Experiments on the Automatic Induction of German Semantic Verb Classes

Sabine Schulte im Walde*
Universität des Saarlandes

This article presents clustering experiments on German verbs: A statistical grammar model for German serves as the source for a distributional verb description at the lexical syntax–semantics interface, and the unsupervised clustering algorithm k-means uses the empirical verb properties to perform an automatic induction of verb classes. Various evaluation measures are applied to compare the clustering results to gold standard German semantic verb classes under different criteria. The primary goals of the experiments are (1) to empirically utilize and investigate the well-established relationship between verb meaning and verb behavior within a cluster analysis and (2) to investigate the required technical parameters of a cluster analysis with respect to this specific linguistic task. The clustering methodology is developed on a small-scale verb set and then applied to a larger-scale verb set including 883 German verbs.

1. Motivation

Semantic verb classes generalize over verbs according to their semantic properties, that is, they capture large amounts of verb meaning without defining the idiosyncratic details for each verb. The classes refer to a general semantic level, and idiosyncratic lexical semantic properties of the verbs are either added to the class description or left underspecified. Examples for semantic verb classes are *Position* verbs such as *liegen* 'lie', *sitzen* 'sit', *stehen* 'stand', and *Manner of Motion with a Vehicle* verbs such as *fahren* 'drive', *fliegen* 'fly', *rudern* 'row'. Manual definitions of semantic verb classes exist for several languages, the most dominant examples concerning English (Levin 1993; Baker, Fillmore, and Lowe 1998) and Spanish (Vázquez et al. 2000). 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 criterion, 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. For example, the English verb classification by Levin (1993) has been used in NLP applications such as word sense disambiguation (Dorr and Jones 1996), machine translation (Dorr 1997), document classification (Klavans and Kan 1998), and subcategorization acquisition (Korhonen 2002). To my knowledge, no comparable German verb classification is available so far; therefore, such a classification would provide a principled basis for filling a gap in available lexical knowledge.

* Department of Computational Linguistics, Saarbrücken, Germany. E-mail: schulte@coli.uni-sb.de.

Submission received: 1 September 2003; revised submission received: 5 September 2005; accepted for publication: 10 November 2005.

How can we obtain a semantic classification of verbs while avoiding tedious manual definitions of the verbs and the classes? Few resources are semantically annotated and provide semantic information off-the-shelf such as *FrameNet* (Baker, Fillmore, and Lowe 1998; Fontenelle 2003) and *PropBank* (Palmer, Gildea, and Kingsbury 2005). Instead, the automatic construction of semantic classes typically benefits from a long-standing linguistic hypothesis that asserts a tight connection between the lexical meaning of a verb and its behavior: To a certain extent, the lexical meaning of a verb determines its behavior, particularly with respect to the choice of its arguments (Pinker 1989; Levin 1993; Dorr and Jones 1996; Siegel and McKeown 2000; Merlo and Stevenson 2001; Schulte im Walde and Brew 2002; Lapata and Brew 2004). Even though the meaning–behavior relationship is not perfect, we can make this prediction: If we induce a verb classification on the basis of verb features describing verb behavior, then the resulting behavior classification should agree with a semantic classification to a certain extent (yet to be determined). The aim of this work is to utilize this prediction for the automatic acquisition of German semantic verb classes.

The verb behavior itself is commonly captured by the diathesis alternation of verbs: alternative constructions at the syntax–semantics interface that express the same or a similar conceptual idea of a verb (Lapata 1999; Schulte im Walde 2000; McCarthy 2001; Merlo and Stevenson 2001; Joanis 2002). Consider example (1), where the most common alternations of the *Manner of Motion with a Vehicle* verb *fahren* 'drive' are illustrated. The conceptual participants are a vehicle, a driver, a passenger, and a direction. In (a), the vehicle is expressed as the subject in a transitive verb construction, with a prepositional phrase indicating the direction. In (b), the driver is expressed as the subject in a transitive verb construction, with a prepositional phrase indicating the direction. In (c), the driver is expressed as the subject in a transitive verb construction, with an accusative noun phrase indicating the vehicle. In (d), the driver is expressed as the subject in a ditransitive verb construction, with an accusative noun phrase indicating the passenger, and a prepositional phrase indicating the direction. Even if a certain participant is not realized within an alternation, its contribution might be implicitly defined by the verb. For example, in the German sentence 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. Verbs in the same semantic class are expected to overlap in their alternation behavior to a certain extent. For example, the *Manner of Motion with a Vehicle* verb *fliegen* 'fly' alternates between (a) such as in *Der Airbus A380 fliegt nach New York* 'The Airbus A380 flies to New York', (b) in marked cases as in *Der ältere Pilot fliegt nach London* 'The older pilot flies to London', (c) as in *Pilot Schulze fliegt eine Boing 747* 'Pilot Schulze flies a Boing 747', and (d) as in *Der Pilot fliegt seine Passagiere nach Thailand* 'The pilot flies his passengers to Thailand'; the *Manner of Motion with a Vehicle* verb *rudern* 'row' alternates between (b) such as in *Anna rudert über den See* 'Anna rows over the lake', (c) such as in *Anna rudert das blaue Boot* 'Anna rows the blue boat', and (d) such as in *Anna rudert ihren kleinen Bruder über den See* 'Anna rows her little brother over the lake'.

Example 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.'

We decided to use diathesis alternations as an approach to characterizing verb behavior, and to use the following verb features to stepwise describe diathesis alternations: (1) syntactic structures, which are relevant for capturing argument functions; (2) prepositions, which are relevant to distinguish, for example, directions from locations; and (3) selectional preferences, which concern participant roles. A statistical grammar model serves as the source for an empirical verb description for the three levels at the syntax–semantics interface. Based on the empirical feature description, we then perform a cluster analysis of the German verbs using *k*-means, a standard unsupervised hard clustering technique as proposed by Forgy (1965). The clustering outcome cannot be a perfect semantic verb classification, since the meaning–behavior relationship on which the clustering relies is not perfect, and the clustering method is not perfect for the ambiguous verb data. However, our primary goal is not necessarily to obtain the optimal clustering result, but rather to assess the linguistic and technical conditions that are crucial for a semantic cluster analysis. More specifically, (1) we perform an empirical investigation of the relationship between verb meaning and verb behavior (that is, Can we use the meaning–behavior relationship of verbs to induce verb classes, and to what extent does the meaning–behavior relationship hold in the experiments?), and (2) we investigate which technical parameters are suitable for the natural language task. The resulting clustering methodology can then be applied to a larger-scale verb set.

The plan of the article is as follows. Section 2 describes the experimental setup with respect to (1) gold standard verb classes for 168 German verbs, (2) the statistical grammar model that provides empirical lexical information for German verbs at the syntax–semantics interface, and (3) the clustering algorithm and evaluation methods. Section 3 performs preliminary clustering experiments on the German gold standard verbs, and Section 4 presents an application of the clustering technique in a large-scale experiment. Section 5 discusses related work, and Section 6 presents the conclusions and outlook for further work.

2. Experimental Setup

2.1 German Semantic Verb Classes

A set of 168 German verbs was manually classified into 43 concise semantic verb classes. The verb 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 left underspecified. The German verbs are provided with a coarse translation into English, given here in brackets; we do not attempt to define subtle differences in meaning or usage. The translated verb senses only refer to the respective semantic class; if the verb translations in one class are too similar to distinguish among them, a common translation is given. Even though the classification is primarily based on semantic intuition and not on facts about syntactic behavior, the verbs grouped in one class share certain aspects of their behavior. (Please note that this overlap does not necessarily transfer to the English translations.) This agreement corresponds to the

long-standing linguistic hypothesis that asserts a tight connection between the meaning components of a verb and its behavior (Pinker 1989; Levin 1993).

The purpose of the manual classification is to evaluate the reliability and performance of the clustering experiments. The following facts refer to empirically relevant properties of the classification: 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 and marked by subscripts. The classes include both high- and low-frequency verbs in order to exercise the clustering technology in both data-rich and data-poor situations: The corpus frequencies of the verbs range from 8 to 71,604 (within 35 million words of a German newspaper corpus, cf. Section 2.2). The class labels are given on two semantic levels: coarse labels such as *Manner of Motion* are subdivided into finer labels, such as *Locomotion, Rotation, Rush, Vehicle, Flotation*. The fine-grained labels are relevant for the clustering experiments, as the numbering indicates. As mentioned before, the classification is primarily based on semantic intuition, not on facts about syntactic behavior. As an extreme example, the *Support* class (23) contains the verb *unterstützen*, which syntactically requires a direct object, together with the verbs *dienen, folgen, and helfen,* which dominantly subcategorize for an indirect object. The classification was checked to ensure lack of bias, so class membership is not disproportionately made up of high-frequency verbs, low-frequency verbs, strongly ambiguous verbs, verbs from specific semantic areas, and so forth.

The classification deliberately sets high standards for the automatic induction process: It would be easier (1) to define the verb classes on a purely syntactic basis, since syntactic properties are easier to obtain automatically than semantic features, or (2) to define larger classes of verbs, so that the distinction between the classes is not based on fine-grained verb properties, or (3) to disregard clustering complications such as verb ambiguity and low-frequency verbs. But the overall goal is not to achieve a perfect clustering on the given 168 verbs but to investigate both the potential and the limits of our clustering methodology that combines easily available data with a simple algorithm. The task cannot be solved completely, but we can investigate the bounds.

The classification is defined as follows:

1. *Aspect*: anfangen, aufhören, beenden, beginnen, enden (start, stop, finish, begin, end)
2. *Propositional Attitude*: ahnen, denken, glauben, vermuten, wissen (guess, think, believe, assume, know)
   - *Desire*
     3. *Wish*: erhoffen, wollen, wünschen (hope, want, wish)
     4. *Need*: bedürfen, benötigen, brauchen (all: need/require)
5. *Transfer of Possession (Obtaining)*: bekommen, erhalten, erlangen, kriegen (all: receive/obtain)
   - *Transfer of Possession (Giving)*
     6. *Gift*: geben, leihen, schenken, spenden, stiften, vermachen, überschreiben (give, borrow, present, donate, donate, bequeath, sign over)
     7. *Supply*: bringen, liefern, schicken, vermitteln$_1$, zustellen (bring, deliver, send, convey, deliver)

- *Manner of Motion*
  8. *Locomotion*: gehen, klettern, kriechen, laufen, rennen, schleichen, wandern (go, climb, creep, walk, run, sneak, wander)
  9. *Rotation*: drehen, rotieren (turn around, rotate)
  10. *Rush*: eilen, hasten (both: hurry)
  11. *Vehicle*: fahren, fliegen, rudern, segeln (drive, fly, row, sail)
  12. *Flotation*: fließen, gleiten, treiben (float, glide, float)

- *Emotion*
  13. *Origin*: ärgern, freuen (be annoyed, be happy)
  14. *Expression*: heulen$_1$, lachen$_1$, weinen (cry, laugh, cry)
  15. *Objection*: ängstigen, ekeln, fürchten, scheuen (frighten, disgust, fear, be afraid)

16. *Facial Expression*: gähnen, grinsen, lachen$_2$, lächeln, starren (yawn, grin, laugh, smile, stare)
17. *Perception*: empfinden, erfahren$_1$, fühlen, hören, riechen, sehen, wahrnehmen (feel, experience, feel, hear, smell, see, perceive)
18. *Manner of Articulation*: flüstern, rufen, schreien (whisper, shout, scream)
19. *Moaning*: heulen$_2$, jammern, klagen, lamentieren (all: wail/moan/complain)
20. *Communication*: kommunizieren, korrespondieren, reden, sprechen, verhandeln (communicate, correspond, talk, talk, negotiate)

