Sabine Schulte im Walde

GermaNet Synsets as Selectional Preferences in Semantic Verb Clustering

Abstract

WordNet and its German version GermaNet have widely been used as source for fine-grained selectional preference information, focusing on but not restricted to verb-object relationships (Resnik 1997; Ribas 1995; Li & Abe 1998; Abney & Light 1999; Wagner 2000; McCarthy 2001; Clark & Weir 2002). In contrast, this paper presents an approach where argument slots of variable verb-frame combinations are refined by coarse selectional preferences as obtained from the top-level GermaNet nodes. The selectional preference information is applied to an alternation-like verb description and successfully utilised for an automatic clustering of German verbs (Schulte im Walde 2003b).

1 Introduction

This work is concerned with the definition and benefit of selectional preferences as used in an alternation-like verb description for the automatic induction of German semantic verb classes. Semantic verb classes are an artificial construct of natural language which generalises over verbs according to their semantic properties; the class labels refer to the common semantic properties of the verbs in a class at a general conceptual level, and the idiosyncratic lexical semantic properties of the verbs are either added to the class description or left underspecified. Examples for the conceptual classes are Position verbs such as *liegen* 'to lie', *sitzen* 'to sit', *stehen* 'to stand', and *Manner of Motion with a Vehicle* verbs such as *fahren* 'to drive', *fliegen* 'to fly', *rudern* 'to row'. On the one hand, verb classes reduce redundancy in verb descriptions, since they encode the common properties of verbs. On the other hand, verb classes can predict and refine properties of a verb that received insufficient empirical evidence, with reference to verbs in the same class; under this aspect, a verb classification is especially useful for the pervasive problem of data sparseness in NLP, where little or no knowledge is provided for rare events.

But how can one obtain a semantic classification of verbs, avoiding a tedious manual definition of the verbs and the classes? A semantic classification demands a definition of semantic properties, but it is difficult to automatically induce semantic features from available resources, both with respect to lexical semantics and conceptual structure. Therefore, the construction of semantic classes typically benefits from a long-standing linguistic hypothesis which asserts a tight connection between the lexical meaning of a verb and its behaviour: To a certain extent, the lexical meaning of a verb determines its behaviour, particularly with respect to the choice of its arguments, cf. Levin 1993. We can utilise this meaning-behaviour relationship in that we induce a verb classification on basis of verb features describing verb behaviour (which are easier to obtain automatically than semantic features) and expect the resulting behaviour-classification to agree with a semantic classification to a certain extent.

A widely used approach to define verb behaviour is captured by the diathesis alternation of verbs (see for example Levin 1993; Dorr & Jones 1996; Lapata 1999; Schulte im Walde 2000; Merlo & Stevenson 2001; McCarthy 2001; Joanis 2002). Alternations are alternative constructions at the syntax-semantic interface which express the same or a similar conceptual idea of a verb. In Example (1), the most common alternations for the *Manner of Motion with a Vehicle* verb *fahren* 'to drive' are illustrated. The participants in the conceptual structure are a driver, a vehicle, a driven person or thing, and a direction. In (a), the vehicle is expressed as subject in a transitive verb construction, with a prepositional phrase indicating the direction of the movement. In (b), the driver is expressed as subject in a transitive verb construction, again with a prepositional phrase indicating the direction. In (c), the driver is expressed as subject in a transitive verb construction, with an accusative noun phrase indicating the vehicle. And in (d), the driver is expressed as subject in a ditransitive verb construction, with an accusative noun phrase indicating a driven person, and a prepositional phrase indicating the direction of the movement. Even if a certain participant is not realised within an alternation, its contribution might be implicitly defined by the verb. For example, in (a) the driver is not expressed overtly, but we know that there is a driver, and in (b) and (d) the vehicle is not expressed overtly, but we know that there is a vehicle.

(1)
(a) *Der Wagen fährt in die Innenstadt.*
'The car drives to the city centre.'
(b) *Die Frau fährt nach Hause.*
'The woman drives home.'
(c) *Der Filius fährt einen blauen Ferrari.*
'The son drives a blue Ferrari.'
(d) *Der Junge fährt seinen Vater zum Zug.*
'The boy drives his father to the train.'

