Particle Verbs in NLP lexicons

Anke LÜDELING, Stuttgart, Germany

Abstract

This paper discusses the problems that arise in NLP lexicons if German particle verbs are analyzed as words. I suggest analyzing them as phrasal constructions consisting of an adverb, an adjective, or a preposition and a verb. It is then not necessary to have separate lexicon entries for compositional particle verbs. Non-compositional particle verbs also do not need separate lexicon entries for tagging and parsing. They need multi-word lexicon entries for all those applications that operate on semantic representations.

1 Introduction

The treatment of the so-called particle verbs in some Germanic languages[1] is notoriously difficult on all levels of linguistic processing. Particle verbs are, roughly speaking, constructions that consist of a verb and a particle (or pre-verb) which is obligatorily verb-adjacent in V-final sentences, as in (1a), and stays clause-final in V-second or V-first sentences, cf. (1b).[2]

(1) a. dass er das Buch anliest
that he the book particle+reads
"that he starts reading the book"

b. er liest das Buch an
he reads the book particle
"he starts reading the book"

c. *er anliest das Buch

The problem is obvious: the separability of the particle and the verb suggests that particle verbs should be analyzed like syntactic constructions consisting of an XP and a verb. However, they seem to behave like words in many other respects and thus many speakers of German feel they should be analyzed and treated as words. The literature reflects this: many researchers argue that particle verbs must be exceptional words ([Neeleman and Weerman 1993],[Stiebels and Wunderlich 1994] and many others) and others argue that they are strangely behaved syntactic constructions ([van Riemsdijk 1978], [Booij 1990], [von Stechow 1993], [Zeller 1997] and others). Because NLP systems typically do not deal with the particle verb problem in a principled manner, the dilemma is reflected implicitly in many NLP systems and their lexicons.

In (computational) lexicography particle verbs are typically treated as words. This is partly due to the fact that particle verbs are spelled without a space in V-final sentences.[3]

The paper is organized as follows. First I will show some examples of problems for NLP applications that arise if particle verbs are analyzed as words (Sections 2 and 3). Then I will sketch a principled approach to particle verbs that avoids these problems.

2 The selection problem: Tagging

The word analysis of particle verbs poses problems for tagging mechanisms since it is not desirable to tag the particle + verb combination as a verb when it occurs V-final and as a verb plus a phrase when it occurs in V-second sentences. Hence, many tagging algorithms have assumed a separate tag for verbal particles. But this poses a problem in the sense that it is not clear which tag the separated particle should receive: typically taggers make a large number of mistakes with respect to the category of the particle, as shown in (2).[4]

(2) a. Es kommt ja auf den Wert ⟨an/PTKVE⟩
it comes modal particle on the value an
"it depends on the value"

b. Auf die Tat jedes einzelnen kommt es ⟨an/APZR⟩
on the action of each separate one comes it an
"it depends on the action of each person"

The problem has two causes: first, particles occur in the same position in sentences that adjectives, adverbs, or prepositional phrases occur in, as shown in (3).

(3) a. daß er die Tasse leer trinkt
that he the cup empty$_{Adj}$ drinks
"that he empties the cup"

b. daß er das Buch langsam liest
that he the book slowly$_{Adv}$ reads
"that he reads the book slowly"

Second, particles are homophonous with elements from at least one of the categories A, Adj, or P. It is therefore not possible to distinguish them from these other elements using statistical methods. It is also not possible to formulate a rule-based tagging algorithm that distinguishes particles from adjectives etc. because there is no theoretical basis for this distinction: in [Lüdeling, to app.] I discuss a number of tests that are commonly cited in the literature to help distinguish particle verbs from resultative or adverbial constructions and show that none of these tests are sufficient.

A different, but related problem arises when one considers particle verbs in V-final positions. There are many which are always spelled as one word but – again due to the fact that it is not clear what a particle verb really is – many others are alternatively spelled as one word or as two words.

(4) a. daß er in das Zimmer hineinkommt
that he into the room particle+comes
"that he enters the room"

b. daß er in das Zimmer hinein kommt

To summarize: it is not possible to clearly distinguish verbal particles from adverbs, adjectives, or prepositions. That principal problem leads to two kinds of tagging problems: first, statistical as well as rule-based taggers will make many mistakes with respect to the classification of the separated particle and second, constructions that are spelled as one word in some instances and as two words in other instances receive different tags.

