BROAD COVERAGE AUTOMATIC MORPHOLOGICAL SEGMENTATION OF GERMAN WORDS

T. PACHUNKE, O. MERTINEIT, K. WOTHKE, R. SCHMIDT

IBM Germany
Heidelberg Scientific Center
Tiergartenstr. 15
D-W-6900 Heidelberg

ABSTRACT

A system for the automatic segmentation of German words into morphs was developed. The main linguistic knowledge sources used by the system are a word syntax and a morph dictionary. The syntax is written in the formalism of right linear regular grammars and comprises approximately 1,400 rules describing the set of those sequences of morph classes which underlie syntactically well formed words. The morph dictionary contains almost 11,000 morphs. Each morph is assigned to up to 6 morph classes. - Statistical evaluations with 6000 test words showed that more than 99% of the segmented words got a correct segmentation.

1 INTRODUCTION

IBM Scientific Center Heidelberg is developing a large vocabulary speech recognition system for German (Wothke et al. 1989). The system needs for each word of its reference vocabulary two types of reference patterns:

- prototypal acoustic reference patterns.
- phonetic transcriptions of the main pronunciation variants of the word.

Up to now the transcriptions were generated for each orthographic word of the reference vocabulary by an automatic procedure having two drawbacks which caused a high amount of manual revision for the generated transcriptions:

- For each word only one transcription was generated. Our speech recognition system, however, needs at least the most significant pronunciation variants of each word.
- The automatic procedure took into account only the letter context of each letter to determine its transcription. In German, however, the transcription of a letter is very often also dependent on its morphological context. - Most of the transcription errors of the former system were a consequence of the fact that the system did not have any information about the morph structure of the words.

To reduce the manual work necessary to revise the transcriptions we currently develop a system with the following new features:

1. An orthographic word is first segmented into its morphs.
2. In a second step one or more phonetic transcriptions are produced for each segmentation of the word using letter-to-phone rules which can refer to the morph structure detected in the first step.

The following paragraphs will deal with the *first step*. We will mainly restrict ourselves to the *linguistic knowledge* incorporated in our current morph segmentation system. The *overall architecture* of the segmentation system and *details of the segmentation algorithm* are described in Wothke/Schmidt (1991).

A morphological segmentation procedure for German has to deal with the following **basic features of German morphology**:

- *Composition.*
- *Derivation.*
- *Inflexion.*
- *Ambiguous morph structure*: Some words can be segmented in several ways.
- *Reduction of consonant triples*: If two lexical morphs are concatenated, where the first morph ends in a vocalic letter and two identical consonantal letters and the second morph starts with the same consonantal letter and a vocalic letter, then the result of the concatenation does not contain the consonantal letter three times but only twice. - The inverse process, i.e. *trebling of double consonants*, has to be carried out, when segmenting such words.

Figure 1 shows the **architecture of the morphological segmentation system**. The interpreter for the segmentation has 5 main input files:

