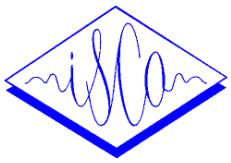
Automatic linguistic processing in a German text-to-speech synthesis system

Betina Schnabel-Le Corre, Harald Roth



The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

AUTOMATIC LINGUISTIC PROCESSING IN A GERMAN TEXT-TO-SPEECH SYNTHESIS SYSTEM

Betina SCHNABEL [1,2] ***& Harald ROTH*** [1]

(1) Centre National d'Etudes des Télécommunications
BP 40 - 22300 LANNION Cedex
(2) Institut de la Communication Parlée
BP 25x - 38040 GRENOBLE Cedex

ABSTRACT

The linguistic processing of the CNET's Text-to-Speech (TTS) synthesis for German has recently been automatized and extended by a series of new algorithms. The modifications concern the following three stages : first, a preprocessing stage was added to convert numerical expressions, abbreviations and diacritics into sequences of graphemes. Then, the morphological analysis has been extended to compound words - one of the major problems of German morphology - and to the grammatical labelling of word classes. Finally, an automatic parser has been developed for the insertion of syntactic-prosodic markers. It uses a set of hierarchized parsing rules which are applied to the extracted sequences of grammatical categories resulting from the upper analysis.

1. INTRODUCTION

In everyday life, written texts very often exhibits ambiguities (for example Telex, abbreviations, logos, ...), which a native speaker can easily solve by his implicit knowledge of the language. For TTS synthesis systems this linguistic competence must be substituted by a more or less sophisticated grammatical analysis in order to ensure correct pronunciation and intonation of a given utterance. Even if German orthography is quite regular compared to English or French, the pronunciation of a word depends on its morphological structure and the position of the stressed syllable. For German speech synthesis, the splitting of words into morphemes and the assignment of lexical stress is thus obligatory.

e.g. *Herbstrose* [hɛrpst // rozə] *Erbstreiter* [ɛrp // ʃtraitɐ]
'übersetzen ("to cross over") *übers'etzen* ("translate")

In this paper, different stages are described which are necessary to transform a German input text into a morphologically analyzed, grammatically labelled and parsed output text which can easily be transcribed into its phonetic representation (figure 1).

2. ARCHITECTURE OF THE SYSTEM

Even if the major linguistic problems are already solved by a given system, room must be provided for future modifications and extensions of the linguistic processing. It is thus essential for such a system that any person, not necessarily familiar with computers (e.g. a linguist or a phonetician) can easily access the files containing the rules. Like many other systems [CARLSON (1986),BAUER (1987),BARBER (1989)], the CNET system aims at defining rule files that are strictly external to the program. Furthermore, this aspect is particularly important for multi-lingual systems because it creates a linguistic independence that allows further applications to other languages.
Sophisticated solutions have been proposed to create special environments for linguistic and phonetic rule sets, such as RULSYS, PROPHON or LIFT [CARLSON (1986),BACKSTRÖM (1989), FRENKENBERGER (1990)]. The CNET System is currently a simple concatenation of the different modules, each transforming a given input format into a different given output format.

2.1. The Preprocessing Stage

The aim of this module is the so-called "normalization" of the input text. It consists first of converting the following symbols into their full orthographic form : numerical expressions, abbreviations, logos, diacritics, special characters, and roman numbers. Secondly, nouns are detected by their

initial capital letter and marked. Subsequently, all capital letters are replaced by small ones. It ensures that the orthographic-phonetic transcription as well as the suprasegmental processing will be performed consistently with all other words.