- *Statement*
  21. *Announcement*: ankündigen, bekanntgeben, eröffnen, verkünden (all: announce)
  22. *Constitution*: anordnen, bestimmen, festlegen (arrange, determine, constitute)
  23. *Promise*: versichern, versprechen, zusagen (ensure, promise, promise)

24. *Observation*: bemerken, erkennen, erfahren$_2$, feststellen, realisieren, registrieren (notice, realize, get to know, observe, realize, realize)
25. *Description*: beschreiben, charakterisieren, darstellen$_1$, interpretieren (describe, characterize, describe, interpret)
26. *Presentation*: darstellen$_2$, demonstrieren, präsentieren, veranschaulichen, vorführen (present, demonstrate, present, illustrate, demonstrate)
27. *Speculation*: grübeln, nachdenken, phantasieren, spekulieren (muse, think about, fantasize, speculate)
28. *Insistence*: beharren, bestehen$_1$, insistieren, pochen (all: insist)
29. *Teaching*: beibringen, lehren, unterrichten, vermitteln$_2$ (all: teach)

- *Position*
  30. *Bring into Position*: legen, setzen, stellen (lay, set, put upright)
  31. *Be in Position*: liegen, sitzen, stehen (lie, sit, stand)

32. *Production*: bilden, erzeugen, herstellen, hervorbringen, produzieren (all: generate/produce)
33. *Renovation*: dekorieren, erneuern, renovieren, reparieren (decorate, renew, renovate, repair)
34. *Support*: dienen, folgen$_1$, helfen, unterstützen (serve, follow, help, support)
35. *Quantum Change*: erhöhen, erniedrigen, senken, steigern, vergrößern, verklenern (increase, decrease, decrease, increase, enlarge, diminish)
36. *Opening*: öffnen, schließen$_1$ (open, close)
37. *Existence*: bestehen$_2$, existieren, leben (exist, exist, live)
38. *Consumption*: essen, konsumieren, lesen, saufen, trinken (eat, consume, read, booze, drink)
39. *Elimination*: eliminieren, entfernen, exekutieren, töten, vernichten (eliminate, delete, execute, kill, destroy)
40. *Basis*: basieren, beruhen, gründen, stützen (all: be based on)
41. *Inference*: folgern, schließen$_2$ (conclude, infer)
42. *Result*: ergeben, erwachsen, folgen$_2$, resultieren (all: follow/result)
43. *Weather*: blitzen, donnern, dämmern, nieseln, regnen, schneien (lightning, thunder, dawn, drizzle, rain, snow)

The evidence used in the class creation process—including the choice of the verbs—was provided by subjective conceptual knowledge, monolingual and bilingual dictionary entries and corpus searches. Interannotator agreement has therefore not been addressed, but the classes were created in close relation to the English classification by Levin (1993) (as far as the English classes have German counterparts) and agree with the German verb classification by Schumacher (1986), as far as the relevant verbs are covered by his semantic 'fields'. To overcome the drawback of a subjective class definition, the classification was accompanied by a detailed class description. This characterization is closely related to Fillmore's scenes-and-frames semantics (Fillmore 1977, 1982), as computationally utilized in *FrameNet* (Baker, Fillmore, and Lowe 1998; Fontenelle 2003); there is no reference to the German *FrameNet* version (Erk, Kowalski, and Pinkal 2003)—as one might expect—just because the German version itself had just started to be developed. 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 the basis of a large German newspaper corpus from the 1990s (cf. Section 2.2). They capture the scene description with idiosyncratic participant names and demarcate major and minor roles. Since a scene might be activated by a number of frame embeddings, the predominant frame variants from the corpus are listed, marked with participating roles, and at least one example sentence for each verb utilizing the respective frame is given. The corpus examples are annotated and illustrate the idiosyncratic combinations of lexical verb meaning and conceptual constructions to capture variations in verb sense. Example 2 presents a verb class description for the class of *Aspect* verbs. For further class descrip-

tions, the reader is referred to Schulte im Walde (2003a, pages 27–103). Verbs allowing a frame variant are marked by "+," verbs allowing the frame variant only in company of an additional adverbial modifier are marked by "$+_{\text{adv}}$," and verbs not allowing a frame variant are marked by "¬." In the case of ambiguity, 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. 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 realized 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). Passivization of a verb–frame combination is indicated by [P]. Appendix 6 lists all possible frame variants with illustrative examples. Note that the corpus examples are given in the old German spelling version, before the spelling reform in 1998.

Semantic verb classes have been defined for several languages, for example, as the earlier mentioned lexicographic resource *FrameNet* for English (Baker, Fillmore, and Lowe 1998; Fontenelle 2003) and German (Erk, Kowalski, and Pinkal 2003); the lexical semantic ontology WordNet for English (Miller et al. 1990; Fellbaum 1998); *EuroWordNet* (Vossen 2004) for Dutch, Italian, Spanish, French, German, Czech, and Estonian, and further languages as listed in *WordNets in the World* (Global WordNet Association, www.globalwordnet.org); syntax–semantics based verb classes for English (Levin 1993), Spanish (Vázquez et al. 2000), and French (Saint-Dizier 1998).

Example 2
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 and Corpus Examples
$\mathbf{n}_E$	+ anfangen, aufhören, beginnen / $+_{adv}$ enden / ¬ beenden
	Nun aber muß [$_E$ der Dialog] **anfangen**. Now though must the dialog begin
	Erst muß [$_E$ das Morden] **aufhören**. First must the killing stop
	[$_E$ Der Gottesdienst] **beginnt**. The service begins
	[$_E$ Das Schuljahr] **beginnt** [$_T$ im Februar]. The school year begins in February
	[$_X$ Für die Flüchtlinge] **beginnt** nun [$_E$ ein Wettlauf gegen die Zeit]. For the fugitives begins now a race against time
	[$_E$ Die Ferien] **enden** [$_R$ mit einem großen Fest]. The vacations end with a big party
	[$_E$ Druckkunst] ... **endet** [$_R$ beim guten Buch]. The art of typesetting ... ends with a good book
	[$_E$ Der Informationstag] ... **endet** [$_T$ um 14 Uhr]. The information day ... finishes at 2pm

$\mathbf{n}_I$	+ anfangen, aufhören / ¬ beenden, beginnen, enden
	... daß [$_I$ er] [$_T$ pünktlich] **anfing**. ... that he in time begins
	Jetzt können [$_I$ wir] nicht einfach **aufhören**. Now can we not just stop
	Vielleicht sollte [$_I$ ich] **aufhören** und noch studieren. Maybe should I stop and yet study
$\mathbf{n}_I$	+ anfangen, beenden, beginnen / ¬ aufhören, enden
$\mathbf{a}_E$	Nachdem [$_I$ wir] [$_E$ die Sache] **angefangen** haben, After we the thing have started
	[$_I$ Die Polizei] **beendete** [$_E$ die Gewalttätigkeiten]. The police stopped the violence
	[$_T$ Nach dem Abi] **beginnt** [$_I$ Jens] [$_L$ in Frankfurt] [$_E$ seine Lehre] ... After the Abitur begins Jens in Frankfurt his apprenticeship ...
$\mathbf{n}_I$	+ anfangen, beenden, beginnen / ¬ aufhören, enden
$\mathbf{a}_E$	Wenn [$_E$ die Arbeiten] [$_T$ vor dem Bescheid] **angefangen** werden ... If the work before the notification is started ...
[P]	Während [$_X$ für Senna] [$_E$ das Rennen] **beendet** war ... While for Senna the race was finished ...
	... ehe [$_E$ eine militärische Aktion] **begonnen** wird ... ... before a military action is begun
$\mathbf{n}_I$	+ anfangen, aufhören, beginnen / ¬ beenden, enden
$\mathbf{i}_E$	[$_I$ Ich] habe **angefangen**, [$_E$ Hemden zu schneidern]. I have started shirts to make
	... daß [$_I$ der Alkoholiker] **aufhört** [$_E$ zu trinken]. ... that the alcoholic stops to drink
	In dieser Stimmung **begannen** [$_I$ Männer] [$_E$ Tango zu tanzen] ... In this mood began men tango to dance
$\mathbf{n}_I$	+ anfangen, aufhören, beginnen / ¬ beenden, enden
$\mathbf{p}_E$: *mit*	Erst als [$_I$ der versammelte Hofstaat] [$_E$ mit Klatschen] **anfing**, Only when the gathered royal household with applause began
	[$_I$ Der Athlet] ... kann ... [$_E$ mit seinem Sport] **aufhören**. The athlete ... can ... with his sports stop
	[$_I$ Man] **beginne** [$_E$ mit eher katharsischen Werken]. One starts with rather catharsic works
$\mathbf{n}_I$	+anfangen, aufhören, beginnen / ¬ beenden, enden
$\mathbf{p}_E$: *mit*	Und [$_E$ mit den Umbauarbeiten] könnte **angefangen** werden. And with the reconstruction work could be begun
[P]	[$_E$ Mit diesem ungerechten Krieg] muß sofort **aufgehört** werden. With this unjust war must immediately be stopped
	[$_T$ Vorher] dürfe [$_E$ mit der Auflösung] nicht **begonnen** werden. Before must with the closing not be started

2.2 Empirical Distributions for German Verbs

We developed, implemented, and trained a statistical grammar model for German that is based on the framework of head-lexicalized, probabilistic, context-free grammars. The idea originates from Charniak (1997), with this work using an implementation by Schmid (2000) for a training corpus of 35 million words from a collection of large German newspaper corpora from the 1990s, including *Frankfurter Rundschau*, *Stuttgarter Zeitung*, *VDI-Nachrichten*, *die tageszeitung*, *German Law Corpus*, *Donaukurier*, and *Computerzeitung*. The statistical grammar model provides empirical lexical information, specializing in but not restricted to the subcategorization behavior of verbs. Details of

the implementation, training, and exploitation of the grammar model can be found in Schulte im Walde (2003a, chapter 3).

The German verbs are represented by distributional vectors, with features and feature values in the distribution being acquired from the statistical grammar. The distributional description is based on the hypothesis that "each language can be described in terms of a distributional structure, that is, in terms of the occurrence of parts relative to other parts" (cf. Harris 1968). The verbs are described distributionally on three levels at the syntax–semantics interface, each level refining the previous level. The first level *D*1 encodes a purely syntactic definition of verb subcategorization, the second level *D*2 encodes a syntactico-semantic definition of subcategorization with prepositional preferences, and the third level *D*3 encodes a syntactico-semantic definition of subcategorization with prepositional and selectional preferences. Thus, the refinement of verb features starts with a purely syntactic definition and incrementally adds semantic information. The most elaborated description comes close to a definition of verb alternation behavior. We decided on this three-step procedure of verb descriptions because the resulting clusters and particularly the changes in clusters that result from a change of features should provide insight into the meaning–behavior relationship at the syntax–semantics interface.

For *D*1, the statistical grammar model provides frequency distributions for German verbs over 38 purely syntactic subcategorization frames (cf. Appendix 6). Based on these frequencies, we can also calculate the probabilities. For *D*2, the grammar provides frequencies for the different kinds of prepositional phrases within a frame type; probabilities are computed by distributing the joint probability of a verb and a PP frame over the prepositional phrases according to their frequencies in the corpus. Prepositional phrases are referred to by case and preposition, such as mit_{Dat}, $für_{Acc}$. The statistical grammar model does not learn the distinction between PP arguments and PP adjuncts perfectly. Therefore, we did not restrict the PP features to PP arguments, but to 30 PPs according to 'reasonable' appearance in the corpus, as defined by the 30 most frequent PPs that appear with at least 10 different verbs. The subcategorization frame information for *D*1 and *D*2 has been evaluated: Schulte im Walde (2002b) describes the induction of a subcategorization lexicon from the grammar model for a total of 14,229 verbs with a frequency between 1 and 255,676 in the training corpus, and Schulte im Walde (2002a) performs an evaluation of the subcategorization data against manually created dictionary entries and shows that the lexical entries have potential for adding to and improving manual verb definitions.

