; as can be seen, jos1M contains less fiction/monographs and newspapers, but more natural science and technical writings / monthly publications.

Type	jos100k	jos1M
fiction (prose)	10,1%	6,7%
nonprofessional	67,6%	66,6%
humanities and soc.sci.	9,6%	9,9%
natural sciences and tech.	6,5%	13,3%
Σ of above	93,8%	96,6%

Table 2: Top text types by percentage of words

Medium	jos100k	jos1M
monograph	28,1%	22,9%
monthly publication	9,6%	16,2%
bi-weekly publication	2,3%	2,0%
weekly publication	7,3%	9,1%
weekly newspaper	9,0%	10,4%
daily newspaper	37,7%	34,3%
Σ of above	94,1%	95,0%

Table 3: Top text media by percentage of words

3. JOS morphosyntactic specifications and tagset

MULTEXT(-East) morphosyntactic specifications are based on work by EAGLES (Calzolari and Monachini, 1996) and set out the grammar and vocabulary of valid morphosyntactic descriptions, MSDs. The specifications determine what, for each language, is a valid MSD and what it means. For instance, they can define that the MSD `Ncmsan` is a valid MSD for Slovene, and that it corresponds to the feature structure `Noun, Type = common, Gender = masculine, Number = singular, Case = accusative, Animate = no`.

It should be noted that the Slovene MULTEXT-East tagset differs substantially from tagsets of inflectionally less rich languages, such as the majority of Western European ones. In Slovene, as in other Slavic languages, words can be marked with a large number of features, and the Slovene MULTEXT-East tagset is correspondingly large, with around 2,000 tags.

While it was felt that the formal basis and principles of the Slovene tagset were adequate – if at times not perfect – for Slovene, there were a number of details that were considered problematic, e.g., certain attributes or their values, allowed combinations of attribute-values, as well as the lexical assignment of MSD to particular words or word groups. Another problematic aspect, this one of the MULTEXT-East specifications as a whole, is the ordering of the attributes in the MSD string; as the specifications cover a large number of quite varied languages, language specific attributes (or those added to the specifications at a later date) wind up at the end of the string, leading to unwieldy MSDs, such as Gppspe--n-----d. It would be better if an individual language had the freedom to reorder attributes, as long the mapping to feature-structure representation was maintained.

These are the reasons why new morphosyntactic specifications were developed for JOS,[5] which will hopefully be able to serve as a standard morphosyntactic tagset for Slovene. To this end, the choices made in MULTEXT-East were re-examined, and the tagset compared and contrasted to other annotation schemes of Slovene, in particular the one used in the LC-Star corpus (Verdonik et al., 2004), and the "Nova beseda" tagset (Jakopin and Bizjak, 1997) which differs from the previous two in its fundamental design (Lönneker, 2005), i.e., it does not use positional attributes and is very closely tied to traditional Slovene grammars. Tagsets of related languages were also studied to compare best practices, in particular the Prague tagset used e.g., in the Czech National Corpus[6] and Prague Dependency Treebank[7].

Compared to MULTEXT-East, the ordering of the attributes was also changed for JOS to make the tagset more natural for Slovene, i.e., lexical attributes were grouped at the start of the MSD string, with inflectional ones, more common first, toward the end.

The resulting JOS specification is still an application of the MULTEXT-East principles, but the procedure to convert between the FidaPLUS/MULTEXT-East corpus MSDs and those of JOS is non-trivial because the mapping has to take into account not only the MSDs but, in general, also the word-form or its lemma.

A synopsis of the JOS features defined for Slovene is given in Table 4. The specifications also include the lexical list of MSDs, i.e., the complete and consistent set of valid MSDs for Slovene, amounting to 1,908. Each MSD is given with its expansion into a feature-structure as well as examples of usage. The rather large number is necessary to distinguish a reasonably complete set of morphosyntactic features, although some combinations are quite rare in practice. For example, the jos100k corpus uses only 1,064 different MSDs for the disambiguated annotations.

