Old Lithuanian Reference Corpus

Jolanta Gelumbeckaitë
Institut für Empirische Sprachwissenschaft
Goethe-Universität Frankfurt am Main
eMail: gelumbeckaite@em.uni-frankfurt.de

The Old Lithuanian Reference Corpus (Germ. Referenzcorpus Altlitauisch), a comprehensive, deeply annotated reference corpus of Old Lithuanian, is being developed in cooperation between the Goethe-University of Frankfurt am Main (Germany), the Institute of Lithuanian Language (Vilnius, Lithuania), and the University of Pisa (Italy). Its ultimate goal is to create a qualitative multilevel electronic retrieval engine for multilateral linguistic research of Old Lithuanian which will allow for reliable results of diachronic Lithuanian language studies. Furthermore, it has to finally enable the implementation of the two biggest desiderata of Baltic linguistics, the Old Lithuanian grammar and the historic dictionary of Lithuanian.

Old Lithuanian covers a period of ca. 300 years, from the 16th to the 19th century. Altogether the Corpus will consist of over 10 million text words. Because of this huge amount and the complex, multilayered structures that are needed for such a diachronic corpus, it seems reasonable to start with a smaller test corpus. Ten Old Lithuanian texts comprising ca. 350 000 tokens were chosen for the test corpus SLIEKKAS (Lith. Senosios lietuvių kalbos korpusas). The main endeavours of it are the following: 1) digitisation of Lithuanian lexica, grammars and word indices, which are relevant for a further lemmatising and glossing of the texts, 2) setting up of a basic-XML-structure, which is relevant for further annotation, and 3) securing a high philological standard as well as a textological and a palaeographic annotation of selected Old Lithuanian texts. Since most of the Old Lithuanian texts are translations from Latin, German or Polish sources, the source texts (in the case of the test corpus ca. 190 000 text words) will be accumulated and annotated in the same way as the Lithuanian ones. This will allow for the alignment of the Old Lithuanian texts with their sources. Furthermore, Old Lithuanian texts of the same genre will be aligned with each other in order to allow for an assessment of possible mutual influences within one and the same genre, as well as across genres. The envisaged annotation scheme of the Corpus embraces the following structural features:

1) a thorough linguistic and textological annotation, including header information, lemmatisation, grammatical information (Part of Speech-Tagging, morphological and basic syntactical information), glossation (in Modern Lithuanian, English and possibly more languages), information about the text structure, palaeographic and textological information. The main purpose is to develop a semi-automatic technique that allows for establishing the core word form in a historical lexicon (lemmatisation), its glossation in Modern Lithuanian and the determination of its actual meanings in a given Old Lithuanian text. More than 50% of the Old Lithuanian word forms are ambiguous as regards their morphological status. A morphological annotation consists of the unalterable morphological categories of the lemmata as well as of the actual word forms in a given text, and of the flexional morphological characteristics of the latter. Moreover a supplementary annotation level for multi-word-items is necessary. Additional annotation levels are required for a taxonomy of both explicit and implicit quotations.

2) a multi-level architecture of the annotations;

3) multi-modality of the corpus through the alignment of the texts with facsimile reproductions of the originals. The Old Lithuanian texts will be aligned to their facsimilia (manuscripts resp. prints) on a word level.

The Old Lithuanian Reference Corpus will be designed to provide an innovative scientific resource for historical and comparative linguistics as well as literary, religious and cultural studies concerning the Baltic countries.