# Separable Verbs in a Reusable Morphological Dictionary for German

Pius ten Hacken[1] & Stephan Bopp[2]

[1]Institut für Informatik / ASW
Universität Basel, Petersgraben 51
CH-4051 Basel (Switzerland)
email: tenhacken@ubaclu.unibas.ch

[2]Lexicologie, Faculteit der Letteren
Vrije Universiteit, De Boelelaan 1105
NL-1081 HV Amsterdam (Netherlands)
email: bopp@let.vu.nl

## Abstract

Separable verbs are verbs with prefixes which, depending on the syntactic context, can occur as one word written together or discontinuously. They occur in languages such as German and Dutch and constitute a problem for NLP because they are lexemes whose forms cannot always be recognized by dictionary lookup on the basis of a text word. Conventional solutions take a mixed lexical and syntactic approach. In this paper, we propose the solution offered by Word Manager, consisting of string-based recognition by means of rules of types also required for periphrastic inflection and clitics. In this way, separable verbs are dealt with as part of the domain of reusable lexical resources. We show how this solution compares favourably with conventional approaches.

## 1. The Problem

In German there exists a large class of verbs which behave like *aufhören* ('stop'), illustrated in (1).

(1) a. Anna glaubt, dass Bernard aufhört.
('Anna believes that Bernard stops')
b. Claudia hört jetzt auf.
('Claudia stops now PRT')
c. Daniel versucht aufzuhören.
('Daniel tries to_stop')

In subordinate clauses as in (1a), the particle *auf* and the inflected part of the verb *hört* are written together. In main clauses such as (1b), the inflected form *hört* is moved by verb-second, leaving the particle stranded. In infinitive clauses with the particle *zu* ('to'), *zu* separates the two components of the verb and all three elements are written together.

In analysis, the problem of separable verbs is to combine the two parts of the verb in contexts such as (1b) and (1c). Such a combination is necessary because syntactic and semantic properties of *aufhören* are the same, irrespective of whether the two parts are written together or not, but they cannot be deduced from the syntactic and semantic properties of the parts. Therefore, a solution to the problem of separable verbs will treat (1b) as if it read (2a) and (1c) as (2b):

(2) a. Claudia aufhört jetzt.
b. Daniel versucht zu aufhören.

The problem arises in a very similar fashion in Dutch, as the Dutch translations (3) of the sentences in (1) show. The only difference is that the infinitive in (3c) is not written together.

(3) a. Anna gelooft dat Bernard ophoudt.
b. Claudia houdt nu op.
c. Daniel probeert op te houden.

On the other hand, the problem of separable verbs in German and Dutch differs from the corresponding one in English, because English verbs such as *look up* are multi-word units in all contexts. A treatment of these cases which is in line with the solution proposed here is described by Tschichold (forthcoming).

As suggested by the English translation, separable verbs in German and Dutch are lexemes. Therefore, an important issue in evaluating a mechanism for dealing with them is how it fits in with the reusability of lexical resources.

Given the importance of the orthographic component in the problem, it is not surprising that it is hardly if ever treated in the linguistic literature.

## 2. Previous Approaches

In existing systems or resources for NLP, separable verbs are usually treated as a lexicographic and syntactic problem. Two typical approaches can be illustrated on the basis of Celex and Rosetta.

Celex (http://www.kun.nl/celex) is a lexical database project offering a German dictionary with 50'000 entries and a Dutch dictionary with 120'000 entries. In these dictionaries separable verbs are listed with a feature conveying the information that they belong to the class of separable verbs and a bracketing structure showing the decomposition into a prefix and a base, e.g. (auf)(hören). Celex dictionaries are reusable, but the rule component for the interpretation of the information on separable verbs, i.e. the mechanism for going from (1b-c) to (2), remains to be developed by each NLP-system using the dictionaries.