PoS	Attributes with No. of values
Noun	Type(2), Gender(3), Number(3), Case(6), Animacy(2)
Verb	Type(2), Aspect(3), Form(7), Person(3), Number(3), Gender(3), Negative(2)
Adjective	Type(3) Degree(3), Gender(3), Number(3), Case(6), Definiteness(2)
Adverb	Degree(3), Participle(2)
Pronoun	Type(9), Person(3), Gender(3), Number(3), Case(6), Owner_Number(3), Owner_Gender(3), Form(2)
Numeral	Form(3), Type(4), Gender(3), Number(3), Case(6), Definiteness(2)
Preposition	Case(6)
Conjunction	Type(2)
Particle	no attributes
Interjection	no attributes
Abbreviation	no attributes
Residual	Type(3)

Table 4: Slovene JOS categories, their attributres and the number of different attribute-values.

Figure 1 shows an example from the JOS morphosyntactic specifications. It is beyond the scope of this paper to discuss the morphosyntactic specifications in detail, however, it should be mentioned, as is evident from the example, that they are available both in Slovene as well as English. This holds both for the accompanying text as for the translations (localisations) of feature names and codes, enabling shifting between Slovene language and English language MSDs and feature structures. For example, with the specifications and an accompanying XSLT stylesheet it is possible to translate the English Ncmsan to Slovene Sometn, and expand either to its English or Slovene feature-structure, the latter being samostalnik, vrsta = občno_ime, spol = moški, število = ednina, sklon = tožilnik, živost = ne. This makes it possible for Slovene speakers to use the annotations in their native language, while also allowing English speakers to understand them.

4. Annotating the corpora

The manual annotation, performed by a supervised team of undergraduate students, consists of correcting the MSDs and lemmas in the two JOS corpora, where the base-line annotations were mapped to the JOS specifications from the Fida+X/MULTEXT-East MSDs and lemmas, automatically assigned by Amebis.

Technically, the manual annotation proceeds via a Web interface, which, for a given corpus and input parameters (e.g., regexp over word-form, lemma or MSD), generates Excel spreadsheets for the annotators. The spreadsheets feature a title sheet, a sheet with the text and annotations to be corrected (via drop-down menus), and guidelines. Upon

[5] The development of JOS specifications and tagset coincided with a major re-design of the MULTEXT-East specifications, coupled with the addition of new languages, in particular Russian (Sharoff et al., 2008).

[6] http://ucnk.ff.cuni.cz/

[7] http://ufal.mff.cuni.cz/pdt/

```
<div type="section" xml:id="msd.N">
<head>SAMOSTALNIK</head>
<table n="msd.cat" xml:id="msd.cat.N">
<head xml:lang="sl">Tabela atributov in vrednosti za
samostalnik</head>
<head xml:lang="en">Attribute-value table for
Noun</head>
<row role="type">
<cell role="position">0</cell>
<cell role="name" xml:lang="sl">samostalnik</cell>
<cell role="code" xml:lang="sl">S</cell>
<cell role="name" xml:lang="en">Noun</cell>
<cell role="code" xml:lang="en">N</cell>
</row>
<row role="attribute">
<cell role="position">1</cell>
<cell role="name" xml:lang="sl">vrsta</cell>
<cell role="name" xml:lang="en">Type</cell>
<cell role="values">
<table>
<row role="value">
<cell role="name" xml:lang="sl">občno_ime</cell>
<cell role="code" xml:lang="sl">o</cell>
<cell role="name" xml:lang="en">common</cell>
<cell role="code" xml:lang="en">c</cell>
</row>
<row role="value">
<cell role="name" xml:lang="sl">lastno_ime</cell>
<cell role="code" xml:lang="sl">l</cell>
<cell role="name" xml:lang="en">proper</cell>
<cell role="code" xml:lang="en">p</cell>
</row>
</table>
</cell>
</row>
...
```

Figure 1: JOS morphosyntactic specifications: start of table for Noun.

correction, the spreadsheets are uploaded and the corpus updated with the new manual annotations. This overall architecture was originally developed for correcting and annotating historical texts (Erjavec, 2007), and was then successfully applied in the JOS project as well.

The annotation process is cyclical, with a mixture of manual and automatic annotation steps, depending on the corpus in question.