Assuming that the verb behaviour can be captured by the diathesis alternation of the verb, which are the relevant syntactic and semantic properties one would have to obtain for a verb description? The verbs are distributionally described on three levels, each of them refining the previous level by additional information. The first level *D1* encodes a purely syntactic definition of verb subcategorisation, the second level *D2* encodes a syntactico-semantic definition of subcategorisation with prepositional preferences, and the third level *D3* encodes a syntactico-semantic definition of subcategorisation with prepositional and selectional preferences. The most elaborated description comes close to a definition of verb alternation behaviour. The benefit of each information level can be determined with respect to the lower levels of information.

This paper concentrates on the definition and benefit of selectional preferences at *D3*, the alternation-like verb description. The selectional preferences are based on the German noun hierarchy in GermaNet (Hamp & Feldweg 1997; Kunze 2000), by specifying a coarse generalisation on the top-level synsets for argument slots of variable verb-frame combinations. Section 2 introduces the alternation-like verb descriptions, and Section 3 describes the automatic induction of semantic verb classes as based on the verb descriptions. Finally, Section 4 discusses the usage of the selectional preference information in semantic verb clustering with respect to the demands of German verbs and verb classes.

2 Alternation-Like Verb Descriptions for Verb Clustering

I have developed a statistical grammar model for German which provides empirical lexical information, specialising on but not restricted to the subcategorisation behaviour of verbs (Schulte im Walde 2002; Schulte im Walde 2003a). The grammar model serves as source for a German verb description at the syntax-semantic interface. For *D1*, it provides frequency distributions of German verbs over 38 purely syntactic subcategorisation frames, which comprise maximally three arguments. Possible arguments in the frames are nominative (n), dative (d) and accusative (a) noun phrases, reflexive pronouns (r), prepositional phrases (p), expletive *es* (x), subordinated non-finite clauses (i), subordinated finite clauses (s-2 for verb second clauses, s-dass for *dass*-clauses, s-ob for *ob*-clauses, s-w for indirect *wh*-questions), and copula constructions (k). For example, subcategorising a direct (accusative case) object and a non-finite clause would be represented by 'nai'.

In addition to a purely syntactic definition of subcategorisation frames, the grammar provides detailed information for *D*2 about the types of PPs within the frames. For each of the prepositional phrase frame types in the grammar, the joint frequency of a verb and the PP frame is distributed over the prepositional phrases, according to their frequencies in the corpus. Prepositional phrases are defined by case and preposition, such as 'mit_{Dat}' and '$für_{Akk}$'. The total number of features on *D*2 is 183.

For *D*3, the verb-frame combinations are refined by selectional preferences, i.e. the argument slots within a subcategorisation frame type are specified according to which 'kind' of argument they require. 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. I.e. the grammar provides frequencies for heads for each verb and each frame type and each argument slot of the frame type. For example, the most frequent nominal argument heads for the verb *verfolgen* 'to follow' and the accusative NP of the transitive frame type 'na' are *Ziel* 'goal', *Strategie* 'strategy', *Politik* 'policy', *Interesse* 'interest', *Konzept* 'concept', *Entwicklung* 'development', *Kurs* 'direction', *Spiel* 'game', *Plan* 'plan', *Spur* 'trace'.

Obviously, we would run into a sparse data problem if we tried to incorporate selectional preferences into the verb descriptions on such a specific level. We are provided with rich information on the nominal level, but we need a generalisation of the selectional preference definition. *WordNet* (MILLER ET AL. 1990; FELLBAUM 1998) and its German version *GermaNet* (HAMP & FELDWEG 1997; KUNZE 2000) have widely been used as source for fine-grained selectional preference information (RESNIK 1997; RIBAS 1995; LI & ABE 1998; ABNEY & LIGHT 1999; WAGNER 2000; MCCARTHY 2001; CLARK & WEIR 2002). In contrast to these approaches, I utilise the German noun hierarchy in GermaNet for a *coarse* generalisation of selectional preferences. The hierarchy is realised by means of synsets, sets of synonymous nouns, which are organised by multiple inheritance hyponym/hypernym relationships. A noun can appear in several synsets, according to its number of senses. Figure 1 illustrates the (slightly simplified) GermaNet hierarchy for the noun *Kaffee* 'coffee', which is encoded with two senses, (1) as a beverage and luxury food, and (2) as expression for an afternoon meal. Both senses are subsumed under the general top-level node *Objekt* 'object'. My approach is as follows. For each noun in a verb-frame-slot combination, the joint frequency is split over the different senses of the noun and propagated upwards the hierarchy. In case of multiple hypernym synsets, the frequency is split again. The sum of frequencies over all top synsets equals the total joint frequency. For example, we assume that the frequency of the noun *Kaffee* 'coffee' with respect to the verb *trinken* 'to drink' and the accusative argument in the transitive frame 'na' is 10. Each of the two synsets containing *Kaffee* is therefore assigned a value of 5, and the node values are propagated upwards, as Figure 1 illustrates. Repeating the frequency assignment and propagation for all nouns appearing in a verb-frame-slot combination, the result defines a frequency distribution of the verb-frame-slot combination over all GermaNet synsets.

