Chapter 5

Clustering Experiments

In the preceding chapters, I have introduced the concept of a German semantic verb classification, a statistical grammar model as a source for verb description, and algorithms and evaluation methods for clustering experiments. This chapter brings together the concept, the data and the techniques, and presents clustering experiments which investigate the automatic induction of semantic classes for German verbs. It is clear from the choice of verbs and verb classes, the available data for feature description and the restricted potential of the clustering algorithm, that the clustering results will not satisfy the semantic definition of the verb classes. But the goal is not to provide the perfect result, but to gain as much insight as possible into the aspects of verb clustering in order to utilise the knowledge in related NLP tasks. Parts of the experiments have already been published by Schulte im Walde and Brew (2002) and Schulte im Walde (2003b).

The first section of the chapter (Section 5.1) introduces the German verbs and the gold standard verb classes from an empirical point of view, and illustrates the verb data and feature choice for the experiments. Section 5.2 describes the clustering setup, process and results, followed by an interpretation of the experiments in Section 5.3. Section 5.4 discusses possibilities to optimise the experiment setup and performance, and Section 5.6 cites and discusses related work to the clustering experiments.

5.1 Clustering Data

Chapter 4 has presented the German verbs as clustering objects, and verb descriptions at the syntax-semantic interface as the object features. This section introduces the clustering objects and the choice of features in more detail (Sections 5.1.1 and 5.1.2), which is relevant for the clustering experiments. Section 5.1.3 illustrates the verbs and their features by various means, to provide the reader with an intuition on the clustering data.

5.1.1 German Verbs and Verb Classes

The hand-constructed German verb classes have been discussed in Chapter 2. The manual classes represent the gold standard classification which on the one hand provides the objects for the clustering experiments and on the other hand defines the basis for evaluating the clustering results. The clustering experiments as described in this chapter first refer to a reduced subset of classes from the existing classification, and later on refer to the entire set. Why experiments on a restricted set of verbs? Main reasons for preliminary experiments on a restricted domain of verbs and verb classes are (i) it is easier to obtain introspective judgements on the value and the interpretation of the automatic verb clusterings, (ii) the dividing line between the classes is more clear-cut, and (iii) it is possible to perform unambiguous evaluations of the clustering results, since I eliminated the ambiguity from the classification. The reduced set of verb classes is listed below. Table 5.1 refers to empirical properties of the full and the reduced set of verb classes.

1. *Aspect*: anfangen, aufhören, beenden, beginnen, enden
2. *Propositional Attitude*: ahnen, denken, glauben, vermuten, wissen
3. *Transfer of Possession (Obtaining)*: bekommen, erhalten, erlangen, kriegen
4. *Transfer of Possession (Supply)*: bringen, liefern, schicken, vermitteln, zustellen
5. *Manner of Motion*: fahren, fliegen, rudern, segeln
6. *Emotion*: ärgern, freuen
7. *Announcement*: ankündigen, bekanntgeben, eröffnen, verkünden
8. *Description*: beschreiben, charakterisieren, darstellen, interpretieren
9. *Insistence*: beharren, bestehen, insistieren, pochen
10. *Position*: liegen, sitzen, stehen
11. *Support*: dienen, folgen, helfen, unterstützen
12. *Opening*: öffnen, schließen
13. *Consumption*: essen, konsumieren, lesen, saufen, trinken
14. *Weather*: blitzen, donnern, dämmern, nieseln, regnen, schneien

	Full Set	Reduced Set
Verbs	168	57
Classes	43	14
Class sizes	2-7	2-6
Average number of verbs per class	3.91	4.07
Verb frequencies (min/max)	8 – 71,604	8 – 31,710
Ambiguous verbs	8	0

Table 5.1: Empirical properties of gold standard verb classes

5.1.2 Feature Choice

One of the most difficult parts of a cluster analysis is the choice of appropriate features to describe the clustering objects. Why is this so difficult?

- The chosen features are supposed to represent a relevant subset of possible features. But what does ‘relevant’ refer to? In our task, does it mean (a) relevant for describing the specific verbs in the manual classification, (b) relevant for a general description of German verbs, or (c) relevant for an optimal clustering result?

- The outcome of a clustering does not necessarily align with expectations as based on the linguistic intuition for the choice and variation of the features. Even if we knew about an optimal feature set to describe the clustering objects, this feature set does not necessarily result in the optimal clustering, and vice versa.

