Automatic Induction of Semantic Classes for German Verbs*

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

July 26, 2004

1 Motivation

The verb is an especially relevant part of the sentence, since it is central to the structure and the meaning of the sentence: The verb determines the number and kind of the obligatory and facultative participants within the sentence, and the proposition of the sentence is defined by the structural and conceptual interaction between the verb and the sentence participants. For that reason, lexical verb information represents the core in supporting computational tasks in Natural Language Processing (NLP) such as lexicography, parsing, machine translation, and information retrieval, which depend on reliable language resources. But especially lexical semantic resources represent a bottleneck in NLP, and methods for the acquisition of large amounts of semantic knowledge with comparably little manual effort have gained importance. In this context, I am concerned with the potential and limits of creating a semantic knowledge base by automatic means, semantic classes for German verbs.

Semantic verb classes generalise over verbs according to their semantic properties. They represent a practical means to capture large amounts of verb knowledge without defining the idiosyncratic details for each verb. 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 conceptual structures 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'. 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 (Pinker, 1989; 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 common approach to define verb behaviour is captured by the diathesis alternation of verbs. 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*

*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.

verb *fahren* 'to drive' are illustrated. The participants in the conceptual structure are a vehicle, a driver, a driven person, and a direction. In (a), the vehicle is expressed as subject in a transitive verb construction, with a prepositional phrase indicating the direction. In (b), the driver is expressed as subject in a transitive verb construction, 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. 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. 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.'

For modelling verb alternation behaviour by automatic means, a statistical grammar model for German provides empirical lexical information, specialising on but not restricted to the subcategorisation behaviour of verbs (Schulte im Walde, 2002, 2003a). The grammar model is utilised for verb descriptions at three levels at the syntax-semantic interface, a purely syntactic definition of verb subcategorisation, a syntactico-semantic definition of subcategorisation with prepositional preferences, and a syntactico-semantic definition of subcategorisation with prepositional and selectional preferences. The most elaborated description comes close to a definition of the verb alternation behaviour. Based on the syntactico-semantic descriptions of the German verbs as empirical verb properties, the standard clustering algorithm k-Means (Forgy, 1965) is applied to induce a semantic classification for the verbs.

What is the usage of semantic verb classes in Natural Language Processing applications? 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. Previous work on verb classes has proven their usefulness: particularly the English verb classification by Levin (1993) has been used for NLP applications such as word sense disambiguation (Dorr and Jones, 1996), machine translation (Dorr, 1997), and document classification (Klavans and Kan, 1998).

The plan of the article is as follows. Section 2 introduces the idea of semantic verb classes and presents a manually defined classification of German verbs. Section 3 describes the empirical verb behaviour and the definition of the clustering methodology. In Section 4, I present and interpret clustering results, and Section 5 closes with examples for using the learned verb classes in NLP applications.

2 German Semantic Verb Classes

Semantic verb classes generalise over verbs according to their semantic properties. They represent a practical means to capture large amounts of verb knowledge without defining the idiosyncratic details for each verb. 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. Semantic verb classes have been defined for several languages: frame-semantic descriptions for the computational lexicographic database resource *FrameNet* for English (Baker

et al., 1998; Johnson *et al.*, 2002) and in early stages also for German (Erk *et al.*, 2003); the lexical semantic ontology *WordNet* for English (Miller *et al.*, 1990; Fellbaum, 1998) and *EuroWordNet* (Vossen, 1999) for Dutch, Italian, Spanish, French, German, Czech and Estonian; and multi-lingual verb classes as based on the syntax-semantic relationship for English (Levin, 1993), Spanish (Vázquez *et al.*, 2000) and French (Saint-Dizier, 1998). To my knowledge, no German verb classification is available for NLP applications. A large-scale German classification would therefore provide a principled basis for filling a gap in available lexical knowledge.

I manually defined 43 German semantic verb classes containing 168 partly ambiguous German verbs. The small-scale manual classification represents a gold standard in order to evaluate the reliability and performance of the clustering experiments. The manual construction of the German verb classes was primarily based on semantic intuition: Verbs are assigned to classes according to similarity of lexical and conceptual meaning, and each verb class is assigned a conceptual class label. The class labels are given on two conceptual levels; coarse labels such as *Manner of Motion* are sub-divided into finer labels, such as *Locomotion, Rotation, Rush, Vehicle, Flotation*. Tthe classification is primarily based on semantic intuition, not on facts about the syntactic behaviour. As an extreme example, the *Support* class (23) contains the verb *unterstützen*, which syntactically requires a direct object, together with the three verbs *dienen, folgen, helfen* which dominantly subcategorise an indirect object. Because of the meaning-behaviour relationship at the syntax-semantic interface, the verbs grouped in one class show a certain agreement in their behaviour. The class size is between 2 and 7, with an average of 3.9 verbs per class. Eight verbs are ambiguous with respect to class membership. The classes include both high and low frequency verbs: the corpus frequencies of the verbs range from 8 to 71,604. I tried to balance the classification not to include any kind of bias, i.e. in the classification are no majorities of high frequent verbs, low frequent verbs, strongly ambiguous verbs, verbs from specific semantic areas, etc. Any bias in the classification could influence the evaluation of clustering methods.

1. *Aspect*: anfangen, aufhören, beenden, beginnen, enden
2. *Propositional Attitude*: ahnen, denken, glauben, vermuten, wissen
3. *Desire*
   (a) *Wish*: erhoffen, wollen, wünschen
   (b) *Need*: bedürfen, benötigen, brauchen
4. *Transfer of Possession (Obtaining)*: bekommen, erhalten, erlangen, kriegen
5. *Transfer of Possession (Giving)*
   (a) *Gift*: geben, leihen, schenken, spenden, stiften, vermachen, überschreiben
   (b) *Supply*: bringen, liefern, schicken, vermitteln$_1$, zustellen
6. *Manner of Motion*
   (a) *Locomotion*: gehen, klettern, kriechen, laufen, rennen, schleichen, wandern
   (b) *Rotation*: drehen, rotieren
   (c) *Rush*: eilen, hasten
   (d) *Vehicle*: fahren, fliegen, rudern, segeln
   (e) *Flotation*: fließen, gleiten, treiben
7. *Emotion*
   (a) *Origin*: ärgern, freuen
   (b) *Expression*: heulen$_1$, lachen$_1$, weinen
   (c) *Objection*: ängstigen, ekeln, fürchten, scheuen
8. *Facial Expression*: gähnen, grinsen, lachen$_2$, lächeln, starren
9. *Perception*: empfinden, erfahren$_1$, fühlen, hören, riechen, sehen, wahrnehmen
10. *Manner of Articulation*: flüstern, rufen, schreien
11. *Moaning*: heulen$_2$, jammern, klagen, lamentieren
12. *Communication*: kommunizieren, korrespondieren, reden, sprechen, verhandeln

