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 >

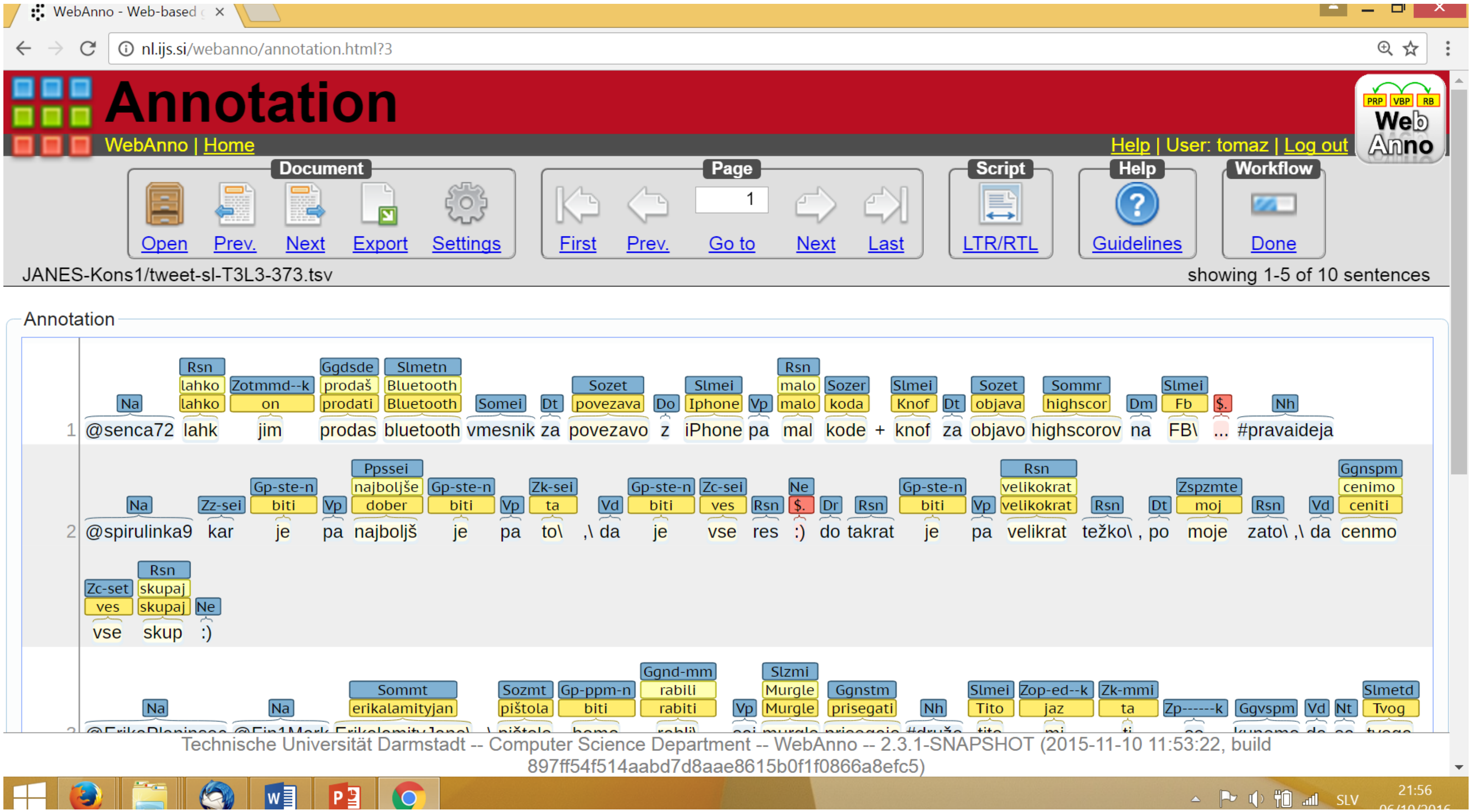
- The Guidelines for Text Encoding and Interchange mostly used in digital humanities (digital editions)
- Less used in HTL, but: PNC, esp. suited for CLARIN!
- De-facto standard for most Slovene corpora: Gigafida, Gos, IMP, Janes… (+ digital editions, lexica…)
- Typical annotation levels:
  - Tokens, 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



