Gold-Standard Datasets for Annotation of Slovene Computer-Mediated Communication

Tomaž Erjavec, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, and Darja Fišer

RASLAN 2016
Karlova Studánka, December 2–4, 2016

Overview

Introduction

Introduction

- Language technologies need hand annotated datasets for training and evaluation
- There are various tools for Slovene, but only for standard langauge
- Computer Mediated Communication (CMC) differs from standard langauge:
  - no diacritics
  - missing spaces
  - non-standard use of punctuation
  - typos
  - phonetic spelling, slang, dialects
- Difficult to search in CMC corpora
- Tools for standard language (e.g. PoS tagging) do not work well on CMC

Janes project

- Janes: "Jezikoslovna analiza nestandardne slovenščine" (Linguistic Analysis of Non-Standard Slovene)
- Slovene basic research project 2014–2017
- Development of a corpus of Slovene CMC
- Performing linguistic analysis on it
- Developing robust tools and *hand-annotated gold-standard datasets for tool training and testing*

Janes corpus

- Current version is Janes 0.4
- 9 million texts, 200 million tokens:
  - 107 mt = tweets
  - 47 mt = forum posts
  - 34 mt = blog articles with user comments
  - 15 mt = user comments on news articles
  - 5 mt = Wikipedia talk pages
- Text metadata:
  - user, time of post, (gender, type of user)
  - sentiment
  - standardness (T1 – T3 + L1 – L3)

Levels of annotation

- Tokenisation, sentence segmentation: Python REs for non-standard langauge
  `:-]`, `:-PPPP`, `^_^`
- Word-standardisation: rediacritisation + CSMT normalisation
  `krizisce` → `križišče`; `jest`, `jst`, `jas`, `js` → `jaz`
- MSD tagging and lemmatisation: new CRF-based tools
- Tools work ok, but could be much better
- Needed hand-annotated datasets for these levels of annotation

Annotation campaign

Preparing the datasets

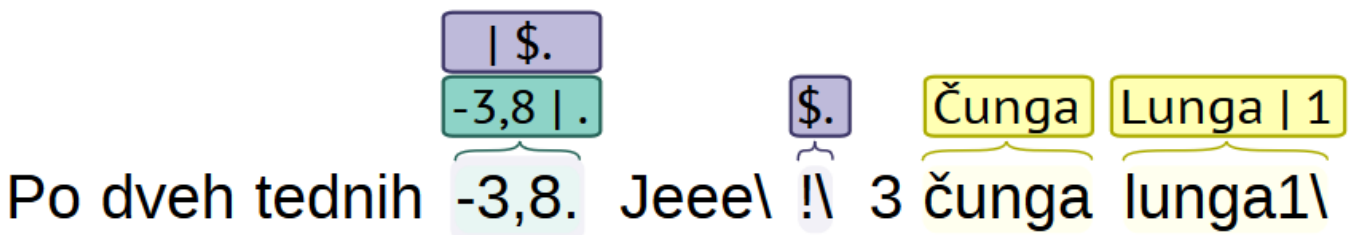
- Datasets sampled from Janes 0.4, (T3L3, T1L3, T3L1, T1L1)
- Kons1: normalisation of 4000 tweets
- Kons2: normalisation of 4000 forum posts and comments on blog posts and news articles
- Kons1-MSD: tagging and lemmatisation of Kons1 (preference to L3)
- Kons2-MSD: tagging and lemmatisation of Kons2 (preference to L3)
- All the texts were first automatically annotated, then imported to the annotation tool

Annotation guidelines

- Followed Guidelines for annotating standard and historical Slovene texts
- Much more complicated than expected:
  - non-standard words without a standard form (e.g. `orng`, `ornk`, `oreng`, `orenk` for `'very'`)
  - foreign language elements (e.g. `updateati`, `updajtati`, `updejtati`, `apdejtati` for `'to update'`)
  - proper names, abbreviations, non-standard use of cases and particles etc.
- A training and testing session was organised for the annotators (a team of cca. 10 students)

Annotation platform

- Annotation was performed in WebAnno
- Difficult to use for correcting tokenisation (multivalued features and special symbols)
- Each text annotated by two annotators and then curated by the team leader.



Format conversion

- Janes corpus is encoded in TEI P5
- WebAnno uses e.g. the tabular TSV format
- TEI2TSV (XSLT)
- TSV2TEI = merge operation:
  exported TSV + source TEI = TEI with corrected annotations

The datasets

Janes-Norm and Janes-Tag

- Janes-Norm (Kons1 and Kons2) = gold-standard dataset for the annotation of tokenisation, sentence segmentation and normalisation
- Janes-Tag is a subset of Janes-Norm (Kons1-MSD and Kons2-MSD) = gold-standard dataset for the annotation of MSDs and lemmas.
- The order of the texts in datasets was randomised

Encoding

```
<ab xml:id="janes.blog.publishwall.4264.3" type="blog" subtype="T1L3">
   <s>
      <w lemma="kaj" ana="#Rgp">Kaj</w><c> </c>
      <w lemma="biti" ana="#Va-r3s-y">ni</w><c> </c>
      <w lemma="ta" ana="#Pd-nsn">to</w><c> </c>
      <choice>
         <orig><w>tazadnje</w></orig>
         <reg>
            <w lemma="ta" ana="#Q">ta</w><c> </c>
            <w lemma="zadnji" ana="#Agpnsn">zadnje</w>
         </reg>
      </choice><c> </c>
      <choice>
         <orig><w>AAjevska</w></orig>
         <reg><w lemma="aa-jevski" ana="#Agpfsn">AA-jevska</w></reg>
      </choice><c> </c>
      <w lemma="molitev" ana="#Ncfsn">molitev</w>
      <pc ana="#Z">?</pc>
   </s>
</ab>
```

Size of datasets

- Janes-Norm:
  - 7,800 texts, 185.000 tokens, 144.000 words
  - split about equally to T3L3, T1L3, T3L1, T1L1
  - 27.3% words normalised, 42% of these normalised at the morphological and lexical levels, 5% of these split/joined
- Janes-Tag:
  - 3,000 texts, 75,000 tokens, 56,000 words
  - 80% are L3 texts

Conclusions

Summary

- Presented first manually annotated CMC datasets
- Lower levels: tokenisation, sentence segmentation, word-normalisation
- Higher levels: MSD tagging, lemmatisation
- More difficult than originally thought

Further work

- Deposit on CLARIN.SI repository
- Re-train the tools with the new datasets