# The application of two-level morphology to non-concatenative German morphology

**Harald Trost**

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)[1]
Stuhlsatzenhausweg 3, D-6600 Saarbrücken 11, FRG
Email: htrost@sbuvax.campus.uni-sb.de

## Abstract

In this paper[2] we describe a hybrid system for morphological analysis and synthesis. We call it hybrid because it consists of two separate parts interacting with each other in a well-defined way. The treatment of morphonology and non-concatenative morphology is based on the two-level approach originally proposed by Koskenniemi (1983). For the concatenative part of morphosyntax (i.e. affixation) we make use of a grammar based on feature-unification. Both parts rely on the same morph lexicon.

Combinations of two-level morphology with feature-based morphosyntactic grammars have already been proposed by several authors (c.f. Bear 1988a, Carson 1988, Görz & Paulus 1988, Schiller & Steffens 1990) to overcome the shortcomings of the continuation-classes originally proposed by Koskenniemi (1983) and Karttunen (1983) for the description of morphosyntax. But up to now no linguistically satisfying solution has been proposed for the treatment of non-concatenative morphology in such a framework. In this paper we describe an extension to the model which will allow for the description of such phenomena. Namely we propose to restrict the applicability of two-level rules by providing them with filters in the form of feature structures. We demonstrate how a well-known problem of German morphology, so-called "Umlautung", can be described in our approach in a linguistically motivated and efficient way.

## Introduction

Conventional morphosyntactic grammars do not allow to describe non-concatenative parts of morphology declaratively. Two-level morphology on the other hand can deal with some of these phenomena like vowel or consonant change, but there is no sound way to transmit information to the morphosyntactic grammar. This leads to quite unnatural solutions like the use of diacritics for the representation of morphosyntactic phenomena.

While German morphology is mainly based on concatenation, some non-concatenative phenomena do exist. The most prominent example is so-called "Umlautung". Umlautung means that in the surface form the original stem vowel is replaced by another vowel in a systematic manner. Possible transformations are a=>ä, au=>äu, o=>ö, u=>ü, and - in some cases - e=>i.

Umlautung in German realizes quite different morphological features. With nouns it can mark the plural either by itself (e.g. Mutter => Mütter) or in combination with an explicit plural-ending (e.g. Mann => Männer), depending on the inflection class. With adjectives it is used to mark comparative forms (groß => größer => am größten), again in combination with an ending, for verbs following strong conjugation it marks the subjunctive II and 2nd and 3rd person singular of the indicative present tense. Umlautung also occurs in derivation in combination with a number of derivational particles, e.g. *-lich* (klagen => kläglich). In contrast to its use in inflection, umlautung provides for no extra morphosyntactic information in derivational forms. At last, it appears in compounding in combination with some "Fugenelement" (joining element) (e.g. *Männerchor* - male chorus).

---

[1] Work on this project has begun while I was working for the Austrian Research Institute for Artificial Intelligence in Vienna, Austria

[2] I want to thank my collegues Greg Dogil, Wolfgang Heinz, Tibor Kiss and Günter Neumann for fruitful discussions and helpful comments on earlier versions of this paper.

There are two common ways to cope with umlautung in conventional morphological components for German. One is to treat all forms created by umlautung as suppletions, i.e. these forms are explicitly entered into the lexicon. This is linguistically inadequate, because it obscures the phonological similarity of the two forms. From a more practical point of view it has the drawback that in a few cases, e.g. forming of the diminutive with the derivational *-chen*, umlautung is still productive, and cannot therefore be lexicalized.

The other solution is a special function replacing (and interpreting) or generating the umlaut in all stems which are marked for umlautung required by the morphosyntactic context (c.f. Trost & Dorffner 1987). This makes umlautung a special case neglecting its status as a regular means of morphosyntactic marking.

