



PHONEME-TO-GRAPHEME CONVERSION SYSTEM FOR UNRESTRICTED GERMAN TEXT

U. Jekosch*

ABSTRACT

In this paper a phoneme-to-grapheme system (PTG) for the German language is introduced. Since this system is based on rules, a dictionary look-up is not required, and thus the conversion of an unrestricted vocabulary is possible. A necessary requirement, however, is a correct segmentation of continuously spoken sentences into isolated words. For the PTG system the phonemic representation of these words must be error-free, i.e., it has to correspond to the standard pronunciation dictionary for German (*Duden* (ref 1)). Nevertheless, an additional system which allows for reconstructing such a required error-free phoneme word from a non-standard phone code word is still in the process of being developed. Due to the definition of the input and output interface the phoneme-to-grapheme conversion system has to be interpreted as part of a blackboard model which collects and uses all information available on the speech-to-text conversion.

PROCESSING STEPS FOR PHONEME-TO-GRAPHEME CONVERSION

Since the phoneme-to-grapheme conversion system is based on rules, any input sentence can be processed, even if such a sentence contains nonsense words. These phoneme words - whether they are error-free or not - are converted to a graphemic equivalent by applying an elaborated rule system. A correct conversion, however, is only guaranteed if the input phoneme word is equivalent to its standardized transcription code. Currently, the system has been implemented to use CPA-coded input strings (cf. ref 2). When such a CPA-coded string is fed to the system, it is segmented into consonant (C) and vowel (V) clusters; blanks are interpreted as markers for word boundaries. After that, each phoneme cluster is matched against the entries of a list containing the basic C and V clusters as well as very specific rules for determining the correct orthographic representation of these clusters.

Basing the whole conversion process on a relatively small list of fundamental clusters turned out to be very efficient. Use can be made of the fact that - by concatenating single phonemes to C or V clusters - the number of units to be processed is highly reduced. Amongst others, this can be explained by articulatory restrictions (cf. ref 3). Compared to the number of all possible C-C or V-V combinations the number of existing clusters is very small. The criterion of the number of entries, however, is not sufficient for justifying the choice of the unit 'cluster' for analysis since, e.g., a method of using single phonemes as the basic entity for the conversion process is in fact based on a very small list of phonemes, but it requires a large rule-system with very specific rules. Thus, in order to assess this approach appropriately, the number of list entries must be correlated to the number of rules which are necessary to determine the correct output.

*Lehrstuhl für allgemeine Elektrotechnik und Akustik,
Ruhr-Universität Bochum, PO Box 102148, D-4630 Bochum 1, FRG

Where the efficiency of the rule system is concerned the C/V cluster approach is advantageous not only because there is a reduction as to the concatenation of single phonemes to a cluster but, additionally, because certain clusters are bound to a certain position within a word, i.e., they occur in an initial, medial, or final position (cf. ref 4). On the one hand this information can be made use of for the determination of cluster boundaries in compound words (e.g., if a final C cluster is followed by an initial C cluster as in the word / hIm@l-blA/ / (sky-blue)); on the other hand the utilization of the information on the actual position of a cluster within a phoneme word again leads to a significant reduction of rules (e.g., the initial C cluster /t/ will never be written as <tt>).

THE RULE FORMAT

The rule system consists of an alphabetically ordered list of phoneme clusters to each of which only one or a set of possible graphemic rewriting(s) is assigned. In general, these rewritings are sorted according to the definiteness of the phonemic context: The most specific case is listed first, the most general is listed last. The rules determine that graphemic rewriting that is adequate to the input phoneme word by evaluating the context of the cluster; they define the direct syntactic context for each cluster. The size of the context is as large as necessary, but as small as possible in order to avoid redundancy. Generally, the system produces only one graphemic representation for each cluster, and thus only one grapheme word is produced. In those cases, however, in which one phoneme word can be written correctly in two different ways (homophones) all these different rewritings are produced (e.g. <ferse> (heel) / <verse> (verses)). For sake of a better understanding the following extract from the rule system may serve as an example. It is cut down to the most interesting cases (the focus here lies on the cluster /f/):

| PHONEME CLUSTER | GRAPHEME CLUSTER | RULE |
|-----------------|------------------|-------------------|
| /ea:/ | <ea> | (;*;) |
| /f/ | <ph> | (!lozo*i;;) |
| | <f> | (&*ER/z@#) |
| | <v> | (&*ER/z@#, ;*i;!) |
| | . | . |
| | <f> | (;*;) |
| /fR/ | <fr> | (;*!) |

