The Slovenian ELTeC corpus

Tomaž Erjavec
Jožef Stefan Institute
Ljubljana, Slovenia

Start of the project

- The first COST partners from Slovenia were Katja Mihurko Poniž from University of Nova Gorica and Marko Juvan & Matija Ogrin (ZRC SAZU)
- I first heard about the project at the presentation of the COST action at the Slovenian conference for Language Technologies and Digital Humanities 2018
- I had experience in compiling TEI corpora of historical texts via the EU IMPACT project and a Google grant in which we made the IMP resources for historical Slovene and was happy to take over the technical aspects of the ELTeC-slv corpus
- The team also involves Andrejka Žejn (ZRC SAZU) and Miran Hladnik (Uni. of Ljubljana & Slovenian Wiki guru)

Selection of texts

Selection by M. Hladnik, M. Juvan, and K. Mihurko Poniž with her student, who made the Excel with:

- selection of the novels with source URL
- ELTeC required metadata
- & I made a table with author VIAF and Wikipedia link, also asked the national library to add missing authors to VIAF (1 still missing…)

Final tally of 100 novels:

- existing novels from the IMP corpus (65)
- existing novels from the Wikisource project "Slovene Literary Classics" (29)
- new novels added to Wikisource esp. for ELTeC (**5**)
- existing novel from the eZISS digital library (1)

Problems

- Older novels are rather short -> lack of novels in the "long" category
- By far the most novels written by men -> lack of novels by female authors
- Wikisource concentrates on key novels -> lack of non-canonical novels
- Difficult to get (digital) source for novels of missing categories (if they even exist), with an expensive and lengthy process to manually correct OCR

Processing for Level 1

Each source (IMP, Wikisource, eZISS) had a dedicated conversion pipeline but all include the following steps:

1. download and rename the novel (all are openly available on the Web);
2. convert source encoding to the ELTeC TEI
   - IMP, eZISS already in TEI -> simple XSLT script
   - Wikisource in Wikipedia and project specific Markdown mixed with direct HTML formatting -> Perl pre-process + TEI MD2TEI XSLT
3. add novel and author metadata from the tables.

NB: the complete pipeline is in the ELTeC-slv GitHub project Orig/ folder, so anybody can re-run it.

Processing for Level 2

1. Tokenisation and sentence segmentation: rule-based ReLDI tokeniser
2. Spelling modernisation of words: character-based statistical machine translation tool CSTMtiser (TM trained on goo300k, LM on the literary portion of Gigafida)
3. PoS tagging & lemmatisation: CLASSLA-StanfordNLP, a fork of StanfordNLP
4. Named entities: Janes-NER tool (PER, PER-DERIV, GEO, ORG, MISC)

State of the corpus

- Level 1: 100 novels
- Level 2: 100 novels annotated with SoA tools
- Corpus available on CLARIN.SI concordancers

i.e. the corpus is finished, and will not change anymore, except:

- possible adjustments to the Level 2 schema
- if anybody does ELTeC NER.

Future plans

- Andrejka Žejn & me are working on an extension of the ELTeC-slv corpus
- Currently collected 45 novels (cca. 1 million words), will be converted to ELTeC & mounted on noSketch Engine @ CLARIN.SI
- Beg. of 2020 Andrejka was on a STSM at the Polish Academy, Institute of Polish language (Maciej Eder) where she used Stylo to perform analyses on a previous version of this corpus
- Beg. of 2021 national research project submission: Application of computer aided tools and methods on the Slovenian ELTeC corpus.