- If the choice of features is optimised with regard to an optimal clustering outcome, we risk to overfit the data for the cluster analysis, i.e. applying the same feature set and the same clustering methodology to a different set of verbs does not necessarily result in the desired optimal clustering.

- Intuitively, one might want to add and refine features ad infinitum, but in practise it is necessary to tune the features to the capability of the clustering algorithm, which must be able to (i) process the features (restrictions on time and space), and (ii) generalise about the features. In addition, there might be a theoretically defined limit on the usefulness of features.

The above discussion demonstrates that when defining an appropriate feature choice for the German verbs, we need to find a compromise between a linguistically plausible verb description and an algorithmically applicable feature set. My strategy is as follows: Since I am interested in a linguistic concern, I specify the verb description in a linguistically appropriate way. Only when it comes to modelling the features in a distribution appropriate for the clustering algorithm, I compromise for practical problems, such as a large number of features causing a sparse data problem. As shown by Schulte im Walde (2000a), a sparse feature vector description destroys valuable clustering results.

This section describes the feature choice as it is used in the clustering experiments. Variations of verb attributes might confuse at this stage of the thesis and will be discussed separately in Section 5.4, which optimises the setup of the clustering experiments and shows that the applied strategy is near-optimal.

In the following, I specify (A) the basic feature description of the German verbs, and then a range of manipulations on the feature distributions: (B) a strengthened version of the original feature values, (C) a variation of the feature values by applying a simple smoothing technique, and (D) artificially introducing noise into the feature values.

A) Basic Feature Description

As said before, the German verbs are described on three levels at the syntax-semantic interface, each of them refining the previous level by additional information. The induction of the features and the feature values is based on the grammar-based empirical lexical acquisition as described in Chapter 3. The first level encodes a purely syntactic definition of verb subcategorisation, the second level encodes a syntactico-semantic definition of subcategorisation with prepositional preferences, and the third level encodes a syntactico-semantic definition of subcategorisation with prepositional and selectional preferences. So the refinement of verb features starts with a purely syntactic definition and step-wise adds semantic information. The most elaborated description comes close to a definition of the verb alternation behaviour. I have decided on this three step proceeding of verb descriptions, because the resulting clusters and even more the changes in clustering results which come with a change of features should provide insight into the meaning-behaviour relationship at the syntax-semantic interface. Further possibilities to extend the verb descriptions by information which helps classify the verbs in a semantically appropriate way (e.g. morphological properties, auxiliary selection, adverbial combinations, etc.) are not realised within the current clustering experiments, but could be added.

Coarse Syntactic Definition of Subcategorisation Chapter 3 has described the definition of subcategorisation frames in the German grammar. The statistical grammar model provides frequency distributions of German verbs over 38 purely syntactic subcategorisation frames. On basis of the frequency distributions, we can define probability distributions, and binary distributions by setting a cut-off for the relevance of a frame type. The cut-off is set to 1%. Table 5.2 presents an example of the distributions for the verb *glauben* ‘to think, to believe’. The reader is reminded of the frame type definitions in Appendix A. The values in the table are ordered by frequency.

Syntactico-Semantic Definition of Subcategorisation with Prepositional Preferences The German grammar also provides information about the specific usage of prepositional phrases with respect to a certain subcategorisation frame type containing a PP (abbreviation: p). On basis of the PP information, I create an extended verb distribution that discriminates between different kinds of PP-arguments. The frequencies can be read from the grammar parameters; the probabilities are created by distributing the joint probability of a verb and the PP frame (np, nap, ndp, npr, xp) over the prepositional phrases, according to their frequencies in the corpus; the binary values are based on a cut-off of 1%, as before.

Prepositional phrases are referred to by case and preposition, such as ‘Dat.mit’, ‘Akk.für’. As mentioned before, the statistical grammar model does not perfectly learn the distinction between PP-arguments and PP-adjuncts. Therefore, I have not restricted the PP features to PP-arguments, but to 30 PPs according to ‘reasonable’ appearance in the corpus. A reasonable appearance is thereby defined by the 30 most frequent PPs which appear with at least 10 different verbs.

- `Akk`: an, auf, bis, durch, für, gegen, in, ohne, um, unter, vgl, über
- `Dat`: ab, an, auf, aus, bei, in, mit, nach, seit, unter, von, vor, zu, zwischen, über
- `Gen`: wegen, während
- `Nom`: vgl

Table 5.3 presents example distributions for the verb *reden* 'to talk' and the frame type `np`, with the joint verb-frame numbers in the first line. The frame combinations are ordered by frequency.