Solutions within the two-level approach have also been proposed. They rely on the idea to represent stem vowels which exhibit umlautung with special characters (diacritics) (e.g. *A*) at the lexical level. These characters are then realized as either the regular vowel (e.g. *a*) or the corresponding umlaut (e.g. *ä*) at the surface level. The idea behind is that these stem vowels are lexically somewhat underspecified. To trigger the appropriate substitution, Görz & Paulus (1988) use a separate data structure to control for each word form which of the two possible rules is applied to create the surface structure. Schiller & Steffens (1989) use still another diacritic symbol for this task. Flexional endings triggering umlautung start with the diacritic $ (realized as 0 at the surface level). The context to the right of the substitution of all umlaut rules requires the occurence of that $. Therefore the umlaut rule would fail if no such affix follows the stem. As a consequence, the null morph must be explicitly represented by $ in lexical strings where morphosyntactic information is expressed by umlautung only (e.g. *Mutter => Mütter*).

Although both solutions certainly do work, at least for flexional morphology, they provide no clean and general solution for the integration of umlautung in the framework of two-level morphology. The use of a separate data structure is contrary to the intuition that umlautung is a regular phenomenon of German morphology, the treatment of which should require no extra mechanism. And the use of the diacritic $ places a burden on morphonology which clearly belongs to morphosyntax.

The handling of non-concatenative morphological phenomena within the two-level approach imposes two new requirements:

- Information about the application of a rule needs to be transferred to the morpho-syntactic grammar.
- It must be possible to restrict the application of two-level rules to certain classes of morphs.

Accordingly, we propose an approach where umlautung requires no extra mechanism at all and where no diacritics are (mis)used to describe morphosyntactic features. The basic idea is to provide two-level rules with a filter in form of a feature structure which controls its applicability. This feature structure has to be unified with the feature structure of the morph found in the lexicon to which the rule applies. In case of failure the two-level rule may not be applied. If unification succeeds information is transferred that way from the two-level part to the associated morphosyntactic grammar. This is crucial for the treatment of umlautung because, as mentioned above, its application conveys morphosyntactic meaning.

In the following we will describe the parts of our system in some detail and explain how umlautung can be handled using that framework. (Basic knowledge of the two-level approach and feature-unification is presupposed.) We will also argue that extending the two-level rules with filters facilitates the description of certain morphonological phenomena as well.

## The Two-Level Part

Our implementation of the two-level part is similar to the one proposed by Bear (1988a, b), i.e. rules are interpreted directly and not compiled into automata. Rules consist of a left context, a right context and a substitution. Left and right contexts are regular expressions over pairs of lexical and surface symbols. A substitution consists of exactly one such pair. Rules may be optional or obligatory (i.e. in contrast to Bear there are no disallowed rules). By definition, all default pairs are regarded as optional rules with empty contexts.

The pair of strings (lexical and surface) is processed from left to right. If more than one

optional rule is applicable at a time this shows an ambiguity, i.e. there are as many continuations as there are different substitutions. Obligatory rules supercede all optional ones (thereby pruning the tree of continuations). If more than one obligatory rule is applicable at the same time (enforcing different substitutions) the whole mapping must be discarded. The same is true if no rule applies at all.

The major difference from other two-level approaches is the possibility to provide the rules with a filter. A filter is an arbitrary feature structure. A rule may only be applied if the filter unifies with the feature-structure of the actual morph, i.e. the morph to which the substitution applies. Filters are used to restrict the application of a rule to certain classes of morphs. This is in contrast to the original view of Koskenniemi that morphonological rules are to be applied over the whole lexicon regardless of morphosyntactic considerations. This is certainly true of post-lexical rules. But there is evidence that it is not even true for all morphonological rules. Take e.g. the verb *senden* (to send), which can form two different past tenses *send-e-te* and *sand-te,* the former being regular weak conjugation, the latter a strong stem with weak inflection ending. The epenthesis of schwa (or *e* in orthography) depends on the morphological class of the stem (weak or strong). Or take the adjective *dunkel,* where the nominalization *im Dunk-e-ln* (in the dark) is different from the attributive use *den dunkl-e-n Mantel* (the dark coat) (c.f. Gigerich 1987). Here nominalization requires schwa epenthesis in the stem, not at the morph boundary like the adjective.

If we want to use two-level rules for the description of non-concatenative morphology, such filters are necessary anyway. Because, as mentioned above, we do need some means to convey information from the two-level part to the morphosyntactic grammar. In the case of umlautung we suppose that it is triggered by the concatenation of a stem which is lexically marked for umlaut (by the occurrence of a diacritical character *A, O, U* or *E*) with an affix allowing for umlautung (i.e. carrying the feature [umlautung: +]). Therefore the filter for all rules concerning umlautung basically contains the feature-value pair which marks affixes [umlautung: +/-].