13. *Statement*
    (a) *Announcement*: ankündigen, bekanntgeben, eröffnen, verkünden
    (b) *Constitution*: anordnen, bestimmen, festlegen
    (c) *Promise*: versichern, versprechen, zusagen
14. *Observation*: bemerken, erkennen, erfahren$_2$, feststellen, realisieren, registrieren
15. *Description*: beschreiben, charakterisieren, darstellen$_1$, interpretieren
16. *Presentation*: darstellen$_2$, demonstrieren, präsentieren, veranschaulichen, vorführen
17. *Speculation*: grübeln, nachdenken, phantasieren, spekulieren
18. *Insistence*: beharren, bestehen$_1$, insistieren, pochen
19. *Teaching*: beibringen, lehren, unterrichten, vermitteln$_2$
20. *Position*
    (a) *Bring into Position*: legen, setzen, stellen
    (b) *Be in Position*: liegen, sitzen, stehen
21. *Production*: bilden, erzeugen, herstellen, hervorbringen, produzieren
22. *Renovation*: dekorieren, erneuern, renovieren, reparieren
23. *Support*: dienen, folgen$_1$, helfen, unterstützen
24. *Quantum Change*: erhöhen, erniedrigen, senken, steigern, vergrößern, verkleinern
25. *Opening*: öffnen, schließen$_1$
26. *Existence*: bestehen$_2$, existieren, leben
27. *Consumption*: essen, konsumieren, lesen, saufen, trinken
28. *Elimination*: eliminieren, entfernen, exekutieren, töten, vernichten
29. *Basis*: basieren, beruhen, gründen, stützen
30. *Inference*: folgern, schließen$_2$
31. *Result*: ergeben, erwachsen, folgen$_2$, resultieren
32. *Weather*: blitzen, donnern, dämmern, nieseln, regnen, schneien

The classification is completed by a detailed class description which is closely related to Fillmore's scenes-and-frames semantics (Fillmore, 1977, 1982), as computationally utilised in *FrameNet* (Baker *et al.*, 1998; Johnson *et al.*, 2002). The frame-semantic class definition contains a prose scene description, predominant frame participant and modification roles, and frame variants describing the scene. The frame roles have been developed on basis of a large German newspaper corpus from the 1990s. They capture the scene description by idiosyncratic participant names and demarcate major and minor roles. Since a scene might be activated by various frame embeddings, I have listed the predominant frame variants as found in the corpus, marked with participating roles, and at least one example sentence of each verb utilising the respective frame. The corpus examples are annotated and illustrate the idiosyncratic combinations of lexical verb meaning and conceptual constructions, to capture the variants of verb senses. Following, I present a verb class description for the class of *Aspect* verbs.[1] Verbs allowing a frame variant are marked by '+', verbs allowing the frame variant only in company of an additional adverbial modifier are marked by '$+_{adv}$', and verbs not allowing a frame variant are marked by '$\neg$'. In the case of ambiguities, frame variants are only given for the senses of the verbs with respect to the class label. The frame variants with their roles marked represent the alternation potential of the verbs, by relating the different syntactic embeddings to identical role definitions. For example, the causative-inchoative alternation assumes the syntactic embeddings $\mathbf{n}_X\mathbf{a}_Y$ and $\mathbf{n}_Y$, indicating that the alternating verbs are realised by a transitive frame type (containing a nominative NP 'n' with role *X* and an accusative NP 'a' with role *Y*) and the corresponding intransitive frame type (with a nominative NP 'n' only, indicating the same role *Y* as for the transitive accusative). Appendix A lists all possible frame variants and illustrative examples. Passivisation of a verb-frame combination is indicated by [P].

[1] For further class descriptions, the reader is referred to Schulte im Walde (2003b, pages 27-103).

Aspect Verbs: *anfangen, aufhören, beenden, beginnen, enden*

Scene: [$_E$ An event] begins or ends, either internally caused or externally caused by [$_I$ an initiator]. The event may be specified with respect to [$_T$ tense], [$_L$ location], [$_X$ an experiencer], or [$_R$ a result].

Frame Roles: I(nitiator), E(vent)

Modification Roles: T(emporal), L(ocal), (e)X(periencer), R(esult)

