MORPHO-ASSISTANT: The Proper Treatment of Morphological Knowledge

Kiril SIMOV, Galia ANGELOVA, Elena PASKALEVA
Linguistic Modelling Laboratory, Center for Informatics
and Computer Technology, Bulgarian Academy of Sciences
Acad. G. Bonchev Str. 25a, 1113 Sofia, Bulgaria

1 Computer Morphology of Slavonic Languages: State of the Art

One of the main features of Slavonic languages is their complex morphological structure. Unfortunately, current industrial software in this field (e.g.morphological components of spell checkers) supports only partial language models (and reflects as a rule the more or less primitive level of the morphological knowledge of its developers). In the authors' opinion, the time has come to develop advanced software products with the following capabilities:

1. An exhaustive set of linguistic data. In this way the kernel of the products could serve as a **computer normative morphology** and could even be built into professional laboratory systems developed at Research Centers and Universities;
2. Support of large and relatively complete machine-readable dictionaries. In this way the user would only have to add some specialized terms - if needed.

The development of industrial products based on exaustive morphological models will stop the stream of linguistically incomplete implementations.

2 A Brief Description of Bulgarian Morphology

Like the other Slavonic languages, Bulgarian is a highly inflexional language – one lexeme produces an average of 10 different wordforms; in the discussed system there exist 58 types of alternation and 102 inflexional types (46 of them concern nouns). The verb paradigm is especially rich – up to 224 forms (synthetic and analytic). The numerous defects of the paradigm create special difficulties: for example, perfective verbs reduce their paradigm by either 10, or 19 members (depending on their transitivity). The numerous non-canonical defects, determined by the lexico-semantic characteristics of the individual lexicon unit can only be described explicitly in the dictionary. The description of Bulgarian morphology becomes even more complicated because of the existence of boundary phenomena between word inflexion and derivation which are treated differently in various descriptions depending on the accepted volume of the paradigm. Such phenomena are the formation of participles in the verb paradigm, the formation of "semianalytical" comparison degrees of adjectives and the formation of aspectual pairs (grammaticalized and lexicalized).

In our opinion, a computer model of such a complex morphology can be elaborated only by qualified linguists, as far as the architecture of morphological knowledge is concerned, and it should work on a sufficiently large lexicon base. Since Bulgarian lexicography does not possess its own Zalizniak [4] so far, the morphological basis of the system is founded on all avalaible normative reference books on morphology (orthographic dictionnaries, normative grammars and works on Bulgarian inflexion). We hope that the MORPHO-ASSISTANT system will serve as a basis for future implementations and research.

3 A Brief Description of MORPHO-ASSISTANT

MORPHO-ASSISTANT is based on a Bulgarian morphology model of a classificational type where the morphological data are organized in three main sets (stem lexicon, endings lexicon and stem alternation patterns). The endings lexicon can be rebuilt as a list of inflexional types as well, where every inflexional type is a list of letter values of the inflexions of each member of the paradigm. The processing of Bulgarian word-inflexion is performed with the help of two basic operations - concatenation of the morphological and the lexical element (in inflexion) and eventual transformation of substrings within the lexical element (in alternation).

MORPHO-ASSISTANT provides five main functions:

1. Analysis (recognition) of an arbitrary wordform (certain types of mistakes, due to incorrect concatenation and/or transformation of the letter substrings being processed; hypotheses about the source of incompleteness of user's knowledge are being built when the causes of the mistakes are analysed).

2. Synthesis of an arbitrary form or a group of forms depending on parameters given by the user.

3. An information-retrieval system in the area of Bulgarian morphology. In fact it reflects the classificational principles which served as a basis for the creation of the morphological model.

4. Support of user dictionaries.

5. Test-creation facility. This function of the package can assist teachers and professors in devising and distributing drills in courses of second-language learning. This function of MORPHO-ASSISTANT is based on the kernel of the system realizing analysis and synthesis of arbitrary forms.

4 MORPHO-ASSISTANT - A Second Language Learning System

Once such an adequate and complete model of the morphology is developed and a big basic dictionary is filled, one is tempted to model the acquisition and the usage of this knowledge for the purposes of teaching. It is on these principles that the teaching part of MORPHO-ASSISTENT is based. It works in two modes:

- mode 1: the user of the system is a teacher preparing a variety of tests for his students learning Bulgarian and Russian;

- mode 2: the user of the system is a student performing ready test elements.

We will consider here the first mode.