Umlautung must only be performed if the stem allows for umlautung and that feature has the value +. Accordingly, all two-level rules substituting a vowel by its umlaut have the filter [umlautung +]. Corresponding rules are needed which keep the original vowel in the surface form. They have the filter [umlautung: -]. All the above-mentioned rules are obligatory, and exactly one of them applies to every occurence of a stem marked for umlaut (see figure 1).

| Rule: | Rule Status: | Rule Filter: |
|---|---|---|
| ⟨A→a⟩ | obligatory | [syn [loc [head [cat: stem] agr [umlautung: -]]]] |
| ⟨A→ä⟩ | obligatory | [syn [loc [head [cat: stem] agr [umlautung: +]]]] |

Fig.1: Two-level rules for lexical *a*

To select exactly one rule (in order to prevent a deadlock between two obligatory rules) the feature *umlautung* must always have a value. But, as stated above, stems are not lexically marked for this feature. The marking is effected by the morphosyntactic grammar, which unifies the agreement features of affix and stem, thereby transferring the feature *umlautung* to the feature structure of the stem (see figure 5).

What is important to note here, is that the two-level rules themselves contain no information about the morphosyntactic interpretation of umlautung. This is only specified in the grammar rules. Therefore the same two-level rules can be used for all the different places, where umlautung occurs. We will now shortly describe the morphosyntactic part of our system.

## Feature-Based Morphosyntax

Concatenative morphology is described in grammar rules following X-bar theory. A head-driven approach is adopted. The basic structure consists of a head, usually some sort of affix, and one or more complements, one of which must be some type of stem. We will not go into any detail concerning the exact format of the grammar rules here, because it is irrelevant for the treatment of umlaut. For the purpose of this paper it suffices to give just one example.

We will describe the overall structure of the grammar using the noun *Mann* (man) as an example. Nouns are constructed from stem, number marker, and case marker. The number marker forms the head and subcategorizes for a stem and a case marker. The relevant syntactic information is collected in the agreement feature which is passed upwards from the daugthers.

Figure 2 shows (a simplified version of) the number markers for [flex-class: er]. We can see that the plural marker triggers umlautung, while the singular marker does not. Both subcategorize for a stem and a case marker.

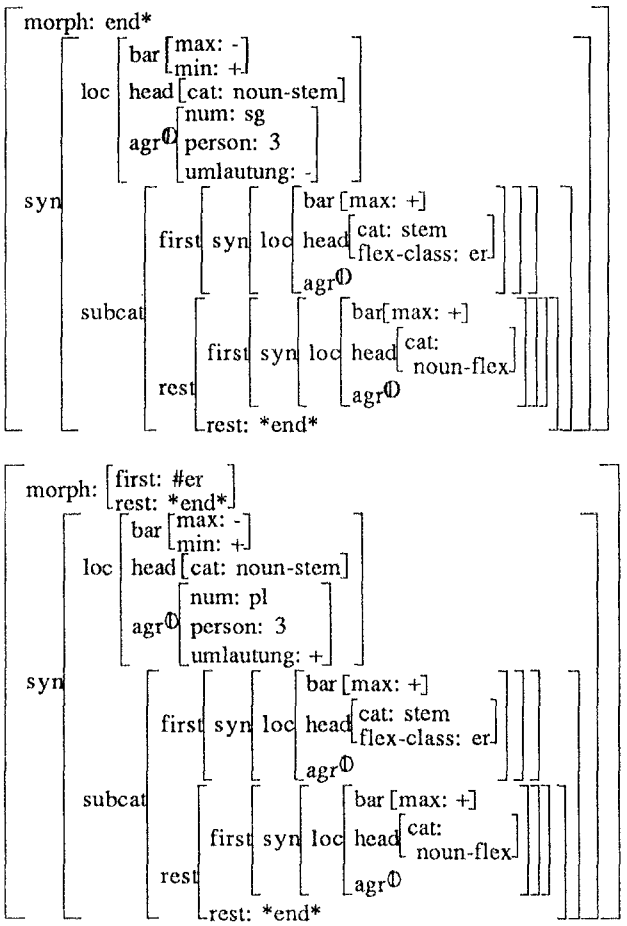


Fig.2: Number markers for nouns with plural -er (unmarked singular and plural #er)

Figure 3 shows the lexical entry for the stem *Mann*, which may take an umlaut (its stem vowel is *A*). The number marker takes that stem as a complement. The agreement featuresare shared between head and complement. As one result, the feature*umlautung* is transferred from the number marker to the stem. It is now locally available to trigger the correct umlaut rule.