When utilising the refined distributions as feature descriptions for verbs, (a) the coarse frame description can either be substituted by the refined information, or (b) the refined information can be given in addition to the coarse definition. With respect to (a), the substitution guarantees in case of probabilities that the distribution values still sum to 1, which is desirable for various similarity measures, while (b) is able to provide frame information on various levels at the same time. For the clustering experiments, I will apply both versions.

Syntactico-Semantic Definition of Subcategorisation with Prepositional and Selectional Preferences A linguistically intuitive extension of the former subcategorisation distributions is a frame refinement by selectional preferences, i.e. the slots within a subcategorisation frame type are specified according to which 'kind' of argument they require. The grammar provides selectional preference information on a fine-grained level: it specifies the possible argument realisations in form of lexical heads, with reference to a specific verb-frame-slot combination. Table 5.4 lists nominal argument heads for the verb *verfolgen* 'to follow' in the accusative NP slot of the transitive frame type `na` (the relevant frame slot is underlined), and Table 5.5 lists nominal argument heads for the verb *reden* 'to talk' in the PP slot of the transitive frame type `np:Akk.über`. The examples are ordered by the noun frequencies. For presentation reasons, I set a frequency cut-off. The tables have been presented before as Tables 3.18 and 3.19, respectively.

Obviously, we would run into a sparse data problem if we tried to incorporate selectional preferences into the verb descriptions on such a specific level. We are provided with rich information on the nominal level, but we need a generalisation of the selectional preference definition. A widely used resource for selectional preference information is the semantic ontology *WordNet* (Miller *et al.*, 1990; Fellbaum, 1998). Within the framework of *EuroWordNet* (Vossen, 1999), the University of Tübingen develops the German version of WordNet, *GermaNet* (Hamp and Feldweg, 1997; Kunze, 2000).

I utilise the German noun hierarchy in GermaNet for the generalisation of selectional preferences. The hierarchy is realised by means of synsets, sets of synonymous nouns, which are organised by multiple inheritance hyponym/hypernym relationships. A noun can appear in several synsets, according to its number of senses. Figure 5.1 illustrates the (slightly simplified) GermaNet hierarchy for the noun *Kaffee* 'coffee', which is encoded with two senses, (1) as a beverage and luxury food, and (2) as expression for an afternoon meal. Both senses are subsumed under the general top level node *Objekt* 'object'.

Frame	Freq	Prob	Bin
ns-dass	1,928.52	0.279	1
ns-2	1,887.97	0.274	1
np	686.76	0.100	1
n	608.05	0.088	1
na	555.23	0.080	1
ni	346.10	0.050	1
nd	234.09	0.034	1
nad	160.45	0.023	1
nds-2	69.76	0.010	1
nai	61.67	0.009	0
ns-w	59.31	0.009	0
nas-w	46.99	0.007	0
nap	40.99	0.006	0
nr	31.37	0.005	0
nar	30.10	0.004	0
nrs-2	26.99	0.004	0
ndp	24.56	0.004	0
nas-dass	23.58	0.003	0
nas-2	19.41	0.003	0
npr	18.00	0.003	0
nds-dass	17.45	0.003	0
ndi	11.08	0.002	0
ndr	2.00	0.000	0
nrs-dass	2.00	0.000	0
nrs-w	2.00	0.000	0
nir	1.84	0.000	0
nds-w	1.68	0.000	0
xd	1.14	0.000	0
nas-ob	1.00	0.000	0
ns-ob	1.00	0.000	0
x	0.00	0.000	0
xa	0.00	0.000	0
xp	0.00	0.000	0
xr	0.00	0.000	0
xs-dass	0.00	0.000	0
nds-ob	0.00	0.000	0
nrs-ob	0.00	0.000	0
k	0.00	0.000	0

