Freitag
10.03.2017
12:30 – 13:00
B3 1, 0.12

A Diacronic Corpus for Romanian (RoDia)

Cătălina Mărănduc[1], Cenel-Augusto Perez[1], Ludmila Malahov[2], Alexandru Colesnicov[2]
[1]*Al. I. Cuza University,* [2]*Academy of Sciences of Moldova*
catalina.maranduc@info.uaic.ro, augusto.perez@info.uaic.ro, lmalahov@gmail.com, acolesnicov@gmx.com

AG4

This paper discusses the evolution of a Romanian corpus of the Dependency Treebank type, built at the Al. I.Cuza University of Iași. The corpus has rich annotation and balanced structure. Having the intention to participate at the PROIEL project, which aligns the oldest Latin, Greek, Slavonic and Armenian New Testaments, we chose to annotate the first printed Romanian NT at Alba Iulia (1648). The print of the book contains a lot of peculiarities difficult to process, described in the paper. We began by the automated processing of its first 2,000 sentences in classical syntactic annotation over the previous morphological annotation. We applied the tools for Contemporary Romanian to a fragment in modern Latin script. But the first edition is written in Old Cyrillic alphabet, used to print old books in Romania and Moldova (where the same language is spoken). A first fragment of the Alba Iulia NT has been transformed in editable Cyrillic text by an OCR program built by the Computer Scientists in Republic of Moldova. The editable text in the Cyrillic script has been checked by the computational linguists from Iași (Romania), comparing it with the printed old book, then it has been wrapped in the checked XML format, and the form with Latin letters, obtained in the second step of the OCR processing, has been compared with the second edition of the book. The entirely annotated and checked book will be used for extracting an old lexicon to be introduced in a POS-tagger able to annotate Old Romanian written with Latin or Cyrillic letters.

References: • Davies M. (2010) Creating Useful Historical Corpora: in *Diacronía de las Lenguas Iberoromances*: 137-166. • Dipper, S. Faulstich, L. Leser, U. Ludeling A. (2010) Challenges in Modelling a Richly Annotated Diachronic Corpus of German. *Proceedings of LREC*. • Haug, D. T. T., Jøhndal, M. L. (2008) Creating a Parallel Treebank of the Old Indo-European Bible Translations. *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data* 27-34.