To restrict the variety of noun concepts to a general level, I consider only the frequency distributions over the top GermaNet nodes. Since GermaNet had not been completed at the point of time I have used the hierarchy, I have manu-

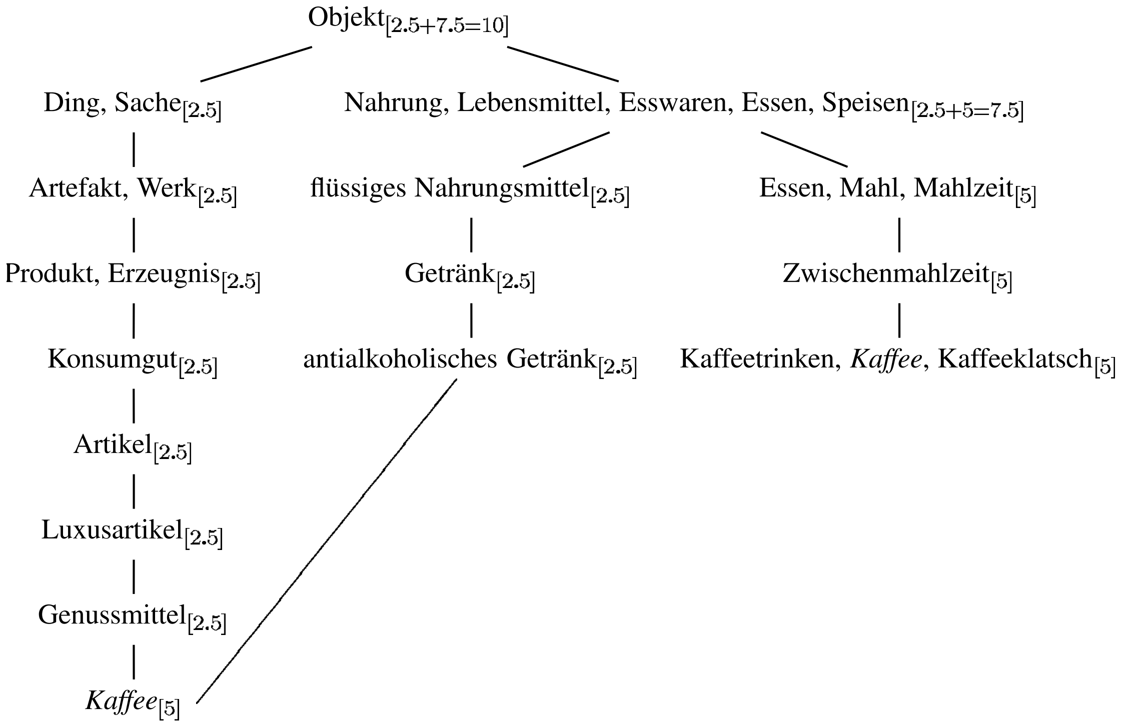

Figure 1: Propagating frequencies through the GermaNet hierarchy

ally added few hypernym definitions, such that most branches are subsumed under the following 15 conceptual top levels. Most of them were already present; the additional links might be regarded as a refinement.

Since the 15 nodes exclude each other and the frequencies sum to the total joint verb-frame frequency, we can use the frequencies to define probability distributions. Therefore, the 15 nodes provide a coarse definition of selectional preferences for a verb-frame-slot combination. Table 1 presents three example verb-frame-slot combinations (the relevant frame slot is underlined) with their preferences. This coarse selectional preference information is provided for each verb-frame-slot combination in the grammar model (trained on 35 million words of German newspaper corpora).

