A PROCEDURE OF AN AUTOMATIC GRAPHEME-TO-PHONEME TRANSFORMATION OF GERMAN

Sabine Koch, Wolfgang Menzel, Ingrid Starke

Zentralinstitut für Sprachwissenschaft, AdW DDR, Berlin, DDR

The automatic transformation of texts graphemically stored to the corresponding phonemic symbols will enable the speech synthesizer Rosy 4000 (developed by VEB Robotron Dresden) to extend its field of application (application in information systems, development of reading machines for the blind). The texts for this kind of application cannot be limited in any way - a fact which had to be taken into account concerning the methods suitable for such a procedure. The use of the dictionary method, that means storing the whole vocabulary needed together with the corresponding phonemic strings was impossible for this purpose.

The procedure presented here can shortly be characterized as a rule system. The transformation is done on the level of word forms not taking into consideration syntactic or semantic criteria.

An important part of the procedure is the analysis of the structure of word forms. The results of this analysis influence the intended high quality of the transformation to a large extent.

The problem of automatically identifying the boundaries between elements of compounds could not be solved having in mind the aim to transform unlimited texts. As it is necessary for a correct phonemic transformation to know these boundar-

ies, all compounds are split by hand when the input text is stored.

The presented procedure identifies graphemic substrings in the word form to be transformed on the basis of an unique deterministic analysis and it also checks if the context of the string or the status of the system fulfil special conditions. In case these tests were successful the substring will be accepted, that means the corresponding phonemic transcription as well as the stress information are added to the substring. In certain cases it is possible to postpone the transformation to one of the following steps.

The graphemic substrings are contained in the information part of the procedure together with the conditions and the results of the transformation. The information part, that means the linguistic part, is strongly separated from the algorithm. This separation was of great advantage when working out the procedure.

The transformation is carried out in six stages, the most important of which are the analysis of the structure of word forms (the prefix and suffix strategy) and the transformation of graphemes by a set of rules.

The analysis of the structure of word forms splits the regarded word form into morphemes and marks the morphemic boundaries on the basis of lists containing prefixes and suffixes together with the corresponding phonemic realizations and the stress information (marking of the stressed syllable or stress shifting to other syllables). These lists also contain exceptions. The exceptions are substrings of certain word forms which are identical with an affix on the graphemic level but they differ in pronounciation or stress or in both of them.

All parts of the word form which are not treated by the prefix or suffix strategy (normally the basis) are to be trans-

formed by transformation rules. These are context sensitive rules which are applied from left to right. The word form is run trough only once. One part of the context conditions result from word structure analysis: That is the marked morphemic boundaries which influence the transformation of graphemic strings with regard to phenomena like the so-called final devoicing and the so-called glottal stop as well as the length of vowels. Classes and subclasses of graphemes and phonemes (classes of consonants, vowels, plosives, etc.) are also used as context conditions for an adequate transformation.

The strategy of stress as the last part in the procedure fixes the main stress in the word form by taking into consideration the stress information supplied from the other strategies. There exist three classes of prefences: the absolute stress information, the conditional stress information (if there is no absolute stress information) and the stress information without preference (if there is no conditional preference information).

For the remaining unstressed word forms the main stress is fixed by stress patterns. The native German vocabulary can be handled by these patterns without large lists of exceptions. Most of the exceptions are foreign words.

The first strategy before these mentioned main parts of the procedure is a look-up in a list containing about 250 of the most frequent German word forms (articles,pronouns) which are transformed as a whole without running through all the strategies of the procedure. This immediate transformation saves a lot of time.

Furthermore there is a list of about 60 homographs, which could be transformed unambigiously only by the aid of syntactic or semantic criteria. The word forms of this list are also immediately transformed to the corresponding variants. The advantage of this method is that the following parts of

the procedure do not have to handle ambiguities.

The paper will contain information concerning the number and kind of transformation mistakes. In general the German vocabulary can be transformed correctly by regularities easily to formulate. Difficulties and a great number of exceptions to the regularities result from foreign words which are very frequent in German. The transformation of foreign words cannot be excluded from the procedure because they are often used in German and sometimes they even have no German equivalent like Ingenieur, Cello, Charta, Chaussee etc.