Frame	Participating Verbs & Corpus Examples
$\mathbf{n}_E$	+ anfangen, aufhören, beginnen / $+_{adv}$ enden / ¬ beenden
	Nun aber muß [$_E$ der Dialog] **anfangen**. ... bevor [$_E$ der Golfkrieg] **angefangen** hatte ... ... damit [$_E$ die Kämpfe] **aufhören**. Erst muß [$_E$ das Morden] **aufhören**. [$_E$ Der Gottesdienst] **beginnt**. [$_E$ Das Schuljahr] **beginnt** [$_T$ im Februar]. [$_X$ Für die Flüchtlinge] **beginnt** nun [$_E$ ein Wettlauf gegen die Zeit]. [$_E$ Sein Zwischenspiel] bei der Wehrmacht **endete** ... [$_R$ glimpflich]. [$_E$ Die Ferien] **enden** [$_R$ mit einem großen Fest]. [$_E$ Druckkunst] ... **endet** [$_R$ beim guten Buch]. [$_E$ Die Partie] **endete** [$_R$ 0:1]. [$_L$ An einem Baum] **endete** in Höchst [$_E$ die Flucht] ... [$_E$ Der Informationstag] ... **endet** [$_T$ um 14 Uhr].
$\mathbf{n}_I$	+ anfangen, aufhören / ¬ beenden, beginnen, enden
	[$_I$ Die Hauptstadt] muß **anfangen**. ... daß [$_I$ er] [$_T$ pünktlich] **anfing**. Jetzt können [$_I$ wir] nicht einfach **aufhören**. Vielleicht sollte [$_I$ ich] **aufhören** und noch studieren.
$\mathbf{n}_I$ $\mathbf{a}_E$	+ anfangen, beenden, beginnen / ¬ aufhören, enden
	Nachdem [$_I$ wir] [$_E$ die Sache] **angefangen** haben, ... [$_I$ er] versucht, [$_E$ ein neues Leben] **anzufangen**. [$_I$ Die Polizei] **beendete** [$_E$ die Gewalttätigkeiten]. [$_T$ Nach dem Abi] **beginnt** [$_I$ Jens] [$_L$ in Frankfurt] [$_E$ seine Lehre] ...
$\mathbf{n}_I$ $\mathbf{a}_E$ $[P]$	+ anfangen, beenden, beginnen / ¬ aufhören, enden
	Wenn [$_E$ die Arbeiten] [$_T$ vor dem Bescheid] **angefangen** werden ... Während [$_X$ für Senna] [$_E$ das Rennen] **beendet** war ... ... ehe [$_E$ eine militärische Aktion] **begonnen** wird ...
$\mathbf{n}_I$ $\mathbf{i}_E$	+ anfangen, aufhören, beginnen / ¬ beenden, enden
	[$_I$ Ich] habe nämlich [$_E$ zu malen] **angefangen**. [$_I$ Ich] habe **angefangen**, [$_E$ Hemden zu schneidern]. [$_I$ Die Bahn] will [$_T$ 1994] **anfangen** [$_E$ zu bauen]. ... daß [$_I$ der Alkoholiker] **aufhört** [$_E$ zu trinken]. ... daß [$_I$ die Säuglinge] einfach **aufhören** [$_E$ zu atmen]. In dieser Stimmung **begannen** [$_I$ Männer] [$_E$ Tango zu tanzen] ... [$_I$ Tausende von Pinguinen] **beginnen** [$_E$ dort zu brüten].
$\mathbf{n}_I$ $\mathbf{p}_E : mit$	+ anfangen, aufhören, beginnen / ¬ beenden, enden
	Erst als [$_I$ der versammelte Hofstaat] [$_E$ mit Klatschen] **anfing**, Aber [$_I$ wir] müssen endlich [$_E$ damit] **anfangen**. [$_I$ Der Athlet] ... kann ... [$_E$ mit seinem Sport] **aufhören**. ... müßten noch [$_I$ viel mehr Frauen] [$_E$ mit ihrer Arbeit] **aufhören** ... Schließlich zog [$_I$ er] einen Trennstrich, **begann** [$_E$ mit dem Entzug] ... [$_I$ Man] **beginne** [$_E$ mit eher katharsischen Werken].
$\mathbf{n}_I$ $\mathbf{p}_E : mit$ $[P]$	+anfangen, aufhören, beginnen / ¬ beenden, enden
	Und [$_E$ mit den Umbauarbeiten] könnte **angefangen** werden. [$_E$ Mit diesem ungerechten Krieg] muß sofort **aufgehört** werden. [$_T$ Vorher] dürfe [$_E$ mit der Auflösung] nicht **begonnen** werden. ... daß [$_E$ mit dem Umbau] ... **begonnen** werden kann.

3 Clustering Methodology

I developed, implemented and trained 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, 2003a). The grammar model serves as source for the German verb description at the syntax-semantic interface. For the purely syntactic definition of subcategorisation frames ($D1$), it provides frequency distributions of German verbs over 38 purely syntactic subcategorisation frames, cf. Appendix A. In addition to $D1$, the grammar provides detailed information for the syntactico-semantic definition of subcategorisation with prepositional preferences ($D2$) 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$_{Akk}$'.

For the syntactico-semantic definition of subcategorisation with prepositional and selectional preferences ($D3$), 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 and Feldweg, 1997; Kunze, 2000) have widely been used as source for fine-grained selectional preference information (Resnik, 1997; Ribas, 1995; Li and Abe, 1998; Abney and Light, 1999; Wagner, 2000; McCarthy, 2001; Clark and Weir, 2002). I utilise the German noun hierarchy in GermaNet for a 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. 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. 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:[2] *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'. 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 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' probability over the different PP types in $D2$, a number of prominent PPs are left, the time indicating

[2] Since GermaNet had not been completed at the point of time I have used the hierarchy, I have manually added few hypernym definitions.

Verb	Distribution					
	$D1$		$D2$		$D3$	
beginnen	np	0.43	n	0.28	n(Situation)	0.12
'to begin'	n	0.28	np:um$_{Akk}$	0.16	np:um$_{Akk}$(Situation)	0.09
	ni	0.09	ni	0.09	np:mit$_{Dat}$(Situation)	0.04
	na	0.07	np:mit$_{Dat}$	0.08	ni(Lebewesen)	0.03
	nd	0.04	na	0.07	n(Zustand)	0.03
	nap	0.03	np:an$_{Dat}$	0.06	np:an$_{Dat}$(Situation)	0.03
	nad	0.03	np:in$_{Dat}$	0.06	np:in$_{Dat}$(Situation)	0.03
	nir	0.01	nd	0.04	n(Zeit)	0.03
	ns-2	0.01	nad	0.03	n(Sache)	0.02
	xp	0.01	np:nach$_{Dat}$	0.01	na(Situation)	0.02
essen	na	0.42	na	0.42	na(Lebewesen)	0.33
'to eat'	n	0.26	n	0.26	na(Nahrung)	0.17
	nad	0.10	nad	0.10	na(Sache)	0.09
	np	0.06	nd	0.05	n(Lebewesen)	0.08
	nd	0.05	ns-2	0.02	na(Lebewesen)	0.07
	nap	0.04	np:auf$_{Dat}$	0.02	n(Nahrung)	0.06
	ns-2	0.02	ns-w	0.01	n(Sache)	0.04
	ns-w	0.01	ni	0.01	nd(Lebewesen)	0.04
	ni	0.01	np:mit$_{Dat}$	0.01	nd(Nahrung)	0.02
	nas-2	0.01	np:in$_{Dat}$	0.01	na(Attribut)	0.02
fahren	n	0.34	n	0.34	n(Sache)	0.12
'to drive'	np	0.29	na	0.19	n(Lebewesen)	0.10
	na	0.19	np:in$_{Akk}$	0.05	na(Lebewesen)	0.08
	nap	0.06	nad	0.04	na(Sache)	0.06
	nad	0.04	np:zu$_{Dat}$	0.04	n(Ort)	0.06
	nd	0.04	nd	0.04	na(Sache)	0.05
	ni	0.01	np:nach$_{Dat}$	0.04	np:in$_{Akk}$(Sache)	0.02
	ns-2	0.01	np:mit$_{Dat}$	0.03	np:zu$_{Dat}$(Sache)	0.02
	ndp	0.01	np:in$_{Dat}$	0.03	np:in$_{Akk}$(Lebewesen)	0.02
	ns-w	0.01	np:auf$_{Dat}$	0.02	np:nach$_{Dat}$(Sache)	0.02

Table 1: Examples of most probable frame types

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. $D3$ illustrates that typical selectional preferences for beginner roles are *Situation*, *Zustand*, *Zeit*, *Sache*. $D3$ has the potential to indicate verb alternation behaviour, e.g. 'na(Situation)' refers to the same role for the direct object in a transitive frame as '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 'n' and 'na' and *Nahrung* for 'na'. *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 'n' and 'na', *Sache* in the inchoative/causative case for 'n' and 'na'.

