Mira Načeva-Marvanová (J. E. Purkyně University)

Electronic Corpora of Slavic Micro-languages at Their Threshold – the State of the Art and its Further Prospects

After the fall of the iron curtain, since the middle of the 1990s, new written and spoken language digital corpora have been developed for almost all of the standard Slavic literary languages. One should add that this quite appropriate, useful, and above all, extremely beneficial approach is the best method of language documentation of micro-languages as well. It has also been applied to a certain extent to some other sociolinguistic varieties of language.

This presentation is focused on the already available current Slavic micro-language corpora, such as: (1) the two monolingual "large size" written text corpora of Upper Sorbian – HOTKO (of 32 million words) and of Lower Sorbian – DOTKO (of 12 million words) - both these corpora are hosted by the Czech National Corpus using its corpus managers and interface; (2) the series of five oral digital multimedia corpora of: Burgenland Croatian, colloquial Upper Sorbian, Molise Slavic (Na-našu) in Southern Italy, Nashta (Liti, Northern Greece) and the language of Edesa (Northern Greece, called by its researcher Bulgaro-Macedonian) – all these corpora are a part of the French-German research Euroslav 2010 and LaCiTo (France); (3) the WORTSCHATZ corpora of Kashubian, Lower Sorbian, Upper Sorbian, Silesian and Transcarpathian Rusyn (a project of the University of Leipzig); (4) the digital archive of Lower Sorbian language data at the DoBeS (Documentation Bedrohter Sprachen = Documentation of Endangered Languages); (5) the Transdanubian electronic corpus of Bulgarian dialects (from 38 locations) in Southern Romania (of the University of Calgary and University of Sofia).

This presentation analyzes the options of the presented corpora, according to their type (corpora of written or spoken language; corpora of literary or non-literary, regional language; monolingual or multilingual corpora) and according to applied linguistic technologies. Not of less importance is the selection and the extent of the corpora attributes, sorting options, parsing and levels of annotation, transcription solutions (e.g. in oral corpora), meta-lingual information, options for collocation and statistical analyses. The application of most of these attributes enables and facilitates the further use of Slavic micro-language corpora for linguistic, sociolinguistic, common cultural and educational purposes.

We can claim that the corpus building of Slavic micro-languages, though in its infancy, is already a reality. One can expect, in the conditions of the current digital world, that the methods for documentation of Slavic micro-languages will proceed in this direction. No doubt EU social politics and the social environment stimulates and supports this approach. However, considering the strategy of the Slavic corpus linguistics and its achievements up to now, it seems appropriate to suggest that it is at least equally important for most of the Slavic micro-languages to seek and find the will to establish mutual internal, as well as trans-border and international, understanding and cooperation. One can only ask whether the creation and funding of the Lithuanian-Latvian-Latgalian corpora project (within the Latvia-Lithuania Cross-Border Cooperation Programme 2007-2013) should not be an inspiration.