

Identification, Quantitative Description, and Preliminary Distributional Analysis of German Particle Verbs

Sabine Schulte im Walde
Computational Linguistics
Saarland University
66041 Saarbrücken, Germany
schulte@coli.uni-sb.de

Abstract

A statistical grammar model is used to identify German particle verbs and induce quantitative lexical information on their subcategorisation frames and selectional preferences. A simple approach to address the semantic class of the particle verb is introduced.

1 Introduction

German particle verbs are morphologically composed of a prepositional affix and a base verb, e.g. *an-fangen*, *ein-führen*, *vor-stellen*. The prepositions are an open set, including such as *ab*, *an*, *auf*, *aus*, *bei*, *durch*, *ein*, *los*, *nach*, *über*, *um*, *unter*, *vor*, *wider*, *zu*, cf. (Lüdeling, 2001). German particle verbs are a challenge for automatic methods in Statistical Natural Language Processing: Both the behaviour and the semantic class of a particle verb are unpredictable with respect to its base verb. Concerning their morpho-syntactic properties, particle and base verb are obligatorily adjacent in V-final sentences but separated in V-second and V-first sentences:

- (1) ..., weil Kai mit den Hausaufgaben *anfängt*.
Kai *fängt* mit den Hausaufgaben *an*.
Fängt Kai mit den Hausaufgaben *an*?

Concerning their semantic properties, particle verbs are either transparent (i.e. the combination of particle and base verb is compositional) such as *ab-holen* ‘to fetch’, or opaque (i.e. the combination of particle and base verb is idiosyncratic) such as *an-fangen* ‘to begin’ vs. *fangen* ‘to catch’, or have both transparent and opaque senses such as *ein-setzen* ‘to insert, to begin’. Particle verbs may change the behaviour of their base verbs, by the particle saturating or adding an argument to the base verb’s argument structure, cf. example (2) as taken from (Lüdeling, 2001). And they may leave the argument structure of the base verb identical, but change the

selectional restrictions, cf. example (3) with *Socken* ‘socks’ and *Geschirr* ‘dishes’.

- (2) *Er *stellt* [NP_{acc} das Glas].
Er *stellt* [NP_{acc} das Glas] [PP auf den Tisch].
Er *stellt* [NP_{acc} das Glas] *ab*.
- (3) Er *wäscht* [NP_{acc} seine Socken].
*Er *wäscht* [NP_{acc} seine Socken] *ab*.
Er *wäscht* [NP_{acc} das Geschirr] *ab*.

Particle verbs constitute a significant part of the verb lexicon, and empirical information is essential for building lexical resources and supporting applications in Natural Language Processing. This paper presents a method to automatically identify German particle verbs and acquire quantitative information on their lexical properties: A German statistical grammar model is designed to identify particle verbs and learn empirical information. The particle verbs are quantitatively described on basis of the grammar parameters, mainly with respect to their subcategorisation behaviour and their selectional preferences. Finally, a simple approach to address the similarity between particle verb and base verb is presented, and the semantic class of the particle verb is approximated.

2 Identification

The identification of German particle verbs is performed by a statistical grammar model for German (Schulte im Walde, 2003). A manually defined context-free grammar is learned by lexicalised parameter estimation as performed by the left-corner parser LoPar (Schmid, 2000), using 35 million words of a large German newspaper corpus from the 1990s. The grammar rules are developed with specific attention towards the identification of the verb complex and the verb head in the free-order German clause structure, and the propagation of the verb head information through the parse trees. When the

verb head is propagated through the sentence analyses, it is related to the subcategorisation frame types (explicitly coded as non-terminal categories) and the argument heads. Because particle verbs might be discontinuous according to the clause type, LoPar provides a distinguished functionality to combine lexical heads, such that complete particle verb heads are identified and related to their argument structure.

3 Quantitative Lexical Description

The trained grammar parameters are used for information extraction on particle verbs.