- Lebewesen ‘creature’
- Sache ‘thing’
- Besitz ‘property’
- Substanz ‘substance’
- Nahrung ‘food’
- Mittel ‘means’
- Situation ‘situation’
- Zustand ‘state’
- Struktur ‘structure’
- Physis ‘body’
- Zeit ‘time’
- Ort ‘space’
- Attribut ‘attribute’
- Kognitives Objekt ‘cognitive object’
- Kognitiver Prozess ‘cognitive process’

Table 2 summarises the verb distributions and presents three verbs from different verb classes and their ten most frequent frame types with respect to the three levels of verb definition, accompanied by the probability values. On *D2* frame types including PPs are specified for the PP type, and on *D3* the frame slot for selectional preference refinement is underlined, and the top-level synset is given in brackets. *D1* for *beginnen* ‘to begin’ defines ‘np’ and ‘n’ as the most probable frame types. Even by splitting the ‘np’

Verb	Frame+Slot	Top-Level Synset	Freq	Prob
verfolgen **'to follow'**	na	Situation	140.99	0.244
		Kognitives Objekt	109.89	0.191
		Zustand	81.35	0.141
		Sache	61.30	0.106
		Attribut	52.69	0.091
		Lebewesen	46.56	0.081
		Ort	45.95	0.080
		Struktur	14.25	0.025
		Kognitiver Prozess	11.77	0.020
		Zeit	4.58	0.008
		Besitz	2.86	0.005
		Substanz	2.08	0.004
		Nahrung	2.00	0.003
		Physis	0.50	0.001
essen **'to eat'**	na	Nahrung	127.98	0.399
		Sache	66.49	0.207
		Lebewesen	50.06	0.156
		Attribut	17.73	0.055
		Zeit	11.98	0.037
		Substanz	11.88	0.037
		Kognitives Objekt	10.70	0.033
		Struktur	8.55	0.027
		Ort	4.91	0.015
		Zustand	4.26	0.013
		Situation	2.93	0.009
		Besitz	1.33	0.004
		Mittel	0.67	0.002
		Physis	0.67	0.002
		Kognitiver Prozess	0.58	0.002
beginnen **'to begin'**	n	Situation	1,102.26	0.425
		Zustand	301.82	0.116
		Zeit	256.64	0.099
		Sache	222.13	0.086
		Kognitives Objekt	148.12	0.057
		Kognitiver Prozess	139.55	0.054
		Ort	107.68	0.041
		Attribut	101.47	0.039
		Struktur	87.08	0.034
		Lebewesen	81.34	0.031
		Besitz	36.77	0.014
		Physis	4.18	0.002
		Substanz	3.70	0.001
		Nahrung	3.29	0.001

Table 1: Selectional preference definition with GermaNet top nodes.