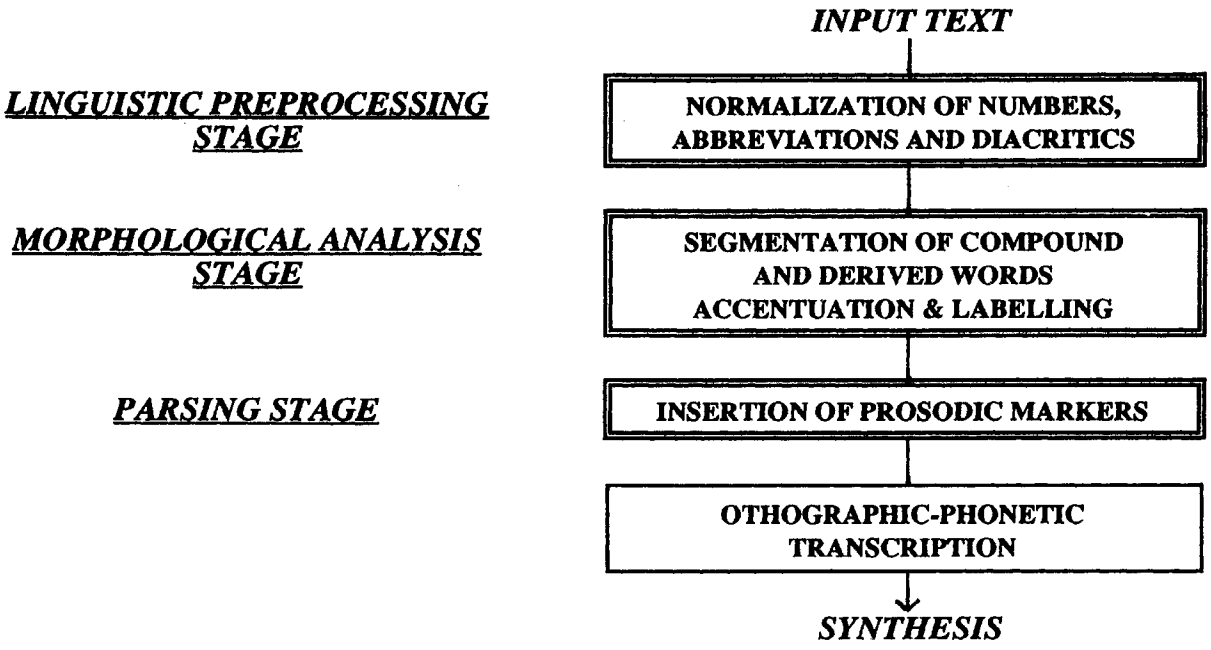


Figure 1 : The different stages to transform German input text to phonetic output text containing prosodic markers. (double lined = stages described in this paper)

Three inventories are used : a dictionary of about 250 abbreviations, logos, diacritics and special characters, and two files in which the normalized graphemic form of the cardinal numbers '0' to '100' and of the ordinal numbers from '0' to '31' are stored as well as words like "tausend", "millionen", and so on. At the present state, the rules to translate the different numerical expressions (date, hour, sum of money, telephone number, command number) are still embedded in the program.

Example : INPUT -> Der 1980 erschienene II. Band der Romantrilogie kostet mit 10% Ermäßigung 110,50 DM.
OUTPUT -> der %0 neun_zehn_hundert-^achzig erschienene zw^eite band + der %0 romantrilogie + kostet mit %1 zehn proz^ent + ermäßigung hundert-z^ehn m^ark f^ünfzig.

2.1. The Morphological and Labelling Stage

This module carries out the segmentation of derived and compound words, the lexical stress assignment and the grammatical labelling of each word. Unlike most other approaches to automatic morphological analysis in TTS synthesis - which are either rule-based [RÜHL (1984), BARBER(1989)] or dictionary-based [BAUER (1987)] - the CNET system combines both. A rather large lexicon of roots (8000 entries) was elaborated as well as 3 dictionaries of about 80 prefixes, 150 suffixes and 370 function words. But beside a marker for verbal roots and the lexical stress in the root lexicon, the dictionaries do not contain any other information. The segmentation is performed by a search graph algorithm guided by linguistic heuristics, and stress positionning is defined by separate rule sets in the following way.