Based on the syntactico-semantic descriptions of the German verbs as empirical verb properties, the clustering of the German verbs is performed by the k-Means algorithm, a standard unsupervised clustering technique as proposed by Forgy (1965). 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).

4 Clustering Examples

This section presents representative parts of a cluster analysis based on the verb description on $D3$. I compare the respective clusters with their pendants under $D1$ and $D2$. For each cluster, the verbs which belong to the same gold standard class are presented in one line, accompanied by the class label.

(a) nieseln regnen schneien – *Weather*

(b) dämmern – *Weather*

(c) beginnen enden – *Aspect*
bestehen$_2$ existieren – *Existence*
liegen sitzen stehen – *Position*
laufen – *Manner of Motion: Locomotion*

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

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

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

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

(h) geben – *Transfer of Possession (Giving): Gift*

The weather verbs in cluster (a) strongly agree in their syntactic expression on $D1$ and do not need $D2$ or $D3$ 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 $D1$ frames already, so *dämmern* is never clustered together with the other weather verbs on $D1$-$D3$. *Manner of Motion, Existence, Position* and *Aspect* verbs are similar in their syntactic frame usage and therefore merged together on $D1$, 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 $D3$: *Manner of Motion* verbs (from the subclasses *Locomotion, Rotation, Rush, Vehicle, Flotation*) are captured well by clusters (d) and (e), since they inhibit strong common alternations, but cluster (c) merges the *Existence, Position* and *Aspect* verbs, since verb-idiosyncratic demands on selectional roles destroy the $D2$ class demarcation. Admittedly, the verbs in cluster (c) 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 $D1$ and $D2$), 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 $D1$-$D3$, 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. *geben* in cluster (h) represents an own cluster. Syntactically, this is caused by being the only verb with a strong preference for `xa`. From the meaning point of view, this specific frame represents an idiomatic expression, only possible with *geben*. The respective frame usage overlaps the *Giving* sense of the verb.

The fact that there are verbs which are clustered semantically on basis of their corpus-based and knowledge-based empirical properties, indicates (i) a relationship between the meaning components of the verbs and their behaviour, and (ii) that the clustering algorithm is able to benefit from the linguistic descriptions and to abstract from the noise in the distributions. Low frequent verbs have been determined as problem in the clustering experiments. Their distributions are noisier than those for more frequent verbs, so they typically constitute noisy clusters. The ambiguity of verbs cannot be modelled by the hard clustering algorithm k-Means. Ambiguous verbs were typically assigned either (i) to one of the correct clusters, or (ii) to a cluster whose verbs have distributions which are similar to the ambiguous distribution, or (iii) to a singleton cluster. The interpretation of the clusterings unexpectedly points to meaning components of verbs which have not been discovered by the manual classification before. An example verb is *laufen* expressing not only a *Manner of Motion* but also a kind of existence when used in the sense of operation. The discovering effect should be larger with an increasing number of verbs, since the manual judgement is more difficult, and also with a soft clustering technique, where multiple cluster assignment is enabled. In a similar way, the clustering interpretation exhibits semantically related verb classes: verb classes which are separated in the manual classification, but semantically merged in a common cluster. For example, *Perception* and *Observation* verbs are related in that all the verbs express an observation, with the *Perception* verbs additionally referring to a physical ability, such as hearing. Related to the preceding issue, the manual verb classes as defined are demonstrated as detailed and subtle. Compared to a more general classification which would appropriately merge several classes, the clustering confirms that I have defined a difficult task with subtle classes. I was aware of this fact but preferred a fine classification, since it allows insight into more verb and class properties. But in this way, verbs which are similar in meaning are often clustered wrongly with respect to the gold standard.

What exactly is the nature of the meaning-behaviour relationship? (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. 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 roles, but they do not belong to the same fine-grained semantic class, e.g. *töten* and *unterrichten*. Why do we encounter an unpredictability concerning the encoding and effect of verb features, especially with respect to selectional preferences? The experiments presented evidence for a linguistically defined limit on the usefulness of the verb features, which is driven by the dividing line between the common and idiosyncratic features of the verbs in a verb class. Recall the underlying idea of verb classes, that the meaning components of verbs to a certain extent determine their behaviour. 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 (a) properties which are general for the respective verb classes, and (b) 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 by step-wise refining the verb description and including lexical idiosyncrasy, the emphasis of the common properties vanishes. From the theoretical point of view, the distinction between common and idiosyncratic features is obvious, but from the practical point

of view there is no unique perfect choice and encoding of the verb features. The feature choice 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.

The goal of my work is to develop a clustering methodology with respect to an automatic acquisition of a high-quality and large-scale German verb classification. I therefore applied the insights on the clustering methodology to a considerably larger amount of verb data. I extracted all German verbs from the statistical grammar model with an empirical frequency between 500 and 10,000 in a newspaper corpus of 35 million words. This selection results in a total of 809 verbs, including 94 verbs from the preliminary set of 168 verbs. I added the remaining verbs of the preliminary set, resulting in a total selection of 883 German verbs. The feature description of the German verbs refers to $D3$, and the number of clusters was set to 100, which corresponds to an average of 8.83 verbs per cluster. As a general characterisation of the cluster analysis, some clusters are extremely good with respect to the semantic overlap of the verbs, some clusters contain a number of similar verbs mixed with semantically different verbs, and for some clusters it is difficult to recognise a common semantic aspect of the verbs. For each kind of result I will present examples. The verbs which I think semantically similar are marked in bold font.

