



Reports from the  
Department of Phonetics  
University of Umeå

May 1990

1

PHONUM

## PHONETIC TRANSCRIPTION OF A SWEDISH MORPHOLOGICAL ANALYZER

*Tina Magnuson, Björn Granström and Rolf Carlson*  
*Dept. of Speech Communication and Music Acoustics, KTH*

*Fred Karlsson*  
*Dept. of General Linguistics, University of Helsinki, Finland*

### INTRODUCTION

This is a status report on our efforts to create an automatic analyser for phonetic transcription of text. There are many obvious uses of such a component and there have been several efforts in that direction. Especially in a text-to-speech environment the needs are clear. Most systems today have a combination of rules and lexica, with a division depending on the strength of the rules, the available size of memory and the needs for parts-of-speech. The first major effort based on a big morph dictionary was made in the 70th at MIT, Allen, Hunnicutt and Klatt 1987. The morph decomposition algorithm was complemented by a set of letter to sound rules. In the KTH text-to-speech system (Carlson & Granström, 1986; Carlson, Granström & Hunnicutt 1990), the rules have played a more important role. In products based on the MIT research the morph decomposition algorithm have been replaced by a straight forward lexical look up procedure. The system that is under development at Bell laboratories relay mainly on a big morph lexicon (Coker, 1985.). At Chalmers an experimental text-to-speech system has been developed using a large dictionary, Hedelin & Jonsson (1986). We have had access to one version of this dictionary in this study.

Our current efforts are not only motivated by the needs in a text-to-speech system. We will also use the morphological component for automatic transcription of text in our speech data base work. The need for more detailed lexical descriptions including parts-of-speech is also increasing in these projects.

We have chosen to base our efforts on a morphological component developed by one of the authors at the University of Helsinki, Department of General Linguistics. This component, based on many different kinds of corpora, has a very good coverage of running text. The morph dictionary is now being transcribed manually based on an automatic procedure.

#### **The Swedish morphological analyzer**

The Swedish morphological analyzer SWETWOL is based on the two-level morphology (TWOL) designed by Kimmo Koskenniemi (1983). TWOL provides a language-independent formalism for dealing with morphological phenomena such as inflection, derivation, compounding, morphophonology, and morphotactics. The main lexicon contains one base form per paradigm, plus inflectional and other information. Inflectional elements are linked to one another in precise ways, morphophonological alternations may be described by appropriate rules.

A central idea in TWOL is to have the rules compiled into finite-state automata to be executed in parallel independently of one another. That is one reason why the model often goes under the name 'finite-state morphology'. TWOL has been successfully applied to more than 20 languages, e.g. English, Russian, Swahili, Arabic, Japanese.

SWETWOL was designed in 1988-89 (Karlsson 1990a). Presently it contains a main lexicon with some 41,000 entries, i.e. the whole central Swedish vocabulary. The bulk

of the words was drawn from Svenska Akademiens Ordlista. This vocabulary was critically evaluated and only lexemes judged to be in current use were included. Additional excerpts were made from several other dictionaries. The coverage of SWETWOL is being continually tested and extended by systematically applying it to more texts (presently > 1,5 million word-forms of running text), thereby spotting relevant neologisms, names, etc. that should be included in a comprehensive analyzer. Every lexical entry has a proper inflectional code and a code describing its behaviour in compounding. Therefore all inflectional forms and all compounds are properly analyzed. If a word-form is ambiguous, all readings are retrieved. Base-forms are also retrieved for inflected words:

bilar  
bil "N NOM SG"  
bila "V PRES SG3"

The present recall of SWETWOL in 'ordinary' running text is 99%. Work on constructing a morphological disambiguator is almost completed (Karlsson 1990b).

#### **Manual transcription based on an automatic procedure**

The morphs in the SWETWOL is divided into several components according to e.g. their parts-of-speech. We have so far transcribed the noun section, which contains about 16000 morphs. This has been done with the help of the KTH text-to-speech system. As a starting point the dictionary is run through the system and the phonetic transcription is made based on the letter-to-sound rules. A small dictionary based on a frequency dictionary (Allén, 1973) is also included in this process. Only the words that are not covered by the rules are included.

In the next phase we replace the lexicon by the dictionary developed at Chalmers, and the text is run through the system again. The output from the two runs is compared and the best transcription is chosen manually and used as a base for the transcription correction phase.

During the correction the text-to-speech system is used to give spoken output of the transcription. This feature has proven to be of big value during this ongoing process. We will in the next section discuss some of the experiences we so far have got from the methodological and practical point of view.

