

## MORPHOLOGICAL SEGMENTATION AND STRESS CALCULUS IN GERMAN WITH AN EXPERT SYSTEM

Henri Zinglé

Université de Nice - BP 369 06007 Nice Cedex  
INRIA - BP 109 06561 Valbonne Cedex

### ABSTRACT

Morphological analysis and stress assignment are important stages of linguistic preprocessing in a text-to-speech system for German. In this paper we present an expert system in PROLOG which is intended to solve both problems as well as its linguistic basis. The underlying idea is that word segmentation in German can give multiple correct solutions - with different possibilities of spelling - and that the expert system methodology is a good framework for such class of problems.

### 1. INTRODUCTION

In 1982 we pointed out that the linguistic preprocessing in a text-to-speech system for German has to take in account morphology and stress. In short the grapheme to phoneme conversion needs morphological informations (for example <e> correspond to [ ] in the prefix be- of "betonen", to [ ] in the prefix er- of "erhalten" and to [e] in the root bet of "beten") and the vowel length depends on stress position (for example [o] is long in "Hof" and short in "hofieren"). The linguistic preprocessing system we had developed at that time showed good results for derivatives, but didn't deal with compound words which were manually marked.

After several tentatives based on left-to-right word parsing associated with linguistic heuristics (Zinglé, 1986) our interest was directed on expert system methodology. Word segmentation is a typical problem of form recognition: the way to get one or more correct solution(s) is not known beforehand and the problem of multiple solutions (cf. <geb-et> or <ge-bet> for "Gebet", <nacht-eile> or <nach-teil-e> for "Nachteile") accords with non-determinism which is inherent in expert-systems.

The prototype presented here is conceived close by linguistic facts. Our paper reviews first the linguistic knowledge about word structure in German and presents afterwards the conception of the expert system itself.

### 2. THE LINGUISTIC FACTS

#### 2.1. The structure of German words

German words are of 3 categories

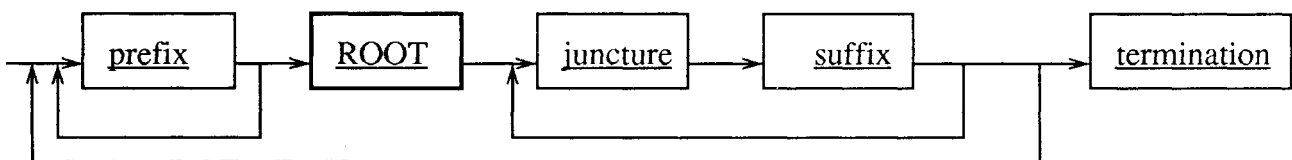
- simple words (SW) based on one root like "dort", "Vater", "Forelle", "Kartoffel"

- derivatives (DW) which are combinations of one root with prefix(es) and/or suffix(es) like "Anbau", "auseinander", "unangenehm", "Armlosigkeit", "unverantwortlich

- compounds (CW) resulting from the combination of SW and DW : "Hauptbahnhof", "Buchhandlung", "Ausfuhrland"

CW can include foreign roots (cf. "Motorboot", "Dialektforschung", "Elektrochemie"). We may consider that a DW is a case of CW and a SW a case of DW. SW, DW and CW can be followed by a termination (T) which is either a case or a conjugation ending. DW and CW sometimes include a juncture (J) : "Ländeskunde", "Nummernschild", "Schweinerei"; a given J is not obligatory bound to the preceding root (cf. "allezeit", "allenfalls", "allerhand", "allenthalben").

In short the structure of german words has to be conform to following diagram



□ optional  
 □ obligatory

It must be pointed out that

- a root is not obligatory bound to a syntactic category (the noun "Gras", the adjective "grasig" and the verb "grasen" has the same root)

- a given segment can represent more than one morphem (<er> is either suffix, juncture, suffix or termination)

-it may exist more than one segmentation for a given word (cf. <be-in-halten> and <beinhalten> for "beinhalten", <ball-laden>, <ball-lad-en> and <ballade-n> for "Balladen")

## 2.2 Rules for stress assignment

We make a difference between ictus (potential stress place) and stress itself (actualization).

- **R1** SW have a unique ictus; if the word is stressed only the syllable marked by ictus will become proeminent ("V.ater" -> "V'ater", "Hol.under" -> "Hol'under")

- **R2** In DW it is necessary to distinguish between prefixes and suffixes with no ictus( be-, ge-, ent-, zer-, -chen, -sel, -ern ...) and these with ictus (an-, aus-, bei-, -bar, -los ...)

