Automatic classification of German *an* particle verbs

Sylvia Springorum, Sabine Schulte im Walde, Antje Roßdeutscher
Universität Stuttgart
{sylvia.springorum,schulte,antje}@ims.uni-stuttgart.de

1 Introduction

German particle verbs (PVs) are a challenge to theoretical and computational linguistics, as both of their parts (i.e., the particles and the base verbs) may be highly ambiguous (cf. Stiebels, 1996; Schulte im Walde, 2005; Lechler and Roßdeutscher, 2009; Springorum, 2009; 2011; among others). The current study works at the interface of theoretical and computational linguistics to explore the semantic properties of *an* particle verbs, i.e., German particle verbs with the particle *an*. Driven by a thorough analysis of the particle *an* from a theoretical point of view (Springorum, 2011a), we identify empirical features to perform an automatic semantic classification of the particle verbs. A focus of the study is on the questions (a) how we could transform the theoretical insights into empirical, corpus-based features, (b) to what extent we could replicate the theoretical classification by a machine learning approach, and (c) whether the computational analysis would in turn deepen our insights to the semantic properties of the PVs.

2 Data

The verb particle *an* has about eleven different readings, according to the detailed analysis by Springorum (2009; 2011a) that modelled the meanings of *an* particle verbs within Discourse Representation Theory (Kamp and Reyle, 1993). For the current study, we chose four of the readings, each represented by 10 verbs.

(i) ***Topological verbs*** describe a contact situation that typically occurs between a direct object of the verb and an implicit background, cf. Example (1). *an* describes a contact situation between the dog (via the leash) and an unmentioned background. In addition to *anketten* (chain), this class includes *anschnallen* (belt on)*, anlehnen* (lean against), *anmalen* (paint [on]), *anstreichen* (brush [on]), *anbauen* (install), *anbinden* (tie sth. up), *ansiedeln* (settle), *anfassen* (touch), *anschließen* (affiliate).

(1) *Maria kettet den Hund an. /* Maria chain the dog [an]
'Maria chains the dog.'

(ii) ***Directional verbs:*** In most cases, the verb event points from the subject to the direct object of the *an* particle verb. This reading has sub readings which in general

express an additional communication attempt, cf. Example (2). Verbs with the simple directional reading are *angucken* (look at)*, anblicken* (gaze at)*, anvisieren* (aim for sth.), *anstreben* (aspire), *anstarren* (stare at), *anpeilen* (locate). Verbs which in addition describe a directional communication attempt are *anreden* (address), *anschreiben* (write to), *anschreien* (scream at), *anlächeln* (smile at).

(2) *Der kleine Junge grinst seine Mutter an. /* The small boy grin his mother [an]
'The small boy grins at his mother.'

(iii) ***Event initiation verbs*** describe an event initiation where the particle triggers a change from a non-progressive state to a progressive state, cf. Example (3). The whistling, together with *an*, is responsible for the starting of the game event. Further verbs in this class are *ankurblen* (boost), *antreiben* (activate), *anheizen* (heat up), *anstimmen* (intone), *anspornen* (cheer on), *anstiften* (incite), *anrichten* (wreak), *anregen* (animate), *anzetteln* (plot).

(3) *Der Schiedsrichter pfeift das Spiel an. /* The referee whistle the game [an]
'The referee starts the game by whistling.'

(iv) ***Partitive verbs:*** The verb event is performed only on parts of the direct object, cf. Example (4), where the sawing event effects only a part of the plank. Verbs with similar particle semantics are *anschneiden* (cut partially), *anbrechen* (broach), *anbraten* (roast partially), *anknabbern* (nibble partially), *anreißen* (scribe), *anrösten* (toast partially), *anbohren* (drill partially), *anzahlen* (deposit), *ansengen* (scorch).

(4) *Der Dachdecker sägt das Brett an. /* The roofer saw the plank [an]
'The roofer partially saws the plank.'

