CLARIN Annual Conference 2016

Aix-en-Provence, 26-28 October 2016

CLARIN.SI
Common Language Resources and Technology Infrastructure, Slovenia

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

♠Faculty of Arts
University of Ljubljana

‡Centre for Language Resources and Technologies
University of Ljubljana

*Trojina, Institute for Applied Slovene Studies

♣Jožef Stefan International Postgraduate School

CLARIN 2016
Aix-en-Provence

CLARIN.SI
Common Language Resources and Technology Infrastructure, Slovenia

Overview

1. Introduction
2. TEI corpus encoding @ CLARIN.SI
3. The WebAnno annotation platform
4. TEI to WebAnno, WebAnno to TEI
5. Use cases
6. Conclusions

Introduction

- Manually annotated corpora are a basic language resource for empirical linguistics and human language technologies:
  - *Linguistics*: 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
  - *HLT*: testing data, training data for supervised ML methods
- How to best to annotate them?
- How to store them in a common format?

TEI < Text Encoding Initiative >

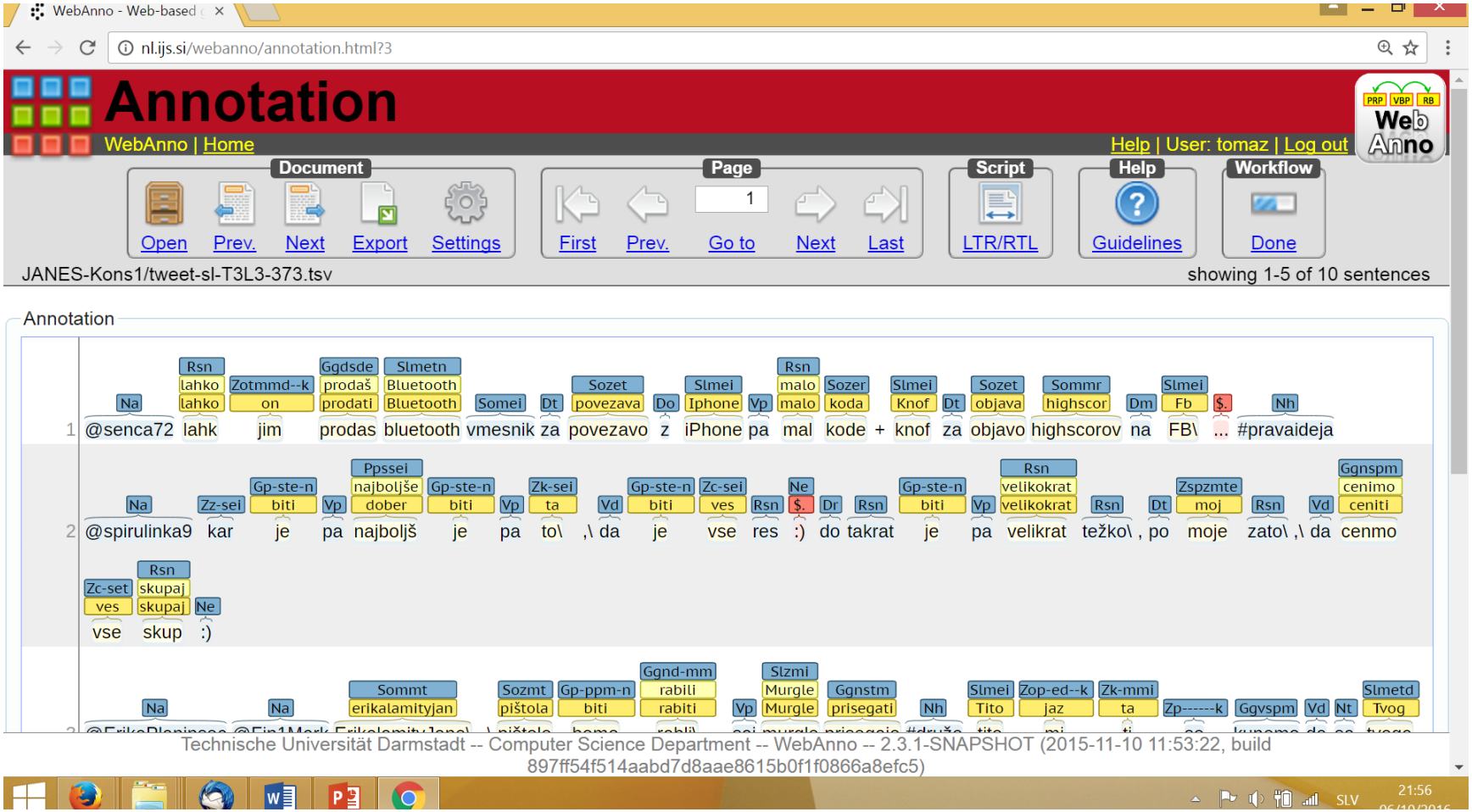
- The Guidelines for Text Encoding and Interchange mostly used in digital humanities (digital editions)
- Less used in HTL, but: PNC, esp. suited for CLARIN!
- De-facto standard for most Slovene corpora: Gigafida, Gos, IMP, Janes… (+ digital editions, lexica…)
- Typical annotation levels:
  - Tokens, sentences
  - Normalised words („phrases“)
  - PoS tags, lemmas
  - Syntactic analysis
  - Named entities
  - etc.