(1) *abschneiden* 'to cut off', *anziehen* 'to dress', *binden* 'to bind', *entfernen* 'to remove', *tunen* 'to tune', *wiegen* 'to weigh'

(2) *aufhalten* 'to detain', *aussprechen* 'to pronounce', *auszahlen* 'to pay off', *durchsetzen* 'to achieve', *entwickeln* 'to develop', *verantworten* 'to be responsible', *verdoppeln* 'to double', *zurückhalten* 'to keep away', *zurückziehen* 'to draw back', *ändern* 'to change'

(3) *anhören* 'to listen', *auswirken* 'to affect', *einigen* 'to agree', *lohnen* 'to be worth', *verhalten* 'to behave', *wandeln* 'to promenade'

(4) ***abholen*** 'to pick up', *ansehen* 'to watch', ***bestellen*** 'to order', ***erwerben*** 'to purchase', ***holen*** 'to fetch', ***kaufen*** 'to buy', ***konsumieren*** 'to consume', *verbrennen* 'to burn', ***verkaufen*** 'to sell'

(5) *anschauen* 'to watch', ***erhoffen*** 'to wish', ***vorstellen*** 'to imagine', ***wünschen*** 'to wish', *überlegen* 'to think about'

(6) ***danken*** 'to thank', *entkommen* 'to escape', ***gratulieren*** 'to congratulate'

(7) *beschleunigen* 'to speed up', ***bilden*** 'to constitute', *darstellen* 'to illustrate', *decken* 'to cover', *erfüllen* 'to fulfil', ***erhöhen*** 'to raise', *erledigen* 'to fulfil', *finanzieren* 'to finance', *füllen* 'to fill', *lösen* 'to solve', *rechtfertigen* 'to justify', ***reduzieren*** 'to reduce', ***senken*** 'to lower', ***steigern*** 'to increase', ***verbessern*** 'to improve', ***vergrößern*** 'to enlarge', ***verkleinern*** 'to make smaller', ***verringern*** 'to decrease', ***verschieben*** 'to shift', ***verschärfen*** 'to intensify', ***verstärken*** 'to intensify', ***verändern*** 'to change'

(8) ***ahnen*** 'to guess', ***bedauern*** 'to regret', ***befürchten*** 'to fear', ***bezweifeln*** 'to doubt', ***merken*** 'to notice', ***vermuten*** 'to assume', *weißen* 'to whiten', ***wissen*** 'to know'

(9) ***anbieten*** 'to offer', *angebieten* is not an infinitive, but a morphologically mistaken perfect participle of 'to offer', ***bieten*** 'to offer', ***erlauben*** 'to allow', ***erleichtern*** 'to facilitate', ***ermöglichen*** 'to make possible', ***eröffnen*** 'to open', ***untersagen*** 'to forbid', *veranstalten* 'to arrange', ***verbieten*** 'to forbid'

(10) ***argumentieren*** 'to argue', ***berichten*** 'to report', ***folgern*** 'to conclude', ***hinzufügen*** 'to add', ***jammern*** 'to moan', ***klagen*** 'to complain', ***schimpfen*** 'to rail', ***urteilen*** 'to judge'

(11) ***basieren*** 'to be based on', ***beruhen*** 'to be based on', ***resultieren*** 'to result from', ***stammen*** 'to stem from'

(12) ***befragen*** 'to interrogate', ***entlassen*** 'to release', ***ermorden*** 'to assassinate', ***erschießen*** 'to shoot', ***festnehmen*** 'to arrest', ***töten*** 'to kill', ***verhaften*** 'to arrest'

(13) ***beziffern*** 'to amount to', ***schätzen*** 'to estimate', ***veranschlagen*** 'to estimate'

(14) *entschuldigen* 'to apologise', ***freuen*** 'to be glad', ***wundern*** 'to be surprised', ***ärgern*** 'to be annoyed'

(15) ***nachdenken*** 'to think about', *profitieren* 'to profit', ***reden*** 'to talk', ***spekulieren*** 'to speculate', ***sprechen*** 'to talk', ***träumen*** 'to dream', ***verfügen*** 'to decree', ***verhandeln*** 'to negotiate'

(16) *mangeln* 'to lack', ***nieseln*** 'to drizzle', ***regnen*** 'to rain', ***schneien*** 'to snow'

Clusters (1) to (3) are example clusters where the verbs do not share meaning aspects. In the overall cluster analysis, the semantically incoherent clusters tend to be rather large, i.e. with more than 15-20 verb members. Clusters (4) to (7) are example clusters where a part of the verbs show overlap in their meaning aspects, but the clusters also contain considerable noise. Cluster (4) mainly contains verbs of buying and selling, cluster (5) contains verbs of wishing, cluster (6) contains verbs of expressing a speech act concerning a specific event, and cluster (7) contains verbs of quantum change. Clusters (8) to (16) are example clusters where most or all verbs show a strong similarity in their conceptual structures. Cluster (8) contains verbs expressing a propositional attitude; the underlined verbs in addition indicate an emotion. The only unmarked verb *weißen* also fits into the cluster, since it is a morphological lemma mistake changed with *wissen* which belongs to the verb class. The verbs in cluster (9) describe a scene where somebody or some situation makes something possible (in the positive or negative sense). Next to a lemmatising mistake (*angebieten* is not an infinitive, but a morphologically mistaken perfect participle of *anbieten*), the only exception verb is *veranstalten*. The verbs in cluster (10) are connected more loosely, all referring to a verbal discussion, with the underlined verbs in addition denoting a negative, complaining way of utterance. In cluster (11) all verbs refer to a basis, in cluster (12) the verbs describe the process from arresting to treating a suspect, and cluster (13) contains verbs of estimating an amount of money. In cluster (14), all verbs except for *entschuldigen* refer to an emotional state (with some origin for the emotion). The verbs in cluster (15) except for *profitieren* all indicate a thinking (with or without talking) about a certain matter. Finally in cluster (16), we can recognise weather verbs.

