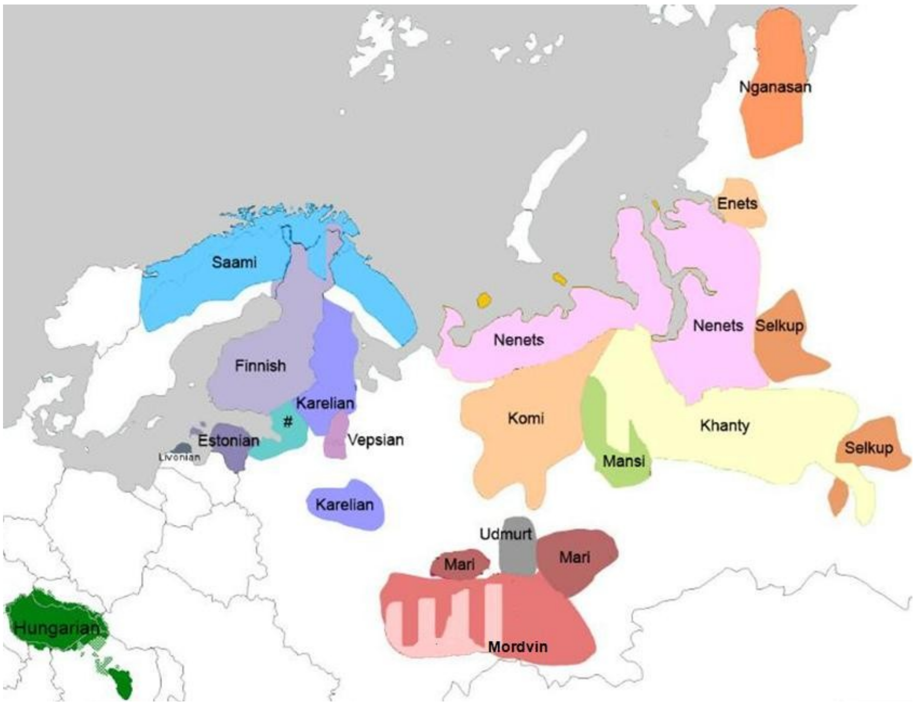
Endangered Uralic Languages and Language Technologies

Gábor Prószéky
MorphoLogic & Pázmány University, Budapest, Hungary
proszeky@morphologic.hu

Language tools and resources for analysis of less-elaborated languages are in the focus of our workshop. There are still research tracks which still do not sufficiently and effectively exploit language technology solutions, and there are many languages for which the available tools and resources still have to be developed to serve as a basis of further applications.

The presentation introduces a set of morphological tools for small and endangered Uralic languages. Various Hungarian research groups specialized in Finno-Ugric linguistics and a Hungarian language technology company (MorphoLogic) have initiated a project with the goal of producing annotated electronic corpora and computational morphological tools for small Uralic languages, like Mordvin, Udmurt (Votyak), Komi (Zyryan), Mansi (Vogul), Khanty (Ostyak), Nenets (Yurak) and Nganasan (Tavgi). Altogether around a dozen Uralic languages totaling some 3.3 million live as scattered minorities in Russia, as shown by the map below:



The morphologies of these languages are complex enough, thus the implementation of the morphological tools was a real challenge. The subprojects concerning the individual languages slightly differed, depending on the special problems these languages raise (how precisely the languages have been described so far, whether there is a standard dialect, what kinds of texts are available, etc.). In the project, we used the morphological analyzer engine called Humor ('High speed Unification MORphology') developed at MorphoLogic, which was first successfully applied to another Finno-Ugric language, Hungarian. We supplemented the analyzer with two additional tools: a lemmatizer and a morphological generator. Creating analyzers for the Samoyed languages involved in the project turned out to be a great challenge. Nganasan from the Northern Samoyed branch is a language on the verge of extinction (the number of native speakers is below 500 by now, most of them are middle-aged or old), so its documentation is an urgent scientific task. Nganasan morphology and especially its phonology is very complex and the available linguistic data and their linguistic descriptions proved to be incomplete and partly contradictory. Thus, using the Humor formalism, which we successfully applied to other languages involved in the project, was not to be feasible in the case of one of the chosen languages, Nganasan. The Humor formalism uses an 'item-and-arrangement' model of morphology where feature-based allomorph adjacency restrictions are the primary device for constraining word structure. Gradation in Nganasan is difficult to formalize as a set of allomorph adjacency restrictions because the segments involved in determining the outcome of the process may belong to non-adjacent morphemes. For Nganasan, we used therefore another tool (xfst of Xerox), mainly because gradation is just a small part of the complicated system of dozens of interacting productive and lexicalized morpho-phonological and phonological alternations.

Besides the annotated corpora and the morphological analyzers, a website was also developed where all of the tools described above are available for a wider public.

Acknowledgments

The projects have been funded by the Hungarian Scientific Research Fund (OTKA) and the National Research and Development Programme (NKFP): A Complex Uralic linguistic database (NKFP-5/135/01), Linguistic databases for Permic languages (OTKA T 048309), Development of a morphological analyzer for Nganasan (OTKA K 60807), Ob Ugric morphological analyzers and corpora (OTKA 71707).