Polish Sejm Corpus – a pilot project

The utterances of Members of Polish Sejm (the lower chamber of the Polish Parliament) are being recorded on video since 1993; for more than 15 years they have been made available online in the form of unannotated PDF transcripts. This document puts forward an idea of creation of a corpus of the annotated text version of the utterances, ready to be further extended to cover alignment with their video sources.

Currently the internal database of utterances is barely a stenographic record of the Sejm sessions, lacking formal representation. The proposed pilot actions intend to convert them into a fully-fledged text corpus, with the plan of including multimedia content at a later stage. From the very beginning the project will apply standards and well-tested formalisms used by large-scale corpus projects, such as 1-bilion National Corpus of Polish (Pol. Narodowy Korpus Języka Polskiego, NKJP, see http://nkjp.pl).

The project will start with definition of corpus format and metadata, methods of record segmentation and level of formal linguistic description. The guidelines of the Text Encoding Initiative will be followed along with the newly proposed ISO standards (MAF, LAF and SynAF). Data will be acquired directly from the shorthand notes database, converted into the corpus output format and automatically amended with basic metainformation (in the form of TEI headers) and internal structure. According to modern practices of maintaining sustainability and interoperability, the wordform segments will be described with stand-off mechanisms of annotation on at least morphosyntactical level. Indexing and search features will be provided by the tools developed in the course of NKJP project (mainly Poliqarp, an efficient search engine and a concordancer).

The pilot project is planned to be further extended in various directions – by incorporating a similar all-cadency database of utterances of the upper house (the Senate), creating a multimedia corpus with aligned text and video or implementing automated tools to create a live version of the corpus, constantly populated with new, automatically annotated data as they are made available after consecutive Sejm sessions. The text data are also intended to be included into NKJP as a subcorpus.

Results of the project will be publicly available online; all implemented tools and resources are planned to be open source. The pilot time frame is 6 months while preparation of converted and aligned videos is highly dependent on additional funding, mainly related to increased storage and processing capacities.

The presented approach aims at providing an interesting, standards-compliant, specialized resource, valuable for researchers dealing with usage of Polish. The planned audio/video version of the aligned corpus might facilitate implementation and training of speech recognition systems, still underdeveloped for Slavic languages.