	Distribution					
Verb	***D*1**		***D*2**		***D*3**	
beginnen	np	0.43	n	0.28	$\underline{n}$(Situation)	0.12
'to begin'	n	0.28	np:um$_{Akk}$	0.16	$\underline{n}$p:um$_{Akk}$(Situation)	0.09
	ni	0.09	ni	0.09	$\underline{n}$p:mit$_{Dat}$(Situation)	0.04
	na	0.07	np:mit$_{Dat}$	0.08	$\underline{n}$i(Lebewesen)	0.03
	nd	0.04	na	0.07	$\underline{n}$(Zustand)	0.03
	nap	0.03	np:an$_{Dat}$	0.06	$\underline{n}$p:an$_{Dat}$(Situation)	0.03
	nad	0.03	np:in$_{Dat}$	0.06	$\underline{n}$p:in$_{Dat}$(Situation)	0.03
	nir	0.01	nd	0.04	$\underline{n}$(Zeit)	0.03
	ns-2	0.01	nad	0.03	$\underline{n}$(Sache)	0.02
	xp	0.01	np:nach$_{Dat}$	0.01	n$\underline{a}$(Situation)	0.02
essen	na	0.42	na	0.42	$\underline{n}$a(Lebewesen)	0.33
'to eat'	n	0.26	n	0.26	n$\underline{a}$(Nahrung)	0.17
	nad	0.10	nad	0.10	n$\underline{a}$(Sache)	0.09
	np	0.06	nd	0.05	$\underline{n}$(Lebewesen)	0.08
	nd	0.05	ns-2	0.02	n$\underline{a}$(Lebewesen)	0.07
	nap	0.04	np:auf$_{Dat}$	0.02	$\underline{n}$(Nahrung)	0.06
	ns-2	0.02	ns-w	0.01	$\underline{n}$(Sache)	0.04
	ns-w	0.01	ni	0.01	$\underline{n}$d(Lebewesen)	0.04
	ni	0.01	np:mit$_{Dat}$	0.01	n$\underline{d}$(Nahrung)	0.02
	nas-2	0.01	np:in$_{Dat}$	0.01	n$\underline{a}$(Attribut)	0.02
fahren	n	0.34	n	0.34	$\underline{n}$(Sache)	0.12
'to drive'	np	0.29	na	0.19	$\underline{n}$(Lebewesen)	0.10
	na	0.19	np:in$_{Akk}$	0.05	$\underline{n}$a(Lebewesen)	0.08
	nap	0.06	nad	0.04	n$\underline{a}$(Sache)	0.06
	nad	0.04	np:zu$_{Dat}$	0.04	$\underline{n}$(Ort)	0.06
	nd	0.04	nd	0.04	$\underline{n}$a(Sache)	0.05
	ni	0.01	np:nach$_{Dat}$	0.04	$\underline{n}$p:in$_{Akk}$(Sache)	0.02
	ns-2	0.01	np:mit$_{Dat}$	0.03	$\underline{n}$p:zu$_{Dat}$(Sache)	0.02
	ndp	0.01	np:in$_{Dat}$	0.03	$\underline{n}$p:in$_{Akk}$(Lebewesen)	0.02
	ns-w	0.01	np:auf$_{Dat}$	0.02	$\underline{n}$p:nach$_{Dat}$(Sache)	0.02

Table 2: Examples of most probable frame types.

probability over the different PP types in *D*2, a number of prominent PPs are left, the time indicating *um*$_{Akk}$ and *nach*$_{Dat}$, *mit*$_{Dat}$ referring to the begun event, *an*$_{Dat}$ as date and *in*$_{Dat}$ as place indicator. It is obvious that adjunct PPs as well as argument PPs represent a distinctive part of the verb behaviour. *D*3 illustrates that typical selectional preferences for beginner roles are *Situation*, *Zustand*, *Zeit*, *Sache*. *D*3 has the potential to indicate verb alternation behaviour, e.g. 'n$\underline{a}$(Situation)' refers to the same role for the direct object in a transitive frame as '$\underline{n}$(Situation)' in an intransitive frame. *essen* 'to eat' as an object drop verb shows strong preferences for both intransitive and transitive usage. As desired, the argument roles are strongly determined by *Lebewesen* for both '$\underline{n}$' and '$\underline{n}$a' and *Nahrung* for 'n$\underline{a}$'. *fahren* 'to drive' chooses typical manner of motion frames ('n', 'np', 'na') with the refining PPs being directional (*in*$_{Akk}$, *zu*$_{Dat}$, *nach*$_{Dat}$) or referring to a means of motion (*mit*$_{Dat}$, *in*$_{Dat}$, *auf*$_{Dat}$). The selectional preferences represent a correct alternation behaviour: *Lebewesen* in the object drop case for '$\underline{n}$' and '$\underline{n}$a', *Sache* in the inchoative/causative case for '$\underline{n}$' and 'n$\underline{a}$'.

3 Induction of Semantic Verb Classes

The selectional preference information is applied to an alternation-like verb description in automatic verb clustering. The clustering of the German verbs is performed by the k-Means algorithm, a standard unsupervised clustering technique as proposed by Forgy 1965. Based on the distributional verb descriptions and standard notions of similarity between distributional vectors, k-Means iteratively re-organises initial verb clusters by assigning each verb to its closest cluster and re-calculating cluster centroids until no further changes take place. For details on the clustering setup and experiments, the reader is referred to Schulte im Walde 2003b.

The clustering experiments are performed on 168 partly ambiguous German verbs. Before the experiments, I manually classified the verbs into 43 semantic classes. The purpose of the manual classification is to evaluate the reliability and performance of the clustering experiments. In the following, I present representative parts of a cluster analysis which uses the alternation-like verb description on *D*3. For each cluster, the verbs which belong to the same gold standard class are presented in one line, accompanied by the class label. I compare the respective clusters with their pendants under *D*1 and *D*2, to demonstrate the effect of the feature refinements.