WebAnno - Web-based

nl.ijs.si/webanno/annotation.html?3

Annotation

WebAnno | Home

Help | User: tomaz | Log out

JANES-Kons1/tweet-sl-T3L3-373.tsv

showing 1-5 of 10 sentences

Annotation

1

Token	Annotations
@senca72	Na
lahk	Rsn lahko lahko
jim	Zotmmd--k on
prodas	Ggdsde prodaš prodati
bluetooth	Slmetn Bluetooth Bluetooth
vmesnik	Somei
za	Dt
povezavo	Sozet povezava
z	Do
iPhone	Slmei Iphone
pa	Vp
mal	Rsn malo malo
kode	Sozer koda
+	
knof	Slmei Knof
za	Dt
objavo	Sozet objava
highscorov	Sommr highscor
na	Dm
FB\	Slmei Fb
...	$.
#pravaideja	Nh

2

Token	Annotations
@spirulinka9	Na
kar	Zz-sei
je	Gp-ste-n biti
pa	Vp
najboljš	Ppssei najboljše dober
je	Gp-ste-n biti
pa	Vp
to\	Zk-sei ta
,\	
da	Vd
je	Gp-ste-n biti
vse	Zc-sei ves
res	Rsn
:)	Ne $.
do	Dr
takrat	Rsn
je	Gp-ste-n biti
pa	Vp
velikrat	Rsn velikokrat velikokrat
težko\	Rsn
,	
po	Dt
moje	Zspzmte moj
zato\	Rsn
,\	
da	Vd
cenmo	Ggnspm cenimo ceniti
vse	Zc-set ves
skup	Rsn skupaj skupaj
:)	Ne

3

Na · Na · Sommt erikalamityjan · Sozmt pištola · Gp-ppm-n biti · Ggnd-mm rabili rabiti · Vp · Slzmi Murgle Murgle · Ggnstm prisegati · Nh · Slmei Tito · Zop-ed--k jaz · Zk-mmi ta · Zp------k · Ggvspm · Vd · Nt · Slmetd Tvog

Technische Universität Darmstadt -- Computer Science Department -- WebAnno -- 2.3.1-SNAPSHOT (2015-11-10 11:53:22, build 897ff54f514aabd7d8aae8615b0f1f0866a8efc5)

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

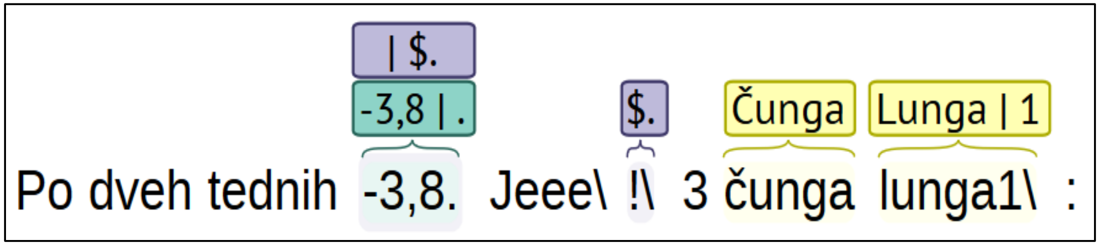
- Source TEI
- Annotation guidelines
- Project set-up in WebAnno:
  annotation layers (with tagsets), annotators, input TSV documents
- One XML configuration file for:
  - TEI2TSV (XSLT)
  - TSV2TEI (Perl + DOM)

Example configuration file

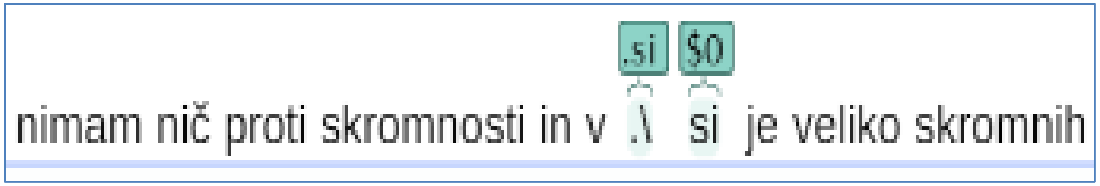
```
<WA-config>
 <verbose>true</verbose>
 <normalised_base>true</normalised_base>
 <null_char>$0</null_char>
 <glue_char/>
 <punct_tag>Z</punct_tag>
 <LayerDefinitions>
  <layer name="webanno.custom.Originals" type="token">
   <feature name="original"/><feature name="original2"/>
   <feature name="original3"/><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



Po dveh tednih -3,8. Jeee\ !\ 3 čunga lunga1\ :



nimam nič proti skromnosti in v .\ si je veliko skromnih

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