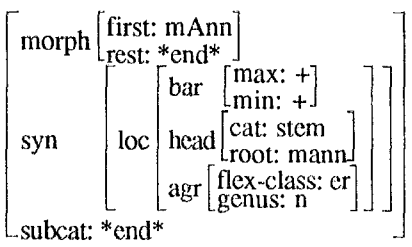


Fig.3: Lexical entry for Mann (man)

Figure 4 shows two different case markers for the unmarked case and for dative plural. After combining with a stem, the number marker may now take a case marker as its second complement.

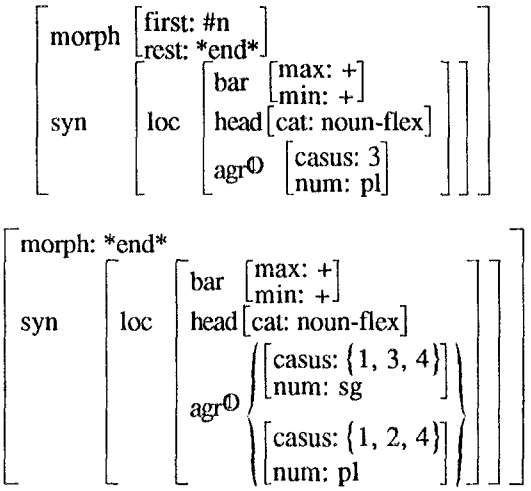


Fig.4: Case morphs for the unmarked case and for dative plural

The grammar fragment sketched in this section must be viewed with care. German inflectional endings often combine different morpho-syntactic information, e.g. with nouns case and number information is sometimes expressed by a single morph. Explaining the unmarked case as a combination of stem with nullmorph is also somewhat problematic. A more relistic grammar would probably collapse parts of the tree into a single structure, i.e. by using case&number markers, which subcategorize only for a stem. This would, for example, eliminate the problem which is posed to the parser by allowing for the occurence of more than one null morphs in one position (as is the case for 1st, 3rd and 4th singular where both number and case marker are realized by the null morph). Nevertheless, with regard to the handling of umlautung in our approach these problems are not relevant.

## The Integration into the Grammar

We will now show how the two parts of our system work together. Take e.g. the dative plural of *Mann* (man), *Männern*. For generation, the grammar part constructs the lexical string *$mAnn#er#n$* (# marks a morph boundary and *$* a word boundary), which is given to the two-level part. The relevant lexical information for the purpose of umlautung is the stem vowel *A* in *mAnn* , and the feature [umlautung: +] in *#er* . As described in the last chapter, by structure sharing this information

has already been enriched by the generation process providing *mAnn* with the feature [umlautung: +]. When reaching the stem vowel *A* the rules try to unify their filters with the feature structure of *mAnn*. Only the umlaut rule succeeds, generating the correct surface form *$männern$*.

Now one can also see why the (incorrect) form *Manner* will not be accepted by the parser. The filter of the obligatory rule *A => a* would add the feature [umlautung -] to the feature structure of *mAnn*. This inhibits the unification with the feature structure of *#er*.