For the refinement of *D*3, the grammar provides selectional preference information at a fine-grained level: It specifies the possible argument realizations in the form of lexical heads, with reference to a specific verb–frame–slot combination. Obviously, we would run into a sparse data problem if we tried to incorporate selectional preferences into the verb descriptions at such a specific level. We are provided with detailed information at the nominal level, but we need a generalization of the selectional preference definition. A widely used resource for selectional preference information is the semantic ontology *WordNet* (Miller et al. 1990; Fellbaum 1998); the University of Tübingen has developed the German version of WordNet, *GermaNet* (Hamp and Feldweg 1997; Kunze 2000). The hierarchy is realized by means of synsets, sets of synonymous nouns, which are organized by multiple inheritance hyponym/hypernym relationships. A noun can appear in several synsets, according to its number of senses. The German noun hierarchy in GermaNet is utilized for the generalization of selectional preferences: For each noun in a verb–frame–slot combination, the joint frequency is divided over the different senses of the noun and propagated up the hierarchy. In case of multiple hypernym

synsets, the frequency is divided 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, only the frequency distributions over the top GermaNet nodes[1] are considered: *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 are mutually exclusive and the node frequencies sum to the total joint verb-frame frequency, we can use their frequencies to define a probability distribution.

Are selectional preferences equally necessary and informative for all frame types? For example, selectional preferences for the direct object are expected to vary strongly with respect to the subcategorizing verb (because the direct object is a highly frequent argument type across all verbs and verb classes), but selectional preferences for a subject in a transitive construction with a nonfinite clause are certainly less interesting for refinement (because this frame type is more restricted with respect to the verbs it is subcategorized for). We empirically investigated which of the overall frame roles may be realized by different selectional preferences and are therefore relevant and informative for a selectional preference distinction. As a result, in parts of the clustering experiments we will concentrate on a specific choice of frame-slot combinations to be refined by selectional preferences (with the relevant slots underlined): 'n', 'na', 'nd', 'nad', 'ns-dass.'

Table 1 presents three verbs from different classes and their 10 most frequent frame types at the three levels of verb definition and their probabilities. *D*1 for *beginnen* 'begin' defines 'np' and 'n' as the most probable frame types. After splitting the 'np' probability over the different PP types in *D*2, a number of prominent PPs are left, the time indicating um_{Acc} and $nach_{Dat}$, mit_{Dat} referring to the begun event, an_{Dat} as date, and in_{Dat} as place indicator. It is obvious that not all PPs are argument PPs, but adjunct PPs also represent a part of the verb behavior. *D*3 illustrates that typical selectional preferences for beginner roles are *Situation* 'situation', *Zustand* 'state', *Zeit* 'time', *Sache* 'thing'. *D*3 has the potential to indicate verb alternation behavior, for example, 'na(Situation)' refers to the same role for the direct object in a transitive frame as "n(Situation)" in an intransitive frame. *essen* 'eat' as an object-drop verb shows strong preferences for both intransitive and transitive usage. As desired, the argument roles are dominated by *Lebewesen* 'creature' for 'n' and 'na' and *Nahrung* 'food' for 'na'. *fahren* 'drive' chooses typical manner of motion frames ('n,' 'np,' 'na') with the refining PPs being directional (in_{Acc}, zu_{Dat}, $nach_{Dat}$) or referring to a means of motion (mit_{Dat}, in_{Dat}, auf_{Dat}). The selectional preferences show correct alternation behavior: *Lebewesen* 'creature' in the object drop case for 'n' and 'na,' *Sache* 'thing' in the inchoative/causative case for 'n' and 'na'.

In addition to the absolute verb descriptions above, a simple smoothing technique is applied to the feature values. The goal of smoothing is to create more uniform distributions, especially with regard to adjusting zero values, but also for assimilating high and low frequencies and probabilities. The smoothed distributions are particularly interesting for distributions with a large number of features, since they typically contain

1 Since GermaNet had not been completed when we used the hierarchy, we manually added a few hypernym definitions.

Table 1
Example distributions of German verbs.

	Distribution					
Verb	*D*1		*D*2		*D*3	
beginnen	np	0.43	n	0.28	n(Situation)	0.12
'begin'	n	0.28	np:um$_{Acc}$	0.16	np:um$_{Acc}$(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
'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
'drive'	np	0.29	na	0.19	n(Lebewesen)	0.10
	na	0.19	np:in$_{Acc}$	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$_{Acc}$(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$_{Acc}$(Lebewesen)	0.02
	ns-w	0.01	np:auf$_{Dat}$	0.02	np:nach$_{Dat}$(Sache)	0.02

persuasive zero values and severe outliers. Chen and Goodman (1998) present a concise overview of smoothing techniques, with specific emphasis on language modeling. We decided to apply the smoothing algorithm referred to as *additive smoothing*: The smoothing is performed simply by adding 0.5 to all verb features, that is, the joint frequency of each verb v and feature x_i is changed by $\text{freq}'(v, x_i) = \text{freq}(v, x_i) + 0.5$. The total verb frequency is adapted to the changed feature values, representing the sum of all verb feature values: $v_{\text{freq}'} = \sum_i \text{freq}'(v, x_i)$. Smoothed probability values are based on the smoothed frequency distributions.

2.3 Clustering Algorithm and Evaluation Techniques

Clustering is a standard procedure in multivariate data analysis. It is designed to allow exploration of the inherent natural structure of the data objects, where objects in the same cluster are as similar as possible and objects in different clusters are as dissimilar as possible. Equivalence classes induced by the clusters provide a means for

generalizing over the data objects and their features. The clustering of the German verbs is performed by the *k*-means algorithm, a standard unsupervised clustering technique as proposed by Forgy (1965). With *k*-means, initial verb clusters are iteratively reorganized by assigning each verb to its closest cluster and recalculating cluster centroids until no further changes take place. Applying the *k*-means algorithm assumes (1) that verbs are represented by distributional vectors and (2) that verbs that are closer to each other in a mathematically defined way are also more similar to each other in a linguistic way. *k*-Means depends on the following parameters: (1) The number of clusters is not known beforehand, so the clustering experiments investigate this parameter. Related to this parameter is the level of semantic concept: The more verb clusters are found, the more specific the semantic concept, and vice versa. (2) *k*-means is sensitive to the initial clusters, so the initialization is varied according to how much preprocessing we invest: Both random clusters and hierarchically preprocessed clusters are used as initial clusters for *k*-means. In the case of preprocessed clusters, the hierarchical clustering is performed as bottom-up agglomerative clustering with the following criteria for merging the clusters: single linkage (minimal distance between nearest neighbor verbs), complete linkage (minimal distance between furthest neighbor verbs), average distance between verbs, distance between cluster centroids, and Ward's method (minimizing the sum of squares when merging clusters). The merging method influences the shape of the clusters; for example, single linkage causes a chaining effect in the shape of the clusters, and complete linkage creates compact clusters. (3) In addition, there are several possibilities for defining the similarity between distributional vectors. But which best fits the idea of verb similarity? Table 2 presents an overview of relevant similarity measures that are applied in the experiments. x and y refer to the verb object vectors, their subscripts to the verb feature values. The *Minkowski metric* can be applied to frequencies and probabilities. It is a generalization of the two well-known instances $q = 1$ (*Manhattan distance*) and $q = 2$ (*Euclidean distance*). The *Kullback–Leibler divergence (KL)* is a measure from information theory that determines the inefficiency of assuming a model probability distribution given the true distribution (Cover and Thomas 1991). The KL divergence is not defined in case $y_i = 0$, so the probability distributions need to be smoothed. Two variants of KL, *information radius* and *skew divergence*, perform a default smoothing. Both variants can tolerate zero values in the distribution because they work with a weighted average of the two distributions compared. Lee (2001) has shown that the skew divergence is an effective measure for distributional similarity in NLP. Similarly to Lee's method, we set the weight w for the skew divergence to 0.9. The *cosine* measures the similarity of the two object vectors x and y by calculating the cosine of the angle between the feature vectors. The cosine measure can be applied to frequency and probability values. For a detailed description of hierarchical clustering techniques and an intuitive interpretation of the similarity measures, the reader is referred to, for example, Kaufman and Rousseeuw (1990).

There is no agreed standard method for evaluating clustering experiments and results, but a variety of evaluation measures from diverse areas such as theoretical statistics, machine vision, and Web-page clustering are generally applicable. We used the following two measures for the evaluation: (1) Hatzivassiloglou and McKeown (1993) define and evaluate a cluster analysis of adjectives, based on common cluster membership of object pairs in the clustering *C* and the manual classification *M*. Recall and precision numbers are calculated in the standard way, with true positives the number of common pairs in *M* and *C*, false positives the number of pairs in *C*, but not *M*, and false negatives the number of pairs in *M*, but not *C*. We use the *f*-score *pairF* (as harmonic mean between recall and precision), which provides an easy to understand

Table 2
Data similarity measures.

Measure	Definition
Minkowski metric / L_q norm	$L_q(x,y) = \sqrt[q]{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^q}$
Manhattan distance / L_1 norm	$L_1(x,y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Euclidean distance / L_2 norm	$L_2(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
KL divergence / relative entropy	$D(x \lvert\lvert y) = \sum_{i=1}^{n} \; x_i \; * \; log \; \frac{x_i}{y_i}$
Information radius	$\text{IRad}(x,y) = D(x \lvert\lvert \frac{x+y}{2}) \; + \; D(y \lvert\lvert \frac{x+y}{2})$
Skew divergence	$\text{Skew}(x,y) = D(x \lvert\lvert w * y \; + \; (1-w) * x)$
Cosine	$\cos(x,y) = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} * \sqrt{\sum_{i=1}^{n} y_i^2}}$

percentage. (2) The adjusted Rand index is a measure of agreement versus disagreement between object pairs in clusterings that provides the most appropriate reference to a null model (Hubert and Arabie 1985); cf. equation (1). The agreement in the two partitions is represented by a contingency table $C \times M$: t_{ij} denotes the number of verbs common to classes C_i in the clustering partition C and M_j in the manual classification M; the marginals $t_{i.}$ and $t_{.j}$ refer to the number of objects in C_i and M_j, respectively; the expected number of common object pairs attributable to a particular cell $\langle C_i, M_j \rangle$ in the contingency table is defined by $\binom{t_{i.}}{2} \binom{t_{.j}}{2} / \binom{n}{2}$. The upper bound for Rand_{adj} is 1, the lower bound is mostly 0, with only extreme cases below zero.

$$\text{Rand}_{adj}(C,M) = \frac{\sum_{i,j} \binom{t_{ij}}{2} \; - \; \frac{\sum_i \binom{t_{i.}}{2} \sum_j \binom{t_{.j}}{2}}{\binom{n}{2}}}{\frac{1}{2} \left(\sum_i \binom{t_{i.}}{2} \; + \; \sum_j \binom{t_{.j}}{2} \right) \; - \; \frac{\sum_i \binom{t_{i.}}{2} \sum_j \binom{t_{.j}}{2}}{\binom{n}{2}}} \tag{1}$$

The above two measures were chosen as a result of comparing various evaluation measures and their properties with respect to the linguistic task (Schulte im Walde 2003a, chapter 4).