4.1. Annotating jos100k

The annotation of the 100k corpus was carried out in parallel with developing the JOS morphosyntactic specifications, tagset and its lexical mapping, along with the guidelines for annotators. This made the process rather complicated but resulted in specifications, tagset, lexicon and corpus which are consistent and made the jos100k corpus as free of errors as possible, given project constraints.

The complete corpus was validated twice by different annotators, and the words where the two manual annotations differed were validated for a third time. When further mistakes were spotted, annotations of certain token types in the corpus were unified or corrected in several subsequent steps to arrive at the gold standard JOS manually annotated corpus.

4.2. Annotating jos1M

As project resources do not allow for manual verification of the complete 1M word JOS, the intent is to validate only “suspicious” MSDs, as well as automatically improve the Fida+X annotations. This work is still in progress (although a preliminary version of jos1M is available), and this section explains the adopted methodology.

We trained the TnT tagger (Brants, 2000) on the manually annotated jos100k, and gave it, as the backup lexicon, the lexicon extracted from Fida+X with its MSD converted to the JOS specification. The word tokens where the TnT and Amebis differ in their MSD assignment were labelled as suspicious, and are being manually validated.

To estimate the accuracy and error separation of the TnT and Amebis taggers we performed an experiment where jos100k was annotated by TnT as above, but with 10-fold cross validation over the corpus, i.e., training on 90% of jos100k and tagging the remaining 10%, with the experiment repeated 10 times, each time taking a different slice of the corpus. We thus arrived at a corpus where each word is annotated with three MSDs - the manually validated one, the one assingned by Amebis, and the one assigned by TnT. The results are shown in Table 5.

The first line gives the complete size of the corpus in words. The second gives an estimate of the accuracy of the TnT tagger (86.6%), and the third of the Amebis tagger, at 85.7%. As mentioned, Amebis does not assign tags to unknown words, which constitute about 2% of the word tokens, whereas manual annotation tagged all words (even if with subtypes of Residual, e.g., foreign), as does the TnT tagger. The fourth line covers cases where both taggers predict the correct MSD, for 78% of the words. The next four lines then cover cases where either one, or the other, or both taggers are wrong. The last line deserves special mention, as in 3.2% of the cases, the taggers agree in their tagging, but are both wrong.

Manual	Amebis	TnT	Words
MSD_1	any	any	100,003
MSD_1	any	MSD_1	86,617
MSD_1	MSD_1	any	85,719
MSD_1	MSD_1	MSD_1	78,011
MSD_1	MSD_1	MSD_2	7,708
MSD_1	MSD_2	MSD_1	8,606
MSD_1	MSD_2	MSD_3	2,440
MSD_1	MSD_2	MSD_2	3,238

Table 5: Amebis / TnT tagging accuracy on jos100k

For validating the MSDs in jos1M we concentrate on the words where the two taggers disagree. For manual annotation, this means that out of 1M words, 7.7% + 8.6% + 2.4% $\sim$ 190,000 need to be validated to achieve an overall word accuracy of 100% - 3.2% = 96.8%. The manual annotation of these words is currently in progress, and we are also investigating ways of automatically combining the outputs of both taggers to increase accuracy over Fida+X.

5. Corpus encoding

The corpora, as well as the morphosyntactic specifications, are encoded in XML, according to the Guidelines for Text Encoding TEI P5 (TEI Consortium, 2007). The corpora therefore come with an XML schema, which allows for validation of the specifications and corpora.

JOS corpora consists of the header and component texts. The header gives, inter alia, the corpus size and tag usage, the list of all bibliographic elements from Fida+X describing the texts in the corpus, the feature structure library defining all the JOS MSDs and their decomposition into features, and the feature library with feature definitions.

Each corpus text is linked to its description in the header and is then composed of a series of the sampled paragraphs, which consists of sentences, and these of punctuation and word tokens, as well as an element to mark whitespace. Each word token is annotated with its disambiguated MSD and lemma.

6. Availability

The JOS resources are available from the homepage of the project, [8] and the corpora can be downloaded subject to the Creative Commons Attribution-Noncommercial 2.5 Slovenia licence.[9] Unfortunately, we cannot allow commercial exploitation, as that is not allowed by the agreement between FIDA and FidaPlus and the text providers, or by the agreement between JOS and the FIDA and FidaPLUS consortium, which includes also commercial partners.