(a) nieseln regnen schneien – *Weather*

(b) dämmern – *Weather*

(c) kriechen rennen – *Manner of Motion*: *Locomotion*
eilen – *Manner of Motion: Rush*
gleiten – *Manner of Motion: Flotation*
starren – *Facial Expression*

(d) klettern wandern – *Manner of Motion*: *Locomotion*
fahren fliegen segeln – *Manner of Motion: Vehicle*
fließen – *Manner of Motion: Flotation*

(e) beginnen enden – *Aspect*
bestehen existieren – *Existence*
liegen sitzen stehen – *Position*
laufen – *Manner of Motion: Locomotion*

(f) festlegen – *Constitution*
bilden – *Production*
erhöhen senken steigern vergrößern verkleinern – *Quantum Change*

(g) töten – *Elimination*
unterrichten – *Teaching*

The weather verbs in cluster (a) strongly agree in their syntactic expression on *D*1 and do not need *D*2 or *D*3 refinements for a successful class constitution. *dämmern* in cluster (b) is ambiguous between a weather verb and expressing a sense of understanding; this ambiguity is idiosyncratically expressed in *D*1 frames already, so *dämmern* is never clustered together with the other weather verbs on *D*1-*D*3. *Manner of Motion, Existence, Position* and *Aspect* verbs are similar in their syntactic frame usage and therefore merged together on *D*1, but adding PP information distinguishes the respective verb classes: *Manner of Motion* verbs primarily demand directional PPs, *Aspect* verbs are distinguished by patient *mit*$_{Dat}$ and time and location prepositions, and *Existence* and *Position* verbs are distinguished by locative prepositions, with *Position* verbs showing more PP variation. The PP information is essential for successfully distinguishing these verb classes, and the coherence is partly destroyed by *D*3: *Manner of Motion* verbs (from the sub-classes *Locomotion, Rotation, Rush, Vehicle, Flotation*) are captured well by clusters (c) and (d), since they inhibit strong common alternations, but cluster (e) merges the *Existence, Position* and *Aspect* verbs, since verb-idiosyncratic selectional preferences destroy the *D*2 class demarcation. Admittedly, the verbs in cluster (e) are close in their semantics, with a common sense of (bringing into vs. being in) existence. *laufen* fits into the cluster with its sense of ‘to function’. Cluster (f) contains most verbs of *Quantum Change*, together with one verb of *Production* and *Constitution* each. The semantics of the cluster is therefore rather pure. The verbs in the cluster typically subcategorise a direct object, alternating with a reflexive usage, 'nr' and 'npr' with mostly *auf*$_{Akk}$ and *um*$_{Akk}$. The selectional preferences help to distinguish this cluster: the verbs agree in demanding a thing or situation as subject, and various objects such as attribute, cognitive object, state, structure or thing as object. Without selectional preferences (on *D*1 and *D*2), the change of quantum verbs are not found together with the same degree of purity. There are verbs as in cluster (g), whose properties are correctly stated as similar on *D*1-*D*3, so a common cluster is justified; but the verbs only have coarse common meaning components, in this case *töten* and *unterrichten* agree in an action of one person or institution towards another.

4 Discussion

Which exactly is the nature of the meaning-behaviour relationship in the constitution of semantic verb classes? And, more specifically, which is the benefit of the selectional preferences in the alternation-like verb description as based on GermaNet top-level nodes?