3. Preliminary Clustering Experiments

The 168 German verbs are associated with distributional vectors over frame types and assigned to initial clusters. Then *k*-means is allowed to run for as many iterations as it takes to reach a fixed point, and the resulting clusters are interpreted and evaluated against the manual classes. The verbs are described by *D*1–*D*3, and each level refers to frequencies and probabilities, with original and smoothed values. The initial clusters for *k*-means are generated either randomly or by a preprocessing cluster analysis, that is, hierarchical clustering as described in Section 2.3. For random cluster initialization the verbs are randomly assigned to a cluster, with cluster numbers between 1 and the number of manual classes. The experiments are performed with the number of *k* clusters being fixed to the number of gold standard classes (43); optimization of the number of clusters is addressed in Section 3.4.

3.1 Baseline and Upper Bound

The experiment baseline refers to 50 random clusterings: The verbs are randomly assigned to a cluster (with a cluster number between 1 and the number of manual classes), and the resulting clustering is evaluated by the evaluation measures. The baseline value is the average value of the 50 repetitions. The upper bound of the experiments (the "optimum") refers to the evaluation values on the manual classification; the manual classification is adapted before calculating the upper bound by randomly deleting additional senses (i.e., more than one sense) of a verb, so as to leave only one sense for each verb, since *k*-means as a hard clustering algorithm cannot model ambiguity. Table 3 lists the baseline and upper bound values for the clustering experiments.

3.2 Experiment Results

The following tables present the results of the clustering experiments. Tables 4 to 7 each concentrate on one technical parameter of the clustering process; Tables 8 to 10 then focus on performing clustering with a fixed parameter set, in order to vary the linguistically interesting parameters concerning the feature choice for the verbs. All significance tests have been performed with $\chi^2, df = 1, \alpha = 0.05$.

Table 4 illustrates the effect of the *distribution units* (frequencies and probabilities) on the clustering result. The experiments use distributions on $D1$ and $D2$ with random and preprocessed initialization, and the cosine as similarity measure (since it works for both distribution units). To summarize the results, neither the differences between frequencies and probabilities nor between original and smoothed values are significant.

Table 5 illustrates the usage of different *similarity measures.* As before, the experiments are performed for $D1$ and $D2$ with random and preprocessed initialization. The similarity measures are applied to the relevant probability distributions (as the distribution unit that can be used for all measures). The tables point out that there is no best-performing similarity measure in the clustering processes. On the larger feature set, the Kullback–Leibler variants information radius and skew divergence tend to outperform all other similarity measures. In fact, the skew divergence is the only measure that shows significant differences for some parameter settings, as compared to all other measures except information radius. In further experiments, we will therefore concentrate on the two Kullback–Leibler variants.

Tables 6 and 7 compare the effects of varying the *initialization* of the *k*-means algorithm. The experiments are performed for $D1$ and $D2$ with probability distributions, using the similarity measures information radius and skew divergence. For random and hierarchical initialization, we cite both the evaluation scores for the *k*-means initial cluster analysis (i.e., the output clustering from the random assignment or the preprocessing hierarchical analysis), and for the *k*-means result. The *manual* columns in the tables refer to a cluster analysis where the initial clusters provided to

Table 3
k-means experiment baseline and upper bound.

Evaluation	Baseline	Optimum
PairF	2.08	95.81
Rand_{adj}	−0.004	0.909

Table 4
Comparing distributions on *D*1 and *D*2.

		Distribution: *D*1				Distribution: *D*2			
		Probability		Frequency		Probability		Frequency	
Eval	Initial	Original	Smoothed	Original	Smoothed	Original	Smoothed	Original	Smoothed
PairF	Random	12.67	12.72	14.06	14.14	14.98	15.37	14.82	15.07
	H-Ward	11.40	11.70	11.56	11.37	10.57	13.71	11.65	9.98
$Rand_a$	Random	0.090	0.090	0.102	0.102	0.104	0.113	0.107	0.109
	H-Ward	0.079	0.081	0.080	0.076	0.065	0.096	0.075	0.056

Table 5
Comparing similarity measures on *D*1 and *D*2.

		Similarity Measure									
		*D*1					*D*2				
Eval	Initial	Cos	L1	Eucl	IRad	Skew	Cos	L1	Eucl	IRad	Skew
PairF	Random	12.67	13.11	13.85	14.19	14.13	14.98	15.20	16.10	16.15	18.01
	H-Ward	11.40	13.65	12.88	13.07	12.64	10.57	15.51	13.11	17.49	19.30
$Rand_a$	Random	0.090	0.094	0.101	0.101	0.105	0.104	0.109	0.123	0.118	0.142
	H-Ward	0.079	0.099	0.093	0.097	0.094	0.065	0.116	0.092	0.142	0.158

k-means are the manual classification, that is, the gold standard. An optimal cluster analysis should realize the "perfect" clustering and not perform any reorganization of the clusters. In the experiments, *k*-means does perform iterations, so the clustering result is suboptimal. This finding is caused by the syntax–semantics mismatches, which we deliberately included in the definition of the gold standard (recall, e.g., that *unterstützen* is syntactically very different compared to the other three *Support* verbs). In addition, the results not only show that the feature sets are suboptimal, but also that the loss in quality is less for the linguistically refined feature level *D*2 compared to *D*1, as we would have hoped. For *random* clustering initialization to *k*-means, the tables present both the best and the average clustering results. The best results are paired with the evaluation of their initial clusters, that is, the random clusterings. As the tables show, the initial clusters receive low evaluation scores. Typically, the clusterings consist of clusters with rather homogeneous numbers of verbs, but the perturbation within the clusters is high, as expected. *k*-means is able to cope with the high degree of perturbation: The resulting clusters improve significantly and are comparable with those based on preprocessed hierarchical clustering; this competitiveness vanishes with an increasing number of features. The average values of the random initialization experiments are clearly below the best ones, but not significantly different. Cluster analyses as based on agglomerative hierarchical clustering with *single-linkage* amalgamation are evaluated as poor compared to the gold standard. This result is probably due to the chaining effect in the clustering, which is characteristic for single linkage; the effect is observable in the analysis, which typically contains one very large cluster and many clusters with few verbs, mostly singletons. *k*-means obviously cannot compensate for this strong bias in cluster sizes (and their respective centroids); the reorganization improves the clusterings, but the result is still worse than for any other initialization. With *average distance* and *centroid distance* amalgamation, both the clusterings and the evaluation results are less extreme than with single linkage since the chaining effect is smoothed.

Table 6
Comparing clustering initializations on *D*1.

		k-means Initialization		
			Random	
Eval	Distance	Manual	Best	Avg
PairF	IRad	18.56	2.16 → 14.19	11.78
	Skew	20.00	1.90 → 14.13	12.17
Rand_a	IRad	0.150	−0.004 → 0.101	0.078
	Skew	0.165	−0.005 → 0.105	0.083

		k-means Initialization				
		Hierarchical				
Eval	Distance	Single	Complete	Average	Centroid	Ward
PairF	IRad	4.80 → 12.73	9.43 → 10.16	10.83 → 11.33	8.77 → 11.88	12.76 → 13.07
	Skew	4.81 → 13.04	11.50 → 11.00	11.68 → 11.41	8.83 → 11.45	12.44 → 12.64
Rand_a	IRad	0.000 → 0.088	0.055 → 0.065	0.067 → 0.072	0.039 → 0.079	0.094 → 0.097
	Skew	0.000 → 0.090	0.077 → 0.072	0.075 → 0.073	0.041 → 0.072	0.092 → 0.094

The overall results are better than for single linkage, but only slightly improved by *k*-means. Hierarchical clusters as based on *complete-linkage* amalgamation are more compact, and result in a closer fit to the gold standard than the previous methods. The hierarchical initialization is only slightly improved by *k*-means; in some cases the *k*-means output is worse than its hierarchical initialization. *Ward's method* seems to work best on hierarchical clusters and *k*-means initialization. The cluster sizes are more balanced and correspond to compact cluster shapes. As for complete linkage, *k*-means improves the clusterings only slightly; in some cases the *k*-means output is worse than its hierarchical initialization. A cluster analysis based on Ward's hierarchical clusters performs best of all the applied methods, especially with an increasing number of features. The similarity of Ward's clusters (and similarly complete linkage clusters)

Table 7
Comparing clustering initializations on *D*2.

		k-means Initialization		
			Random	
Eval	Distance	Manual	Best	Avg
PairF	IRad	40.23	1.34 → 16.15	13.37
	Skew	47.28	2.41 → 18.01	14.07
Rand_a	IRad	0.358	0.001 → 0.118	0.093
	Skew	0.429	−0.002 → 0.142	0.102

		k-means Initialization				
		Hierarchical				
Eval	Distance	Single	Complete	Average	Centroid	Ward
PairF	IRad	5.06 → 11.12	15.37 → 14.44	10.50 → 10.64	9.16 → 12.90	17.86 → 17.49
	Skew	5.20 → 10.64	15.21 → 13.81	10.02 → 10.02	9.04 → 10.91	15.86 → 15.23
Rand_a	IRad	0.003 → 0.063	0.114 → 0.105	0.059 → 0.060	0.045 → 0.082	0.145 → 0.142
	Skew	0.004 → 0.063	0.115 → 0.102	0.054 → 0.054	0.042 → 0.064	0.158 → 0.158

and *k*-means is not by chance, since these methods aim to optimize the same criterion, the sum of distances between the verbs and their respective cluster centroids. Note that for *D*2, Ward's method actually significantly outperforms all other initialization methods, complete linkage significantly outperforms all but Ward's. Between single linkage, average and centroid distance, there are no significant differences. For *D*1, there are no significant differences between the initializations.

The low scores in the tables might be surprising to the reader, but they reflect the difficulty of the task. As mentioned before, we deliberately set high demands for the gold standard, especially with reference to the fine-grained, small classes. Compared to related work (cf. Section 5), our results achieve lower scores because the task is more difficult; for example, Merlo and Stevenson (2001) classify 60 verbs into 3 classes, and Siegel and McKeown (2000) classify 56 verbs into 2 classes, as compared to our clustering, which assigns 168 verbs to 43 classes. The following illustrations should provide an intuition about the difficulty of the task:

1. In a set of additional experiments, a random choice of a reduced number of 5/10/15/20 classes from the gold standard is performed. The verbs from the respective gold standard classes are clustered with the optimal parameter set (see Table 8), which results in a pairwise f-score *PairF* of 22.19%. The random choice and the cluster analysis are repeated 20 times for each reduced gold standard size of 5/10/15/20 classes, and the average *PairF* is calculated: The results are 45.27/35.64/30.30/26.62%, respectively. This shows that the clustering results are much better (with the same kind of data and features and the same algorithm) when applied to a smaller number of verbs and classes.

2. Imagine a gold standard of three classes with four members each, for example, {{a, b, c, d}, {e, f, g, h}, {i, j, k, l}}. If a cluster analysis of these elements into three clusters resulted in an almost perfect choice of {{a, b, c, d, e}, {f, g, h}, {i, j, k, l}} where only *e* is assigned to a "wrong" class, the pairwise precision is 79%, the recall is 83%, and *pairF* is 81%, so the decrease of *pairF* with only one mistake is almost 20%. If another cluster analysis resulted in a choice with just one more mistake such as {{a, b, c, d, e, i}, {f, g, h}, {j, k, l}} where *i* is also assigned to a "wrong" class, the result decreases by almost another 20%, to a precision of 57%, a recall of 67%, and *pairF* of 62%. The results show how much impact a few mistakes may have on the pairwise *f*-score of the results.