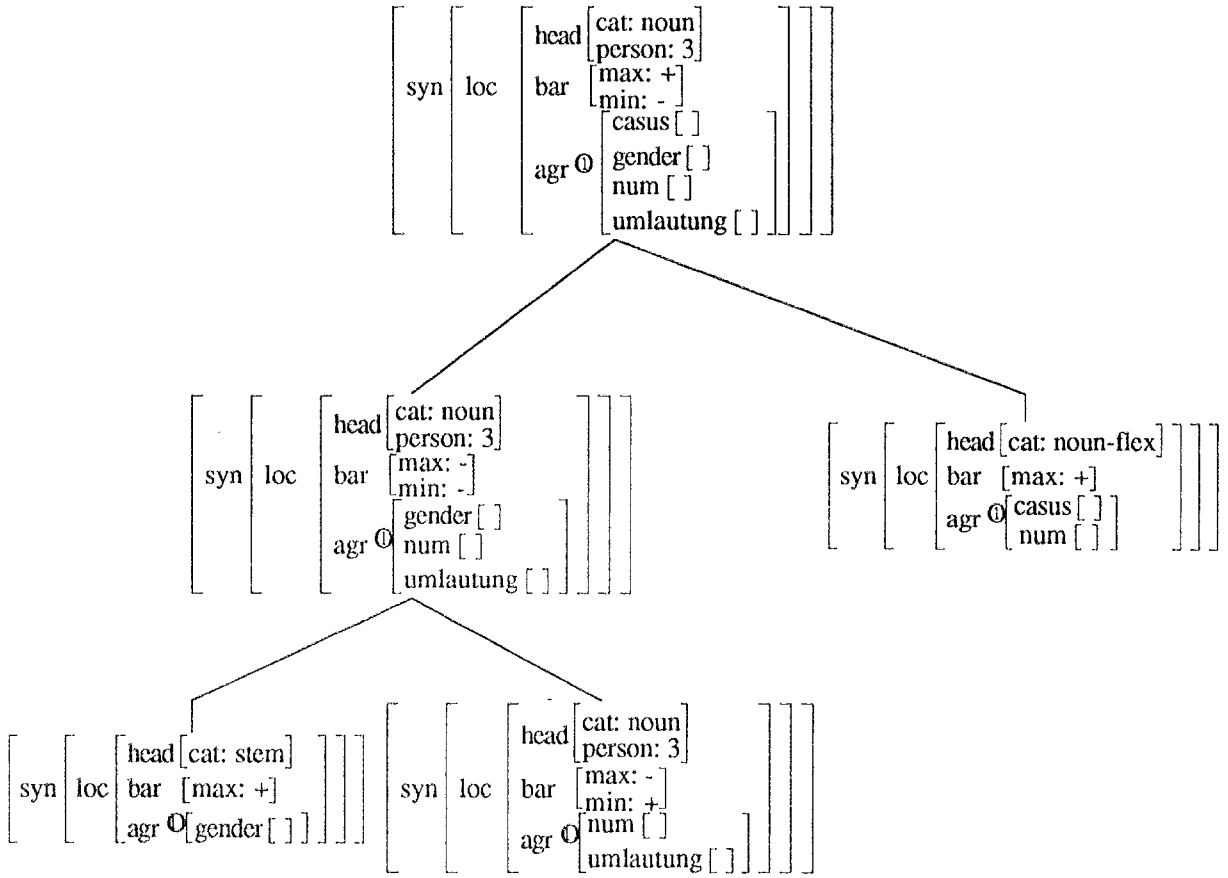


Figure 5: Basic tree structure created by the grammar for nouns

The analysis starts with the surface form *$männern$*. Because no morph has been recognized yet, both *mAnn* and *männ* are derivable (because of the default rule *ä => ä*). At that point the morph *mAnn* is found in the lexicon. The filter is successfully unified with the lexical entry, transferring the information that umlautung has taken place. Now a morph boundary must be created. A *0* is inserted in the surface form which is mapped to # in the lexical form. (There still remains the other possibility to look for a longer morph. This hypothesis can only be discarded when the end of the surface form is reached without finding a corresponding morph in the lexicon.) Next *#er* is found in the lexicon. Since that ending allows for umlautung (having the feature [umlautung: +]), the morphosyntactic grammar can combine it with the stem which has already been unified with the rule filter. Next the morph *#n* is recognized, completing the analysis path.

At this point we want to remind you of the fact that the morphosyntactic grammar uses the notion of null morphs for all unmarked forms (e.g. in our example the singular marker). Null morphs are not included in the lexical string though. They operate solely at the level of the morphosyntactic grammar. Take e.g. the generation of *Mann* (nom sg). Although only *$mAnn$* is created as lexical string, the null morph has enriched the associated feature structure with [umlautung: -] enforcing the generation of the surface string *Mann* by blocking the umlaut rule.