To provide an example of the formal semantic descriptions of the *an* classes, the DRS in Figure 1 describes the directed communication attempt in Example (2). There is a presupposition which claims that *x* (the boy) believes that *y* (the mother) is an experiencer of the grinning event *(e')* of which he is the agent. The vector *v* is the direction of *e'* and is defined from *x* to *y.* Such DRSs exist for all classes.

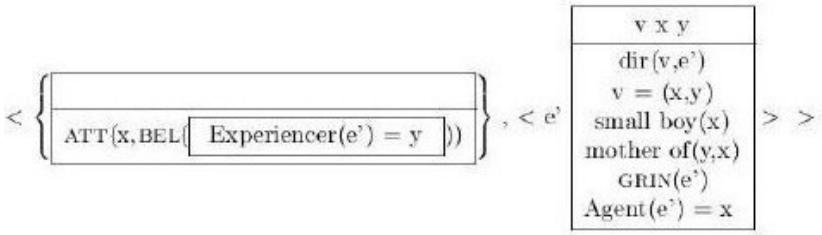

Figure 1: DRS for Example (2).

The classification serves as the gold standard for the classification experiments. In addition, we collected human judgements on the semantic classes: eight linguists were asked to classify the 40 verbs into four classes with 10 verbs each. The proportion of agreement $p_0=0.79$ serves as upper bound for the classification.

The empirical features for the classification experiments were derived from the *SdeWac* Corpus (Faaß et al., 2010), a German web corpus. The corpus was preprocessed by the Tree Tagger (Schmid, 1994) and a parser (Schiehlen, 2003).

3 Classification experiments

The verbs in a common class are expected to share semantic properties, and we were interested in how we could transform the theoretical properties into empirical, corpus-based features. For example, particle verbs with the topological reading are very likely to go along with a prepositional phrase (PP) that makes the implicit background explicit, cf. Example (5). In contrast, event initiation verbs appear more frequently than others with a PP headed by *zu.*

(5) *Maria kettet den Hund am Zaun an. /* Maria chain the dog at the fence [an].
'Maria chains her dog at the fence.'

In sum, we performed classification experiments relying on PPs, direct objects, and adverbs as verb features. To reduce the data sparseness, we added experiments where the nominal heads of the PPs and the direct objects were generalized by GermaNet. The baseline uses subjects as classification features, which are expected to provide little support for the semantic classification. Below are the results from the best experiments; more details and other experiments will be discussed in the talk.

The classification was carried out using the WEKA tool (Hall et al., 2009) with the J48 decision tree algorithm. The best results came out by mixed feature configurations using PPs and direct objects generalized by GermaNet (Exp 1), and just the *an* PP and direct objects generalized by GermaNet (Exp 2). The experiments reach an agreement of 67.5% and 70%, respectively. The distribution of the verbs over the classes is visualized in Table 1. Figure 1 presents the decision tree of Exp 2, where each branch stands for a decision rule, and the leaves represent the classes. Table 2 shows the results in comparison to the baseline and the human judgments.

A	B	C	D	
anbauen **anschließen** **anketten** **anlehnen**	ansiedeln	anmalen anbinden anfassen	anstreichen anschnallen	A= Top.
	anpfeifen **anrichten** **antreiben** **anzetteln** **anregen** **ankurbeln** **anheizen**	anspornen anstiften anstimmen		B= Ev. I.
anstreben	angucken anvisieren	**anschreiben** **anpeilen** **anschreien** **anblicken** **anstarren** **anlächeln** **anreden**		C= Dir.
			all partitive verbs	D= Par.

Table 1: Distribution of verbs over classes (Exp 2).

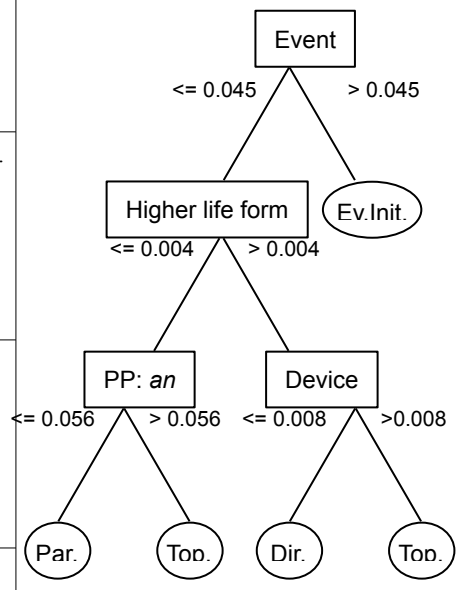

Figure 2. Decision tree (Exp 2).

