LiLe project: Database as 'dynamic corpus'

Zoltán Bódis, Judit Kleiber, Éva Szilágyi, Anita Viszket

Department of Linguistics, University of Pécs

Our research team has been engaged in developing a linguistic lexicon in the form of an MS-SQL database.

The technology provides the opportunity to store lexical units (morphemes) as well as applying underlying representations, and also using the well-known features of unification morphology; in a dynamically expandable structure. In our system not only the phonological, morphological, syntactic and semantic features of a morpheme are stored as records, but also the rules operating within or between the lexical items. Consequently, the set of rules can be dynamically expanded as well. The theory behind the definition-method of the rules is GASG, a totally lexicalist grammar by Gábor Alberti.

The structure of rules for describing the certain language is determinated by the describers, and the grammars are linked through a semantic representation. So in describing lexical elements only the semantic features are common (universal), other features are freely formed; in this way we can provide useable lexicon for any grammatical model.

By using our database, we can build up a corpus which is called 'dynamic', because it doesn't contain existing (ever existed) wordforms, but deducted elements and rules, so the possible words, expressions or even sentences of the given state of language can be generated – consequently, it models our competency.

This supports several purposes, like developing computational linguistic applications at our department, and we also wish to support teaching Hungarian language in public or higher education or as a foreign language with our lexicon as a teaching device.

Due to the structure of the system, our program not only decides between correct and incorrect morpheme strings, but it is able to name the rules used as the basis of the decision, and that is a useful help in langage teaching or developing language consciousness; besides, non-native speakers lacking competency can be supplied or helped out. Even, the generating algorhythm can operate on selected set of elements of the database (morphemes or rules) and this gives an opportunity to demonstrate or practice certain phenomenon of the language in an interesting way, for example, by switching rules on and off.

The current version of the program is developed in an object-oriented environment (Delphi), because we have found this to be the easiest way of building user-friendly interfaces but we plan to develop web-based surfaces as well, applying MS-SQL's built-in procedures which make possible to retrieve data in XML-format (which is generally used for data-storing in corpus linguistics).