Analysis works in a similar way. When *$mann$* is input as surface string, it is mapped to the lexical string unchanged. It is now associated to the morph *mann* the feature structure of which has been unified with the rule filter providing it with [umlautung: -]. To create a legal word form it must now be

combined with number and case markers. These can only be null morphs and their agreement features must unify, which leads to the correct interpretation.

Another example from derivation shall demonstrate how well the marking of stem vowels and the feature *umlautung* work together to define the occurence of umlautung. The verb *klagen* (to mourn) shows no umlaut in any of its forms. The same is true for the nominalization *Klage*. But the derived adjective *kläglich* surprisingly exhibits an umlaut. A closer look shows that this behaviour conforms to what our system predicts. The morph *klAg* is marked as a stem which may take umlaut. Since all endings of weak verb conjugation are marked with [umlautung -] no umlautung takes place for any of the verb forms. The same is true for the noun plural ending *#n*. But *#lich* comes with the feature [umlautung +] triggering the umlaut-rule to produce the surface form *kläglich*.

Unfortunately in derivation and composition there are exceptions to the rule. Contrary to our expectations we find the adjective *handlich* derived from the noun *Hand*. Since the plural form of the noun is *Hände* the morph must clearly be stored as *hAnd* in the lexicon which would yield *händlich* which is incorrect. There are two solutions to this problem. One can take the stance that in such cases derivation is no longer transparent and that these words should be entered into the lexicon as a whole.

The other solution would be to introduce exception markers with such morphs which block the application of the umlaut rule (say [flex-uml-poss: -] for flexion and [deriv-uml-poss: -] for derivation). Instead of the single feature *umlautung* for all affixes we then need to mark flexional endings and derivational endings with the features *flexional-umlautung* and *derivational-umlautung* respectively. The rule filters become more complex too. Umlaut rules are equipped with the following filter: $\{$[flexional-umlaut: + ] [flex-uml-poss: +] / [derivational-umlaut: +] [deriv-uml-poss: +] $\}$, the corresponding no-umlaut rules get an according one. All morphs not explicitly marked will behave like before, i.e. take umlautung in both cases.

## Conclusion

We have shown a hybrid system for morphological analysis and synthesis, based on two-level morphology and unification-based grammar rules. By providing two-level rules with a filter in the form of a feature structure the application of these rules can be controlled by the morphosyntactic grammar in a consistent way. The filters are also used to transfer morphosyntactic information from the two-level part to the grammar. This allows the description of non-concatenative morphological phenomena using such rules without the use of (phonologically) unmotivated diacritics.

As an example, we have shown how our system can handle German umlautung in a linguistically satisfactory manner. Translation of the umlaut is performed by a two-level rule which is filtered by a feature *umlautung*. The morphosyntactic interpretation of the umlaut is only performed at the level of the grammar rules.

The proposed method can be applied to other non-concatenative phenomena as well. The idea of filters seems also to be a promising solution for morphonological phenomena which are restricted to certain classes of morphs (or words).

## References:


Bear J. (1988a): A Morphological Recognizer with Syntactic and Phonological Rules, COLING-86, Bonn, BRD.

Bear J. (1988a): Generation and Recognition of Inflectional Morphology, in: H.Trost (ed.), 4.Österreichische Artificial Intelligence-Tagung, Springer, Berlin, 3-7.

Carson J. (1988): Unification and transduction in Computational Phonology, COLING-88, Budapest, 106-111.

Gigerich H. (1987): Zur Schwa-Epenthese im Standarddeutschen, Linguistische Berichte 112, 449-469.

Görz G., Paulus D. (1988): A Finite State Approach to German Verb Morphology, COLING-88, Budapest, 212-215.

Karttunen L. (1983): KIMMO: A General Morphological Processor, Texas Linguistic Forum 22, 167-186.

Koskenniemi K. (1983): Two-level Model for Morphological Analysis, IJCAI-83, Karlsruhe, BRD, 683-685.

Schiller A., Steffens P. (1990): A Two-Level Morphology for a German natural language understanding system, IBM Stuttgart, manuscript.

Trost H., Dorffner G. (1987): A System for Morphological Analysis and Synthesis of German Texts, in: D. Hainline (ed.), New Developments in Computer-Assisted Language Learning, Crooms Helm, London.