The characters in the column 'RULE' which do not belong to the CPA have the following meaning:

- ; there may be none, one, or more phoneme(s) either before or after the cluster or context
- * marks the cluster which has to be converted
- ! at this position there must be at least one additional phoneme which, however, has not to be specified exactly
- # marks a word boundary

& indicates that the following string belongs to an ambiguous word and that searching has to be continued.

This extract shows that there are exception rules as well as general rules for each cluster. The general rule is applied in all those cases where exception rules do not fit. This guarantees a conversion of any input phoneme word. The risk here, however, is that not all of these grapheme words are necessarily correct, but experience shows that only in very rare cases an incorrect output is produced. If a cluster has been converted incorrectly, an additional exception rule can easily be inserted.

A SAMPLE CONVERSION PROCESSING

For sake of a better understanding of the segmentation and conversion of a phoneme word, the following description of the successive processing steps may serve as an example:

- input: error-free phoneme word

```
/ ?a/z@nba:nkno:t@npUNkt /      (<Eisenbahnknotenpunkt>)  
                                (railway junction)
```

- segmentation: determination of V and C cluster boundaries (marked by '-')

```
/ ? - a/ - z - @ - nb - a: - nkn - o: - t - @ - np - U - Nkt /
```

- matching of each cluster against the rule system

- assignment of the graphemic equivalent corresponding to each phoneme cluster

```
/ ?- a/ - z - @ - nb - a: - nkn - o: - t - @ - np - U - Nkt /  
< ei s e nb ah ... o t e np u nkt >
```

If a cluster match cannot be done (in the example above, the cluster /nkn/ was not found), this string is segmented into smaller sub-units by inserting an additional sub-cluster boundary: Starting out from the end of the cluster the last phoneme is cut off and then these two new sub-clusters are examined again; if still no matching can be done, the original string is segmented again until the remaining left cluster is reduced to the size of 1 phoneme.

```
/ nkn /  
/ nk - n /  
/ n - kn /
```

If this searching process is not successful, the string is segmented into its constituting phonemes; their occurrence in the rule system is guaranteed:

```
/ n - k - n /
```

- output: grapheme word

```
< eisenbahnknotenpunkt >
```

GENERAL REMARKS

The PTG rule system is based on a text corpus of about 25.000 words (text type 'office communication'). Currently, 651 clusters have been defined, and 3.600 rules are needed to convert them correctly. Generally, the analysis of another corpus will not lead to a proportional increase of the number of rules; the number of additional rules for new corpora is decreasing asymptotically to a minimum. This is, of course, dependent on the text type; if the text contains many technical terms, e.g., medical terms, possibly many exception rules have to be inserted since the general rules for these words being mainly of Latin origin are not always the same as those for the German language. Nevertheless, any input phoneme word, even a foreign word, can be converted after having formulated rules in this format.

Currently, an additional rule system (a 'text normalizer') is being developed. The task of this module is to adequately represent numbers, abbreviations, time and date entries etc. The PTG system converts phonemically transcribed numbers (e.g., / dRa/C/e:n / (thirteen)) to a character string (<dreizehn>). The task of the text normalizer is to write down such character strings as digits <13>. A similar procedure must be applied to abbreviations, which have to be converted to sequences of single characters, and other special entries.

1. Duden Aussprachewörterbuch, 2. Auflage, Mannheim (1974)
2. Kugler-Kruse, M.: Computer Phonetic Alphabet, unpublished manuscript Bochum (1985)
3. Scholes, R.J.: Phonotactic Grammaticality (1966)
4. Mangold, M.: Anlaute und Auslaute im Deutsch, in: *Phonetica Sarraviensia* (1976), pp 1-33