Subcategorisation Frames: The verbs are described by frequency and probability distributions over 38 frame types. Possible arguments in the frames are nominative (n), dative (d) and accusative (a) noun phrases, reflexive pronouns (r), prepositional phrases (p), expletive *es* (x), non-finite clauses (i), finite clauses (s-2 for V-second clauses, s-dass for *dass*-clauses, s-ob for *ob*-clauses, s-w for indirect *wh*-questions), and copula constructions (k). In addition to the syntactic frame information, the frame types distinguish prepositional phrase types by distributing the probability mass of pp-frames over prepositional phrases, according to their frequencies in the corpus. We consider the 30 most frequent PPs, referred to by case and preposition such as ‘Dat.mit’, ‘Akk.für’. Table 1 presents the five most probable subcategorisation frame types for a choice of four German particle verbs. Next to the frames types, pp-frames are listed in case they have a probability of more than 1%. The particle verbs show a concentration on a specific choice of frame types, which is more restricted than for German verbs in general, cf. (Schulte im Walde, 2003): transitive ‘na’ or intransitive ‘n’, (with reflexives ‘nr’ and ‘nar’), ‘nad’ adding a (free) dative or ‘nap’ adding a pp-adjunct to transitive ‘na’. This choice is realised for transparent verbs such as *ankommen*, *aufstellen*, and for opaque particle verbs such as *einsetzen*, *umbringen*. The pp-types differ for the verbs and do not provide a common picture.

Selectional Preferences: The grammar provides selectional preference information on a fine-grained level: it specifies the possible argument realisations in form of lexical heads, with reference to a specific verb-frame-slot combination. Table 2 lists the most frequent nominal argument heads for the verb *einsetzen* and the

argument slots in the probable frame types ‘n’, ‘na’ and ‘nr’; the relevant frame slot is underlined. The particle verb has multiple meanings, ‘to begin’ when used as intransitive (frame: n), ‘to stand up for something’ when used as an intransitive reflexive (frame: nr), and ‘to insert, to install’ when used with a direct object (frame: na). The arguments for the verb differ with respect to the frame-slot combination, e.g. as subject in the aspectual intransitive construction, events and movements dominate the nominal choice, vs. in the causative transitive construction we find groups and organisations as subjects and means as direct objects.

<i>ankommen</i> (freq: 1,831)			
n	38.82	np:Dat.bei	1.50
x	16.12	nap:Dat.in	1.40
na	10.56	np:Akk.auf	1.02
ns-w	5.76		
ns-2	4.63		
<i>aufstellen</i> (freq: 1,353)			
na	64.63	nap:Dat.in	6.90
nap	21.54	nap:Akk.für	3.23
nad	7.54	nap:Dat.an	2.30
nai	1.03	nap:Dat.auf	2.18
n	1.01	nap:Dat.zu	1.61
<i>einsetzen</i> (freq: 3,390)			
na	40.29	nap:Dat.in	4.33
nap	16.44	nap:Dat.bei	2.81
nr	15.87	npr:Akk.für	2.63
n	10.86	nap:Dat.zu	1.76
nad	4.71	nap:Akk.für	1.35
<i>umbringen</i> (freq: 683)			
na	53.60	nap:Dat.in	5.43
nr	19.36	nap:Dat.nach	1.54
nap	12.23	npr:Dat.in	1.23
nad	3.20		
nas-2	1.97		

Table 1: Particle verb subcategorisation

4 Comparison and Semantic Class

The quantitative information enables us to compare the properties of particle verbs and base verbs and approach the semantic class of the particle verbs. Table 3 presents the five most probable frame types for the base verbs of Table 1, followed by examples of nominal fillers in dominating frames and argument slots. The verbs *ankommen* and *kommen* agree in the verb sense ‘to arrive’; both verbs use the frame type ‘n’ to express this meaning; adjunctive PPs differ, being locative for the particle verb and directional for the base verb. Table 4 illustrates the strong overlap in nominal fillers and confirms the transparency of the particle verb sense: For the most frequent intransitive sub-