Table 5.2: Frame distributions for *glauben*

Frame	Freq	Prob	Bin
np	1,427.24	0.455	1
np:Akk.über	479.97	0.153	1
np:Dat.von	463.42	0.148	1
np:Dat.mit	279.76	0.089	1
np:Dat.in	81.35	0.026	1
np:Nom.vgl	13.59	0.004	0
np:Dat.bei	13.10	0.004	0
np:Dat.über	13.05	0.004	0
np:Dat.an	12.06	0.004	0
np:Akk.für	9.63	0.003	0
np:Dat.nach	8.49	0.003	0
np:Dat.zu	7.20	0.002	0
np:Dat.vor	6.75	0.002	0
np:Akk.in	5.86	0.002	0
np:Dat.aus	4.78	0.002	0
np:Dat.auf	4.34	0.001	0
np:Dat.unter	3.77	0.001	0
np:Akk.vgl	3.55	0.001	0
np:Akk.ohne	3.05	0.001	0
np:Dat.seit	2.21	0.001	0
np:Akk.gegen	2.13	0.001	0
np:Akk.an	1.98	0.001	0
np:Gen.wegen	1.77	0.001	0
np:Akk.um	1.66	0.001	0
np:Akk.bis	1.15	0.000	0
np:Gen.während	0.95	0.000	0
np:Dat.zwischen	0.92	0.000	0
np:Akk.durch	0.75	0.000	0
np:Akk.auf	0.00	0.000	0
np:Akk.unter	0.00	0.000	0
np:Dat.ab	0.00	0.000	0

Table 5.3: Frame+PP distributions for *reden* and frame type `np`

Noun		Freq
Ziel	‘goal’	86.30
Strategie	‘strategy’	27.27
Politik	‘policy’	25.30
Interesse	‘interest’	21.50
Konzept	‘concept’	16.84
Entwicklung	‘development’	15.70
Kurs	‘direction’	13.96
Spiel	‘game’	12.26
Plan	‘plan’	10.99
Spur	‘trace’	10.91
Programm	‘program’	8.96
Weg	‘way’	8.70
Projekt	‘project’	8.61
Prozeß	‘process’	7.60
Zweck	‘purpose’	7.01
Tat	‘action’	6.64
Täter	‘suspect’	6.09
Setzung	‘settlement’	6.03
Linie	‘line’	6.00
Spektakel	‘spectacle’	6.00
Fall	‘case’	5.74
Prinzip	‘principle’	5.27
Ansatz	‘approach’	5.00
Verhandlung	‘negotiation’	4.98
Thema	‘topic’	4.97
Kampf	‘combat’	4.85
Absicht	‘purpose’	4.84
Debatte	‘debate’	4.47
Karriere	‘career’	4.00
Diskussion	‘discussion’	3.95
Zeug	‘stuff’	3.89
Gruppe	‘group’	3.68
Sieg	‘victory’	3.00
Räuber	‘robber’	3.00
Ankunft	‘arrival’	3.00
Sache	‘thing’	2.99
Bericht	‘report’	2.98
Idee	‘idea’	2.96
Traum	‘dream’	2.84
Streit	‘argument’	2.72

Table 5.4: Nominal arguments for *verfolgen* in na

Noun		Freq
Geld	‘money’	19.27
Politik	‘politics’	13.53
Problem	‘problem’	13.32
Thema	‘topic’	9.57
Inhalt	‘content’	8.74
Koalition	‘coalition’	5.82
Ding	‘thing’	5.37
Freiheit	‘freedom’	5.32
Kunst	‘art’	4.96
Film	‘movie’	4.79
Möglichkeit	‘possibility’	4.66
Tod	‘death’	3.98
Perspektive	‘perspective’	3.95
Konsequenz	‘consequence’	3.90
Sache	‘thing’	3.73
Detail	‘detail’	3.65
Umfang	‘extent’	3.00
Angst	‘fear’	3.00
Gefühl	‘feeling’	2.99
Besetzung	‘occupation’	2.99
Ball	‘ball’	2.96
Sex	‘sex’	2.02
Sekte	‘sect’	2.00
Islam	‘Islam’	2.00
Fehler	‘mistake’	2.00
Erlebnis	‘experience’	2.00
Abteilung	‘department’	2.00
Demokratie	‘democracy’	1.98
Verwaltung	‘administration’	1.97
Beziehung	‘relationship’	1.97
Angelegenheit	‘issue’	1.97
Gewalt	‘force’	1.89
Erhöhung	‘increase’	1.82
Zölle	‘customs’	1.00
Vorsitz	‘chair’	1.00
Virus	‘virus’	1.00
Ted	‘Ted’	1.00
Sitte	‘custom’	1.00
Ressource	‘resource’	1.00
Notwendigkeit	‘necessity’	1.00

Table 5.5: Nominal arguments for *reden über*$_{Akk}$ ‘to talk about’