Stand-off vs. in-line

- Stand-off:
  - Arbitrary relations
  - Simple to implement
  - Tool friendly
- In-line: none of the above! But:
  - The annotated text can still be corrected
  - Easier to validate

```
<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>
```

WebAnno



CLARIN.SI
Common Language Resources and Technology Infrastructure, Slovenia

WebAnno

- CLARIN-D, GitHub
- Various types of annotations:
  - token, span, link
  - stacked, multivalued
- Supports many annotators + curation
- Is fairly well maintained (but for how long?)
- I/O: XML stand-off and tabular formats, e.g. TSV

- Decided to use and promote it @ CLARIN.SI
- Organised tutorials end of 2015:
  Janes Express: Ljubljana – Zagreb – Belgrade

Annotation workflow

1. Convert TEI to TSV:
   - what is a WA „sentence“?
   - do we show spaces?
   - what to export?
2. Upload TSV to WebAnno
3. Manual annotation
4. Download TSV
5. Merge TSV with source TEI

Project (meta)data

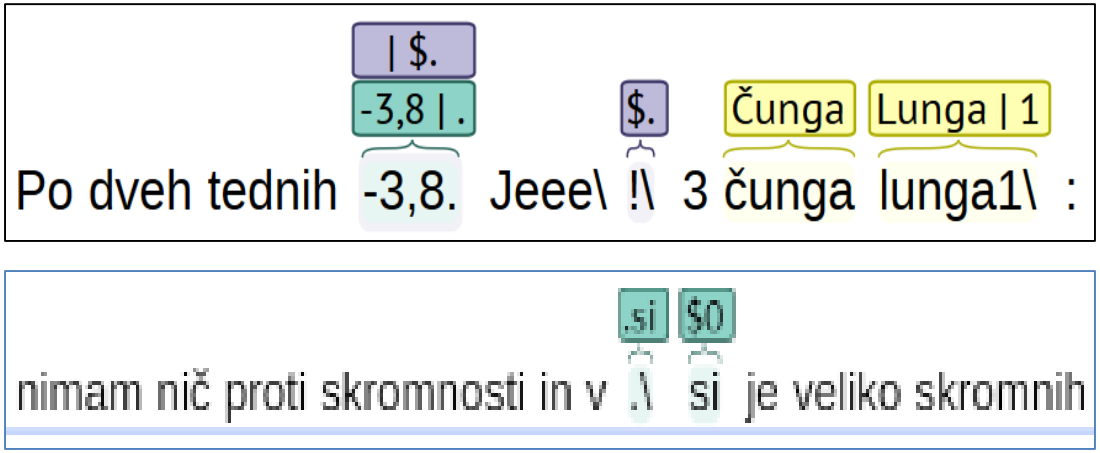
- Source TEI
- Annotation guidelines
- Project set-up in WebAnno:
  annotation layers (with tagsets), annotators, input TSV documents
- One XML configuration file for:
  - TEI2TSV (XSLT)
  - TSV2TEI (Perl + DOM)

Example configuration file

```
<WA-config>
 <verbose>true</verbose>
 <normalised_base>true</normalised_base>
 <null_char>$0</null_char>
 <glue_char/>
 <punct_tag>Z</punct_tag>
 <LayerDefinitions>
  <layer name="webanno.custom.Originals" type="token">
   <feature name="original"/><feature name="original2"/>
   <feature name="original3"/><feature name="original4"/>
  </layer>
  <layer name="webanno.custom.Lemma" type="token"><feature name="lemma"/></layer>
  <layer name="webanno.custom.jos.MSD.en" type="token"><feature name="msd"/></layer>
  <layer name="webanno.custom.jos.Dep.en" type="relation">
   <feature name="label"/>
   <feature name="AttachTo" value="webanno.custom.jos.MSD.en"/>
  </layer>
 </LayerDefinitions>
</WA-config>
```

Bells and whistles

Correction, tokenisation, normalisation, sentences



Use cases

- Essential annotation of Slovene CMC:
  - 4,000 tweets + 4,000 user comments and forum posts
  - tokens, sentences and normalisations
  - MSD tags and lemmas
- Multi-level annotation of speech transcriptions
  - 30,000 tokens
  - lemmas, UD PoS tags, morphological features, dependency syntax
  - semantic annotation of multi-word discourse structuring devices
- Named entities
  - Standard Slovene + CMC + historical Slovene
- Comma placement, shortening strategies, …

Conclusions

- Presented WebAnno + TEI @ CLARIN.SI
- WebAnno problems:
  - lacking search + metadata / links
  - various small bugs and crashes
  - unclear future?
- Merge step
  - can be, in general, quite complex (tokenisation, whitespace)
- Workflow
  - too many manual steps
- GitLab (2 GitHub, to share?)

Hvala!

CLARIN.SI

Common Language Resources and Technology Infrastructure, Slovenia