RoDia: Fostering Language Diversity in One Corpus

Victoria Bobicev   Cătălina Mărănduc

We are working on a corpus with rich morphological, syntactic and partially semantic annotation. Its main characteristics are the large variety of non-standard texts and several types of annotation.

The creation of this corpus pursues several objectives:

(1) a better coverage of linguistic diversity of Romanian language;
(2) diachronic analysis of Romanian;
(3) creation of a gold standard annotation for various types of Romanian texts which permits:
(4) creation of robust machine learning models for various types of annotation.