2.1.1. *Segmentation*

For each word - not already identified and marked as function word (§ 2.1.3), number or abbreviation by the upper preprocessing stage - a graph-tree is created which splitts the word into morphemes. Six classes of morphemes are distinguished : nominal roots, verbal roots, prefixes, suffixes,

junction-'s'[1] and an unknown element 'X'.
The segmentation proceeds word by word from left to right. Starting from the beginning of a word, all possible constituents are searched. For every detected morpheme, a node of the graph-tree is created and the rest of the word is reexamined up to the end of the word. To avoid impossible combinations of morphemes or redundant segmentations, linguistic heuristics limit the choice of morphemes and thus the creation of new nodes.
* *Explicit heuristics* are linguistic criteria which determine the combination of morphemes. For example, it is impossible that a word begins with a suffix, or that a prefix is followed by a junction-'s'.
* *Implicit heuristics* set the order in which the different morphemes are searched. After skipping the eventual prefix of a word, the search is always ordered in the following way : a root has priority over a prefix, a prefix over a suffix, a suffix over a junction or an unknown element. Those priorities are governed by the order of the search in the dictionaries. Furthermore a long element is prefered to a shorter one of the same class, by ranking the morphemes in dictionaries according to their initial character and according to their length for the same initial character.
After the validation of a possible segmentation, morphological markers are inserted between the different morphemes and the word can now be stressed in an adequate manner (figure 2).

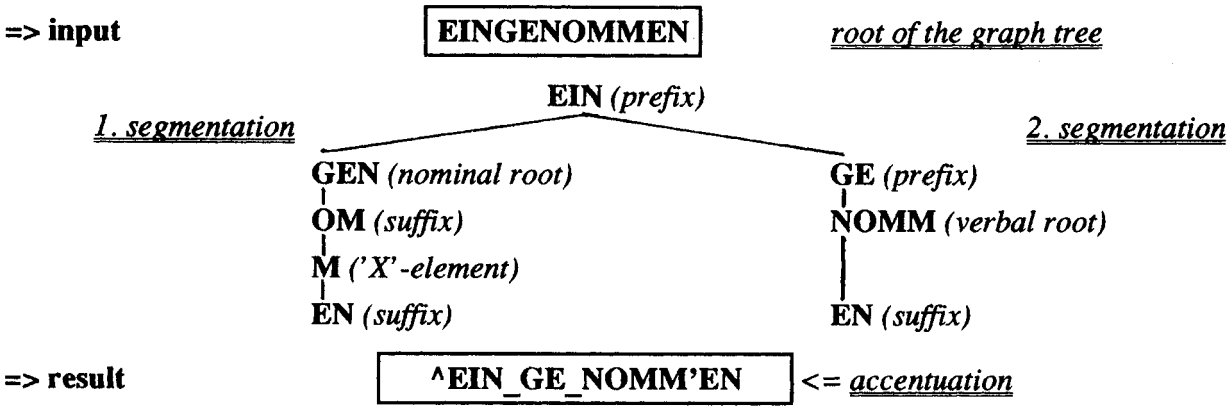


Figure 2 : Example for the choice between possible segmentations, insertion of the morphological markers, and stress assignment (the second segmentation is considered to be correct, because the first one contains the 'X'-element "M").

2.1.2. *Stress Assignment*

This function is based on two separate files of accentuation rules for prefixes and suffixes (constituting an extension of an earlier work by ZINGLE [ZINGLE (1982)]) and on the position of the stressed syllable of the root stored in a buffer during the search in the root lexicon. The rules are ranked according to the alphabetical order of the affix inventory ; each general stress rule is preceded by hierarchized context depending exception rules. Rules are applied by first examining the prefixes and then the suffixes. Each time a rule can be applied, the stress is assigned to the affix, the search is stopped, and the stress on the root is cancelled. If no stress rule for an affix in its given context is found, the stress remains on the stressed syllable of the root.

2.1.3. *Labelling*

At present, seven labels are applied to distinguish the different word classes. Nouns (1 label) have already been labelled in the preprocessing stage. Function words (4 labels) and some auxilaries (1 label) are detected before segmentation by consulting the dictionary of function words including the conjugated form of some highly frequent verbs (e.g. *'ist', 'hat'*) to speed up the procedure.
Verbs (1 label) are identified after segmentation : a word is labelled as verb if the following

1 A junction is a morphological element, not obeying any systematic rule, inserted between two roots of a compound word, e.g. *Rat's keller.*

conditions are met : (1) the word contains a verbal root, (2) the root is followed by a verbal suffix, (3) this suffix is the last constituent of the word. Subsequently, all remaining non-classified words are labelled as adjectives and adverbs.