<i>einsetzen</i>							
<u>n</u>		<u>na</u>		<u>na</u>		<u>nr</u>	
Run	13.00	Polizei	65.47	Gas	41.70	Regierung	16.97
Regen	12.93	Regierung	16.69	Mittel	33.11	Minister	13.21
Prozeß	11.04	Wehr	14.80	Kommission	32.01	SPD	9.90
Welle	7.72	Bahn	12.53	Waffe	29.97	Partei	6.99
Kampf	6.00	Seite	9.91	Stock	21.00	FDP	6.58
Kritik	5.84	Polizist	7.95	Zeug	17.69	Senator	5.91
Fall	5.50	Armee	6.04	Gerät	16.06	Senat	5.00
Wanderung	5.00	Gesellschaft	5.99	Zug	15.94	Demokrat	5.00
Kraft	4.99	Kraft	5.94	Ausschuß	15.05	CDU	4.97
Aufschwung	4.97	Verband	5.81	Kraft	14.81	Beirat	4.55

Table 2: Selectional preferences for specified arguments of *einsetzen*

jects for *ankommen* the base verb provides frequencies as well. The verbs *aufstellen* and *stellen* agree in the verb sense ‘to put up’; both verbs use the frame type ‘na’, plus a free dative and similar locative pp-ad adjuncts to express that meaning. The overlap is less strong than for *(an)kommen* but still obvious. Presenting the most probable nominal fillers for the direct objects of opaque *umbringen* and *bringen* in Table 5 illustrates that *umbringen* is dominated by individual persons, while *bringen* is dominated by inanimate objects. We can demonstrate such strong difference between argument categories for arbitrary combinations of frame-slot types for the two verbs. Summarising, even though the base verbs show more variety in the usage of subcategorisation frames than the particle verbs (especially high-frequency verbs such as *kommen*), we can again identify a high percentage of frame probabilities for ‘n’, ‘na’, ‘nad’ and ‘nap’. In addition, transparent particle verb senses agree in the conceptual idea of relevant frame arguments, and even in the specific nominal choices. The relevant frame arguments are not predictable by the particle or by the transparency vs. opacity of the particle verb.

Based on the comparison of particle and base verbs, we associate German verbs with their frequency distributions over the 38 subcategorisation frame types excluding pp-specification, and frequencies for nominal fillers for ‘n’, ‘nr’, ‘na’, ‘nad’, and ‘nar’ as relevant frames and argument slots. The cosine distance measure compares particle verbs with their base verbs, plus synonyms as defined by (Bulitta, 2003). The most similar verbs provide an indication of the semantic class of the particle verbs, and the similarity between the particle verbs and the base verbs provides an indication of the degree of transparency. Two examples are presented in

<i>kommen</i> (freq: 43,270)			
np	34.18	np:Dat.zu	11.21
n	28.78	np:Akk.in	5.35
na	8.05	xp:Dat.zu	5.33
x	5.65	np:Dat.aus	4.77
xp	5.53	np:Dat.von	2.67
<i>stellen</i> (freq: 11,233)			
na	35.97	nap:Dat.in	5.96
nap	20.13	nap:Dat.zu	3.98
nad	11.37	nap:Akk.in	1.76
npr	6.17		
ndr	5.85		
<i>setzen</i> (freq: 7,545)			
na	25.87	np:Akk.auf	13.66
nap	19.09	nap:Dat.in	5.03
np	16.50	nap:Akk.auf	2.80
nad	12.28	nap:Dat.mit	1.86
n	10.13	nap:Akk.in	1.66
<i>bringen</i> (freq: 12,249)			
na	42.71	nap:Akk.in	9.58
nap	31.65	nap:Dat.zu	4.86
nad	11.47	nap:Akk.auf	4.44
n	3.21	nap:Dat.mit	4.17
nd	2.05	nap:Dat.in	3.41

Table 3: Base verb subcategorisation

Table 6: For the transparent verb *ankommen*, the base verb *kommen* has rank one in the list of similar verbs; for the ambiguous verb *einsetzen* the base verb *setzen* has only rank 21. The table illustrates that this simple approach can be used to get a grip on the semantic class of particle verbs, in order to refine the respective lexical entry. Future work will elaborate on the semantic content of the particle verbs.

