What is good Humor like?

Attila Novák

Morphologic Ltd., Budapest
novak@morphologic.hu

Keywords: morphological analysis, linguistic database

Computational processing of highly inflectional languages, like Hungarian, relies upon an efficient morphological analysis. The program most commonly used for the morphological analysis of Hungarian texts is the analyzer called Humor, developed by a Hungarian language technology company, MorphoLogic. Various versions of Humor have been in use for over a decade now. Although the program itself proved to be an efficient tool, the original database format turned out to be problematic, because it was hard to create and maintain. This paper describes how this problem was solved.

The Humor analyzer analyses the input word as a sequence of morphs. It is segmented into parts which have a surface form, a lexical form and a category label. While the program performs a search on the input word form for possible analyses, two kinds of checks are performed at every step: it checks the local compatibility between adjacent morphs and it examines whether the morphemes in the analysis instantiate a possible word construction in the given language.

The operations that the analyzer uses when analyzing the input word must be very simple so that processing can be efficient. This requires that the data structures it uses contain much redundant data (so that they do not have to be calculated on the fly during analysis). The most important problem with the Humor analyzer was that MorphoLogic had no tools for creating and maintaining these redundant data structures. The data structures optimized for efficient manipulation by the analyzer were hardly readable for humans, and they were very hard to modify in a consistent way. This resulted in many errors and inconsistencies in the descriptions, which were very difficult to find and correct.

To solve this problem an environment was created which facilitates the creation of the database. In the new environment, the linguist has to create a high level human readable description which contains no redundant information and which is thus easy to keep consistent. This high level representation makes the maintenance of the lexicon very easy.

This representation is transformed by the system in a consistent way to a redundant, but still readable description using rules defined by the creator of the database. The rules describe allomorphy patterns and implicational relations between morphological properties. At this level of representation, it is easy to catch errors in the rule system. In the next step, the low level representation used by the analyzer is created.

Using the development environment, a completely new version of the Hungarian analyzer was created, which contains less errors and is much easier to maintain than the previous one. A project is under way in which morphological analyzers for various Finno-Ugric and other Uralic languages are created using the system.