#### **Accuracy of the automatic procedure**

It can be of interest to see how much help the automatic procedure has given us in the project. In general the rules in the text-to-speech system are not created to handle the type of material that can be found in a morph dictionary. Normally most words have some type of ending that give valuable hints on how the word should be transcribed. The accuracy of the letter to sound rules in transcribing morphs get reduced for this reason. Another reason is that many of the forms are very unusual and thus do not follow the general rules of pronunciation. This is of course one of the main reasons to adopt a morphologically based approach. We find that the rules yield 41% errors for the noun corpus. This should be compared to our estimated accuracy of 24% on the Allén whole word corpus (not frequency adjusted). By adding the small lexicon (2400 words) that usually goes with the text-to-speech system only 5% of the morphs get a different pronunciation. If the larger lexicon (around 30000 base forms) developed at Chalmers is added the accuracy improves, but only 22% of the morphs have a different pronunciation, compared to the original run without a lexicon. It should be noted that we only had an early version of the Chalmers lexicon available. After the first automatic phase of the transcription project we still had to correct 34% of the morphs.

### Transcription problems and strategies

During the process there have been issues not totally evident and certain decisions have had to be made. As an example there is no audible difference between /e/ and /ä/ in the Central Swedish dialect, the orthography has in this case been guiding whether the transcription should have an /e/ or an /ä/. The Gothenburg dialect, as found in the Chalmers lexicon, has a slightly different choice of allophones. This constitutes about 3% of the corrections that have to be made. Another example of difference can be found in the transcription philosophy. We have left the /nk/ as a phonemic representation and assumed that the /n/ will get velarized in the phonetic rule component that has to operate on the output of the morph analysis. Many manual changes have been made due to stress errors in the automatic transcription. As the morph decomposition separates the suffixes from the roots the rule system will lack the information about the complete word pattern. This will cause many stress errors to occur.

The thorough testing of our letter to sound rules has given us new ideas about missing regularities, that so far have not been modelled in our system. This is specially true for some stress assignment rules. Other experiences concern the segmental realization rules as well, but these findings are outside the scope of this paper.

### Merge of the morphs into the phonetic transcription

The morph decomposition module has as output the pronunciation of the different individual morphs. These morphs have to be merged together according to the normal compounding rules in the language. Stress has to be adjusted and phonological rules operating across morph boundaries have to be applied. Processes like this can be described in terms of transformation rules and can in our case be implemented as part of the phonetic component in the text-to-speech system. This component takes input from other modules as well. Irrespective of origin the same rule processes have to apply.

The transcription phase is still in progress and the work has given valuable feedback. In a text-to-speech context this component will give a high precision of the pronunciation of the individual words. At the same time valuable information for the syntax analysis will be supplied. This information will improve the possibilities to generate a better prosody. It should also be emphasized that the use of this component is not restricted to text-to-speech applications. Generation of phonetic transcriptions in dictionaries and text corpora in general are obvious alternative applications.

### ACKNOWLEDGEMENTS

This work has been supported by grants from the Swedish National Board for Technical Development and the Swedish Telecom.

### REFERENCES

- Allén, S. (1973): *Nusvensk frekvensordbok* (Frequency Dictionary of Present-Day Swedish), Almqvist & Wiksell, Stockholm.
- Allen, J., Hunnicutt, M.S., & Klatt, D. (1987): *From Text to Speech*. The MITalk System, Cambridge University Press, Cambridge, England.
- Carlson, R. & Granström, B. (1986): "Linguistic processing in the KTH multi-lingual text-to-speech system", *Proc. ICASSP 86*, Vol. 4, Tokyo, pp. 2403-2406.

Carlson, R., Granström, B. & Hunnicutt, S. (1990): "Multilingual text-to-speech development and applications", in A.W. Ainsworth (ed), *Advances in speech, hearing and language processing*, JAI Press, London

Coker, C. H.(1985): "A dictionary intensive letter-to-sound program", *J. Acoust. Soc. Amer.*, vol. 78 S7.

Hedelin, P. & Jonsson, A. (1986): "Svenskt uttalslexikon", Technical report No 4; Department of Information Theory, Chalmers, Göteborg.

Hunnicutt, S. (1980): "Grapheme-to-phoneme rules: a review", *STL-QPSR* 2-3/1980.

Karlsson, F. (1990a): "A Comprehensive Morphological Analyzer for Swedish", manuscript, University of Helsinki, Department of General Linguistics.

Karlsson, F. (1990b): "The Constraint Grammar Parser CGP", manuscript, University of Helsinki, Department of General Linguistics.

Kimmo Koskenniemi (1983): "Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production", University of Helsinki, Department of General Linguistics, Publications No. 11.