Addressing the nature of the meaning-behaviour relationship in the clustering, (a) already a purely syntactic verb description allows a verb clustering clearly above the baseline. The result is a successful (semantic) classification of verbs which agree in their syntactic frame definitions, e.g. most of the *Support* verbs *dienen*, *helfen*, *folgen*. The clustering fails for semantically similar verbs which differ in their syntactic behaviour, e.g. *unterstützen* which does belong to the *Support* verbs but demands an accusative instead of a dative object. In addition, it fails for syntactically similar verbs which are clustered together even though they do not exhibit semantic similarity, e.g. many verbs from different semantic classes subcategorise an accusative object, so they are falsely clustered together. (b) Refining the syntactic verb information by prepositional phrases is helpful for the semantic clustering, not only in the clustering of verbs where the PPs are obligatory, but also in the clustering of verbs with optional PP arguments. The improvement underlines the linguistic fact that verbs which are similar in their meaning agree either on a specific prepositional complement (e.g. *glauben/denken an*$_{Akk}$) or on a more general kind of modification, e.g. directional PPs for manner of motion verbs. (c) Defining selectional preferences for arguments once more improves the clustering results, but the improvement is not as persuasive as when refining the purely syntactic verb descriptions by prepositional information. For example, the selectional preferences help demarcate the *Quantum Change* class, because the respective verbs agree in their structural as well as selectional properties. But in the *Consumption* class, *essen* and *trinken* have strong preferences for a food object, whereas *konsumieren* allows a wider range of object types. On the contrary, there are verbs which are very similar in their behaviour, especially with respect to a coarse definition of selectional preferences, but they do not belong to the same fine-grained semantic class, e.g. *töten* and *unterrichten*.

The description of the clustering examples has shown that the dividing line between the common and idiosyncratic features of verbs in a verb class defines the level of verb description which is relevant for the class constitution. The meaning components of verbs to a certain extent determine their behaviour, but this does not mean that all properties of all verbs in a common class are similar and we could extend and refine the feature description endlessly. The meaning of verbs comprises both (i) properties which are general for the respective verb classes, and (ii) idiosyncratic properties which distinguish the verbs from each other. As long as we define the verbs by those properties which represent the common parts of the verb classes, a clustering can succeed. But step-wise refining the verb description by including lexical idiosyncrasy, the emphasis of the common properties vanishes. Some verbs and verb classes are distinctive on a coarse feature level, some need fine-grained extensions, some are not distinctive with respect to any combination of features. There is no unique perfect choice and encoding of the verb features; the feature choice rather depends on the specific properties of the desired verb classes.

The usage of selectional preference information in semantic verb clustering is a particular challenge for the verb description. On the one hand, one would want a selectional preference description as fine-grained as possible, to e.g. distinguish the verbs *töten* and *unterrichten* which are similar on a coarse selectional preference level (agreeing in an action of one person or institution towards another), but distinguished on a fine-grained level: in a transitive construction, *töten* appears with subjects such as *Soldat* 'soldier', *Angreifer* 'attacker', *Schütze* 'shooter', *Terrorist* 'terrorist', *Jäger* 'hunter' and direct objects such as *Soldat* 'soldier', *Zivilist* 'civilian', *Rebell* 'rebel', *Nebenbuhler* 'rival', *Tier* 'animal', and *unterrichten* appears with subjects such as *Lehrerschaft* 'community of teachers', *College* 'college', *Professor* 'professor' and direct objects such as *Kind* 'child', *Schüler* 'pupil', *Klasse* 'class', *Fach* 'subject', *Grammatik* 'grammar'. Assuming that we use GermaNet as source for the preference definition, in the example case we would need an algorithm comparable to those by RESNIK 1997; RIBAS 1995; LI & ABE 1998; ABNEY & LIGHT 1999; WAGNER 2000; MCCARTHY 2001; CLARK & WEIR 2002 which is able to filter selectional preferences of arbitrary depth in the hierarchy. On the other hand, one would want a selectional preference description on a more general level. Consider the most specific conceptual level of semantic classes, a classification with classes of verb synonyms.[1] But even the verb behaviour of synonyms does not overlap perfectly, since e.g. selectional preferences of synonyms vary. For example, the German verbs *bekommen* and *erhalten* 'to get, to receive' are synonymous, but they cannot be exchanged in all contexts, cf. *einen Schnupfen bekommen* 'to catch a cold' vs. *einen Schnupfen erhalten*. This means that even for synonyms a fine-grained definition of selectional preferences would not provide a perfect overlap of the distributional features and that some generalisation is desirable.

In addition to the linguistic conflict in clustering when defining selectional preferences for verbs, a clustering algorithm has to pay attention to the technical issue of feature encoding. We would run into a sparse data problem if we tried to incorporate selectional preferences into the verb descriptions on a fine-grained level. Again, this means that some generalisation level of selectional preferences is adequate.

