Morphosyntactic Annotation of Historical Texts

The Making of the Baroque Corpus of Polish

W. Kieraś, D. Komosińska, E. Modrzejewski & M. Woliński

Institute of Computer Science & Institute of Polish Language
Polish Academy of Sciences

Abstract

We present some technical issues concerning processing 17th & 18th century texts for the purpose of building a corpus of that period. We describe a chain of procedures leading from transliterated source texts to morphological annotation of text samples that was implemented for building the Baroque Corpus of Polish. The procedure consists of: automatic transcription from original spelling to modern one, morphological analysis including the construction of an inflectional dataset for Baroque Polish and a tool for manual morphosyntactic annotation. The toolchain is being used to create a small manually validated subcorpus, which will serve as training data for a stochastic tagger. Then a larger corpus will be annotated automatically and made available via the Poliqarp corpus search tool.

Automatic transcription

• The source texts of the Baroque Corpus of Polish divide into three types of editions: (1) original, (2) 19th century, and (3) contemporary.
• Types (1), (2) require automatic transcription to normalize the texts to make them better suited for morphological analysis.
• The converter contains nearly 4,000 substitution rules, based on regular expressions and the context of the sequence of characters’ appearance.
• To check and correct the transcription, corpus annotators have a possibility to work on both transliterated and transcribed versions:

<table>
<thead>
<tr>
<th>token occurrences</th>
<th>transliteration</th>
<th>transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>12,832,214</td>
<td>646,410</td>
<td>476,733</td>
</tr>
</tbody>
</table>

unrecognized token types: 5.4% of tokens

Number of tokens in the two representations of the corpus

Morphological Analysis

• Automatic morphological analysis is performed using Morfeusz analyzer with a dictionary based on the following sources of data:
  – inflectional information from the Electronic Dictionary of 17th & 18th century Polish (e-SXVII, http://sxvii.pl/)
  – contemporary data of Grammatical Dictionary of Polish (SGJP) modified (“aged”) to fit into the tagset for Baroque morphosyntax.
• Some automatically obtained extensions of inflectional paradigms from e-SXVII was also used. The extension procedure enriched the data set by over 210,000 new forms.

Example of partial reconstruction of inflectional paradigm from e-SXVII

<table>
<thead>
<tr>
<th>stem</th>
<th>“…ings”</th>
<th>“…ed”</th>
<th>“…-ed”</th>
</tr>
</thead>
</table>
| celnictwo | subst|sg|gen|n
| celnictwem | subst|sing|inst|n
| celnictwile | subst|sing|loc|n
| bognacto | subst|sg|gen|n
| bognactwem | subst|sing|inst|n
| bognactwile | subst|sing|loc|n
| bognactwa | subst|pl|gen|n
| bognactwem | subst|pl|inst|n
| bognactwile | subst|pl|loc|n
| bognactwa | subst|pl|pin|n
| bognactwem | subst|pl|pin|n
| bognactwile | subst|pl|pin|n
| bognactwa | subst|pl|pin|n
| bognactwem | subst|pl|pin|n
| bognactwile | subst|pl|pin|n

<table>
<thead>
<tr>
<th>endings + tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>“…ing”</td>
</tr>
</tbody>
</table>

Conflicts to be resolved as seen by an adjudicator

Manual Annotation

• A corpus of 500,000 tokens of BCP will be manually annotated (currently ca. 150,000).
• All tokens marked as punctuation, foreign elements or structural markers were excluded for the purpose of the evaluation.
• The percentages of manual changes introduced by annotators are as follows:
  – modification of transcription – 2.66% of tokens, of tokenization – 1.59%,
  – morphological interpretation modified by the annotator – 8.6%.
• Each text sample is annotated independently by two annotators. In case of conflict the adjudicator has to resolve it by selecting one of the existing annotations or providing a new one.
• Currently annotators generate conflicts on 12.6% of tokens.

Conclusions & further work

• The BCP is a work in progress since the manual annotation of the subcorpus is ongoing.
• All components involved in processing text samples are being constantly enhanced according to the feedback from annotators and adjudicators, but it seems that the tools are already quite stable.
• When the manually annotated corpus is ready, a stochastic tagger will be trained on the annotations and the rest of the BCP corpus will be automatically annotated. We also plan to check whether machine learning techniques can be used to train a better tagger based on transcription data manually corrected by the annotators.
• The presented toolchain will also be used to prepare a similarly organized corpus of 19th century Polish texts.

Acknowledgements

The work being reported was co-financed by a National Science Centre, Poland grant DEC-2014/15/B/HS2/03119 and a Ministry of Science and Higher Education National Programme for the Development of Humanities grant 0036/NPRH2/H11/81/2012.