The user is provided with two basic tools: the complete computer model of morphology realized by the first two basic functions of the system MORPHO-ASSISTANT (organized in an already determined part of knowledge) and special test-generation oriented tools which use undetermined knowledge in the designing. The former provide the basic building material and the latter provide the rules for its assembly when different types of tests are built according to its conception of the acquisition and testing of the morphological knowledge. Such are for example:

- the simplest classical tasks on morphological analysis or synthesis of a given form or a string of forms (which can be a sentence as well). This type of drills examines the student's ability to analyse or synthesize wordforms comparing directly the results with those of the same operation, realized by the first two functions of MORPHO-ASSISTANT. For the testing of knowledge (in mode 2, which will not be discussed here), also used are the created diagnostic tools for pointing out the types of mistakes when wordforms are generated incorrectly (see above III). Thus the computer has actually accepted the functions of a tester of knowledge. In the design of the testing part these functions can be enriched only by entering grammatical information. We have in mind that the user may to some extent change the name of the grammatical categories (e.g. "present tense", "pres.tense", "present", "pres." or "pr.") and the order of their coocurence in the set of grammatical features for a given wordform.

- a second type of drills tests the morphological knowledge(already situated in the framework of the given sentence). Thestandard method in this kind of testing is to give an inputsentence, in which certain wordforms are replaced by the basicforms. At the output these wordforms should be restored. These arethe well-known tasks of the type "Put the given words into thenecessary form." Here, as well, the computer entirely replaces the teacher in checking the student's work, because it restores thealready given wordforms in mode 1. To this kind of drills alsobelong the drills in which the student corrects the spellingmistakes deliberately made by the teacher.

- in a third type of drills certain morphological characteristics are replaced correctly by others (for example, achange in the grammatical person or number of the narrator, achange of the verb tense in a given story). Similarly to the previous type of drills, the organisation of knowledge goes beyond the framework of the first two functions of MORPHO-ASSISTANT, although an intelligent help-system for diagnostics and correction of mistakes is provided, which can be considered as a superstructure over the diagnostics of MORPHO-ASSISTANT.

- the fourth type of drills differs in principle from the others in the degree of grammatical abstraction - while the first three types deal with concrete lexical units and their grammatical characteristics, the fourth type of drills works with grammatical models of characteristics of a certain series of words in the sentence connected through a certain mechanism. These series form a so- called unification string consisting of two elements - unifying (input) and unified (output) element (they can consist of more than one word). The pattern which forms new unification strings is given by concrete wordforms and the test generating system itself finds the elements subjected to unification. Such a word string can express phenomena such as syntactic agreement of any type, correct generation of a pronoun from its antecedent etc. Such drills are usually formulated as follows: "Fill in the blanks with the suitable words". The drill can provide a list of lexemes in their basic forms or give the opportunity to choose freely the lexical units (relying on the rich lexicon of the system).

In fact, the mechanism connecting the elements of the unification string models phenomena from linguistic levels higher than morphology - syntax (as in the case with agreement) and even semantics. In the latter we can imagine for example the connection between the verb tense and a given type of temporal adverbs (under the condition that the entry of the adverb includes its type), the connection between some verbs and pronouns pointing near or distant objects. With the extension of the volume of the phenomena in the unification string it becomes more difficult to formulate them with the help of the grammatical categories of the wordform. The extension of the linguistic scope of these tasks depends to a great extent on the user's skill to formulate e.g. the syntactic dependencies through the morphological categories.

Providing the user only with the "shell" of the test elements, the system works with a kind of undetermined morphological knowledge, tending to get adapted in this way to the cognitive model of the user who generates the test. Thus its teaching effects directly depend on the teacher's ability to find the correspondence among the elements of the various levels of language.

5 Future Developments

MORPHO-ASSISTANT will be offered on the software market with a basic dictionary of 60 000 Bulgarian lexemes [3]. Our future aim is to develop a series of products assisting the use and the study of Slavonic morphology. An analogous product is being developed for Russian in cooperation with the Institute of Russian language of the Academy of Sciences of the USSR; the basic dictionary of the system, based on Zalizniak's dictionary [4], contains nearly 100 000 lexemes. A possibility for simultaneous work with both languages will also be supported.

The test generation system doesn't require an orientation towards a concrete language. The only restriction on its reasonable application is that the processed language must have a sufficiently developed morphology of the inflexional type.

We also plan to develop a computer model of the derivational morphology of Bulgarian based on the lexicon of MORPHO-ASSISTANT.

References

1. **Paskaleva, E.** Bulgarian Morphology in Logic Programming.- *Studies in Honor of Bernard Vauquois. Linguistica Computazionale*, Pisa, 1989 (to appear).

2. **Avgustinova, T. and Paskaleva, E.** Computational Modelling of Inflexional Morphology (Based on Bulgarian and Russian Data).- *Slavica Helsingiensia*, 1989, Vol. 10.

3. MORPHO-ASSISTANT (brief description). 1989 Humanities Computing Yearbook, Clarendon Press, Oxford.

4. **Зализняк А.А.** Грамматический словарь русского языка: Словоизменение. М., 1977.