In addition to defining a difficult task, we also chose strong evaluation measures: Evaluating pairs of objects results in lower numbers than evaluating the individual objects. For example, the *accuracy/purity* measure (Stevenson and Joanis 2003; Korhonen, Krymolowski, and Marx 2003) evaluates whether a verb is assigned to a correct cluster with respect to the gold standard class of the majority of cluster members. That is, in a first step each induced verb cluster is assigned a gold standard class according to which class captures the majority of the cluster members. In a second step, each verb in a cluster is evaluated as correct or wrong with respect to its gold standard class, and accuracy/purity of the whole clustering is calculated as the proportion of correct verbs divided by the total number of verbs. If we applied this measure to our optimal clustering with a pairwise *f*-score *PairF* of 22.19%, we achieve an accuracy of 51.19%; if we applied the measure to the above random choices of gold standard classes with 5/10/15/20 classes, we achieve accuracies of 68.20/60.73/57.82/55.48%.

The last series of experiments applies the algorithmic insights from the previous experiments to a linguistic variation of parameters (cf. Schulte im Walde 2003b). The verbs are described by probability distributions on different levels of linguistic information (frames, prepositional phrases, and selectional preferences). A preprocessing hierarchical cluster analysis is performed by Ward's method, and *k*-means is applied to re-organize the clusters. Similarities are measured by the skew divergence. Table 8 presents the first results comparing *D*1, *D*2, and *D*3, either on specified frame slots ('n,' 'na,' 'nd,' 'nad,' 'ns-dass'), on all noun phrase slots (NP), or on all noun phrase and prepositional phrase slots (NP–PP). The number of features in each experiment is given in square brackets. The table demonstrates that a purely syntactic verb description gives rise to a verb clustering clearly above the baseline. Refining the coarse subcategorization frames with prepositional phrases considerably improves the verb clustering results. Adding selectional preferences to the verb description further improves the clustering results, but the improvement is not as persuasive as in the first step, when refining the purely syntactic verb descriptions with prepositional information. The difference between *D*1 and *D*2 is significant, but neither the difference between *D*2 and *D*3 (in any variation) nor the differences between the variants of *D*3 are significant. In the case of adding role information to all NP (and all PP) slots, the problem might be caused by sparse data, but for the linguistically chosen subset of argument slots we assume additional linguistic reasons are directly relevant to the clustering outcome.

In order to choose the most informative frame roles for *D*3, we varied the selectional preference slots by considering only single slots for refinements, or small combinations of argument slots. The variations should provide insight into the contribution of slots and slot combinations to the clustering. The experiments are performed on probability distributions for *D*3; all other parameters were chosen as above. Table 9 shows that refining only a single slot (the underlined slot in the respective frame type) in addition to the *D*2 definitions results in little or no improvement. There is no frame-slot type that consistently improves results, but success depends on the parameter instantiation. The results do not match our linguistic intuitions: For example, we would expect the arguments in the two highly frequent intransitive 'na' and transitive 'na' frames with variable semantic roles to provide valuable information with respect to their selectional preferences, but only those in 'na' actually improve *D*2. However, a subject in a transitive construction with a non-finite clause 'ni', which is less variable with respect to verbs and roles, does work better than 'n'. In Table 10, selected slots are combined to define selectional preference information, for example, n/na means that the nominative slot in 'na', and both the nominative and accusative slot in 'na' are refined by selectional preferences. It is obvious that the clustering effect does not represent a sum of its parts, for example, both the information in 'na' and in 'na' improve Ward's clustering based on *D*2 (cf. Table 9), but it is not the case that 'na' improves the clustering, too.

Table 8
Comparing feature descriptions.

	Distribution				
Eval	*D*1 [38]	*D*2 [183]	*D*3 [288]	*D*3_NP [906]	*D*3_NP–PP [2,726]
PairF	12.64	18.81	22.19	19.29	21.11
$Rand_{adj}$	0.094	0.151	0.182	0.158	0.176

Table 9
Comparing selectional preference slot definitions.

	Distribution						
				*D*3			
Eval	*D*2	n	na	na	nad	nad	nad
PairF	18.81	16.22	21.15	20.19	17.82	15.13	19.48
Rand_{adj}	0.151	0.125	0.176	0.164	0.144	0.115	0.161

	Distribution							
					*D*3			
Eval	*D*2	nd	nd	np	ni	nr	ns-2	ns-dass
PairF	18.81	18.88	17.92	16.77	18.26	17.22	15.55	19.29
Rand_{adj}	0.151	0.152	0.143	0.133	0.148	0.136	0.121	0.156

As in Table 9, there is no combination of selectional preference frame definitions that consistently improves the results. On the contrary, some additional *D*3 information makes the result significantly worse, for example, 'nad'. The specific combination of selectional preferences as determined preexperimentally actually achieves the overall best results, better than any other slot combination, and better than refining all NP slots or refining all NP and all PP slots in the frame types (cf. Table 8).

3.3 Experiment Interpretation

For illustrative purposes, we present representative parts of the cluster analysis as based on the following parameters: The clustering initialization is obtained from a hierarchical analysis of the German verbs (Ward's amalgamation method), the number of clusters being the number of manual classes (43); the similarity measure is the skew divergence. The cluster analysis is based on the verb description on *D*3, with selectional roles for

Table 10
Comparing selectional preference frame definitions.

	Distribution					
				*D*3		
Eval	*D*2	n	na	n/na	nad	n/na/nad
PairF	18.81	16.22	17.82	17.00	13.36	16.05
Rand_{adj}	0.151	0.125	0.137	0.128	0.088	0.118

	Distribution				
				*D*3	
Eval	*D*2	nd	n/na/nd	n/na/nad/nd	np/ni/nr/ns-2/ns-dass
PairF	18.81	18.48	16.48	20.21	16.73
Rand_{adj}	0.151	0.150	0.124	0.161	0.131

'n,' 'na,' 'nd,' 'nad,' 'ns-dass.' We compare the clusters with the respective clusters by *D*1 and *D*2.

(a) nieseln regnen schneien – *Weather*

(b) dämmern – *Weather*

(c) beginnen enden – *Aspect*
bestehen 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 *D*1 and do not need *D*2 or *D*3 refinements for an improved class constitution. *dämmern* in cluster (b) is ambiguous between a weather verb and expressing a sense of understanding; this ambiguity is already idiosyncratically expressed in *D*1 frames, so *dämmern* is never clustered together with the other weather verbs by *D*1–*D*3. *Manner of Motion, Existence, Position,* and *Aspect* verbs are similar in their syntactic frame usage and therefore merged together by *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 distinguishing these verb classes, and the coherence is partly destroyed by *D*3: *Manner of Motion* verbs (from the subclasses *Locomotion, Rotation, Rush, Vehicle, Flotation*) are captured well by clusters (d) and (e), since they use common alternations, but cluster (c) merges *Existence, Position,* and *Aspect* verbs because verb-idiosyncratic demands on selectional roles destroy the *D*2 class demarcation. Still, the verbs in cluster (c) are close in their (more general conceptual) semantics, with a common sense of (bringing into versus being in) existence. *laufen* fits into the cluster with its sense of "function." Cluster (f) contains most verbs of *Quantum Change*, together with one verb of *Production* and *Constitution* each. The common conceptual level of this cluster therefore refers to a quantum change including the quantum change from zero to something (as for the two verbs *festlegen*, 'constitute,' and *bilden*, 'found'). The verbs in this cluster typically subcategorize for a direct object, alternating with a reflexive usage, "nr" and "npr" with mostly auf_{Acc} and um_{Acc}. 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 by *D*1–*D*3, so a common cluster is justified, but the verbs only have coarse common meaning components; in this case *töten* 'kill' and *unterrichten* 'teach' agree in an action of one person or institution towards another. *geben* in cluster (h) represents a singleton. 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*.

An overall interpretation of the clustering results gives insight into the relationship between verb properties and clustering outcome. (1) The fact that there are verbs that are clustered semantically on the basis of their corpus-based and knowledge-based empirical properties indicates (a) a relationship between the meaning components of the verbs and their behavior and (b) that the clustering algorithm is able to benefit from the linguistic descriptions and to abstract away from the noise in the distributions. (2) Low-frequency verbs were a problem in the clustering experiments. Their distributions are noisier than those for more frequent verbs, so they typically constitute noisy clusters. (3) As known beforehand, verb ambiguity cannot be modeled by the hard clustering algorithm *k*-means. Ambiguous verbs were typically assigned either (a) to one of the correct clusters or (b) to a cluster whose verbs have distributions that are similar to the ambiguous distribution, or (c) to a singleton cluster. (4) The interpretation of the clusterings unexpectedly points to meaning components of verbs that have not been discovered by the manual classification. 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 discovery effect should be more impressive with an increasing number of verbs, since manual judgement is more difficult, and also with a soft clustering technique, where multiple cluster assignment is enabled. (5) In a similar way, the clustering interpretation exhibits semantically related verb classes, that is, verb classes that 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. (6) Related to the preceding issue, the manual verb classes as defined are demonstrated as detailed and subtle. Compared to a more general classification that would appropriately merge several classes, the clustering confirms that we defined a difficult task with subtle classes. We were aware of this fact but preferred a fine-grained classification, since it allows insight into verb and class properties. In this way, verbs that are similar in meaning are often clustered incorrectly with respect to the gold standard.

To come to the main point, what exactly is the nature of the meaning–behavior relationship? (1) Already a purely syntactic verb description allows a verb clustering clearly above the baseline. The result is a (semantic) classification of verbs that agree in their syntactic frame definitions, for example, most of the *Support* verbs. The clustering fails for semantically similar verbs that differ in their syntactic behavior, for example, *unterstützen*, which belongs to the *Support* verbs but demands an accusative rather than a dative object. In addition, it fails for syntactically similar verbs that are clustered together even though they do not exhibit semantic similarity; for example, many verbs from different semantic classes subcategorize for an accusative object, so they are falsely clustered together. (2) Refining the syntactic verb information with 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 that are similar in their meaning agree either on a specific prepositional complement (e.g., *glauben/denken an*$_{Acc}$) or on a more general kind of modification, for example, directional PPs for *Manner of Motion* verbs. (3) Defining selectional preferences for arguments improves the clustering results further, but the improvement is not as persuasive as when refining the purely syntactic verb descriptions with prepositional information. For example, 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. In contrast, there are verbs that are very similar in their behavior, especially with respect to a coarse definition of selectional roles, but they do not belong to the same fine-grained semantic class, for example, *töten* and *unterrichten*. The effect could be due to (a) noisy or (b) sparse data, but the basic verb descriptions appear reliable with respect to their desired linguistic content, and Table 8 illustrates that even with little added information the effect exists (e.g., refining few arguments by 15 selectional roles results in 253 instead of 178 features, so the magnitude of feature numbers does not change). Why do we encounter an indeterminism concerning the encoding and effect of verb features, especially with respect to selectional preferences? The meaning of verbs comprises both properties that are general for the respective verb classes, and idiosyncratic properties that distinguish the verbs from each other. As long as we define the verbs by those properties that represent the common parts of the verb classes, a clustering can succeed. But by stepwise refining the verb description and including lexical idiosyncrasy, the emphasis on the common properties vanishes. From a theoretical point of view, the distinction between common and idiosyncratic features is obvious, but from a practical point of view there is no perfect choice for the encoding of 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 behavioral properties of the verb classes might differ. Still, for a large-scale classification of verbs, we need to specify a combination of linguistic verb features as a basis for the clustering. Which combination do we choose? Both the theoretical assumption of encoding features of verb alternation as verb behavior and the practical realization by encoding syntactic frame types, prepositional phrases, and selectional preferences seem promising. In addition, we aimed at a (rather linguistically than technically based) choice of selectional preferences that represents a useful compromise for the conceptual needs of the verb classes. Therefore, this choice of features best utilizes the meaning–behavior relationship and will be applied in a large-scale clustering experiment (cf. Section 4).