Summarising, both the theoretical assumption of encoding features of verb alternation as verb behaviour and the practical realisation by encoding syntactic frame types, prepositional phrases and selectional preferences have proven successful. But the exact feature choice for verb descriptions in verb clustering depends on the specific properties of the desired verb classes. And even if classes are perfectly defined on a common conceptual level, the relevant level of behavioural properties of the verb classes might differ. This insight is especially problematic for the definition of selectional preferences, since numerous variations for their encoding are possible, but each choice would present advantages for some verb classes and disadvantages for others. This work has presented evidence for the usefulness of GermaNet top levels nodes as coarse generalisation of selectional preferences, but the issue of improving the level of GermaNet preference definitions is subject to further work.

Acknowledgements

The work reported here was performed while the author was a member of the DFG-funded PhD program 'Graduiertenkolleg' *Sprachliche Repräsentationen und ihre Interpretation* at the Institute for Natural Language Processing (IMS), University of Stuttgart, Germany.

Note

[1] In this context, synonymy refers to 'partial synonymy' where synonymous verbs cannot necessarily be exchanged in all contexts, as compared to 'total synonymy' where synonymous verbs can be exchanged in all contexts – if anything like 'total synonymy' exists at all (Bussmann 1990).

References

Abney, S.; Light, M. (1999). "Hiding a Semantic Class Hierarchy in a Markow Model." In: Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing, College Park, MD, 1-8.

Bussmann, H. (1990^2). Lexikon der Sprachwissenschaft. Stuttgart: Alfred Kröner Verlag.

Clark, S.; Weir, D. (2002). "Class-Based Probability Estimation using a Semantic Hierarchy." In: Computational Linguistics, 28(2) (2002), 187-206.

Dorr, B. J.; Jones, D. (1996). "Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues." In: Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, August 1996, 322-327.

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Forgy, E. W. (1965). "Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications." In: Biometrics 21 (1965), 768-780.

Hamp, B.; Feldweg, H. (1997). "GermaNet - a Lexical-Semantic Net for German." In: Vossen, P. et al. (eds.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Buliding of Lexical-Semantic Resources for NLP Applications, 9-15.

Joanis, E. (2002). Automatic Verb Classification using a General Feature Space. Master's thesis, University of Toronto, Department of Computer Science.

Kunze, C. (2000). "Extension and Use of GermaNet, a Lexical-Semantic Database." In: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, May 2000, 999-1002.

Lapata, M. (1999). "Acquiring Lexical Generalizations from Corpora: A Case Study for Diathesis Alternations." In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99), College Park, MD, 397-404.

Levin, B. (1993). English Verb Classes and Alternations. Chicago, IL: The University of Chicago Press.

Li, H.; Abe, N. (1998). "Generalizing Case Frames Using a Thesaurus and the MDL Principle." In: Computational Linguistics 24(2) (1998), 217-244.

McCarthy, D. (2001). Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. Ph.D. thesis, University of Sussex, Brighton, UK.

Merlo, P.; Stevenson, S. (2001). "Automatic Verb Classification Based on Statistical Distributions of Argument Structure." In: Computational Linguistics 27(3) (2001), 373-408.

Miller, G. A. et al. (1990). "Introduction to Wordnet: An On-line Lexical Database." In: International Journal of Lexicography 3(4) (1990), 235-244.

Resnik, P. (1997). "Selectional Preference and Sense Disambiguation." In: Proceedings of the ACL SIGLEX / ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, Washington, D. C., April 1997.

Ribas, F. (1995). "On Learning More Appropriate Selectional Restrictions." In: Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL'95), Dublin, March 1995.

Schulte im Walde, S. (2000). "Clustering Verbs Semantically According to their Alternation Behaviour." In: Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), Saarbrücken, Germany, August 2000, 747-753.

Schulte im Walde, S. (2002). "A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG." In: Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, May/June 2002, vol. IV, 1351-1357.

Schulte im Walde, S. (2003a). "A Collocation Database for German Nouns and Verbs." In: Proceedings of the 7th Conference on Computational Lexicography and Text Research (COMPLEX 2003), Budapest, April 2003.

Schulte im Walde, S. (2003b). Experiments on the Automatic Induction of German Semantic Verb Classes. Ph.D. thesis, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung [= AIMS Report 9(2)].

Wagner, A. (2000). "Enriching a Lexical Semantic Net with Selectional Preferences by Means of Statistical Corpus Analysis." In: Proceedings of the ECAI-2000 Workshop on Ontology Learning, Berlin, August 2000, 37- 42.