AN INTEGRATED SYSTEM FOR MORPHOLOGICAL ANALYSIS OF THE SLOVENE LANGUAGE

Tomaž Erjavec, Peter Tancig

NLU Lab., Department of Computer Science and Informatics
Jožef Stefan Institute
Jamova 39, 61000 Ljubljana
Yugoslavia

ABSTRACT: The paper presents an integrated environment for morphological analysis of word-forms of the Slovene language. The system consists of a lexicon input and maintenance module, a lexicon output module for accessing lexical word forms, a two-level rule compiler and a two-level morphological analysis/synthesis unit. The basic paradigms and lexical alternations of word forms are handled by the lexicon system, while the two-level component takes care of phonologically induced alternations.

1. Introduction

We present an integrated environment for morphological analysis of (written) word-forms of the Slovene language. The language belongs to the Slavic family of languages, but exhibits some very idiosyncratic properties (e.g. having also a "dual" number and very rich inflection). Our project of writing a morphological analyzer and synthesizer (MAS) for the Slovene language has had primarily two aims. First, to write a useful MAS, which could serve as a front-end to other Slovene language processing systems, and second, to implement a model general enough to allow us to facilitate the study of Slovene morphology.

The work on the project itself is split into two parts much along the same lines. First of the task of selecting and implementing a model versatile enough to cover the quirks of Slovene morphology, and second, the task of writing down the rules of Slovene morphology (Toporišič 84) in the formalism of the chosen model.

The two-level model of Kimmo Koskeniemmi (Karttunen 88, Koskenniemi 84,85,86) was selected as the basic scheme for our MAS, our choice being influenced - among other things - by its prevalence in current (computer) morphological studies. This makes the system well documented and thus easy to implement, as well as simplifying the task of writing the rules for (phonologically) induced alternations of Slovenian word-forms (Erjavec 89).

2. Structure

The system was implemented on VAX/VMS in Quintus Prolog and consists of the following parts:

(1) The *compiler*, which takes as its input two-level rules and produces final state automata (transducers).

(2) The *lexicon module* which provides a user interface for the creation and updating of the lexicon - the *lexicon input module*. This module embodies that part of morphological knowledge of Slovene inflectional morphology which cannot be (elegantly) covered by two-level rules. It is also the part of the system responsible for passing lexical word forms to (3) - the *lexicon output module*.

(3) The *MAS* itself, which, having access to the transducers and (indirectly) to the lexicon, is able to analyze Slovene word forms into their lexical counterparts, and to synthesize word forms from lexical data.

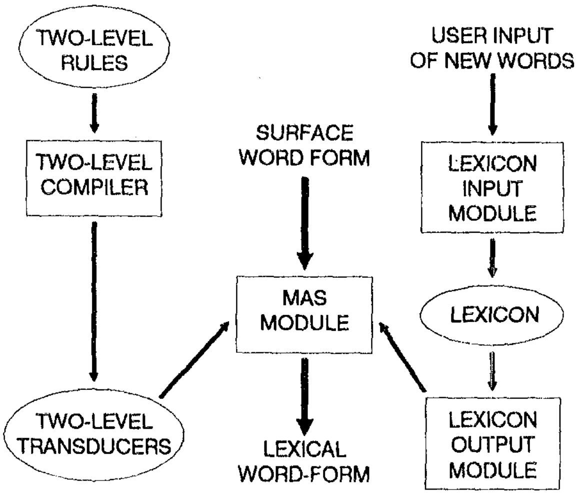

Figure 1: System structure

As we can see, the MAS with its knowledge of phono-morphological alternations embodied in the transducers guides the lexicon module in choosing the correct lexical word from the lexicon. The MAS module is of course also able to synthesize words, given their lexical representation. The "feeding" of lexical words to the MAS is however application dependent, and will thus not be dealt with further in this paper. The workings of the compiler will also not be discussed, as this is not its first implementation (Karttunen 87).

3. Lexicon module

A basic part of our MAS system is the lexicon. The structure of the lexicon accords with the two-level model type lexicon; that is, the lexicon is composed of letter-tree sub-lexicons (tries), consisting of morphemes with a common property. We can have, for instance, a sub-lexicon for stems, another for endings of male noun declension, another for conjugative endings of certain verbs, etc. A set of sub-lexicons is marked as initial, meaning that a (recognizable) word can only start with a member of these sub-lexicons. The other sub-lexicons are connected to initial sub-lexicons through pointers, typically making them inflectional paradigms of various word classes.

An entry in a sub-lexicon consists of three parts:

(1) the "morpheme", which, in stem sub-lexicons (two-level rules aside), is the invariant part of the stem lexeme, written in the symbols of the lexical alphabet;

(2) the continuation lexicon(s) of the morpheme;

(3) morpho-syntactic features of the morpheme.

To illustrate:

```
 bolezEn decl_subst_f2 / bv=subst gen=fem;
      (1)          (2)              (3)
```

(1) - the stem of the lexeme "illness"; the lexical symbol "E" denotes an unstressed "e" (schwa sound), deleted in word forms with non-null endings ("bolezen" - nom. sg., but "bolezni" - gen. sg.);

(2) - the name of the lexicon with endings of second female declension;

(3) - inherent morpho-syntactic properties of the lexeme (noun, female gender).