Rosetta is a machine translation system which includes Dutch as one of the source and target languages. Rosetta (1994:78-79) describes how separable verbs are treated. For the verb *ophouden* illustrated in (3), there are three lexical entries, *ophouden* for the continuous forms as in (3a), and *houden* and *op* for the discontinuous forms as in (3b-c). When a form of *houden* is found in a text, it is multiply ambiguous, because it can be a form of the simple verb *houden* ('hold') or of one of the separable verbs *ophouden* ('stop'), *aanhouden* ('arrest'), *afhouden* ('withhold'), etc. The entry for *houden* as part of *ophouden* contains the information that it must be combined with a particle *op*. At the same time, *op* is ambiguous between a reading as preposition or particle. In syntax, there is a rule combining the two elements in a sentence such as (3b). It is clear that, while this approach may work, it is far from elegant. It creates ambiguity and redundancies, because *ophouden* written together is treated in a different entry from *op* + *houden* as a discontinuous unit. These properties make the resulting dictionaries less transparent and do not favour reusability.

It should be pointed out that Celex and Rosetta were not chosen because their solution to the problem of separable verbs is worse than others. They are representative examples of currently used strategies, chosen mainly because they are relatively well-documented.

## 3. The Word Manager Approach

Word Manager™ (WM) is a system for morphological dictionaries. It includes rules for inflection and derivation (WM proper) and for clitics and multi-word units (Phrase Manager, PM). We will use WM here as a name for the combination of the two components. A general description of the design of WM, with references to various publications where the formalism is discussed in more detail, can be found in ten Hacken & Domenig (1996).

The German WM dictionary consists of a comprehensive set of inflectional and word formation rules describing the full range of morphological processes in German. In the last two years we have specified more than 100'000 database entries by classification of lexemes in terms of inflection rules (for morphologically simple entries) and by the application of word formation rules (for morphologically complex entries). In addition, the PM module contains a set of rules for clitics and multi-word units which covers German periphrastic inflection patterns and separable verbs.

The rule types invoked in the treatment of separable verbs in WM include Inflection Rules (IRules), Word Formation Rules (WFRules), Periphrastic Inflection (PIRules), and Clitic Rules (CRules). We will describe each of them in turn.

## 3.1. Inflection

In inflection, *aufhören* is treated as a verb with a detachable prefix *auf*. The detachable prefix is defined as an underspecified IFormative. This means that, in the same way as for stems, its specification is distributed over a class specification and a

```
RIRule  V_Detachable-Prefix
citation-forms
(ICat  Detachable-Prefix)    (ICat  V-Stem)    (ICat  V-Suffix)(Mod  Inf)
word-forms
(ICat  Detachable-Prefix)    (ICat  V-Stem)    (ICat  V-Suffix)
(ICat  Detachable-Prefix)    (ICat  V-Prefix.ge)    (ICat  V-Stem)  ...
                                               ...  (ICat  V-Suffix)(Mod  PaPa)
```

Fig. 1: Inflection rule for separable verbs in WM. The dots in the last line mark the absence of a line break in the actual code. Feature specifications separated by tabs refer to sets of formatives in paradigmatic variation. Each line thus generates one or more word forms.

```
target
(RIRule  V_Detachable-Prefix)  separable
      1     (ICat  Detachable-Prefix)
      2     (ICat  V-Stem)
```

Fig. 2: Target specification of the WFRule for separable verbs in WM.

specification of the individual string. The class is defined by the linguist in the specification of inflection processes. The specification of the string is part of the lexicographic specification, i.e. the string specification is the result of the application of the word formation rule the lexicographer chooses for the definition of an individual entry. In the IRules, detachable prefixes are referred to as formatives in the formulae generating the word forms. Fig. 1 gives the relevant rule of the database for otherwise regular separable verbs, such as *aufhören*.

## 3.2. Word Formation

Word Formation Rules consist of a source definition and a target definition. The source definition determines what (kind of) formatives are taken to form a new word. The target definition specifies how the source formatives are combined, and which inflection rule the new word is assigned to.

