Annotating CLARIN.SI TEI corpora with WebAnno

Tomaž Erjavec[†♣], Špela Arhar Holdt[♠*], Jaka Čibej[♠],
Kaja Dobrovoljc[*], Darja Fišer[♠†], Cyprian Laskowski[‡], Katja Zupan[†♣]

†Dept. for Knowledge
Technologies
Jožef Stefan Institute

‡Centre for Language
Resources and Technologies
University of Ljubljana

∗Trojina, Institute for
Applied Slovene Studies

♠Faculty of Arts
University of Ljubljana

Ljubljana, Slovenia

♣Jožef Stefan International
Postgraduate School

tomaz.erjavec@ijs.si, spela.arharholdt@ff.uni-lj.si,
jaka.cibej@ff.uni-lj.si, kaja.dobrovoljc@trojina.si,
darja.fiser@ff.uni-lj.si, cyp@trojina.si, katja.zupan@ijs.si

Abstract

The abstract presents the CLARIN.SI supported WebAnno platform for manual annotation of corpora. We concentrate on the conversion of the corpus TEI encoding to the WebAnno format and the merge of WebAnno export into the original TEI. We also overview some annotation campaigns over Slovene corpora.

1 Introduction

Manually annotated corpora are a basic language resource for empirical linguistics and human language technologies. Linguists want to compile and use such corpora for research into particular phenomena of language, esp. where automatic methods produce results of insufficient quality for subsequent analyses, or where automatic annotation methods do not even exist. Manually annotated corpora are even more crucial for the development of human language technologies for particular languages, as the prevalent method of annotation tool development is now supervised machine learning, where the approaches are largely language independent, but they do need corpora with high-quality annotations for training their models. Furthermore, regardless of the method, gold standard corpora are needed for evaluating the quality of the developed tools.

In this paper we concentrate on the platform, standard and workflow we are promoting and facilitating in the scope of CLARIN.SI and, so far, mostly for the Slovene language. However, we believe that the methodology is largely language independent and could serve as a stepping-stone for others in need of similar functionality.

In Slovenia in general, and CLARIN.SI in particular, the main encoding standard used and promoted is the TEI and the reasons for this are described in Section 2. The annotation platform we use is WebAnno, which we introduce in Section 3, with particular emphasis on converting TEI corpora into a WebAnno format and back into the TEI. In Section 4 we illustrate the use of the developed platform on several on-going projects, while Section 5 gives some conclusions and directions for further research.

2 Using in-line TEI for linguistic annotation

The Guidelines for Text Encoding and Interchange (TEI) have become a de-facto standard in digital humanities, esp. for complex digital editions. They are used much less in natural language processing, where many teams have developed their own XML annotation schemas, such as the German TCF (Hinrichs at al., 2010) used in WebLicht or FoLiA (van Gompel and Reynaert, 2014) used for the annotation of most Dutch corpora. Nevertheless, the TEI also has a following in corpus encoding, in particular for annotating the Polish National Corpus (Bański and Przepiórkowski, 2009) or the proposal for annotating corpora of Computer Mediated Communication (CMC), which has been established as a TEI SIG (Beißwenger et al., 2012).

In Slovenia, the TEI has also been used to annotate most large publicly available corpora, such as the sampled corpus of contemporary Slovene ccGigafida (http://hdl.handle.net/11356/1035), the reference speech corpus Gos (http://hdl.handle.net/11356/1040) and the corpus of historical Slovene IMP (http://hdl.handle.net/11356/1031).

The difference between the mentioned schemas and the Slovene one is the preferred placement of linguistic annotations. The others tend to use stand-off annotation, whereby the basic text is unmodified, with annotation layers stored separately and linked to the base text with pointers. The stand-off approach can encode arbitrary relations, is conceptually simple and annotation-tool friendly, but also brings problems, esp. for hand-annotated corpora. First, it becomes impossible to change the base text or dependent annotations as this breaks the pointers. Such changes are common when annotating historical texts, where leftover errors in transcriptions are often noticed only when linguistic annotation is already well underway, or with annotating CMC corpora where spaces are often left out of the text so tokenisation needs to be corrected. Second, validation of stand-off annotation, such as proper nesting of annotations, is difficult because the pointers must be resolved and their scope compared.

For these reasons we use, as much as possible, in-line TEI annotations: if necessary, they are simple to down-convert to stand-off, but with the relationship between various annotation layers explicit they allow changing selected parts of the text and direct validation with the TEI XML schema.

