The first morphological analyzer for Nganasan

Attila Novák

MorphoLogic Ltd. 1126 Budapest Orbánhegyi út 5.,
novak@morphologic.hu

This article presents a morphological analyzer for *Nganasan*, a small language belonging to the Northern Samoyed branch of the Finno-Ugrian language family. Creating this analyzer is part of a project the aim of which is to create annotated corpora and other electronically available linguistic resources for a number of small members of the Uralic language family. The project was initiated by Various Hungarian research groups specialized in Finno-Ugric linguistics and a Hungarian language technology company, MorphoLogic.[1]

Ngansan turned out to be especially interesting among the languages involved in the project. On the one hand, it 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. On the other hand, its morphology and especially its phonology is so complex that the implementation of the analyzer turned out to be a real challenge. Using the formalism of the morphological analyzer engine called *Humor*, which we successfully applied to other languages involved in the project, turned out not to be feasible in the case of Nganasan. Finally, we used the regular relation calculus based toolset of Xerox Corp. (namely, *xfst*) to create the analyzer.

In Nganasan, a quite morphology-independent surface phonology plays an important role in shaping the form of words. The very productive gradation processes are governed by a complicated set of constraints on surface syllable structure. Gradation is a systematic alternation of obstruents in syllable onsets governed in the case of Nganasan by various factors such as vowel length and the presence of a coda in the preceding syllable, the presence of a coda in the current syllable, and whether the syllable is in an odd or even position within the word. The syllabification of certain segments or clusters is exceptional and there are also apparent lexical exceptions to the general gradation patterns.

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. Moreover, gradation is just a small part of the complicated system of dozens of interacting productive and lexicalized morphophonological and phonological alternations. The Xerox finite-state calculus (specifically *xfst*), which fortunately became freely available and easily accessible for non-commercial purposes in 2003, proved to be more easily applicable to this language.

[1] Complex Uralic Lingustic Database, NKFP 5/135/2001.