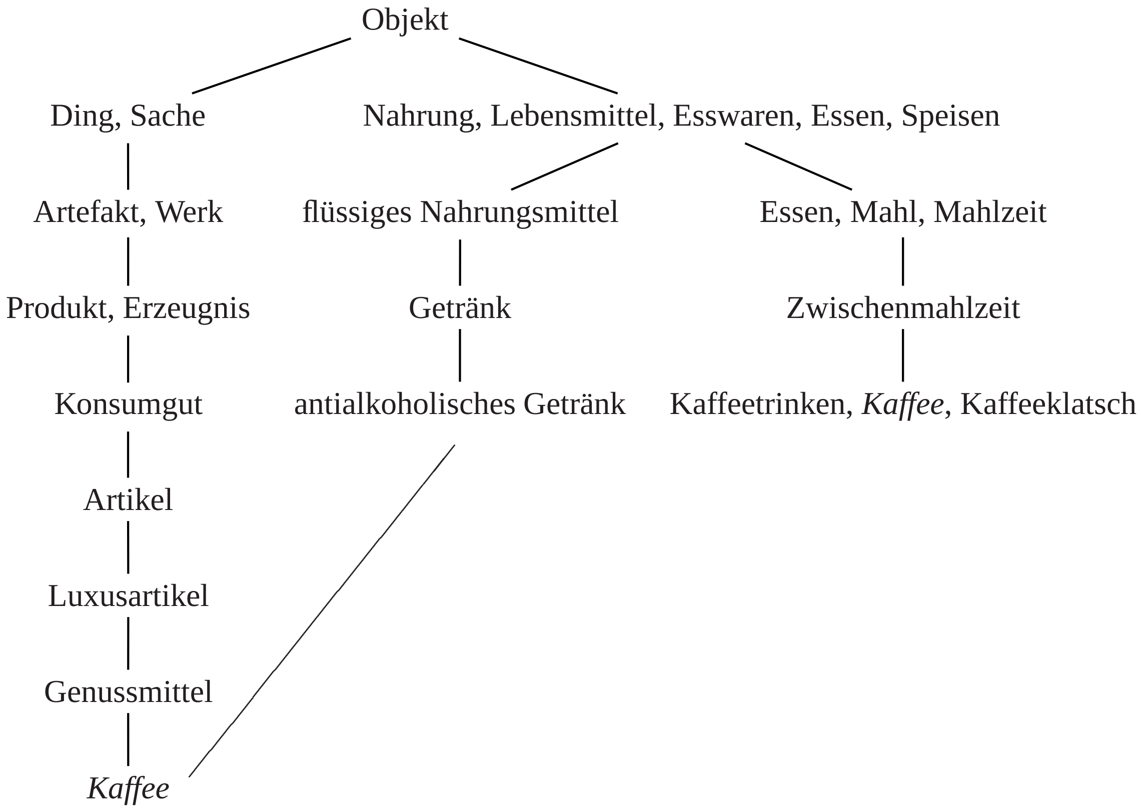

Figure 5.1: GermaNet hierarchy for noun *Kaffee* ‘coffee’

For each noun in a verb-frame-slot combination, the joint frequency is split over the different senses of the noun and propagated upwards the hierarchy. In case of multiple hypernym synsets, the frequency is split again. The sum of frequencies over all top synsets equals the total joint frequency. For example, we assume that the frequency of the noun *Kaffee* ‘coffee’ with respect to the verb *trinken* ‘to drink’ and the accusative slot in the transitive frame na is 10. Each of the two synsets containing *Kaffee* is therefore assigned a value of 5, and the node values are propagated upwards, as Figure 5.2 illustrates.

Repeating the frequency assignment and propagation for all nouns appearing in a verb-frame-slot combination, the result defines a frequency distribution of the verb-frame-slot combination over all GermaNet synsets. For example, Table 5.6 lists the most frequent synsets (presentation cut-off: 7) for the direct object of *essen* ‘to eat’. As expected, the more general synsets appear at the top of the list, since they subsume the frequencies of all subordinated synsets in the hierarchy. In addition, the algorithm tends to find appropriate synsets according to the specific frame-noun combination, such as *Fleisch* ‘meat’, *Backware* ‘pastry’ in the example.