To illustrate, we give in Figure 1 an example from the Janes corpus of Slovene CMC (Fišer et al., 2015, Fišer et al. 2016), where the words are first normalised, and the normalised words then PoS tagged and lemmatised with models for standard Slovene. As can be seen, we separate the original word(s) from the normalised (regularised) one(s), with the former receiving linguistic annotations.

```
TEI:    <w lemma="operater" ana="#Somei">operater</w><c> </c>
        <w lemma="vedeti" ana="#Ggnste">ve</w><c> </c>
        <choice>
          <orig><w>dab</w></orig>
          <reg><w lemma="da" ana="#Vd">da</w><c> </c><w lemma="biti" ana="#Gp-g">bi</w></reg>
        </choice><c> </c>
        <w lemma="ob" ana="#Dt">ob</w><c> </c>
        <w lemma="#soocenje" ana="#Nh">#soocenje</w>
        …
```

Original:	operater	ve	dab	ob	#soocenje	popizdu
Normalised:	operater	ve	da bi	ob	#soocenje	popizdil
English translation:	the_operator	knows	that he_would	at	#faceoff	go_cunt_mad

Figure 1: Example of an annotated sentence from the Janes corpus.

3 WebAnno and TEI

From the platforms for manual annotation we chose WebAnno (Eckart de Castilho, 2014) as the one installed and supported by CLARIN.SI: it is open source, supports various types of annotations (token, span, link), allows for annotation campaigns involving many annotators also over the same texts, has a curation step where conflicting annotations are resolved, and is fairly well maintained.

However, WebAnno does not support TEI, except in a limited sense for import. Given our wide-ranging requirements for annotation we developed scripts to import TEI into the WebAnno tabular TSV format and to merge exported files of this format back into the source TEI, thereby adding new annotations to existing ones in the TEI documents.

3.1 Importing documents to WebAnno

To import texts into WebAnno, the TEI sample to be annotated needs to be split into files of convenient size for annotation, and these converted to TSV. The TEI to TSV conversion is, for the most part, straightforward and is implemented as an XSLT stylesheet, which makes use of XML configuration files specifying which existing (typically automatically assigned) annotation should be exported to

TSV for a particular project. The complications in the stylesheet come from the bells and whistles that we also wanted to support. For example, and as illustrated in the Figure 1 above, we want to allow many-to-one and one-to-many mappings between tokens and their normalised versions, and allow correcting the base tokens or the tokenisation. Probably the most complex part of the conversion comes from the ability to specify that the original and normalised token layers should be switched, so that the latter become the base annotation layer as this allows syntactic annotation of the normalised tokens, which is otherwise impossible due to the non-bijective mapping between the two.

3.2 Exporting documents from WebAnno

Once an annotation campaign is completed, the annotations are exported in TSV and merged into the original TEI document. This is indeed a merge rather than a conversion, since TEI contains more information than TSV, e.g., various metadata and other data that might not concern annotations, such as the presence (or lack) of spaces between tokens, XML identifiers associated with different kinds of elements, or previous linguistc annotations not exported to TSV for the current annotation campaign.

As our WebAnno layers allow for deeper changes to the TEI document on the level of sentences or tokens, we take the skeleton from the original TEI document, but recreate the TEI sentences (or other elements that were taken as the base for WebAnno “sentence” units) from scratch using a combination of the data from the original TEI sentences and the exported TSV documents.

This is done with a Python script which expects command-line parameters specifying the configuration file (the same one that was used in the TEI to TSV conversion), the original TEI document, the exported TSV document, the name of the merged TEI document to be created, and some logging-related parameters.

4 Use cases

At CLARIN.SI we have organised two tutorials on WebAnno (Čibej, 2015) and set-up and completed several annotation campaigns, with others still on-going. The typically workflow follows the MATTER framework (Pustejovsky and Stubbs, 2013) and consists of identifying the annotation task, formalising it in terms of TEI annotations and WebAnno structure, writing the annotation guidelines, creating the sample to be annotated, a small test annotation which shows up conversion errors or inconsistencies, training the annotators, and then the actual annotation campaign, usually with one curator and double annotations by the student annotators. Each campaign also has a dedicated mailing list, while the new files to be annotated are distributed on a regular basis. In the rest of this section we overview several annotation campaigns concentrating on their more interesting aspects.