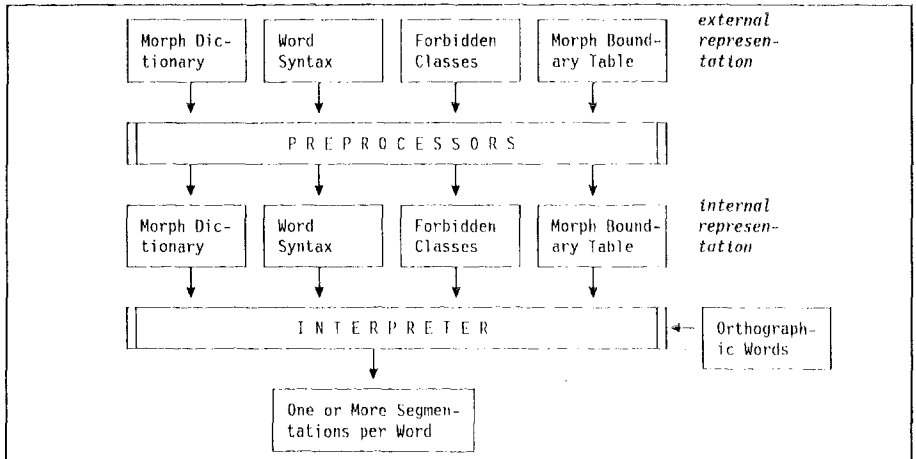

Figure 1. Architecture of the Morphological Segmentation System

- A morph dictionary containing information about the morph class/es each morph belongs to.
- A word syntax represented in the formalism of right linear regular grammars. It has to describe the set of those sequences of morph classes which underlie words.
- A morph boundary table, where the user can specify the symbols used by the interpreter to mark the different kinds of morph boundaries. We specified that
  - \+ is inserted before a prefix,
  - = is inserted before a lexical morph,
  - % is inserted before an infix, a derivational, or an inflexional suffix,
  - _ is inserted before a Latin or Greek derivational suffix,
  - ~ is inserted before a French or English derivational suffix.
- A table of forbidden classes, where the user can enter the names of those morph classes which may not attract either of the three identical consonantal letters arising from consonant trebling (i.e. infix classes, suffix classes, and prefix classes).
- A file containing the orthographic words to be segmented into morphs.

The linguistic knowledge in the first 4 files exists in 2 representations:

- An external representation which is created by the user of the system and which is human readable.
- An internal representation which is automatically generated by a preprocessor from the external representation and which is more suitable for the processing by the interpreter.

The interpreter loads the internal representations of the 4 files and segments orthographic words according to the knowledge in the files. If a word is morphologically ambiguous, several segmentations are generated.

2 THE LINGUISTIC KNOWLEDGE

The main linguistic knowledge sources of the system are the morph dictionary, which contains information about the morph class/es each morph belongs to, and the word syntax. We developed a classification scheme for German morphs and a suitable word syntax. The significant step to our current version was the classification of an extensive German morph list based on about 9,000 morphs compiled by the Institut für deutsche Sprache in Mannheim (Germany). We merged these morphs with an experimental list of about 2,200 morphs which we used in the former versions of our system. Additionally, we increased the resulting list up to almost 11,000 entries by many foreign morphs.

It turned out that for the *manual* development of the syntax the formalism of finite state networks is easier to handle than a right linear regular grammar. So we first represented the syntax with a finite state network, which finally was translated into a functionally equivalent right linear grammar.

Syntax and Classification Scheme used		Third	Fourth
Syntax	Number of states	129	289
	Number of arcs	1,050	1,368
Classification Scheme	Number of morphs	2,200	10,784
	Number of morph-classes	183	198

Table 1. Overview of the Extent of the Syntaxes and Classification Systems Developed

So far, we have developed and tested successively four classification schemes, each with a new, better syntax. We describe the third and fourth scheme, which are of actual interest (cf. table 1).

The substructures of the entire transition net dealing with the word classes verb, adjective, noun etc. will be called *verb net, adjective net, noun net* etc. We should stress that these substructures are not independent automata with any separate input. Nevertheless we call them nets; parts of these nets will be called *subnets*. - Although the word formation of the different word classes is not fully distinct and does share some substructures, it was not possible to design the entire net in such a way that the nets for these word classes physically share some subnets. Instead physical copies of common subnets had to be created for each occurrence of such a subnet in each of the nets. This is since we used a finite state network for the representation of the word syntax. This formalism does not allow to activate from different points one common subnet and afterwards to return to the appropriate activation point.

We will limit the following description to the nets for those word classes with productivity in word formation.

Verbs

Our Verb Net is responsible for the segmentation of *finite verbs*. Those of its subnets containing stem-labelled arcs refer to different combinations of mood, tense, and weakness vs strongness of the verb stem. Each of these combinations demands specific inflexional endings. - Weak stems are tense-invariant, strong stems can vary - according to tense - by vowel gradation (Umlaut or Ablaut). As a consequence, the classification of strong verb stems is oriented towards their suitability for certain tenses. For example, <=ging> is an imperfect tense form of <=geh%en> (engl. to go). In our morph dictionary the two morphs <geh> and <ging> are two independent entries, each with its own tense-oriented classification. Weak stems are classified according to prefixation and derivation needs. We took into account three groups of derivational suffixes: 1) <-el>, <-er>, 2) <-ig>, <-lich>, and 3) <-ier>. In the area of verb prefixes, one problem solved in our current version was to avoid the splitting of particular prefixes, e.g. *<+her+unter=geh%en> apart of the correct segmentation which is <+herunter=geh%en> (engl.: to go down).