Separable verbs are the result of WFRules which are remarkable because of their target. The target specification is as in Fig. 2. This specification departs from the usual specification of a target in a WFRule in two respects. First, instead of concatenating the source formatives, the rule lists them, leaving concatenation to the IRule. This is necessary to form the past participle *aufgehört*, where the two formatives are separated by the prefix *ge-* (cf. last line of Fig. 1). Separable verbs are specified by the lexicographer by linking a word to a WFRule having a target specification as in Fig. 2. In the case of *aufhören*, this is a rule for prefixing in which "1" in Fig. 2 matches a closed set of predefined prefixes. The IRules and WFRules described so far cover the non-separated occurrences as in (1a).

The second special property of the specification in Fig. 2 is the system keyword "`separable`" in the second line. It assigns the result of the WFRule to the predefined class `%separable`. This class, whose name is defined in the WM-formalism, can be used to establish a link between the result of word formation and the input to the periphrastic inflection mechanism used to recognize occurrences such as in (1b).

## 3.3. Periphrastic Inflection

The mechanism for periphrastic inflection in WM consists of two parts. PIClasses are used to identify the components and PIRules to turn them into a single word form. The PIRule for separable verbs in German is given in Fig. 3. The rule in Fig. 3 consists of a name and a body, which in turn consists of input and output specifications separated by "=". The input specifies a finite verb form (infinitive and participles are excluded by "^") and a detachable prefix. The output combines them in the position of the verb, with the form prefix + verb, and with the features percolated from the verb (person,

```
Separable
(Cat V)^(Mod Inf)^(Mod Part) + %separable = ...
                                ... (POS 1)(FORM 2+1)(PERC 1)(Cat V)
```

Fig. 3: Periphrastic Inflection Rule for separable verbs in WM.

```
%separable + (CElement zu) + (Cat V)(Mod Inf)(Temp Pres) = ...
       ... (CElement zu), %separable + (Cat V)(Mod Inf)(Temp Pres)
```

Fig. 4: CRule for the infinitive of separable verbs in WM.

number, etc.). This yields (2a) as a step in the analysis of (1b).

The possibilities for specifying the relative position of the two elements to be combined are the same as the possibilities for multi-word units in general. In the PIClass for German it is specified that the finite verb always precedes the particle when the two are separated. In Dutch this is not the case, as illustrated by (3c), so that a different specification is required.

## 3.4. Clitic Rules

The clitic rule mechanism is used to analyse *aufzuhören* in (1c) and produce *zu aufhören* as in (2b). The CRule used is given in Fig. 4. Again input and output are separated by "=". The input consists of the concatenation of three elements: a detachable prefix, infinitival *zu*, and an infinitive. Graphic concatenation is indicated by "+". The CElement *zu* is defined elsewhere as a form of the infinitival *zu*, rather than the homonymous preposition, in order not to lose information. The output consists of two words, as indicated by the comma, the second of which concatenates the prefix and the verb.

## 3.5. Recognition and Generation

In recognition, the input is the largest domain over which components of multi-word units (MWUs) can be spread. In practice, this coincides with the sentence. Since WM does not contain a parser, larger chunks of input will result in spurious recognition of potential MWUs. Let us assume as an example that the sentences in (1) are given as input.

The first component to act is the clitics component. It leaves everything unchanged except (1c), which is replaced by (2b): *aufzuhören* => *zu aufhören*. Then the rules of WM proper are activated. They replace each word form by a set of analyses in terms of a string and feature set. In (1a), *aufhört* is analysed as third person singular or second person plural of the present tense of *aufhören*, in (1b) *hört* and *auf* are analysed separately, and in (1c) *aufhören*, which was given the feature infinitive by the CRule in Fig. 4, only as infinitive, not as any of the homonymous forms in the paradigm. The next step is periphrastic inflection. It applies to (1a) and (1c) vacuously, but combines *hört* and *auf* in (1b), producing the feature description corresponding to (2b): *hört auf* => *aufhört*. Finally, the idiom recognition component (not treated here) applies vacuously.