5 Related Work

As to my knowledge, no quantitative analysis of German particle verb been performed so far. Recent work is devoted to theoretical investigations such as (Lüdeling, 2001) addressing a coherent class of particle verbs; and (Dehé et

<u>n</u>		
	<i>ankommen</i>	<i>kommen</i>
Botschaft	17.17	10.22
Zug	11.14	14.78
Flüchtling	7.13	28.07
Film	5.18	13.32
Spende	5.00	6.61
Brief	4.72	22.18
Buch	4.60	13.78
Geld	4.48	114.19
Sache	4.46	12.94
Kunst	3.82	9.57

<u>na</u>		
	<i>aufstellen</i>	<i>stellen</i>
Rekord	50.33	0.00
Kandidat	35.29	4.32
Liste	32.08	0.00
Plan	26.12	0.00
Forderung	24.68	85.96
Schild	21.26	0.00
Zelt	12.95	2.05
Behauptung	10.84	1.46
Programm	9.72	2.95
Container	9.16	3.01

Table 4: Transparent particle verb fillers

<u>na</u>			
<i>umbringen</i>		<i>bringen</i>	
Mensch	13.63	Ergebnis	111.86
Frau	10.24	Erfolg	73.39
Kind	9.54	Geld	54.29
Mann	7.46	Problem	52.44
Vater	5.97	Vorteil	47.58
Million	4.91	Opfer	44.34
Leute	4.87	Entscheidung	39.92
Tausend	3.00	Entlastung	39.52
Freundin	3.00	Licht	39.33
Geisel	2.97	Klarheit	38.05

Table 5: Opaque particle verb fillers

al., 2002) with articles on morphological, syntactic and semantic properties of particle verbs. For English, (Baldwin and Villavicencio, 2002) propose techniques to identify English particle verbs from unannotated corpus data. (Villavicencio, 2003) investigates the characteristics of English particle verbs and applies the insights to improve the coverage of existing resources. (Baldwin et al., 2003) use a latent semantic analysis in order to find the semantically most similar (base) verbs.

References

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the Unextractable: A Case Study on Verb Particles. In *Proceedings of the Sixth Conference on CoNLL*.

	<i>ankommen</i>		<i>einsetzen</i>	
1.	kommen	0.50	benutzen	0.37
2.	erscheinen	0.42	verwenden	0.36
3.	daherkommen	0.41	engagieren	0.32
4.	anrollen	0.40	handeln	0.31
5.	herkommen	0.38	anwenden	0.27
6.	einlaufen	0.24	schicken	0.26
7.	landen	0.22	eintreten	0.25
8.	nähern	0.21	einschalten	0.24
9.	eintreffen	0.19	entfalten	0.23
10.	einfinden	0.14	kümmern	0.21
...
21.	setzen	0.10

Table 6: Semantic classes of particle verbs

- Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL Workshop on Multiword Expressions*.
- Erich and Hildegard Bulitta. 2003. *Wörterbuch der Synonyme und Antonyme*. Fischer Taschenbuch Verlag.
- Nicole Dehé, Ray Jackendoff, Andrew McIntyre, and Silke Urban, editors. 2002. *Verb-Particle Explorations*. Mouton de Gruyter.
- Anke Lüdeling. 2001. *On German Particle Verbs and Similar Constructions in German*. CSLI Publications.
- Helmut Schmid. 2000. LoPar: Design and Implementation. Arbeitspapiere des SFB 340, Universität Stuttgart.
- Sabine Schulte im Walde. 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes*. Ph.D. thesis, Universität Stuttgart.
- Aline Villavicencio. 2003. Verb-Particle Constructions and Lexical Resources. In *Proceedings of the ACL Workshop on Multiword Expressions*.