In German, each infinitive can take the role of a noun, and each participle can do the same after being inflected. As a consequence, the part of our transition net related to *infinite verbs* is integrated into the Noun Net.

The set of verb stem classes had to be expanded for our current version to implement composition restrictions concerning verb stems as *parts of nouns*. For example, we had to cope with missegmentations such as *<+Er=find+er=schon%ung> apart of the correct segmentation <+Er=find%er=schon%ung> (engl. careful treatment of inventors). At least two restrictions exist: Firstly, the verb stem <find> is not allowed before a noun (which the word *<Erschonung> would be, if existing) but, e.g., the originally identically classified verb stem <bind> is, as in <Bindfaden> (engl. string). Secondly, the morph <er> is no suitable prefix for the verb stem <schon> but, e.g., for the originally identically classified verb stem <schein>, leading to <erscheinen> (engl. to appear). Verb-stem-related restrictions like these, which we implemented in our system by adding morph classifications to the existing ones, are only relevant for nouns. In the first case mentioned above, this is obvious. The second restriction does not concern finite verbs, because missegmentations only occur when the morph <er> is positioned between two stems. At the beginning of a word, the morph <er> can be seen as a prefix without any restrictions.

Adjectives

The adjective net consists of three subnets, each representing a possible way of adjectival derivation in German.

1. Simple adjectives like <schnell> (engl.: fast), <schön> (engl.: beautiful) etc.

These stems can be compared and inflected. Some stems occur only in a certain degree of comparison like <bess> (stem of engl.: better), <bes> (stem of engl.: best). They have obligatory comparative or superlative suffixes while the corresponding stems of the positive degree must not be followed by these suffixes, like e.g. <gut> (engl.: good).

2. Adjectives derived from verbs or verbal stems like <+be=geh%bar> (engl.: passable)
3. Adjectives derived from nouns. Example: <=held%en%haft> (engl.: heroic).

As a peculiarity of German word formation, a past participle may be compared and inflected like an adjective stem. Example: The past participle of <gelingen> (engl.: to succeed) has the comparative forms <+ge=lung%en>, <+ge=lung%en%er>, <(am) +ge=lung%en%st%en>. These may be translated as 'successful, more successful, most successful'.

Roughly speaking, the concept of the adjective net is to allow an adjective stem to be substituted by more complex constructions, like the ones described above. Special subnets are existing for adverbs and for adjectives with non-German stems. The latter is needed for marking foreign suffixes like in <=parall_el> because these suffixes attract the word accent, which in German causes a vowel to be pronounced long.

Nouns

A very productive feature of German word formation in the area of nouns is composition: New nouns may be formed by concatenation of lexical morphs, optionally interspersed with prefixes, suffixes, and infixes. In our noun net, this feature is modelled by loops over lexical morphs which can be left by inflexion modules to reach a final state and which cross infix modules (including zero-infix), prefix modules, and (derivational) suffix modules.

Inflexional suffixes occurring in compound nouns between lexical morphs are treated by us the same way as infixes.

Noun stems are classified according to the features umlaut, etymology (German vs not German), obligatory affix, inflexional suffix, and composition needs.

Foreign Words

Each of the described nets contains subnets dealing with the formation of foreign words which are involved in German word formation, e.g. <=mum_if_iz%ier%en> (engl. to mumify), <=Bas_is> (engl. basis), <=Port~ier (French; engl. porter).

Foreign words without connection to German word formation and names are not intended to be segmented by our system. So an unsegmented word is not necessarily a system failure but can be a required rejection (cf. Table 3).