3.4 Optimizing the Number of Clusters

It is not a goal of this article to optimize the number of clusters in the cluster analysis. We are not interested in the question of whether, for example, 40, 42, 43, or 45 clusters represent the best semantic classification of 168 verbs. But there are two reasons why it is interesting and relevant to investigate the properties of clusterings with respect to different numbers of clusters. (1) The clustering methodology should basically work the way we expect it to work, that is, the evaluation of the results should show deficiencies for extreme numbers of clusters, but (possibly several) optimal values for various numbers of clusters in between. (2) Even if we do not check for an exact number of clusters, we should check the magnitude of the number of clusters, since the clustering

methodology might be successful in capturing a rough verb classification with few verb classes but not a fine-grained classification with many subtle distinctions.

Figure 1 illustrates the clustering results for the series of cluster analyses as performed by *k*-means with hierarchical clustering initialization (Ward's method) on probability distributions, with skew divergence as the similarity measure. The feature description refers to *D*2. The number of clusters is varied from 1 through the number of verbs (168), and the results are evaluated by Rand_{adj}. A range of numbers of clusters is determined as optimal (71) or near-optimal (approx. 58–78). The figure demonstrates that having performed experiments on the parameters for clustering, it is worthwhile exploring additional parameters: The optimal result is 0.188 for 71 clusters as compared to 0.158 for 43 clusters reported previously.

4. Large-Scale Clustering Experiments

So far, all clustering experiments were performed on a small scale, preliminary set of 168 manually chosen German verbs. One goal of this article was to develop a clustering methodology with respect to the automatic acquisition of a large-scale German verb classification. We therefore apply the insights on the theoretical relationship between verb meaning and verb behavior and our findings regarding the clustering parameters to a considerably larger amount of verb data.

We extracted all German verbs from our statistical grammar model that appeared with an empirical frequency of between 500 and 10,000 in the training corpus (cf. Section 2.2). This selection resulted in a total of 809 verbs, including 94 verbs from the preliminary set of 168 verbs. We added the missing verbs from the preliminary set, resulting in a total of 883 German verbs. The feature description of the German verbs refers to the probability distribution over the coarse syntactic frame types, with

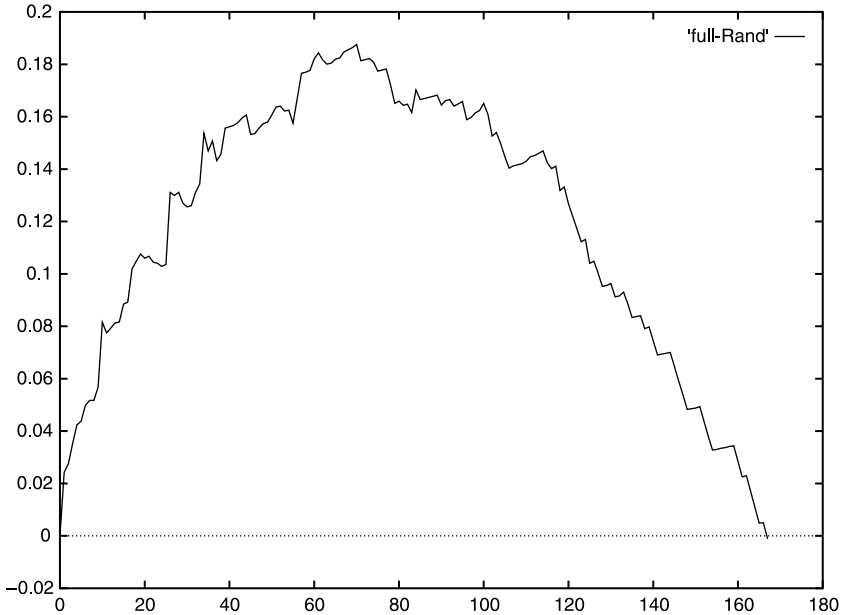

Figure 1
Varying the number of clusters (evaluation: Rand_{adj}).

prepositional phrase information on the 30 chosen PPs and selectional preferences for our empirically most successful combination 'n,' 'na,' 'nd,' 'nad,' and 'ns-dass.' As in previous clustering experiments, the features are stepwise refined. *k*-means is provided hierarchical clustering initialization (based on Ward's method), with the similarity measure being skew divergence. The number of clusters is set to 100, which corresponds to an average of 8.83 verbs per cluster, that is, not too fine-grained clusters but still possible to interpret. The preliminary set of 168 verbs is a subset of the large-scale set in order to provide an "auxiliary" evaluation of the clustering results: Considering only the manually chosen verbs in the clustering result, this partial cluster analysis is evaluated against the gold standard of 43 verb classes. Results were not expected to match the results of our clustering experiments using only the preliminary verb set, but to provide an indication of how different cluster analyses can be compared with each other.

Tables 11 to 13 present the clustering results for the large-scale verb set for *D*1–*D*3 in the rightmost columns, citing the evaluation scores of the initial (hierarchical) clusters and the resulting *k*-means clusters. The subset of the 168 gold standard verbs is scattered over 72 of the 100 resulting clusters. The results are compared to our previous results for the 168 verbs in 43 clusters, and to the case where those 168 verbs are clustered into 72 hierarchical classes. The large-scale clustering results once more confirm the general insights (1) that the stepwise refinement of features improves the clustering and (2) that Ward's hierarchical clustering is seldom improved by the *k*-means application. In addition, several of the large-scale cluster analyses were quite comparable with the clustering results using the small-scale set of verbs, especially when compared to 72 clusters.

In the following, we present example clusters from the optimal large-scale cluster analysis (according to the above evaluation): Ward's hierarchical cluster analysis based on subcategorization frames, PPs, and selectional preferences, without running *k*-means on the hierarchical clustering. 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 recognize common elements of meaning. The verbs that we think are semantically similar are marked in bold face.

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

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

Table 11
Large-scale clustering on *D*1.

	Small-Scale		Large-Scale
Eval	43 Clusters	72 Clusters	72 Clusters
PairF	12.44 → 12.64	10.83 → 11.73	12.15 → 12.88
Rand_{adj}	0.092 → 0.094	0.084 → 0.091	0.094 → 0.102

Table 12
Large-scale clustering on *D*2.

	Small-Scale		Large-Scale
Eval	43 Clusters	72 Clusters	72 Clusters
PairF	18.64 → 18.81	17.56 → 18.81	18.22 → 16.96
$Rand_{adj}$	0.148 → 0.151	0.149 → 0.161	0.152 → 0.142

Table 13
Large-scale clustering on *D*3 with n/na/nd/nad/ns-dass.

	Small-Scale		Large-Scale
Eval	43 Clusters	72 Clusters	72 Clusters
PairF	22.86 → 22.19	19.47 → 20.48	19.92 → 15.06
$Rand_{adj}$	0.190 → 0.182	0.165 → 0.174	0.170 → 0.115

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

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

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

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

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

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

(9) ***anbieten*** 'offer', ***bieten*** 'offer', ***erlauben*** 'allow', ***erleichtern*** 'facilitate', ***ermöglichen*** 'make possible', ***eröffnen*** 'open', ***untersagen*** 'forbid', *veranstalten* 'arrange', ***verbieten*** 'forbid'

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

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

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

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

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

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

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

Clusters (1) to (3) are examples where the verbs do not share elements of meaning. In the overall cluster analysis, such semantically incoherent clusters tend to be rather large, that is, with more than 15–20 verb members. Clusters (4) to (7) are examples of clusters where some of the verbs show overlap in meaning, but 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 approval, and cluster (7) contains verbs of quantum change. Clusters (8) to (16) are examples of clusters where most or all verbs show a strong similarity in their semantic concept. Cluster (8) contains verbs expressing a propositional attitude; the underlined verbs, in addition, indicate an emotion. The only unmarked verb *weißen* can be explained, since some of its inflected forms are ambiguous with respect to their base verb: either *weißen* or *wissen*, a verb that belongs to the *Aspect* verb class. The verbs in cluster (9) describe a scene where somebody or some situation makes something possible (in the positive or negative sense); 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 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 releasing 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 thought (with or without talking) about a certain matter. Finally in cluster (16), we can recognize the same weather verb cluster as in the previously discussed small-scale cluster analyses.

We experimented with two variations in the clustering setup: (1) For the selection of the verb data, we considered a random choice of German verbs in approximately the same magnitude of number of verbs (900 verbs plus the preliminary verb set), but without any restriction on the verb frequency. The clustering results are—both on the basis of the evaluation and on the basis of a manual inspection of the resulting clusters—much worse than in the preceding cluster analysis, since the large number of low-frequency verbs destroys the clustering. (2) The number of target clusters was set to 300 instead of 100, that is, the average number of verbs per cluster was 2.94 instead of 8.83. The resulting clusters are numerically slightly worse than in the preceding cluster analysis, but easier for inspection and therefore a preferred basis for a large-scale resource. Several of the large, semantically incoherent clusters are split into smaller and more coherent clusters, and the formerly coherent clusters often preserved their constitution. To present one example, the following cluster from the 100-cluster analysis

> *anzeigen* 'announce', *aufklären* 'clarify', *beeindrucken* 'impress', *befreien* 'free', *begeistern* 'inspire', *beruhigen* 'calm down', *enttäuschen* 'disappoint', *retten* 'save', *schützen* 'protect', *stören* 'disturb', *überraschen* 'surprise', *überzeugen* 'persuade'

is split into the following four clusters from the 300-cluster analysis:

(a) ***anzeigen*** 'announce', ***aufklären*** 'clarify'

(b) ***beeindrucken*** 'impress', ***enttäuschen*** 'disappoint', ***überraschen*** 'surprise', ***überzeugen*** 'persuade'

(c) ***befreien*** 'free', ***beruhigen*** 'calm down', ***retten*** 'save', ***schützen*** 'protect', ***stören*** 'disturb'

(d) begeistern

where cluster (a) shows a loose semantic coherence of declaration, the verbs in cluster (b) are semantically very similar and describe an emotional impact of somebody or a situation on a person, and the verbs in cluster (c) show a protective (and the negation: nonprotective) influence of one person towards another.

Summarizing, the large-scale clustering experiment results in a mixture of semantically coherent and incoherent verb classes. Semantically incoherent verb classes and clustering mistakes need to be split into finer and more coherent clusters, or to be filtered from the classification. Semantically coherent verb classes need little manual correction as a lexical resource. Interestingly, the coherence in verb classes refers to different criteria on meaning coherence, such as synonymy (e.g., *reduzieren* 'reduce' and *verringern* 'decrease'), antonymy (e.g., *reduzieren* 'reduce' and *erhöhen* 'raise'), situational overlap (e.g., emotional state containing *freuen* 'be glad' and *ärgern* 'be annoyed'), and participation in a common process/script (e.g., *bestellen* 'order', *kaufen* 'buy', *verkaufen* 'sell', and *abholen* 'pick up').

5. Related Work

The following paragraphs describe related classification and clustering experiments on the automatic induction of verb classes. The classifications refer to different class criteria, for example, aspectual properties (Siegel and McKeown 2000), syntactic categories (Merlo and Stevenson 2001; Merlo et al. 2002; Tsang, Stevenson, and Merlo 2002), and—most similar to my approach—semantic categories (Schulte im Walde 2000; Joanis 2002). The soft clustering approaches indicate how we might extend our hard clustering to verb ambiguity, now that we have determined the relevant set of verb features.