A general remark on recognition is in order here. The rule components of PM, i.e. clitics, periphrastic inflection and idiom recognition add their results to the set of intermediate representations available at the relevant point. Thus, after the clitic component, *aufzuhören* continues to exist alongside *zu aufhören* in the analysis of (1c). Since the former cannot be analysed by WM proper, it is discarded. Likewise, *hört* will survive in (1b) after periphrastic inflection and indeed as part of the final result. This is necessary in examples such as (4):

(4) Der Hund hört auf den Namen Wurzel.
('The dog answers to the name [of] Wurzel')

Since rules in WM are not inherently directional, it is also possible to generate all forms of a lexeme such as *aufhören* in the way they may occur in a text. The client application required for this task can also include codes indicating places in the string where other material may intervene, because this information is available in the relevant PIClass of the database.

## 4. Conclusion

Separable verbs in German and Dutch constitute a problem in NLP because they are lexemes whose recognition is not simply a matter of dictionary lookup. Therefore, a reusable lexical database such as Celex does not offer a comprehensive solution to the problem. On the other hand, treating them as a problem of syntactic recognition, as implemented in, for instance, Rosetta, fails to account for the lexeme character of separable verbs. As a consequence, spurious ambiguities and redundancies are created. Ambiguities arise between a simple verb such as *hören* ('hear') and the same form functioning as part of a separable verb such as *aufhören*. Redundancies emerge between the two different entries for *aufhören*, one for the continuous and one for the discontinuous occurrences.

In Word Manager, the recognition of separable verbs is entirely within the reusable lexical domain. A client application can start from an input which resembles (2) rather than (1b-c). An indication of the type of input is given in (5) and (6). For (1b), (5a) and (5b) are offered as alternatives. For (1c), (6) is offered as the only analysis (modulo syncretism of *versucht*).

| | | | |
|---|---|---|---|
| (5) | a. | claudia | (Cat Noun) |
| | | aufhören | (Cat Verb)(Tense Pres)<br>(Pers Third)(Num SG) |
| | | jetzt | (Cat Adv) |
| | b. | claudia | (Cat Noun) |
| | | hören | (Cat Verb)(Tense Pres)<br>(Pers Third)(Num SG) |
| | | jetzt | (Cat Adv) |
| | | auf | (Cat Prep) |

| | | |
|---|---|---|
| (6) | daniel | (Cat Noun) |
| | versuchen | (Cat Verb)(Tense Pres)<br>(Pers Third)(Num SG) |
| | zu | (Cat Inf-marker) |
| | aufhören | (Cat Verb)(Mode Inf) |

The task of the client application in the recognition of separable verbs in (1) is reduced to the choice of (5a) rather than (5b).

Finally, two points deserve to be emphasized. First, the entire WM-formalism for separable verbs has been implemented as described here. The rules for German have been formulated and a large dictionary for German (100'000 entries) including separable verbs is available. Moreover, the only provision in the WM-formalism specifically geared towards the treatment of separable verbs is the keyword *separable* in WFRules (cf. Fig. 2) and the corresponding class name *%separable*. Otherwise the entire formalism used for separable verbs is available as a consequence of general requirements of morphology and multi-word units.

## References

ten Hacken, Pius & Domenig, Marc (1996), 'Reusable Dictionaries for NLP: The Word Manager Approach', *Lexicology* 2: 232-255.

Rosetta, M.T. (1994), *Compositional Translation*, Kluwer Academic, Dordrecht.

Tschichold, Cornelia (forthcoming), *English Multi-Word Units in a Lexicon for Natural Language Processing*, Ph.D. dissertation, Universität Basel (Dec. 1996), to appear at Olms Verlag, Hildesheim.

Word Manager:
http://www.unibas.ch/LIlab/projects/wordmanager/wordmanager.html

Fig. 5: URL for Word Manager.