The JOS corpora are available in the source XML TEI P5 encoding, as well as several derived formats, more suitable for immediate processing and exploitation. In particular, we offer the corpora in the file format for the Corpus Workbench[10] (Christ, 1994), where each line contains either a structural tag or a tab separated line containing the wordform, lemma, MSD, and the MSD decomposed into separate features. This allows for searching on particular attributes of the token, regardless of the part-of-speech. For instance, to return all tokens marked as feminine genitive, the CWB query would be: `[gender="feminine" & case="genitive"]`

7. Conclusions

The paper presented the initial results of the JOS project, with the focus on the finished jos100k corpus, and the jos1M corpus that is currently being finalised. The paper discusses the FidaPLUS source for the corpora, the cleanup and sampling procedure, the morphosyntactic specifications, the annotation procedure, the encoding of the corpus and its availability. The presented corpora are the first such publicly available resources for Slovene, and should significantly advance part-of-speech tagging and lemmatisation research for the language.

Future work in the JOS project concerns the next two levels of linguistic annotation. In particular, the next couple of years should see, taking jos100k as the basis, the annotation of a tree-bank and a sense-disambiguated corpus with the Slovene WordNet (Erjavec and Fišer, 2006) used for the sense inventories.

Acknowledgements

The work described in this paper was supported in part by Slovenian Research Agency grant J2-9180 “Linguistic annotation of Slovene language: methods and resources” and by the EU 6FP-033917 project SMART “Statistical Multilingual Analysis for Retrieval and Translation”.

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).

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*. http://www.coli.uni-sb.de/˜thorsten/tnt/.

Nicoletta Calzolari and Monica Monachini, editors. 1996. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: A common proposal and applications to european languages. EAGLES Report EAG—CLWG—MORPHSYN/R, ILC, Pisa. http://www.ilc.cnr.it/EAGLES96/morphsyn/.

Oliver Christ. 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of COMPLEX ’94*, pages 23–32, Budapest, Hungary. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/.

Tomaž Erjavec and Darja Fišer. 2006. Building Slovene WordNet. In *LREC’06*.

Tomaž Erjavec and Bence Sárossy. 2006. Morphosyntactic tagging of Slovene legal language. *Informatica*, 30(4):483–488.

Tomaž Erjavec, Vojko Gorjanc, and Marko Stabej. 1998. Korpus FIDA (The FIDA Corpus). In *Proceedings of the Conference ’Language Technologies for the Slovene Language’*, pages 124–127, Ljubljana, Slovenia.

Tomaž Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *LREC’04*.

Tomaž Erjavec. 2007. An architecture for editing complex digital documents. In *Proceedings of INFuture2007: “Digital Information and Heritage”*, Zagreb, Croatia.

Primož Jakopin and Aleksandra Bizjak. 1997. O strojno podprtem oblikoslovnem označevanju slovenskega besedila (On Machine Assited Morphosyntactic Tagging of Slovene Texts). *Slavistična Revija*, 45(3-4):513–532.

Birte Lönneker. 2005. Strojno oblikoslovno označevanje slovenskih besedil: Kako daleč smo? (Part-of-speech tagging of Slovenian texts: How far did we get?) *Slavistična Revija*, (2):193–210.

Serge Sharoff, Mikhail Kopotev, Tomaž Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a Russian tagset. In *LREC’08*.

C. M. Sperberg-McQueen and Lou Burnard, editors. 1999. *Guidelines for Electronic Text Encoding and Interchange, Revised Reprint*. The TEI Consortium.

C. M. Sperberg-McQueen and Lou Burnard, editors. 2002. *Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines*. The TEI Consortium.

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

Darinka Verdonik, Matej Rojc, and Zdravko Kačič. 2004. Creating Slovenian Language Resources for Development of Speech-to-Speech Translation Components. In *LREC’04*.

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

[9] http://creativecommons.org/licenses/by-nc/2.5/si/

[10] http://cwb.sf.net