Siegel and McKeown (2000) used three supervised and one unsupervised machine-learning algorithm to perform an automatic aspectual classification of English verbs. (1) For the supervised classification, 97,973 parsed sentences from medical discharge summaries were used to extract frequencies for verbs on 14 linguistic indicators, such as manner adverb, duration *in* PP, past tense, and perfect tense. Logistic regression, decision tree induction, and genetic programming were applied to the verb data to distinguish states and events. Comparing the ability of the learning methods to combine the linguistic indicators was claimed to be difficult, as they rank differently depending on the classification task and evaluation criteria. Decision trees achieved an accuracy of 93.9%, as compared to the uninformed baseline of 83.8%. (2) For the unsupervised clustering, 14,038 distinct verb–object pairs of varying frequencies were extracted from 75,289 parsed novel sentences. A random partition of the set of verbs was improved by a hill-climbing method, which improved the partition by moving a verb to the cluster that decreases the sum of distances most. For a small set of 56 verbs whose frequency in the verb–object pairs was larger than 50, Siegel and McKeown (2000) claimed on the basis of an evaluation of 19 verbs that their clustering algorithm discriminated

event verbs from stative verbs. Overall, they performed a comparably simpler task than presented in this article, since the aspectual class criteria can be defined more objectively and more clearly than semantic criteria based on situational similarity. Their choice of features delimited their class criteria well, and they were able to achieve excellent results.

In previous work on English, Schulte im Walde (2000) clustered 153 verbs into 30 verb classes taken from Levin (1993), using unsupervised hierarchical clustering. The verbs were described by distributions over subcategorization frames as extracted from maximum-probability parses using a robust statistical parser, and completed by assigning WordNet classes as selectional preferences to the frame arguments. Using Levin's verb classification as a basis for evaluation, 61% of the verbs were classified correctly into semantic classes. The clustering was most successful when utilizing syntactic subcategorization frames enriched with PP information; selectional preferences decreased the performance of the clustering approach. The detailed encoding and therefore sparse data made the clustering worse with the selectional preference information.

Merlo and Stevenson (2001) presented an automatic classification of three types of English intransitive verbs, based on argument structure and crucially involving thematic relations. They selected 60 verbs with 20 verbs from each verb class, comprising unergatives, unaccusatives, and object-drop verbs. The verbs in each verb class show similarities with respect to their argument structure, in that they all can be used both as transitives and intransitives. Therefore, argument structure alone does not distinguish the classes, and subcategorization information is refined by thematic relations. Merlo and Stevenson defined verb features based on linguistic heuristics that describe the thematic relations between subject and object in transitive and intransitive verb usage. The features included heuristics for transitivity, causativity, animacy, and syntactic features. For example, the degree of animacy of the subject argument roles was estimated as the ratio of occurrences of pronouns to all subjects for each verb, based on the assumption that unaccusatives occur less frequently with an animate subject compared to unergative and object-drop verbs. Each verb was described by a five-feature vector, and the vector descriptions were fed into a decision tree algorithm. Compared with a baseline performance of 33.9%, the decision trees classified the verbs into the three classes with an accuracy of 69.8%. Further experiments demonstrated the contribution of the different features within the classification. Compared to the current article, Merlo and Stevenson (2001) performed a simpler task and classified a smaller number of 60 verbs into only three classes. The features of the verbs were restricted to those that should capture the basic differences between the verb classes, in line with the idea that the feature choice depends on the specific properties of the desired verb classes. But using the same classification methodology for a large-scale experiment with an enlarged number of verbs and classes faces more problems. For example, Joanis (2002) reported an extension of their work that used 802 verbs from 14 classes from Levin (1993). He defined an extensive feature space with 219 core features (such as part of speech, auxiliary frequency, syntactic categories, and animacy as above) and 1,140 selectional preference features taken from WordNet. As in Schulte im Walde (2000), the selectional preferences did not improve the clustering. In recent work, Stevenson and Joanis (2003) compared their supervised method for verb classification with semisupervised and unsupervised techniques. In these experiments, they enlarged the number of gold standard English verb classes to 14 classes related to Levin classes, with a total of 841 verbs. Low-frequency and ambiguous verbs were excluded from the classes. They found that a semisupervised approach where the classifier was trained with five seed verbs from each verb class outperformed both a manual selection of features and the unsupervised

approach of Dash, Liu, and Yao (1997), which used an entropy measure to organize data into a multidimensional space.

The classification methodology from Merlo and Stevenson (2001) was applied to multilinguality by Merlo et al. (2002) and Tsang, Stevenson, and Merlo (2002). Merlo et al. (2002) showed that the classification paradigm is applicable in languages other than English by using the same features as defined by Merlo and Stevenson (2001) for the respective classification of 59 Italian verbs empirically based on the Parole corpus. The resulting accuracy is 86.4%. In addition, they used the content of Chinese verb features to refine the English verb classification, explained in more detail by Tsang, Stevenson, and Merlo (2002). The English verbs were manually translated into Chinese and given part-of-speech tag features, passive particles, causative particles, and sublexical morphemic properties. Verb tags and particles in Chinese are overt expressions of semantic information that is not expressed as clearly in English, and the multilingual set of features outperformed either set of monolingual features, yielding an accuracy of 83.5%.

Pereira, Tishby, and Lee (1993) describe a hierarchical soft clustering method that clusters words according to their distribution in particular syntactic contexts. They used an application of their method to nouns appearing as direct objects of verbs. The clustering result was a hierarchy of noun clusters, where every noun belongs to every cluster with a membership probability. The initial data for the clustering process were frequencies of verb–noun pairs in a direct object relationship, as extracted from parsed sentences from the Associated Press news wire corpus. On the basis of the conditional verb–noun probabilities, the similarity of the distributions was determined by the Kullback–Leibler divergence. The EM algorithm (Baum 1972) was used to learn the hidden cluster membership probabilities, and deterministic annealing performed the divisive hierarchical clustering. The resulting class-based model can be utilized for estimating information for unseen events (cf. Dagan, Lee, and Pereira 1999).

Rooth et al. (1999) produced soft semantic clusters for English that represent a classification on verbs as well as on nouns. They gathered distributional data for verb–noun pairs in specific grammatical relations from the British National Corpus. The extraction was based on a lexicalized probabilistic context-free grammar (Carroll and Rooth 1998) and contained the subject and object nouns for all intransitive and transitive verbs in the parses—a total of 608,850 verb–noun types. Conditioning of the verbs and the nouns on each other was done through hidden classes, and the joint probabilities of classes, verbs, and nouns were trained by the EM algorithm. The resulting model defined conditional membership probabilities for each verb and noun in each class; for example, the class of communicative action contains the most probable verbs *ask, nod, think, shape, smile* and the most probable nouns *man, Ruth, Corbett, doctor, woman*. The semantic classes were utilized for the induction of a semantically annotated verb lexicon.

6. Conclusion and Outlook

This article presented a clustering methodology for German verbs whose results agreed with a manual classification in many respects and should prove useful as automatic basis for a large-scale clustering. Without a doubt the cluster analysis needs manual correction and completion, but represents a plausible foundation. Key issues of the clustering methodology concern linguistic criteria on the one hand, and technical criteria on the other hand.

Linguistic Criteria: The strategy of utilizing subcategorization frames, prepositional information, and selectional preferences to define the verb features seems promising,

since the experiments illustrated a relation between the induced verb behavior and the membership of the semantic verb classes. In addition, each level of representation generated a positive effect on the clustering and improved upon the less informative level. The experiments presented evidence for a linguistic limit on the usefulness of the verb features: The meaning of verbs comprises both (1) properties that are general for the respective verb classes and (2) idiosyncratic properties that distinguish the verbs from each other. As long as we define the verbs by those properties that represent the common parts of the verb classes, a clustering can succeed. But by stepwise refining the verb description and including lexical idiosyncrasy, emphasis on 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, feature choice then depends on the definition of the verb classes, and this definition might vary according to the conceptual level and also according to the kind of semantic coherence captured by the class. So far, we have concentrated on synonymy, but the large-scale experiment, in particular, discovered additional semantic relations within a verb class, such as participation in a process/script. However, the investigated feature combination within this article seems to be a useful starting point for verb description.

Technical Criteria: We investigated the relationship between clustering idea, clustering parameters, and clustering result in order to develop a clustering methodology that is suitable for the demands of natural language. The clustering initialization played an important role: *k*-means needed compact, similarly-sized clusters in order to achieve a linguistically meaningful classification. The linguistically most successful initial clusters were therefore based on hierarchical clustering with complete linkage or Ward's method, as the resulting clusters are comparable in size and correspond to compact cluster shapes. The hierarchical clustering achieved more similar clustering outputs than *k*-means, which is due to the similarity of the clustering methods with respect to the common clustering criterion of optimizing the sum of distances between verbs and cluster centroids. The similarity measure used in the clustering experiments proved to be of secondary importance, since the differences in clustering due to varying the measure were negligible. For larger object and feature sets, Kullback–Leibler variants tended to outperform other measures, confirming language-based results on distributional similarity (Lee 2001). Both frequencies and probabilities represented a useful basis for the verb distributions. The number of clusters played a role concerning the magnitude of numbers: Inducing fine-grained clusters as given in the manual classification proved to be an ambitious goal because the feature distinction for the classes was also fine-grained. Inducing coarse clusters provided a coarse classification that was subject to less noise and easier to manually correct. The "optimal" number of clusters is always a compromise and depends on the purpose of the classes, for example, as a fine-grained lexical resource, or for an NLP application. In the latter case, the optimal number should be determined by automatic means, that is, by trying different magnitudes of cluster numbers, because the level of generalization depends on the purpose for the abstraction.

There are various directions for future research. (1) The manual definition of the German semantic verb classes will be extended in order to include a greater number and a larger variety of verb classes. An extended classification would be useful as a gold standard for further clustering experiments, and more generally as a resource for NLP applications. (2) Low-frequency verbs require a specific handling in the clustering procedure: Both the small-scale and the large-scale experiments showed that the low-frequency verbs have a negative impact on the cluster coherence. An alternative model for the low-frequency verbs might, for example, first take out of the cluster analysis

those verbs below a certain frequency cutoff, and then assign the left-out verbs to the nearest clusters. The cluster assignment should also be special, for example, using verb features restricted to the reliable features, that is, above a certain frequency threshold. For example, if we consider the *D*2 features of the low-frequency verb *ekeln* 'disgust' (frequency: 31) with a minimum feature frequency of 2, we get a strong overlap with the distinguishing features of the verb *fürchten* 'fear'. Future work will address these issues. (3) Possible features for describing German verbs will include any kind of information that helps to classify the verbs in a semantically appropriate way. Within this article, we concentrated on defining the verb features with respect to alternation behavior. Other features that are relevant for describing the behavior of verbs are their auxiliary selection and adverbial combinations. In addition, if we try to address additional types of semantic verb relations such as script-based relations, we will need to extend our features. For example, Schulte im Walde and Melinger (2005) recently showed that nouns in co-occurrence windows of verbs contribute to verb descriptions by encoding scene information, rather than intrasentential functions. They proposed the integration of window-based approaches into function-based approaches, a combination that has not yet been applied. (4) Variations in 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. For example, the verbs *töten* and *unterrichten* require a finer version of selectional preferences in order to be distinguished. It is worthwhile either to find a more appropriate level of selectional preferences in WordNet or to apply a more sophisticated approach towards selectional preferences such as that of Li and Abe (1998), in order to determine a more flexible choice of selectional preferences. (5) With respect to a large-scale classification of verbs, it will be interesting to apply classification techniques to the verb data. This would require more data manually labeled with classes in order to train a classifier. But the resulting classifier might abstract better than *k*-means over the different requirements of the verb classes with respect to the feature description. (6) As an extension of the existing clustering, a soft clustering algorithm will be applied to the German verbs. Soft clustering enables us to assign verbs to multiple clusters and therefore address the phenomenon of verb ambiguity. These clustering outcomes should be even more useful for discovering new verb meaning components and semantically related classes, compared with the hard clustering technique. (7) The verb clusters as resulting from the cluster analysis will be used within an NLP application in order to prove the usefulness of the clusters. For example, replacing verbs in a language model by the respective verb classes might improve the language model's robustness and accuracy, as the class information provides more stable syntactic and semantic information than the individual verbs.