4.1 Essential annotation of Slovene CMC

In the scope of the Janes project we have compiled a large corpus of Slovene CMC, which is automatically annotated for tokens, sentence boundaries, normalised forms of words, their PoS tags and lemmas, with all the processing steps except the first two relying on machine learning methods (Fišer et al., 2015, Fišer et al., 2016). While the tools produce reasonably good results, there is still significant room for improvement. For this reason we sampled two datasets from the corpus, one of 4,000 tweets and the other of 4,000 posts of user comments and forums. All the listed annotation layers were taken into consideration, but split between two campaigns – in the first the tokens, sentences and normalisations were corrected, while the second one treats MSD tagging and lemmatisation and is currently still on-going. Non-bijective normalisation and tokenisation corrections are also catered for with a combination of multi-valued features and special symbols used as their values.

4.2 Syntactic annotation of CMC

Slovene currently has one treebank, the ssj500k (http://hdl.handle.net/11356/1052), of which about 200,000 tokens are dependency annotated. We wanted to open this and other corpora to collective annotation. In particular, we also wish to syntactically annotate non-standard texts, where the problem of annotating the normalised tokens is faced and solved as explained above. Currently we have automatically treebanked the Tweet sample and are in the processes of manually correcting these annotations.

4.3 Multi-level annotation of speech transcriptions

WebAnno was also used for manual annotation of a representative sample of the Gos corpus of spoken Slovene (http://hdl.handle.net/11356/1040), a collection of manually segmented and normalized transcripts of spontaneous speech in different everyday situations. First, the sampled subset of the corpus (approx. 30,000 tokens) was additionally annotated for lemmas, PoS tags, morphological features and dependency syntax using the speech-adapted Universal Dependencies annotation scheme (Dobrovoljc and Nivre, 2016). In the second phase, an additional span layer was introduced for semantic annotation of multi-word discourse structuring devices (Dobrovoljc, 2016). Although WebAnno proved to be a highly flexible and user-friendly annotation tool for the annotation of all six linguistic layers (including some normalization corrections), its usage in speech-related annotation projects would be significantly improved if it enabled hyperlinking or importing the original audio recordings.

4.4 Named entities

About one third of the ssj500k corpus is also annotated for named entities (NE), and this is currently the only available NE annotated corpus of Slovene. We plan to extend the NE annotated portion of ssj500k and also annotate the goo300k corpus of historical Slovene (http://hdl.handle.net/11356/1025).

As there were no explicit guidelines for the current ssj500k NER annotation, we first plan to write them, possibly changing decisions and correcting inconsistencies in the current annotations. We have currently defined the layers, opened a project for these corrections, and uploaded the annotated as well as the not yet NER-annotated portion of the ssj500k.

4.5 Comma placement and shortening strategies

We also undertook two annotation campaigns oriented towards a linguistic analysis rather than tool development. The first concerned the comma, which is the most notoriously difficult punctuation to use correctly in Slovene. The annotation task consisted of annotating a set of tweets with misplaced or missing commas, according to a typology of 35 classes. The study (Popič et al., 2016) showed that in Slovene CMC comma use is problematic mostly in regard to the missing comma, especially after and before small clauses and between dependent and independent clauses. While the reason for the former can be attributed to the nature of the medium, the latter is a universal problem of Slovene speakers and has nothing to do with the interactivity and informality of the Twitter platform.

The other campaign concerned the strategies that Slovene Twitter users employ to shorten their tweets. Here the spans that are shortened were marked up with a typology of 34 classes, with one span possibly exhibiting several classes. 800 tweets were annotated in this manner, and the subsequent study (Goli et al., 2016) showed that shortening strategies were present in most analysed tweets and are much more common in non-standard ones. The highest number and widest range of shortening strategies arise at the orthographic level, relatively few were identified at the lexical level, and very few at the syntactic level. This is understandable as omissions of spaces, punctuation, etc. has the lowest impact on the understandability of the message, compared to lexical and syntactic reductions.

5 Conclusions

The abstract presented the CLARIN.SI supported WebAnno platform, its interface to our TEI corpora and gave some examples of use. The development of the conversion scripts is done on our installation of GitLab, which includes continuous integration testing in order not to introduce bugs in the development process.

