Ivan Šimko
Slavisches Seminar, University of Zurich

Browsing the Digital Damaskini

This presentation describes a new diachronic corpus of Balkan Slavic literature. While there exists a significant amount of resources for the study of Church Slavonic and Middle Bulgarian literature, the texts from the era of Ottoman rule are underrepresented and mostly available only as raw transcripts or scans, unsuitable for quantitative studies. Given the conservative nature of Church Slavonic literature, these texts are also hardly useful for such studies, as they only scarcely reflect language shifts in the vernacular.

From the 15th century onward, the vernacular is better reflected in the literature. One of the reasons is the competition between the local Orthodox clergy with Catholic and Muslim missionaries. The so-called damaskini, collections of homilies and stories in a vernacular-based language, were meant to be accessible to a broader audience, to foster the awareness of confessional (later also national) identity in the population. Another reason were the limited resources of the Slavic clergy, what led to the decay of the education system. Thus even Church Slavonic texts start to show more influences of the vernacular. Thus, our corpus focuses on this era while collecting the material, on which we can observe the development of linguistic features.

Our corpus contains short (under 5000 tokens) samples of texts from this era in two formats: Unicode text tables and CoNLL-U, both of which are available at the repository. The texts are selected to represent a wide array of dialects and periods. Some of the samples have been chosen as a parallel corpus: they are multiple versions of the same text, which can be compared while minimizing the influence of genre and contents. Using the morphological and syntactic annotation, various aspects of Balkan-specific language shifts can be observed. The annotation scheme is customized, in order to reflect both archaic and innovative features.

The presentation will demonstrate three methods of work with the corpus. First, we can measure the frequency of individual features which were spreading (e.g. use of postposed articles) or vanishing (e.g. synthetic infinitive forms) in the given period across the texts and periods/areas they represent. Another method is to compare the texts according to the differences by a given set of features. Third, the texts can be processed with scans of originals for a better overview and used for closer philologic research or didactic purposes.