KNOWLEDGE ENGINEERING APPROACH TO MORPHOLOGICAL ANALYSIS*

Harri Jäppinen[1], Aarno Lehtola, Esa Nelimarkka[2], and Matti Ylilammi
Helsinki University of Technology
Helsinki, Finland

ABSTRACT

Finnish is a highly inflectional language. A verb can have over ten thousand different surface forms - nominals slightly fewer. Consequently, a morphological analyzer is an important component of a system aiming at "understanding" Finnish. This paper briefly describes our rule-based heuristic analyzer for Finnish nominal and verb forms. Our tests have shown it to be quite efficient: the analysis of a Finnish word in a running text takes an average of 15 ms of DEC 20 CPU-time.

I INTRODUCTION

This paper briefly discusses the application of rule-based systems to the morphological analysis of Finnish word forms. Production systems seem to us a convenient way to express the strongly context-sensitive segmentation of Finnish word forms. This work demonstrates that they can be implemented to efficiently perform segmentations and uncover their interpretations.

For any computational system aiming at interpreting a highly inflectional language, such as Finnish, the morphological analysis of word forms is an important component. Inflectional suffixes carry syntactic and semantic information which is necessary for a syntactic and logical analysis of a sentence.

In contrast to major Indo-European languages, such as English, where morphological analysis is often so simple that reports of systems processing these languages usually omit morphological discussion, the analysis of Finnish word forms is a hard problem.

A few algorithmic approaches, i.e. methods using precise and fully-informed decisions, to a morphological analysis of Finnish have been reported. Brodda and Karlsson (1981) attempted to find the most probable morphological segmentation for an arbitrary Finnish surface-word form without a reference to a lexicon. They report surprisingly high success, close to 90 %. However, their system neither transforms stems into a basic form, nor finds morphotactic interpretations. Karttunen et al. (1981) report a LISP-program which searches in a root lexicon and in four segment tables for adjacent parts, which generate a given surface-word form. Koskenniemi (1983) describes a relational, symmetric model for analysis, as well as for production of Finnish word forms. He, too, uses a word-root lexicon and suffix lexicons to support comparisons between surface and lexical levels.

Our morphological analyzer MORFIN was planned to constitute the first component in our forthcoming Finnish natural-language database query system. We therefore rate highly a computationally efficient method which supports an open lexicon. Lexical entries should carry the minimum of morphological information to allow a casual user to add new entries.

We relaxed the requirement of fully informed decisions in favor of progressively generated and tested plausible heuristic hypotheses, dressed in production rules. The analysis of a word in our model represents a multi-level heuristic search. The basic control strategy of MORFIN resembles the one more extensively exploited in the Hearsay-II system (Erman et al.,1980).

II FINNISH MORPHOTACTICS

Finnish morphotactics is complex by any ordinary standard. Nouns, adjectives and verbs take numerous different forms to express case, number, possession, tense, mood, person and other morpheme categories. The problem of analysis is greatly aggravated by context sensitivity. A word stem may obtain different forms depending on the suffixes attached to it. Some morphemes have stem-dependent segments, and some segments are affected by other segments juxtaposed to it.

Due to lack of space, we outline here only the structure of Finnish nominals. The surface form of a Finnish nominal may be composed of the following constituents (parentheses denote optionality):

(1) root + stem_ending + number + case
+ (possessive) + (clitic)

The stem endings comprise a large collection of highly context-sensitive segments which link the word roots with the number and case suffixes in phonologically sound ways. The authorative Dictionary of Contemporary Finnish classifies nomi-

*This research is being supported by SITRA (Finnish National Fund for Research and Development)
P.O. Box 329, 00121 Helsinki 12, Finland

1Digital Systems Laboratory
2Institute of Mathematics

nals into 85 distinct paradigms based on the variation in their stem endings in the nominative, genetive, partitive, essive, and illative cases. The plural in a nominal is signaled by an 'i', 'j', 't', or the null string (ø) depending on the context. The fourteen cases used in Finnish are expressed by one or more suffix types each. Furthermore, consonant gradation may take place in the roots and stem endings with certain manifestations of 'p', 't' or 'k'.

As an example, consider the word 'pursi' (=yacht). The dictionary representation 'pur| si^{42}' indicates the root 'pur', the stem ending 'si' in the nominative singular case, and the paradigm number 42. Among others, we have the inflections

(2) pur + re + ø + lla + mme + kin
(=also on our yacht)
pur + s + i + lla + mme + ko
(=on our yachts?)

