Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene

Tomaž Erjavec

Dept. of Knowledge Technologies,
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Motivation & Background

Goal:

- modernise, PoS tag, and lemmatise historical texts

Applications:

- full-text search, improved comprehension, further analysis

Slovene language:

- highly inflecting (difficult tagging and lemmatisation)
- late standardisation of written language
- alphabet change in ~1850 (*nezhiſte* → *nečiste*)

Basic approach

- Modernise words, then use existing models for modern language for tagging and lemmatisation
- Program ToTaLe:
  - Tokenise: mlToken, hand-written rules
  - Tag: TnT, trained on JOS corpora
  - Lemmatise: CLOG, trained on JOS-derived lexicon
- New steps for historical language:
  - Re-tokenise (*po noči → ponoči; nemore → ne more)*
  - Transcribe (nezhiſte → nečiste; *kakšniga → kakšnega)*

Tool chain

Perl program ToTrTaLe comprises the following modules:

1. Extract processing chunks from source TEI
2. Tokenise, taking into account compounds/splits
3. Extract text to be annotated
4. Transcribe to modern word-forms
5. Tag for Part-of-Speech
6. Lemmatise
7. Produce TEI output

Transcription

- Words are modernised in spelling only
- Use fixed lexicon for high-frequency and unpredictable words
  - Lexicon also blocks use of inappropriate modern day words (*sim* → *sem,* not *SIM*)
- For the rest, use transcription patterns:
  - *iga@ → ega@ (kakšniga → kakšnega)*
  - Vaam (Variant aware approximate matching) FSA library
  - use all patterns to try and match historical word against modern day lexicon
  - problems of ambiguity
  - currently hand-crafted rule-set

Processing TEI

- ToTrTaLe takes TEI P5 encoded document as input
- Parameter to set top level element for processing
- Processes only contiguous text bearing elements
- Outputs TEI P5 document:
  - original annotation
  - added in-place token level annotations
  - in case of conflicts splits source elements
- Last stage: TEI validity check (Relax NG schema)
- Solution caters for the type of documents present in our corpus
- Not general enough

Output example

```
<w subtype="lexicon" nform="nekiga" mform="nekega"
   lemma="nek" ctag="Pi-msg">Nekiga</w><c> </c>
<w subtype="pattern" pattern="[ega@←iga@]" mform="bogatega"
   lemma="bogat" ctag="Agpmsg">bogatiga</w><c> </c>
<w lemma="knez" ctag="Npmsg">kneza</w><c> </c>
<lb/>
<pb n="93" facs="#FPG00012.097" xml:id="pb.97"/>
<w type="multiw" subtype="pattern" pattern="[@v←@v_]"
   mform="vmes" lemma="vmes" ctag="Rgp"
   n="mw_jeGx2">v</w><c> </c>
<w type="multiw" subtype="pattern" pattern="[@v←@v_]"
   mform="vmes" lemma="vmes" ctag="Rgp"
   n="mw_jeGx2">mes</w><c> </c>
<w type="split" mform="ne_more" lemma="ne_moči"
   ctag="Q_Vmpr3s">nemore</w>
<gap/>
```

Conclusions

- ToTrTaLe performs basic annotation on historical texts
- Current work: producing a hand-annotated corpus
- Problem of transcription ambiguity
- Need a more general approach to TEI encoding
- Develop a better methodology of transcription pattern development and use:
  - automatic induction of transcriptions patters
  - pattern profiles, depending on age of text

Presented work was support by projects EU IMPACT „Improving Access to Text“ and Google award „Developing language models for historical Slovene“