3 EVALUATION

The experimental morph list used during the development of versions 1, 2, and 3 of our syntax and classification scheme consisted of 2,200 morphs mainly selected from Ortmann (1985). In the fourth system the morph list was extended to almost 11,000 morphs (cf. Table 1). To evaluate the different versions (each consisting of syntax and classification scheme), two word sets were used each containing 3,000 words. The first set consisted of rank 1-1,000, 300,001-301,000, and 600,001-601,000 of a frequency list sorted in descending order which was created from a corpus containing articles of a German business newspaper with about 31,000,000 running words. This set was used for the iterative improvement of our system. It is called *test set*. The second set, serving as a *control set*, contained rank 1,001-2,000, 200,001-201,000, and 400,001-401,000 of a corpus obtained from a common newspaper with about 13,200,000 running words. The control set was necessary because of the risk that the later versions were designed in such a way as to cope only with those errors which arose when applying the earlier versions to the test set.

Table 2 shows the improvement in coverage mainly achieved by extending the morph list: While - referring to the control set only - the third system segmented only 1,425 of the input words (=47.5%), the fourth system segmented 2,492 words (=83%). The quality of the segmentations of the fourth system is a little worse. The reason for this effect is the larger morph list allowing more nonsense concatenations of morphs. Although we made grammar and classification system more restrictive, it would have been too costly to strive for equal or better segmentation quality.

Table 3 gives an overview of the words which were not segmented by the fourth system. Many of them are proper names.

Syntax and Classification Scheme used	Third				Fourth			
Corpus used	'business newspaper' (test set)		'common newspaper' (control set)		'business newspaper' (test set)		'common newspaper' (control set)	
Number of words segmented and percentage related to total number of 3,000	1,955	65.2%	1,425	47.5%	2,715	90.5%	2,492	83.1%
Number of segmentations and ratio of segmentations obtained per segmented word on average	2,026	1.04	1,533	1.08	2,960	1.09	2,815	1.13
Number of correct segmentations and percentage related to total number of segmentations	1,991	98.3%	1,489	97.1%	2,854	96.4%	2,656	94.4%
Number of wrong segmentations and percentage related to total number of segmentations	35	1.7%	44	2.9%	106	3.6%	159	5.6%
Number of words with at least one correct segmentation and percentage related to number of segmented words	not evaluated		not evaluated		2,710	99.8%	2,482	99.6%

Table 2. Results for Words Segmented by the Third and Fourth Version of the Segmentation Procedure

Corpus used	'business newspaper' (test set)		'common newspaper' (control set)		both (test + control set)	
Number of words rejected	285 (of 3,000)		508 (of 3,000)		793 (of 6,000)	
- words which should be segmented	33	11.6%	215	42.3%	248	31.3%
- words which cannot be segmented	252	88.4%	293	57.7%	545	68.7%
- words with spelling errors	19	6.7%	8	1.6%	27	3.4%
- foreign words which are not used in German	55	19.3%	17	3.3%	72	9.1%
- names	178	62.5%	268	52.8%	446	56.2%

Table 3. Results for Words Rejected by Fourth Version of the Segmentation Procedure

4 CONCLUSION

A morph classification scheme and a syntax for the segmentation of German words were developed, with which more than 99% of the segmented words were correctly segmented, i.e. at least one resulting segmentation was correct.

With the extended morph list used about 13.2% of the words (= 793 out of 6,000) were not segmented. Out of these unsegmented words, 68.7% a priori could not be segmented, i.e. no conceivable segmentation procedure can segment these words.

ACKNOWLEDGEMENTS

We thank Georg Walch, who put at our disposal the frequency lists mentioned in chapter 3. We would also like to express our gratitude for the encouragement received from our manager Eric Keppel.

REFERENCES

Ortmann, W.D. (1985): Wortbildung und Morphemstruktur hochfrequenter deutscher Wortformen. Teil I. München.

Wothke, K. / Bandara, U. / Kempf, J. / Keppel, E. / Mohr, K. / Walch, G. (1989): The SPRING Speech Recognition System for German. In: Eurospeech 89. European Conference on Speech Communication and Technology. Paris - September 1989. Edinburgh. Vol. II. pp. 9-12.

Wothke, K. / Schmidt, R. (1991): A Morphological Segmentation Procedure for German. In: Proceedings of the International Conference on Current Issues in Computational Linguistics. Penang (Malaysia). pp. 137-147.