Appendix A: Subcategorization Frame Types

The syntactic part of the German verb behavior is captured by 38 subcategorization frame types. The frames comprise maximally three arguments. Possible arguments 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 A.1, accompanied by annotated verb–second example clauses. The German examples are provided with English glosses; in cases where the glosses are difficult to understand, an English translation is added.

Table A1
Subcategorization frame types.

Frame Type	Example
`n`	Natalie$_n$ schwimmt. Natalie swims
`na`	Hans$_n$ sieht seine Freundin$_a$. Hans sees his girlfriend
`nd`	Er$_n$ glaubt den Leuten$_d$ nicht. He believes the people not
`np`	Die Autofahrer$_n$ achten besonders auf Kinder$_p$. The drivers watch out especially for children
`nad`	Anna$_n$ verspricht ihrem Vater$_d$ ein tolles Geschenk$_a$. Anna promises her father a great present
`nap`	Die Verkäuferin$_n$ hindert den Dieb$_a$ am Stehlen$_p$. The saleslady keeps the thief from stealing
`ndp`	Der Moderator$_n$ dankt dem Publikum$_d$ für sein Verständnis$_p$. The moderator thanks the audience for their understanding
`ni`	Mein Freund$_n$ versucht immer, pünktlich zu kommen$_i$. My friend tries always in time to arrive 'My friend always tries to arrive in time.'
`nai`	Er$_n$ hört seine Mutter$_a$ ein Lied trällern$_i$. He hears his mother a song sing 'He hears his mother singing a song.'
`ndi`	Helene$_n$ verspricht ihrem Großvater$_d$ ihn bald zu besuchen$_i$. Helene promises her grandfather him soon to visit
`nr`	Die Kinder$_n$ fürchten sich$_r$. The children are afraid themselves
`nar`	Der Unternehmer$_n$ erhofft sich$_r$ schnellen Fortschritt$_a$. The businessman hopes himself quick progress
`ndr`	Sie$_n$ schließt sich$_r$ nach 10 Jahren wieder der Kirche$_d$ an. She associates herself after 10 years again the church with
`npr`	Der Pastor$_n$ hat sich$_r$ als der Kirche würdig$_p$ erwiesen. The pastor has himself to the church worthy proven
`nir`	Die alte Frau$_n$ stellt sich$_r$ vor, den Preis zu gewinnen$_i$. The old women imagines herself the price to win
`x`	Es$_x$ blitzt. It lightenings
`xa`	Es$_x$ gibt viele Bücher$_a$. There exist many books
`xd`	Es$_x$ graut mir$_d$. It terrifies me

Table A1
(cont.)

Frame Type	Example
xp	Es$_x$ geht um einen tollen Preis für mein Sofa$_p$. It is about a great price for my sofa
xr	Es$_x$ rechnet sich$_r$. It calculates itself 'It is worth it.'
xs-dass	Es$_x$ heißt, dass Thomas sehr schlau ist$_{s-dass}$. It says, that Thomas very intelligent is
ns-2	Der Professor$_n$ hat gesagt, er halte bald einen Vortrag$_{s-2}$. The professor has said, he gives soon a talk
nas-2	Der Chef$_n$ schnauzt ihn$_a$ an, er sei ein Idiot$_{s-2}$. The chef bawls him out, he is an idiot
nds-2	Er$_n$ sagt seiner Freundin$_d$, sie sei zu krank zum Arbeiten$_{s-2}$. He tells his girlfriend, she is too ill to work
nrs-2	Der Kleine$_n$ wünscht sich$_r$, das Mädchen bliebe bei ihm$_{s-2}$. The boy wishes himself, the girl stays with him
ns-dass	Der Winter$_n$ hat schon angekündigt, dass er bald kommt$_{s-dass}$. Winter has already announced, that it soon arrives
nas-dass	Der Vater$_n$ fordert seine Tochter$_a$ auf, dass sie verreist$_{s-dass}$. The father requests his daughter that she travels
nds-dass	Er$_n$ sagt seiner Geliebten$_d$, dass er verheiratet ist$_{s-dass}$ He tells his lover, that he married is
nrs-dass	Der Kleine$_n$ wünscht sich$_r$, dass seine Mutter bleibt$_{s-dass}$. The boy wishes himself, that his mother stays
ns-ob	Der Professor$_n$ hat gefragt, ob die neue Forscherin interessant sei$_{s-ob}$ The professor has asked, whether the new researcher interesting is
nas-ob	Anton$_n$ fragt seine Frau$_a$, ob sie ihn liebt$_{s-ob}$. Anton asks his wife, whether she him loves
nds-ob	Der Nachbar$_n$ ruft der Frau$_d$ zu, ob sie verreist$_{s-ob}$ The neighbor shouts the woman whether she travels
nrs-ob	Der Alte$_n$ wird sich$_r$ erinnern, ob das Mädchen dort war$_{s-ob}$. The old man will himself remember, whether the girl there was
ns-w	Der Kleine$_n$ hat gefragt, wann die Tante endlich ankommt$_{s-w}$ The boy has asked, when the aunt finally arrives
nas-w	Der Mann$_n$ fragt seine Freundin$_a$, warum sie ihn liebt$_{s-w}$. The man asks his girlfriend, why she him loves
nds-w	Der Vater$_n$ verrät seiner Tochter$_d$ nicht, wer zu Besuch kommt$_{s-w}$. The father tells his daughter not, who for a visit comes
nrs-w	Das Mädchen$_n$ erinnert sich$_r$, wer zu Besuch kommt$_{s-w}$. The girl remembers herself, who for a visit comes
k	Der neue Nachbar$_k$ ist ein ziemlicher Idiot. The new neighbor is a complete idiot

Acknowledgments

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. Many thanks to Helmut Schmid, Stefan Evert, Frank Keller, Scott McDonald, Alissa Melinger, Chris Brew, Hinrich Schütze, Jonas Kuhn, and the two anonymous reviewers for their valuable comments on previous versions of this article.

References

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. 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*, pages 86–90, Montreal, Canada.

Baum, Leonard E. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, III:1–8.

Carroll, Glenn and Mats Rooth. 1998. Valence induction with a head-lexicalized PCFG. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, Granada, Spain.

Charniak, Eugene. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference on Artificial Intelligence*, Menlo Park, CA.

Chen, Stanley and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Center for Research in Computing Technology, Harvard University.

Cover, Thomas M. and Joy A. Thomas. 1991. *Elements of Information Theory*. Telecommunications. John Wiley & Sons, New York.

Dagan, Ido, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69. Special Issue on Natural Language Learning.

Dash, Manoranjan, Hua Liu, and Jun Yao. 1997. Dimensionality reduction for unsupervised data. In *Proceedings of the 9th International Conference on Tools with Artificial Intelligence*, pages 532–539, Newport Beach, CA.

Dorr, Bonnie J. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–322.

Dorr, Bonnie J. and Doug Jones. 1996. 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.

Erk, Katrin, Andrea Kowalski, and Manfred Pinkal. 2003. A corpus resource for lexical semantics. In *Proceedings of the 5th International Workshop on Computational Semantics*, Tilburg, The Netherlands.

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

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

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

Fontenelle, Thierry, editor. 2003. *FrameNet and Frame Semantics*, volume 16(3) of *International Journal of Lexicography*. Oxford University Press.

Forgy, Edward W. 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics*, 21:768–780.

Hamp, Birgit and Helmut Feldweg. 1997. 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.

Harris, Zellig. 1968. Distributional structure. In Jerold J. Katz, editor, *The Philosophy of Linguistics*, Oxford Readings in Philosophy. Oxford University Press, pages 26–47.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1993. Towards the automatic identificaton of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 172–182, Columbus.

Hubert, Lawrence and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218.

Joanis, Eric. 2002. Automatic verb classification using a general feature space. Master's thesis, Department of Computer Science, University of Toronto.

Kaufman, Leonard and Peter J. Rousseeuw. 1990. *Finding Groups in Data—An Introduction to Cluster Analysis*. Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Klavans, Judith L. and Min-Yen Kan. 1998. The role of verbs in document analysis. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 680–686, Montreal, Canada.

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

Korhonen, Anna, Yuval Krymolowski, and Zvika Marx. 2003. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 64–71, Sapporo, Japan.

Kunze, Claudia. 2000. 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.

Lapata, Maria. 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*, pages 397–404, College Park, MD.

Lapata, Mirella and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–73.

Lee, Lillian. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics*, pages 65–72.

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

Li, Hang and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

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

Merlo, Paola and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Merlo, Paola, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. 2002. A multilingual paradigm for automatic verb classification. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 207–214, Philadelphia, PA.

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

Palmer, Martha, Dan Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, OH.

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

Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, MD.

Saint-Dizier, Patrick. 1998. 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.

Schmid, Helmut. 2000. LoPar: Design and implementation. Arbeitspapiere des Sonderforschungsbereichs 340 *Linguistic Theory and the Foundations of Computational Linguistics* 149, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schulte im Walde, Sabine. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 747–753, Saarbrücken, Germany.

Schulte im Walde, Sabine. 2002a. Evaluating verb subcategorisation frames learned by a German statistical grammar against manual definitions in the *Duden* Dictionary. In *Proceedings*

of the 10th EURALEX International Congress, pages 187–197, Copenhagen, Denmark.
Schulte im Walde, Sabine. 2002b. 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.
Schulte im Walde, Sabine. 2003a. *Experiments on the automatic induction of German semantic verb classes*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. Published as AIMS Report 9(2).
Schulte im Walde, Sabine. 2003b. Experiments on the choice of features for learning verb classes. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 315–322, Budapest, Hungary.
Schulte im Walde, Sabine and Chris Brew. 2002. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 223–230, Philadelphia, PA.
Schulte im Walde, Sabine and Alissa Melinger. 2005. Identifying semantic relations and functional properties of human verb associations. In *Proceedings of the joint Conference on Human Language Technology and Empirial Methods in Natural Language Processing*, pages 612–619, Vancouver, Canada.
Schumacher, Helmut. 1986. *Verben in Feldern*. de Gruyter, Berlin.
Siegel, Eric V. and Kathleen R. McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and Revealing Linguistic Insights. *Computational Linguistics*, 26(4):595–628.
Stevenson, Suzanne and Eric Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 71–78, Edmonton, Canada.
Tsang, Vivian, Suzanne Stevenson, and Paola Merlo. 2002. Crosslinguistic transfer in automatic verb classification. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1023–1029, Taipei, Taiwan.
Vázquez, Gloria, Ana Fernández, Irene Castellón, and María Antonia Martí. 2000. *Clasificación verbal: Alternancias de diátesis*. Number 3 in *Quaderns de Sintagma*. Universitat de Lleida.
Vossen, Piek. 2004. EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *International Journal of Lexicography*, 17(2):161–173.