3 The productivity problem: transparent vs. non-transparent particle verbs

Many particle verbs are non-transparent. There are, however, many other constructions that are often called particle verbs that are transparent and – what's more important for lexicographic purposes – productive. Some examples of productive particle verb patterns are given in (5).

(5) a. anlesen "to start reading", andiskutieren "to start discussing", andenken "to start thinking (about)", ...

b. hineinlaufen "to walk into", hineinrennen "to run into", hineinschwimmen "to swim into", ...

c. weglaufen "to walk away", wegrennen "to run away", wegschwimmen "to swim away", ...

It becomes clear from these examples that NLP applications that refer to semantic properties of their input have to distinguish between compositional (transparent) and the non-compositional (non-transparent) particle verbs.

Analyses that simply list particle verbs as words in a lexicon typically do not make this distinction.[5] Without this distinction it is not possible to assume productive operations for the compositional cases. One example is the machine translation project *Verbmobil* where the examples in (5) are all listed separately – as complex words – in the lexicon and thus have to be treated by separate rules. It is not possible to write a rule like (6) which captures the productivity of *hinein*-formations.

(6) German *hinein* with a manner of motion verb should always be translated as English *into*

4 No particle verbs

In [Lüdeling, to app.] I show that particle verbs do not form a linguistically distinguishable class of constructions: it is not possible to determine a set of properties that all particle verbs have and that other secondary predicate constructions or adverbial constructions do not have. I discuss the consequences of this approach for the syntactic and morphological treatment of particle verbs. The solution that I will sketch in this section is based on the analysis presented there and focuses on the consequences of that approach for computational lexicography: particle verbs are to be treated in the same manner as secondary predicate or adverbial constructions.

It is neither necessary nor desirable to have a class of particle verbs or verbal particles in the lexicon.

For tagging, particle verbs should not be treated different from secondary predicate and adverbial constructions. This means that it is not necessary to list a (not delimitable) class of particles or to introduce the pos-tag "verbal particle". Rather, particles have to be tagged as Adv, Adj, or P.[6] It is necessary to re-analyze those particle verbs that are spelled in one word - the particle has to be separately tagged. Many taggers already do such a kind of re-analysis in order to find out whether a verb is inflected or uninflected: in German, the infinitive is often marked by *zu* "to". For simplex verbs, *zu* is separated, as shown in (7a). In particle verbs, *zu* attaches to the base verb (another piece of evidence to show that the particle is not part of the verb), as in (7b). Taggers that distinguish between finite and infinitival verbs (as TreeTagger for example does) already separate the particle and the *zu* from the verb. The desired representation would be the one in (7c). The tag for *hinein* would have to be ADV.

(7) a. er bat sie, zu kommen
he asked her to$_{infinitive-marker}$ come
"He asked her to come"

b. er bat sie, hineinzukommen
he asked her into + to$_{infinitive-marker}$ + come
"he asked her to come in"

c. er bat sie, hinein zu kommen
he asked her into to$_{infinitive-marker}$ come

For transparent particle verbs the story ends here: they can be parsed like all other secondary predicate constructions or adverbial constructions. The semantic representation is built up by regular semantic principles.

Non-compositional particle verbs can also be parsed regularly. The semantic represenation of such particle verbs, however, must be listed in a semantic lexicon. It is not necessary to introduce special mechanisms for this since mechanisms that deal with non-compositional input are needed in any case if an application operates on semantic representation. This is summarized in Table 1.