Consonant gradation takes place, for instance, in the word 'takk| i^{4}' (=coat) as follows:

(3) tak + i + ø + ssa + ni (=in my coat)
takk + e + i + hi + ni (=into my coats)

III DESCRIPTION OF THE HEURISTIC METHOD

A. Control Structure

Our heuristic method uses the hypothesis-and-test paradigm used in many AI systems. A global database is divided into four distinct levels. Productions, which carry local heuristic knowledge, generate or confirm hypotheses between two levels as shown in the figure.

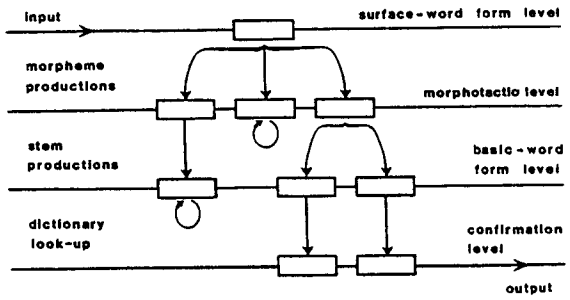

Figure. The control structure of MORFIN.

B. Morpheme Productions

Morpheme productions recognize legal morphological surface-segment configurations in a word, and slice and interprete the word accordingly. We use directly the allomorphic variants of the morphemes. Since possible segment configurations overlap, several mutually exclusive hypotheses are usually produced on the morphotactic level. All valid interpretations of a homographic word form are among them.

The extracted rules were packed and compiled into a network of 33 distinct state-transition automata (3 for clitic, 1 for person, 6 for tense, 3 for case, 2 for number, 5 for adjective comparation, 3 for passive, 5 for participle, and 5 for infinitive segments). These automata were generated by 204 morpheme productions of the form:

(4) name: (2nd_context)(1st_context)segment -->
POSTULATE(interpretation,next)

'Segment' exhibits an allomorph; the optional '1st' and '2nd contexts' indicate 0 to 2 left-contextual letters. The operation POSTULATE separates a recognized segment, attaches an interpretation to it, and proceeds to the indicated automata ('next'). For example, the production

(5) [V]n --> POSTULATE([gen,sg,...],
[NUM1,NUM2,PAR1,PAR4,PAR5,INF3,INF4,COMP4])

recognizes the substring 'n', if preceeded by a vowel, as an allomorph for the singular genetive case, separates 'n', and proceeds in parallel to two automatons for number, three for participles, two for infinitive, and one for comparation.

C. Stem Productions

Stem productions are case- and number-specific heuristic rules (genus-, mood- and tense-spesific for verbs) postulating nominative singular nouns as basic forms (1st infinitive for verbs) which, under the postulated morphotactic interpretation, might have resulted in the observed stem form on the morphotactic level. They may reject a candidate stem-form as an impossible transformation, or produce one or more basic-form hypotheses.

The Reverse Dictionary of Finnish lists close to 100 000 Finnish words sorted backwards. For each word the dictionary tags its syntactic category and the paradigm number. From that corpus we extracted heuristic information about equivalence classes of stem behavior. This knowledge we dressed into productions of the following form:

(6) condition --> POSTULATE(cut,string,shift)

If the condition of a production is satisfied, a basic-form hypothesis is postulated on the basic word-form level by cutting the recognized stem, adding a new string (separated by a blank to indicate the boundary between the root and the stem ending), and possibly shifting the blank. These operations are indicated by the arguments 'cut', 'string', and 'shift'. A well-formed condition (WFC) is defined recursively as follows. Any letter in the Finnish alphabet is a WFC, and such a condition is true if the last letter of a stem matches the letter. If $\&_1,\&_2,\ldots,\&_n$ are WFCs, then the following constructions are also WFCs:

(7) (I) $\&_2\&_1$
(II) $<\&_1,\&_2,\ldots,\&_n>$

(I) is true if $\&_1$ and $\&_2$ are true, in that order, under the stipulation that the recognized letters in a stem are consumed. (II) is true if $\&_1$ or $\&_2$ or ... or $\&_n$ is true. The testing in (II) proceeds from left to right and halts if recognition occurs. The recognized letters are consumed. A capital letter can be used as a macro name for a WFC. For example, a genetive 'n'-specific production

(8) <Ka,y>hde --> POSTULATE(3,'ksi',0)

('K' is an abbreviation for <d,f,g,h...> - the consonants) recognizes, among other stems, the genetive stem 'kahde' and generates the basic form hypothesis 'ka ksi' (= two).