In further work we would like to simplify TEI import and export, i.e. make the job of the TEI curator easier, most probably through a Web interface or plug-in for WebAnno. A functionality which the annotators would find very convenient is the ability to search through the already annotated texts in order to see how particular phenomena were annotated in the campaign. While it is probably unrealistic to expect that this utility will be added to WebAnno, we are considering introducing the option of automatically creating a searchable corpus under noSketchEngine (Rychlý, 2007) as part of the process of TEI curation of a (partially completed) WebAnno annotation.

While the complete conversion platform is still under development, it has already been tested in practice on several complex annotation scenarios. It is also language independent, and we are happy to share it with other interested CLARIN centres.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their helpful comments. The work presented in this paper was supported by the Slovenian Programme “European Research Infrastructures”, the Slovenian ARRS research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (no. J6-6842), the programme “Knowledge Technologies” (no. P2-103) and the Young Researchers programme (36491, 37487).

References

[Bański and Przepiórkowski 2009] Piotr Bański and Adam Przepiórkowski. (2009). Stand-off TEI annotation: the case of the National Corpus of Polish. In *Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 64-67.

[Beißwenger et al. 2012] Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. (2012). A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative, (3)* (doi: 10.4000/jtei.476). [http://jtei.revues.org/476]

[Čibej 2015] Jaka Čibej. (2015). Delavnice JANES Ekspres za promocijo korpusnih in spletnih virov za slovenščino (JANES Express Tutorials for the Promotion of Corpus and Web resources for Slovene). *Slovenščina 2.0, 3 (2)*: 63–66.

[Dobrovoljc 2016] Kaja Dobrovoljc. (2016). Annotation of Multi-Word Discourse Markers in the Spoken Slovenian Treebank. *3rd International Conference on Linguistic & Psycholinguistic Approaches to Text Structuring (LPTS 2016).*

[Dobrovoljc and Nivre 2016] Kaja Dobrovoljc and Joakim Nivre. (2016). The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC'16)*, 1566–1573.

[Eckart de Castilho 2014] Richard Eckart de Castilho, Chris Biemann, Iryna Gurevych, and Seid Muhie Yimam. (2014): WebAnno: a flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands.

[Fišer et al. 2015] Darja Fišer, Nikola Ljubešić, and Tomaž Erjavec. (2015). The JANES corpus of Slovene user generated content: construction and annotation. *International Research Days: Social Media and CMC Corpora for the eHumanities*. October 23rd – 24th, 2015, Rennes, France.

[Fišer et al. 2016] Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. (2016). JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin (JANES 04: Corpus of Slovene User-Generated Content). *Slovenščina 2.0, 4 (2)*: 67–100.

[Goli, Osrajnik, and Fišer 2016] Teja Goli, Eneja Osrajnik, and Darja Fišer. (2016). Strategije krajšanja slovenskih sporočil na družbenem omrežju Twitter (Strategies of Slovene message shortening on the Twitter social network). *Proceedings of the Language Technlogies and Digital Humanities Conference*, Sept. 29th – October 1st, 2016, Faculty of Arts, Ljubljana [http://www.sdjt.si/jtdh-2016/].

[Hinrichs at al. 2010] Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. (2010). WebLicht: web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations (ACLDemos '10)*, 25–29.

[Popič, Fišer, Zupan and Logar 2016] Damjan Popič, Darja Fišer, Katja Zupan, and Polona Logar. (2016). Raba vejice v uporabniških spletnih vsebinah (The use of the comma in user-generated web content). *Proceedings of the Language Technologies and Digital Humanities Conference,* Sept. 29th – October 1st, 2016, Faculty of Arts, Ljubljana [http://www.sdjt.si/jtdh-2016/].

[Pustejovsky and Stubbs 2013] James Pustejovsky, and Amber Stubbs. (2013). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. Sebastopol,* CA: O’Reilly.

[Rychlý 2007] Pavel Rychlý. (2007). Manatee/Bonito – A Modular Corpus Manager. *Proceedings of the Workshop on Recent Advances in Slavonic Natural Language Processing.* Brno: Masaryk University, 65–70.

[TEI] TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. [http://www.tei-c.org/Guidelines/P5/].

[van Gompel and Reynaert 2014] Maarten van Gompel and Martin Reynaert. (2014). FoLiA: A practical XML format for linguistic annotation – a descriptive and comparative study; *Computational Linguistics in the Netherlands Journal; 3*:63–81; 2013.