**R21** If elements of the first category are added to the root R1 is applied ("b.et" -> "Geb'et")

**R22** If elements of the second category are used the word has several ictus; only one of them can be stressed; ictus-syllables left from stress become unmarked (Philip,1970;

Zinglé,1977), contrary to ictus-syllables under stress or right from it (which has for consequence that tense vowels in that position are lengthened) : ".um"+"z.iehen" -> "'umz.iehen", ".um"+"s.onst"->"ums'onst", "r.at"+"l.os"-> "r'atl.os", "fr.oh"+"l.ocken" ->"frohl'ocken". Some suffixes like -ant, -ie, -ei, -ur etc.. induce systematically the lost of the ictus on the root (cf. "Lagerist" and "Glasur").

-R3 For CW the situation is near to that of R22 ("v.iel"+"m.ehr"->"vielm'ehr", "v.iel"+"s.eitig"-> "v'iels.eitig").

The consequence of the preceding rules is that in all cases the first ictus of a word will be stressed. Adding a termination has no effect on the syllable hierarchy established in this way except for any foreign words (cf, "D'oktor" vs. "Dokt'oren").

### 3. THE EXPERT SYSTEM

Our expert system, written in PROLOG, follows a backward chaining strategy and is classically structured in facts, rules and metaknowledge.

#### 3.1. The facts

In difference to Zinglé(1989) all facts are regrouped now in an external database, where are defined all possible segments (prefixes, suffixes, roots, junctures and terminations) and their status in regard to ictus.

The choice of an external database was made because the number of segments is too important and because the classical fact assertion in PROLOG not a good solution is here for memory management problems. In order to have a quick lookup the facts are accessed via keys ordered in a BTree.

#### 3.2 The rules

There are 2 types of rules

- rules to check word structure : these rules define in application of 2.1 what is a compound, a derivative a prefixed derivative and a suffixed derivative. These rules are not used to build up possible segmentations but to select only correct segmentations proposed by 3.2 below
- rules for stress calculus which apply 2.2. Examples of stress rules :

a root lose its ictus

if it is followed by one suffix of the set [-ei, -ist ...]

a root ending with -or has its ictus on -or

if it is followed by a termination -en

else the ictus is on the preceding syllable

Rules are marked if they are exclusive of others or not (in the last case more than one solutions can be build).

#### 3.4 Metaknowledge

This level defines the global strategy of problem solving and the processes bound to it. The strategy is expressed as follows

a word is segmented and hierarchized  
if all possible segments occurring are looked up  
and if all type segment lists have been build which may be unified with the word  
and if correct solutions have been checked  
and if stress rules have been applied

This process is written in manner to obtain the better yield of the natural backtracking mechanism of the inference engine of PROLOG.

Other rules define lower tasks as trigraph resolution like in "Bettuch" and database lookup.

Examples of segmentations

GEBET

[p(ge,0),l(bet,2)]

[l(geb,2),t(et,0)]

NACHTEILE

[l(nacht,2),l(eile,1)]

[p(nach),l(teil,2),t(e,0)]

## CONCLUSION

With this system we obtain largely better results as with the left-to-right parsing with linguistic heuristics. However the expert system produce solutions that are linguistic correct in the that sense that they conduct to possible words with possible meanings (cf. our example "Balladen" in 2.1). Our aim is now to valid the segmentation in regard to an analyzed sentence in the context of an multi-expert system, in which the present expert-system will take its place.

## BIBLIOGRAPHY

PHILIP M. Phonologie de l'allemand, P.U.F, 1970

ZINGLE H. L'accent en allemand, Thèse 3ème cycle, Université de Strasbourg II, 1977

ZINGLE H. Traitement de la prosodie allemande dans un systeme de synthese de la parole  
These d'Etat, Université de Strasbourg II, 1982

ZINGLE H. Définition d'une méthode d'identification des segments linguistiques : application à l'allemand, Travaux du Laboratoire Langues et informatique de l'Université de Savoie, 1986, 104-119

ZINGLE H. Analyse morpho-accentuelle de l'allemand à l'aide d'un système expert,  
Mélanges P. SIMON, Publications de l'Institut de Phonétique de Strasbourg, 1989,  
833-853