2.3. The Syntactic-Prosodic Parsing Stage

This stage allows to determine the boundaries and types of prosodic groups in the sentence, and to insert the 8 syntactic prosodic markers. The syntactic unit to be considered is the clause, which is defined as the unit between two punctuation marks (or beginning and end of file). Primarily, the parser operates by comparing the sequences of labels according to a clause (resulting from the upper analysis) to the sequences of labels defined within the hierarchized parsing rules, similarly structured to those for French [LARREUR (1989)]. More precisely, the sequences of labels are extended by several parameters : the number of syllables of each word in the clause, the punctuation mark that precedes and follows the clause, as well as general symbols such as 'anything but verb', 'any function word'.
The rules are divided into several classes and hierarchized within classes. They are compared sequentially to the parameters of the clause until an adequate rule is found and applied. The same procedure is repeated for each class and the markers are placed into the output text as shown in the example :

INPUT -> der %0 neun_zehn_hundert-^achzig erschienene zw^eite band + der %0 romantrilogie + kostet mit %1 zehn proz^ent + ermäßigung hundert-z^ehn m^ark f^ünfzig.

-> OUTPUT der %0 neun_zehn_hundert-^achzig er_sch^ien'en'e (/) zw^eite b^and + der %0 rom^an-tri-log'ie + (=) k^ost'et ($) mit %1 zehn proz^ent + er_m^äß'ig'ung (*) hundert-z^ehn m^ark f^ünfzig. *(Brackets indicate syntactic-prosodic markers)*

CONCLUSION

The first tests prove that the text-to-speech synthesis generated with this automatic linguistic processing yields the same quality as with manual insertion of prosodic markers [SCHNABEL (1988)]. One of the major problems of German morphology, the splitting of compound words, as well as the 'normalization' of the text and the lexical stress assignment have been solved in a reliable and simple manner. It is nevertheless obvious that the generation of natural sounding prosody requires very accurate and detailed parsing rules in order to insert well-defined prosodic markers. Our main interest in the future is thus to extend the modules for parsing and prosodic generation and to carry out formal tests permitting to obtain official error rates of the whole linguistic processing and of each of its separate stages.

ACKNOWLEDGEMENT

The authors wish to thank Isabelle METAYER for her precious help in the development of the programs and the elaboration of statistical tools for checking the quality of the linguistic processing and thus allowing us to improve it.

REFERENCES

[1] BARBER S., CARLSON R., COSI P., DI BENEDETTO M.G., GRANSTRÖM B.& VAGGES K. (1989). ***A Rule Based Italien Text-to-Speech System.*** Europ. Conf. on Speech Comm. and Techn., Paris, pp. 517-520.
[2] BAUER S., KOMMENDA M. & POUNDER A. (1987). ***Graphem-Phonem-Umsetzung : Lexikon versus Regelkatalog.*** Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung e.V., Bonn, pp. 18-25.
[3] BÄCKSTRÖM M., CEDERK. & LYERG B. (1989). ***Prophon - An Interactive Environment for Text-to-Speech Conversion.*** Europ. Conf. on Speech Communication and Technology, Paris, pp. 144-148.
[4] CARLSON R. & GRANSTRÖM B. (1986). ***Linguistic Processing in the KTH Multi-Lingual Text-to-Speech System.*** Int. Conf. on Acoustic, Speech, and Signal Processing, Japan, pp.2403-2407.
[5] FRENKENBERGER S. & KOMMENDA M. (1990). ***Die mehrdimensionale Datenstruktur LIFT angewendet in der Sprachsynthese.*** DAGA, 9-12 April, Wien (in print).
[6] LARREUR D., EMERARD F. & MARTY F.(1989). ***Linguistic and Prosodic Processing for a Text-to-Speech Synthesis System.*** EUROSPEECH, Paris, pp 510-514.
[7] RÜHL H.-W. (1984). ***Sprachsynthese nach Regeln für unbeschränkten deutschen Text.*** PhD-Thesis, Ruhr-Universität Bochum.
[8] SCHNABEL B. & CHARPENTIER F. (1988). ***Multilinguale Sprachsynthese.*** BIGTECH Berlin, pp 97-105.
[9] ZINGLE H. (1982). ***Traitement de la prosodie en Allemand dans un système de synthèse de la parole.*** Thèse d'Etat, Université de Straßbourg II.