5 Application of Semantic Verb Classes

The previous section has presented clustering examples which agree with the manual classification in many respects. Without any doubt the cluster analysis needs manual correction and completion, but represents a plausible basis for a semantic lexicon resource. Following, I discuss possible NLP applications for a verb classification. (a) Providing lexical information about verbs by verb class labels serves two purposes in *parsing*: (i) On the one hand the class information restricts possible parses and decreases parse ambiguities, since the class labels implicitly define the range of possible syntactic and semantic verb environments. (ii) On the other hand the class information supplies additional information on the syntax-semantic embedding for verbs which are defined vaguely. Combining both uses, the parsing quality might be improved. (b) Replacing verbs in a *language model* by the respective verb classes might improve a language model's robustness and accuracy, since the class information provides more stable syntactic and semantic information than the individual verbs. For example, the probability of the preposition *nach* following any manner of motion verb is comparably high, since (among other senses) it indicates a path. Nevertheless, the model might provide less reliable information on the individual manner of motion verbs, especially in low frequent cases such as *rasen* 'to speed'. The verb class information contributes this missing information by generalising over the verbs within one class, and is therefore able to predict a *nach*-PP for *rasen*. (c) A user query which requires *information extraction* on documents can be extended with respect to its syntactic and especially semantic information (e.g. complement realisation) by adding the existing class information of the query predicate to the query description, in addition to the individual verb information. (d) Assuming that a similar system of verb classes exists in various languages, the problem in *machine translation* that the translation of a verb from one language into another activates several verbs in the target language can be solved by filtering the correct translation with respect to the source verb class. For example, the verb *bestehen* has at least four different senses, each coupled with a preferred subcategorisation behaviour: (i) *bestehen* meaning 'to insist' subcategorises np with auf_{Dat}, (ii) *bestehen* meaning 'to consist' subcategorises np with aus_{Akk}, (iii) *bestehen* meaning 'to exist, to survive' subcategorises n or np with in_{Akk},

and (iv) *bestehen* meaning 'to pass' (e.g. of an exam) subcategorises `na`. With respect to the source context (the syntactico-semantic embedding), the verb class of the source verb is determined, and based on the source class the target verb is filtered. In addition, missing information concerning the source or target verb with respect to its syntactic and semantic embedding might be added by the respective class and refine the translation.

Because these ideas might seem speculative, the following sections provide examples of verb class usage which have already been performed. Most of them are based on the Levin classes, some on German soft-clustering approaches. I should add that there are multiple uses of the WordNet classes, but I do not provide a picture of them within the scope of this thesis. The reader is referred to the WordNet bibliography at `http://engr.smu.edu/~rada/wnb/`.

Parsing-Based Word Sense Disambiguation Dorr and Jones (1996) show that the Levin classes can be used for word sense disambiguation. They describe the English verbs in the Levin classes by their syntactic descriptions, based on parsing patterns on the example sentences for the verb classes. The approach distinguishes positive and negative examples by 1 and 0, respectively. For example, the parsing pattern for the sentence *Tony broke the vase to pieces* would be `1-[np,v,np,pp(to)]`. The syntactic description of a verb consists of the set of parsing patterns which are assigned to the verb according to its class affiliations.

Dorr and Jones determine the overlap on the sets of verbs (a) in the semantic Levin classes, and (b) as based on the agreement on syntactic descriptions. The comparison is performed within two experiments: (i) The syntactic patterns of the example sentences within a Levin class are assigned to all verbs within the class, disregarding the different verb senses the verbs might have. The syntactic description of a verb might therefore contain syntactic patterns of several verb classes, according to its number of class affiliations. (ii) The syntactic patterns of class examples are only assigned to the verb senses activated by the specific class. The overlap of (a) the 'semantic' and (b) the 'syntactic' sets of verbs are (i) 6.3% accuracy, because there are far more syntactic descriptions than semantic classes, vs. (ii) 97.9% accuracy, because the semantic classes agree with the disambiguated syntactic descriptions. The experiments validate the strong relation between the syntactic and the semantic information in the verb classes, and show that this relation can be utilised for word sense disambiguation, because the classification can disambiguate verb senses according to syntactic descriptions.

Machine Translation Dorr (1997) uses Levin's verb class approach to construct a large-scale dictionary for machine translation. Dorr defines Lexical Conceptual Structures (LCSs) (Jackendoff, 1983, 1990) as a means for the language-independent lexicon representation of verb meaning components. She presents possibilities of how to obtain the LCS representations, ranging from manual to fully-automatic approaches. The following automatic approach is based on the Levin classes.

Assuming as in (Dorr and Jones, 1996) that basic verb meaning components can be systematically derived from information about the syntactic realisation, Dorr utilises and extends Levin's classes for the lexicon construction. The syntax and semantics of the verb classes are captured by a matrix relating the existence of alternations with the definition of the semantic classes. Verbs are then assigned to a semantic class according to which alternations they undergo within a large corpus. The classes are decomposed into primitive units of meaning which are captured in an LCS representation. Even though neither the syntactic constructions nor the class system is expected to hold cross-linguistically, the meaning components underlying two translationally related verbs are expected to overlap. The language-independent LCS lexicon entries for machine translation are therefore constructed via the syntactic and semantic definitions in Levin's classification.

Document Classification Klavans and Kan (1998) use Levin's verb classes to discriminate article types within the news domain of the *Wall Street Journal (WSJ)* corpus. They consider the nouns in a document

as the conceptual entities, and the verbs as the conceptual events and actions within the documents. The paper focuses on the role of verbs in document analysis.

Klavans and Kan place their investigation on the 100 most frequent and 50 additional verbs in the WSJ, covering a total of 56% of the verb tokens in the corpus. They select 50 out of 1,236 articles, with each article containing the highest percentage of a particular verb class. The investigation reveals that each verb class distinguishes between different article types, e.g. manner of motion verbs are typically found in posted earnings and announcements, communication verbs in issues, reports, opinions, and editorials. The work shows that the verb classes can be used as type labels in information retrieval.