Let me sketch what this would mean for Verbmobil and its lexicons. Consider particle verbs with *hinein* "into" that combine transparently and productively with manner of motion verbs, as seen in (5). Recall that at the moment the lexicon contains separate entries for all verbs occurring with *hinein* and that therefore a generalizing transfer rule such as (6) cannot be stated. Instead one has to write a separate translation rule for each new entry. In my conception, there would only be one entry for *hinein* as an adverb. For the transparent cases one would then be able to write a rule like (6). The nontransparent cases are treated like idioms: They can either receive multi-word entries in the lexicon itself (e.g. *in+ sich+hinein+lachen* "to chuckle to oneself" or *sich+ hinein+steigern* "to work oneself up", analogous to *in+den+sauren+Apfel+beissen*, lit: to bite into the sour apple, "to swallow the bitter pill") or (as it is done in Verbmobil for phrasal idioms, see [Emele *et al.* 2000]) they are translated as units in the transfer module. The lexicon

	tagger lexicon	parser lexicon	semantic lexicon (e.g. machine translation, inference systems, text understanding)
transparent particle verbs	no separate entry	no separate entry	no separate entry
non-transparent particle verbs	no separate entry	no separate entry	multi-word entry

Table 1: Lexical entries for particles

is then freed of redundant entries and the system is able to deal with the productivity of particle verb formation.

5 Summary

To sum up: to analyze particle verbs as a special class of words yields many problems for NLP lexicons. Since there is also no theoretical basis for such a class of constructions these problems can be avoided by treating particles as regular adverbs, adjectives, or prepositions for tagging lexicons and to treat non-compositional particle verbs as idioms in a semantic lexicon. The advantage of such an approach is that unnecessary ambiguities are avoided and the productivity of transparent particle verb formation can be dealt with.

Notes

[1]Particle verbs can be found in German, Dutch, the Scandinavian languages and (arguably) English. Constructions that are similar to the Germanic particle verbs exist in Hungarian. In the following I will speak about German examples only. I suspect that many of the conclusions hold for Dutch as well. I will not make any claims about the other languages containing particle verbs.

[2]According to [Stiebels and Wunderlich 1994] particles in particle verbs can be of any major syntactic category. In this paper I will deal only with particle verbs with adverbial, adjectival, or prepositional particles. The problems described below are even worse when verbal and nominal particles are considered as well. See [Lüdeling, to app.] for reasons why nominal and verbal particles should be treated differently.

[3]It becomes clear that the orthographic rules cannot be a criterion of what counts as a particle verbs when one considers that after the German orthographic reform of 1998 there are suddenly fewer particle verbs than there were before the reform since many constructions that were spelled in one word are now spelled in two.

[4]The example stems from the TreeTagger which was developed at the University of Stuttgart ([Schmid 1994, 1995]). This is not meant as criticism of this particular tagger – other taggers give comparable results. PTKVE stands for 'verbal particle', APZR stands for 'circumposition right'.

[5]One might argue that productive word formation should be part of any lexicon. If it were then this problem would disappear even in a word analysis of particle verbs. However, most computational lexicons do not contain enough word formation information to handle this.

[6]It has to be noted that there are a number of elements that have adverbial, adjectival, and prepositional readings. The selection between these three pos-tags remains difficult.

References

Booij, G. (1990). The boundary between morphology and syntax: Separable complex verbs in Dutch. In G. Booij and J. v. Marle, editors, *Yearbook of Morphology*, volume 1, pages 45 – 63. Foris, Dordrecht.

Emele, M. C., Dorna, M., Lüdeling, A., Zinsmeister, H., and Rohrer, C. (2000). Semantic-based transfer. Unpublished Manuscript, University of Stuttgart.

Lüdeling, A. (to app.). *On Particle Verbs and Similar Constructions in German*. CSLI, Stanford.

Neeleman, A. and Weerman, F. (1993). The balance between syntax and morphology: Dutch particles and resultatives. *Natural Language and Linguistic Theory*, **11**, 433–475.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Stiebels, B. and Wunderlich, D. (1994). Morphology feeds syntax: The case of particle verbs. *Linguistics*, **32**(6), 913 – 968.

van Riemsdijk, H. (1978). *A Case Study in Syntactic Markedness. The Binding Nature of Prepositional Phrases in Dutch*. Foris, Dordrecht.

von Stechow, A. (1993). Grundlagen. Foundations. In *Syntax. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of Contemporary Research*. Walter de Gruyter.

Zeller, J. (1997). Against overt particle incorporation. *Penn Working Papers in Linguistics (Proceedings of the 21st Annual Penn Linguistics Colloquium)*, **4**.