

The aim of the project is to adapt existing technologies and develop completely new procedures for the use of artificial intelligence in the automatic processing of natural spoken language in its regionally diverse forms, specifically on the model of Czech dialects. It will be the first project of its kind – automatic speech recognition has so far been applied exclusively to speech with a majority of standard (canonical) language. Such a unique project will significantly streamline work with authentic sound material, the manual processing of which is (not only for linguists) nowadays extremely time-consuming. At the same time, it will enable easier and more effective documentation of the Czech language dialects and an interactive presentation of this disappearing linguistic heritage. The goal will be achieved by (1) creating an innovative Methodology for the transfer of structured knowledge from dialectology to machine learning, (2) intensive work on data sets, especially building a Dialect Speech Database for the professional public, whose fulfillment by representative audio recordings will be ensured by one of the project partners (an owner of the largest archive of this kind in the Czech Republic), (3) creation of several software packages, namely The dialect detector based on audio recording, Speech recognizer adapted to generate dialectological transcription from audio recordings, and Multilingual recognizer of West Slavic languages for generating folklore transcription from audio recordings, (4) creation of three specialized Maps of dialect differential phonetic phenomena with interactive and multimedia elements, and especially (5) the creation of the application Language Memory of the Regions of the Czech Republic. It will be open to the general public for community contributions, map browsing and listening to dialect recordings. The project publications will include a book and a number of journal and conference papers.