Word Sense Disambiguation in Target Word Selection Prescher, Riezler, and Rooth (2000) present an approach for disambiguation in target word selection. Given a translation produces multiple equivalences of a source word, a disambiguation model selects the target word. The core part of the disambiguation system is represented by a probabilistic class-based lexicon, which is induced in an unsupervised manner (via the EM algorithm) from unannotated newspaper corpus data. The lexicon provides estimated frequencies for English verb-noun pairs with respect to a grammatical relationship. For example, Table 2 presents the 10 most frequent nouns which are learned as direct objects of the verb *to cross*. Given that in a translation process a decision has to be made concerning which of a set of alternative target nouns is the most appropriate translation of an ambiguous source noun, the target nouns are looked up in the probabilistic lexicon with respect to the grammatical relationship to the (already translated) target verb. For example, in *eine Grenze überschreiten* possible English target nouns for the German source noun *Grenze* are *border, frontier, boundary, limit, periphery, edge*. But with respect to the direct object relationship to the verb *to cross* which is the translation of *überschreiten*, the lexicon determines *border* as the most probable translation, cf. Table 2.

cross ¡subj,***obj***¿	Freq
mind	74.2
road	30.3
line	28.1
bridge	27.5
room	20.5
border	17.8
boundary	16.2
river	14.6
street	11.5
atlantic	9.9

Table 2: Class-based estimated frequencies of direct object nouns

Subcategorisation Acquisition Korhonen (2002b) uses Levin's verb classes for the hypothesis filtering in an automatic acquisition of subcategorisation frames for English verbs. Her work is based on the framework of (Briscoe and Carroll, 1997) who automatically induce a subcategorisation lexicon for English verbs. Since automatic subcategorisation lexica in general show a lack in accuracy, the lexical acquisition is typically followed by a filtering on the frame definitions.

Korhonen suggests a filter that smoothes the statistical subcategorisation frame information with back-off estimates on the verbs' semantic Levin classes: Provided with a probabilistic distribution of the verbs over subcategorisation frame types as obtained from (Briscoe and Carroll, 1997), each verb is assigned via WordNet classes to the Levin class representing its dominant sense (Korhonen, 2002a). From each Levin class, 4-5 verbs are manually chosen to represent the semantic class. The verbs' distributions are merged to obtain back-off estimates with respect to the class, and the back-off estimates are then used to smooth

the subcategorisation distributions of the verbs within that class. Setting an empirically defined threshold on the smoothed distributions filters out the unreliable hypotheses.

6 Conclusions and Outlook

This article has presented an automatic induction of German semantic verb classes. A statistical grammar model has provided verb descriptions at three levels at the syntax-semantic interface, with the most elaborated description being close to a definition of the verb alternation behaviour. Based on the syntactico-semantic descriptions, the standard clustering algorithm k-Means (Forgy, 1965) was applied to induce a semantic classification for the verbs. I presented clustering examples which agree with a manual classification in many respects. Without any doubt the cluster analysis needs manual correction and completion, but represents a plausible basis for a semantic lexicon resource. I closed the article with examples for using the learned verb classes in NLP applications.

The strategy of utilising subcategorisation frames, prepositional information and selectional preferences to define the verb features has proven successful, since the experiments illustrated a tight connection between the induced verb behaviour and the constitution of the semantic verb classes. In addition, each level of representation has generated a positive effect on the clustering and improved the less informative level. The experiments present evidence for a linguistically defined limit on the usefulness of the verb features, which is driven by the dividing line between the common and idiosyncratic features of verbs in a verb class. The feature choice therefore depends on the specific properties of the desired verb classes.

There are various directions for future research. (i) The manual definition of the German semantic verb classes might be extended in order to include a larger number and a larger variety of verb classes. An extended classification would be useful as gold standard for further clustering experiments, and more general as manual resource in NLP applications. (ii) Possible features to describe German verbs might include any kind of information which helps classify the verbs in a semantically appropriate way. Within this article, I have concentrated on defining the verb features with respect to the alternation behaviour. Other features which are relevant to describe the behaviour of verbs are e.g. their auxiliary selection and adverbial combinations. (iii) Variations of the existing feature description are especially relevant for the choice of selectional preferences. The experiment results demonstrated that the 15 conceptual GermaNet top levels are not sufficient for all verbs. (iv) As an extension of the existing clustering, I might apply a soft clustering algorithm to the German verbs. The soft clustering enables us to assign verbs to multiple clusters and therefore address the phenomenon of verb ambiguity.

A Subcategorisation Frame Types

The syntactic aspect of the German verb behaviour is captured by 38 subcategorisation frame types in the context-free German grammar, according to standard German grammar definitions such as Helbig and Buscha (1998). The subcategorisation frame types 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). The resulting frame types are listed in Table 3, accompanied by annotated verb second example clauses.

Frame Type	**Example**
n	*Natalie$_n$ schwimmt.*
na	*Hans$_n$ sieht seine Freundin$_a$.*
nd	*Er$_n$ glaubt den Leuten$_d$ nicht.*
np	*Die Autofahrer$_n$ achten besonders auf Kinder$_p$.*
nad	*Anna$_n$ verspricht ihrem Vater$_d$ ein tolles Geschenk$_a$.*
nap	*Die kleine Verkäuferin$_n$ hindert den Dieb$_a$ am Stehlen$_p$.*
ndp	*Der Moderator$_n$ dankt dem Publikum$_d$ für sein Verständnis$_p$.*
ni	*Mein Freund$_n$ versucht immer wieder, pünktlich zu kommen$_i$.*
nai	*Er$_n$ hört seine Mutter$_a$ ein Lied singen$_i$.*
ndi	*Helene$_n$ verspricht ihrem Großvater$_d$ ihn bald zu besuchen$_i$.*
nr	*Die kleinen Kinder$_n$ fürchten sich$_r$.*
nar	*Der Unternehmer$_n$ erhofft sich$_r$ baldigen Aufwind$_a$.*
ndr	*Sie$_n$ schließt sich$_r$ nach 10 Jahren wieder der Kirche$_d$ an.*
npr	*Der Pastor$_n$ hat sich$_r$ als der Kirche würdig$_p$ erwiesen.*
nir	*Die alte Frau$_n$ stellt sich$_r$ vor, den Jackpot zu gewinnen$_i$.*
x	*Es$_x$ blitzt.*
xa	*Es$_x$ gibt viele Bücher$_a$.*
xd	*Es$_x$ graut mir$_d$.*
xp	*Es$_x$ geht um ein tolles Angebot für einen super Computer$_p$.*
xr	*Es$_x$ rechnet sich$_r$.*
xs-dass	*Es$_x$ heißt, dass Thomas sehr klug ist$_{s-dass}$.*
ns-2	*Der Abteilungsleiter$_n$ hat gesagt, er halte bald einen Vortrag$_{s-2}$.*
nas-2	*Der Chef$_n$ schnauzt ihn$_a$ an, er sei ein Idiot$_{s-2}$.*
nds-2	*Er$_n$ sagt seiner Freundin$_d$, sie sei zu krank zum Arbeiten$_{s-2}$.*
nrs-2	*Der traurige Vogel$_n$ wünscht sich$_r$, sie bliebe bei ihm$_{s-2}$.*
ns-dass	*Der Winter$_n$ hat schon angekündigt, dass er bald kommt$_{s-dass}$.*
nas-dass	*Der Vater$_n$ fordert seine Tochter$_a$ auf, dass sie verreist$_{s-dass}$.*
nds-dass	*Er$_n$ sagt seiner Geliebten$_d$, dass er verheiratet ist$_{s-dass}$.*
nrs-dass	*Der Junge$_n$ wünscht sich$_r$, dass seine Mutter bleibt$_{s-dass}$.*
ns-ob	*Der Chef$_n$ hat gefragt, ob die neue Angestellte den Vortrag hält$_{s-ob}$.*
nas-ob	*Anton$_n$ fragt seine Frau$_a$, ob sie ihn liebt$_{s-ob}$.*
nds-ob	*Der Nachbar$_n$ ruft der Frau$_d$ zu, ob sie verreist$_{s-ob}$.*
nrs-ob	*Der Alte$_n$ wird sich$_r$ erinnern, ob das Mädchen dort war$_{s-ob}$.*
ns-w	*Der kleine Junge$_n$ hat gefragt, wann die Tante endlich ankommt$_{s-w}$.*
nas-w	*Der Mann$_n$ fragt seine Freundin$_a$, warum sie ihn liebt$_{s-w}$.*
nds-w	*Der Vater$_n$ verrät seiner Tochter$_d$ nicht, wer zu Besuch kommt$_{s-w}$.*
nrs-w	*Das Mädchen$_n$ erinnert sich$_r$, wer zu Besuch kommt$_{s-w}$.*
k	*Der neue Nachbar$_k$ ist ein ziemlicher Idiot.*