Experiment	Features	+	%	-	%	Top.	Ev.I.	Dir.	Par.
Baseline	Subject	13	32.50	27	67.50	0	3	1	9
Human jugdgment			79.06		20.94				
Exp 1	PP + Obj. Class	27	67.50	13	32.50	5	5	7	10
Exp 2	*an* + Obj. Class	28	70.00	12	30.00	4	7	7	10

Table 2: Classification results across experiments and in the context of the classes.

4 Discussion

The optimal classification is performed with a combination of features that largely correspond to the linguistic intuitions based on our former linguistic studies. Thus, we succeeded in transforming our theoretical insights about the semantic verb classes into empirical, corpus-based features, and replicated the semantic classification by 70%. The machine learning model creates a classification whose agreement with the gold standard is twice as good as the baseline and only 9% below the upper bound by the human judgments.

The decision trees provide insight into the most indicative features of the experiments. For example, at the top of the tree in Figure 2, the semantic class *Event* is identified as an effective feature for the event initiation class. This corresponds to the theoretical observation that the semantics of the verbs operates on events which are often introduced through a direct object. Table 1 shows that the verbs *anspornen, anstiften* and *anstimmen* are wrongly classified as directional verbs because they usually take *Higher life form* as an object, which however is used as a main feature for the directional class. The event in these cases is then expressed by a PP with *zu.* The talk will have a focus on the analyses of the decision trees underlying the experiment decisions, and inspect the actual class assignments of the particle verbs comprehensively and in detail.

We also succeeded in the second goal, to deepen our insights to the semantic properties of the PVs through the computational analysis. For example, the verb *anspornen* was characterized as event initiation, but we found that this classification is not sufficient because it also has a directional component, cf. *'Der Chef spornt seine Mitarbeiter zu Höchstleistungen an'* (The boss incites his employees to work more efficiently). A continuative usage of our classification model is to disambiguate the most common meaning in a domain of verbs with more than one reading.

References

Gertrud Faaß, Ulrich Heid, and Helmut Schmid (2010). Design and application of a gold standard for morphological analysis: SMOR as an example of morphological evaluation. In *Proceedings of LREC.*

Mark Hall, Eibe Frank, Georey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11(1).

Hans Kamp and Uwe Reyle (1993). From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. *Kluwer, Dordrecht.*

Andrea Lechler and Antje Roßdeutscher (2009). German particle verbs with *auf*. Reconstructing composition in a DRT-based framework. *Linguistische Berichte*, 220:439478.

Michael Schiehlen (2003). A cascaded finite-state parser for German. In *Proceedings of the 10th EACL.*

Helmut Schmid (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the 1st International Conference on New Methods in Language Processing.*

Sabine Schulte im Walde (2005). Exploring features to identify semantic nearest neighbours: A case study on German particle verbs. In *Proceedings of RANLP.*

Sylvia Springorum (2009). Zur Semantik der Partikelverben mit *an*. Eine Studie zur Konstruktion ihrer Bedeutung im Rahmen der Diskursrepräsentationstheorie. *Studienarbeit. Universität Stuttgart*.

Sylvia Springorum (2011a). DRT-based analysis of the German verb particle *an*. *Leuvense Bijdragen 97.*

Sylvia Springorum (2011b). Untersuchungen zur automatischen Klassifikation von Partikelverben mit *an*. *Diplomarbeit. Universität Stuttgart.*

Barbara Stiebels. (1996). Lexikalische Argumente und Adjunkte: Zum semantischen Beitrag von verbalen Präfixen und Partikeln (Studia Grammatica 39). *Berlin: Akademie Verlag.*