We can see that the lexicon system can take care of regular paradigms of inflecting words of the language (at least for suffixing languages, such as Slovene), while the two-level rules handle phono-morphological alternations. The Slovene language, however, abounds in alternations that are lexically conditioned. This is not to say that no rules can be constructed to cover these alternations, but rather that they are not (purely) phonologically conditioned. There is for instance an alternation that affects only nouns of male gender which have the "animate" property, and another one which pertains only to the plural and dual of certain Slovene nouns. Since two-level rules are sensitive only to the form of the word (string) they proces, they are insufficient for expressing such alternations.

To handle lexically conditioned types of alternations, we have concentrated on the linking mechanism between the sub-lexicons. The "continuation" information belonging to an entry can also, along with a pointer to another sub-lexicon, include a list of *lexical alternations*. When accessing word forms from the lexicon, these alternations tell the lexicon output module how to modify the continuation sub-lexicon to express the desired changes. The rules governing such modifications of the continuation sub-lexicon can perform a certain number of primitive "transformational" operations on the sub-lexicon in question.

To make the point clearer, we give a simple case of an alternation that affects certain nouns of male gender. The alternation "j epenthesis" inserts a "j" in the stem final position in word forms with a non-null ending; e.g. "krompir" -potato, but "krompirja" for the singular genitive form. The lexicon entry looks like this:

```
krompir decl_subst_m(pod_j) /
        bv=subst gen=mas -anim;
```

When the lexicon output module "jumps" to the continuation lexicon, the "pod_j" item will trigger the corresponding alternation in the morphological rule base of the system. The alternation procedure then takes as its input the continuation lexicon, modifies it, and returns the modified lexicon (with "j" prefixed to the non-null gramatemes). Analysis then proceeds with entries of the modified lexicon.

4. Input Module

If new entries are to be added to our lexicon by persons not acquainted with implementation details of the system (lexical alphabet and alternations), an input module with a friendly user interface is of prime importance. In our system the user is therefore expected to enter only the base form of the new word (e.g. nom. sg. for nouns) along with inherent morpho-syntactic properties of the word (e.g. noun, male, animate), and another "comparative" word form of the same word (e.g. gen. pl.). Both word-forms are entered in the surface alphabet.

With this information at its disposal, the input module must, in order to store the entry into the lexicon, do the following:

- extract form the word its lexical stem;
- transcribe it from surface into lexical characters;
- determine the continuation lexicon(s) (paradigms) and lexical alternations;

Extracting the lexical stem and assigning lexical alternations are performed by comparing the (base and comparative) word forms entered. For example the comparison of "ladja" (ship) and "ladij" (gen. pl.) shows an insertion of "i" into the stem, so the name of the lexical alternation for "i" epenthesis is added to the entry.

The "lemmatization" of words, especially the mapping from surface to lexical symbols, is basically nondeterministic; i.e. the input module "guesses" the correct lemmatization of the word, produces the lexical word form of the comparative word, and synthesizes its surface word form. If the synthesized word-form matches the one entered by the user, the lemmatization is correct; if not, the module tries again, with a different mapping.

5. Conclusion

Our system is in a certain sense redundant, since it has two ways with dealing with alternations - two-level rules and lexicon rules. The two-level rules are more linguistically justified, while the lexicon rules offer greater power in expressing alternations. This (partial) overlap is to a large extent intentional, as Slovene morphology has been - to say the least - insufficiently studied and our aim was to have at our disposal a variety of tools for expressing various morphological processes which occur in our language.

We have also tried to simplify the input of new words into the lexicon; a problem which in our opinion, has received insufficient attention in the two-level framework.

6. References:

Erjavec T., Tancig P. (1989) *Dvo-nivojska pravila za alternacije slovenskih samostalniških sklanjatev (Two-Level Rules for Alternations of Slovene Noun Declensions)* / V. kongres zveze društev za uporabno jezikoslovje Jugoslavije; Ljubljana

Karttunen L., Koskenniemi K.,Kaplan R. (1987) *A Compiler for Two-Level Phonological Rules* / in "Tools for Morphological Analysis", Center for the Study of Language and Information, Report No. CLSI-87-108

- (1988) *Unpublished lectures* / PreCOLING '88, Morphological analysis; Computational Linguistics Conference

Koskenniemi K. (1984) *A General Computational Model for Word- Form Recognition and Production* / Computational Linguistics, Conference Proceedings; COLING '84

- (1985) *A General Computational Model for Word Form Recognition and Production* / Computational Morphosyntax, Report on Research 1981-84; University of Helsinky, Dept. of General Lingusistics Publications No. 13

- (1986) *Compilation of Automata from Morphological Two-Level Rules* / Papers form the fifth Scandinavian Conference of Computational Linguistics 1985; University of Helsinky, Dept. of General Lingusistics Publications No. 15

Toporišič J. (1984) *Slovenska slovnica* / Založba Obzorja Maribor