Table 3: Subcategorisation frame types

References

Steven Abney and Marc Light. Hiding a Semantic Class Hierarchy in a Markow Model. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 1–8, College Park, MD, 1999.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Canada, 1998.

Ted Briscoe and John Carroll. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC, 1997.

Stephen Clark and David Weir. Class-Based Probability Estimation using a Semantic Hierarchy. *Computational Linguistics*, 28(2):187–206, 2002.

Bonnie J. Dorr. Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation*, 12(4):271–322, 1997.

Bonnie J. Dorr and Doug Jones. Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 322–327, Copenhagen, Denmark, 1996.

Katrin Erk, Andrea Kowalski, and Manfred Pinkal. A Corpus Resource for Lexical Semantics. In *Proceedings of the 5th International Workshop on Computational Semantics*, Tilburg, The Netherlands, 2003.

Christiane Fellbaum, editor. *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, 1998.

Charles J. Fillmore. Scenes-and-Frames Semantics. In Antonio Zampolli, editor, *Linguistic Structures Processing*, volume 59 of *Fundamental Studies in Computer Science*. North Holland Publishing, Amsterdam, 1977.

Charles J. Fillmore. Frame Semantics. *Linguistics in the Morning Calm*, pages 111–137, 1982.

Edward W. Forgy. Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications. *Biometrics*, 21:768–780, 1965.

Birgit Hamp and Helmut Feldweg. GermaNet – a Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP Applications*, Madrid, Spain, 1997.

Gerhard Helbig and Joachim Buscha. *Deutsche Grammatik*. Langenscheidt – Verlag Enzyklopädie, 18th edition, 1998.

Ray Jackendoff. *Semantics and Cognition*. MIT Press, Cambridge, MA, 1983.

Ray Jackendoff. *Semantic Structures*. MIT Press, Cambridge, MA, 1990.

Christopher R. Johnson, Charles J. Fillmore, Miriam R. L. Petruck, Collin F. Baker, Michael Ellsworth, Josef Ruppenhofer, and Esther J. Wood. *FrameNet: Theory and Practice*. ICSI Berkeley, 2002. URL `http://www.icsi.berkeley.edu/ framenet/book/book.html`.

Judith L. Klavans and Min-Yen Kan. The Role of Verbs in Document Analysis. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 680–686, Montreal, Canada, 1998.

Anna Korhonen. Assigning Verbs to Semantic Classes via WordNet. In *Proceedings of the COLING Workshop on Building and Using Semantic Networks*, Taipei, Taiwan, 2002a.

Anna Korhonen. *Subcategorization Acquisition*. PhD thesis, University of Cambridge, Computer Laboratory, 2002b. Technical Report UCAM-CL-TR-530.

Claudia Kunze. Extension and Use of GermaNet, a Lexical-Semantic Database. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 999–1002, Athens, Greece, 2000.

Beth Levin. *English Verb Classes and Alternations*. The University of Chicago Press, 1993.

Hang Li and Naoki Abe. Generalizing Case Frames Using a Thesaurus and the MDL Principle. *Computational Linguistics*, 24(2):217–244, 1998.

Diana McCarthy. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex, 2001.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to Wordnet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244, 1990.

Steven Pinker. *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, MA, 1989.

Detlef Prescher, Stefan Riezler, and Mats Rooth. Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 649–655, Saarbrücken, Germany, 2000.

Philip Resnik. Selectional Preference and Sense Disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, 1997.

Francesc Ribas. On Learning More Appropriate Selectional Restrictions. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, Ireland, 1995.

Patrick Saint-Dizier. Alternations and Verb Semantic Classes for French: Analysis and Class Formation. In Patrick Saint-Dizier, editor, *Predicative Forms in Natural Language and in Lexical Knowledge Bases*. Kluwer Academic Publishers, Dordrecht, 1998.

Sabine Schulte im Walde. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas de Gran Canaria, Spain, 2002.

Sabine Schulte im Walde. A Collocation Database for German Nouns and Verbs. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research*, Budapest, Hungary, 2003a.

Sabine Schulte im Walde. *Experiments on the Automatic Induction of German Semantic Verb Classes*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2003b. Published as AIMS Report 9(2).

Gloria Vázquez, Ana Fernández, Irene Castellón, and María Antonia Martí. *Clasificación Verbal: Alternancias de Diátesis*. Number 3 in Quaderns de Sintagma. Universitat de Lleida, 2000.

Piek Vossen. EuroWordNet General Document. Technical Report LE2-4003, LE4-8328, University of Amsterdam, 1999.

Andreas Wagner. Enriching a Lexical Semantic Net with Selectional Preferences by Means of Statistical Corpus Analysis. In *Proceedings of the ECAI Workshop on Ontology Learning*, pages 37–42, Berlin, Germany, 2000.