We collected 12 sets of productions for nominal and 6 for verb stems. On average, a set has about 20 rules. These sets were compiled into 18 efficient state-transition automata.

We could also apply productions to consonant gradation. However, since a Finnish word can have at most two stems (weak and strong), MORFIN trades storage for computation and stores double stems in the lexicon.

D. Dictionary Look-up

The dictionary look-up procedure confirms or rejects the basic-word form hypotheses that have proliferated from the previous stages by matching them against the lexicon. Thus in MORFIN the only morphological information a dictionary entry carries is the boundary between the root and the stem ending in the basic-word form and grade. All other morphological knowledge is stored in MORFIN in an active form as rules.

In MORFIN, input words are totally analyzed before a reference to the lexicon happens. Consequently, also words not existing in the lexicon are analyzed. This fact and the simple lexical form make it easy to add new words in the lexicon: a user simply chooses the right alternative(s) from postulated basic-word form hypotheses.

IV DISCUSSION

MORFIN has been fully implemented in standard PASCAL and is in the final stages of testing. The lexicons contain nearly 2000 most frequent Finnish words. In addition to one lexicon for nominals, and one for verbs, MORFIN has two "front" lexicons for unvarying words, and words with slight variation (pronouns, adverbs etc. and those with exceptional forms).

Currently MORFIN does not analyze compound nouns into parts (as Karttunen et al. (1981) and Koskenniemi (1983) do). By modifying our system slightly we could do this by calling the system recursively. We rejected this kind of analysis because the semantics of many compounds must be stored as separate lexical entries in our database interface anyway. MORFIN does not produce word forms as the other two systems do.

With respect to the goals we set, our tests rate MORFIN quite well (Jäppinen et al., 1983). Lexical entries are simple and their addition is easy. On average, only around 4 basic-word form hypotheses are produced on the basic-word form level. The analysis of a word in randomly selected newspaper texts takes about 15 ms of DEC 2060 CPU-time. Karttunen et al. (1981) report on their system that "It can analyze a short unambiguous word in less than 20 ms [DEC-2060/Interlisp] ... a long word or a compound ... can take ten times longer." Koskenniemi (1983) writes that "with a large lexicon it [his system] takes about 0.1 CPU seconds [Burroughs B7800/PASCAL] to analyze a reasonably complicated word form."

Both Karttunen et al. (1981) and Koskenniemi (1983) proceed from left to right and compare an input word with forms generated from lexical entries. It is not clear how such models explain the phenomenon that a native speaker of Finnish spontaneously analyzes also grammatical but meaningless word forms. Most Finns would probably agree that, for instance, 'vimpuloissa' is a plural inessive form of a meaningless word 'vimpula'. How can a model based on comparison function when there is no lexical entry to be compared with? Our model encounters no problems with new or meaningless words. 'Vimpuloissa', if given as an input, would produce, among others, the hypothesis 'vimpul a' with correct interpretation. It would be rejected only because it is a non-existent Finnish word.

ACKNOWLEDGEMENTS

Lauri Carlson has given us helpful linguistic comments. Vesa Yläjääski and Panu Viljamaa have implemented parts of MORFIN. We greatly appreciate their help.

REFERENCES

Brodda, B. and Karlsson, F., An experiment with automatic morphological analysis of Finnish. Un. of Stockholm, Insitute of Linguistics, Publication 40, 1981.

Erman, L.D. et al., The Hearsay-II speech-understanding system: integrating knowledge to resolve uncertainty. Computing Surveys, Vol. 12, No 2, (June, 1980), 213-253.

Jäppinen H., Lehtola, A., Nelimarkka, E., and Ylilammi, M., Morphological analysis of Finnish: a heuristic approach. Helsinki University of Technology, Digital Systems Laboratory, 1983 (forthcoming report).

Karlsson, F., Finsk Grammatik. Suomalaisen Kirjallisuuden Seura, 1981.

Karttunen, L., Root, R., and Uszkoreit, H., TEXFIN: Morphological analysis of Finnish by computer. The 71st Ann. Meeting of the SASS, Albuquerque, 1981.

Koskenniemi, K., Two-level model for morphological analysis. IJCAI-83, 1983, 683-685.