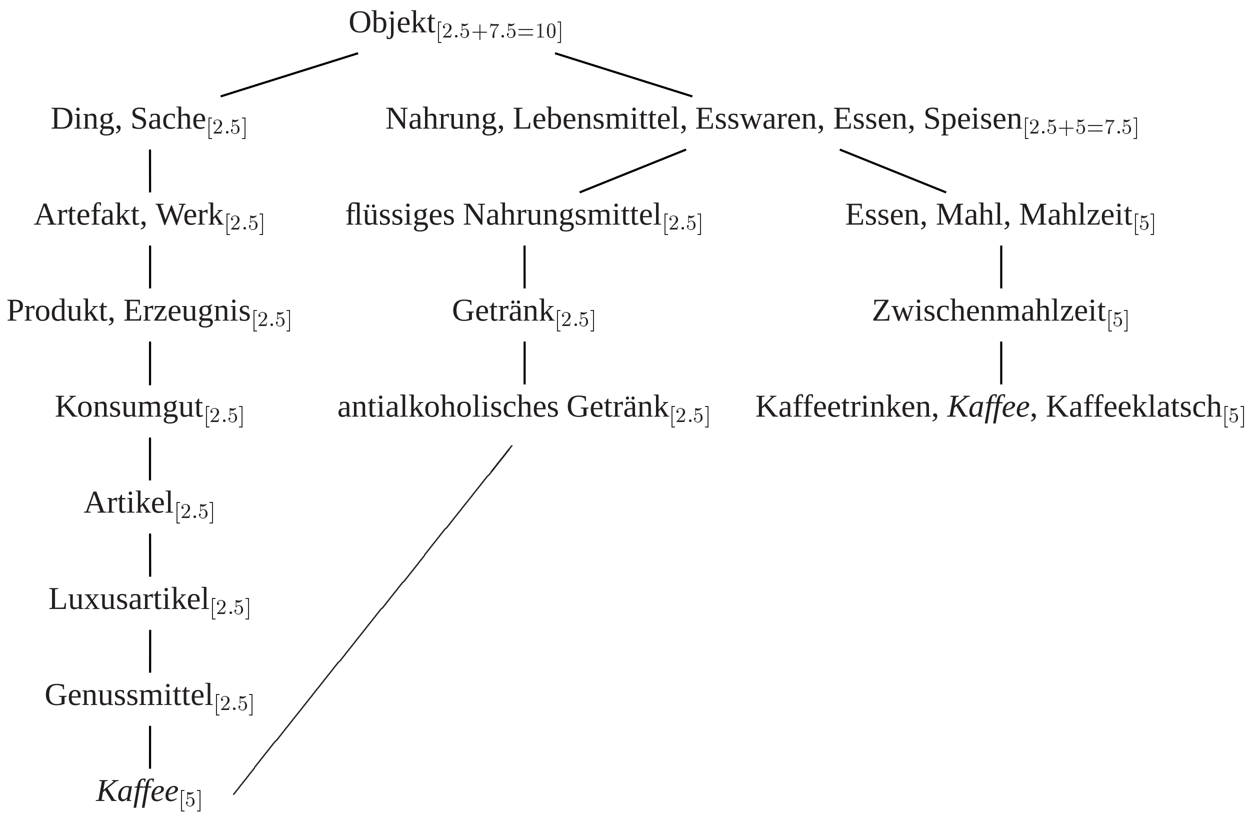
Objekt$_{[2.5+7.5=10]}$
Ding, Sache$_{[2.5]}$
Nahrung, Lebensmittel, Esswaren, Essen, Speisen$_{[2.5+5=7.5]}$
Artefakt, Werk$_{[2.5]}$
flüssiges Nahrungsmittel$_{[2.5]}$
Essen, Mahl, Mahlzeit$_{[5]}$
Produkt, Erzeugnis$_{[2.5]}$
Getränk$_{[2.5]}$
Zwischenmahlzeit$_{[5]}$
Konsumgut$_{[2.5]}$
antialkoholisches Getränk$_{[2.5]}$
Kaffeetrinken, *Kaffee*, Kaffeeklatsch$_{[5]}$
Artikel$_{[2.5]}$
Luxusartikel$_{[2.5]}$
Genussmittel$_{[2.5]}$
Kaffee$_{[5]}$

Figure 5.2: Propagating frequencies through GermaNet hierarchy

Synset		Freq
Objekt	‘object’	261.25
Nahrung, Lebensmittel, Esswaren, Essen, Speisen	‘food’	127.98
festes Nahrungsmittel	‘solid food’	100.28
Ding, Sache, Gegenstand, Gebilde	‘thing’	66.24
Lebewesen, Kreatur, Wesen	‘creature’	50.06
natürliches Lebewesen, Organismus	‘organism’	49.14
Fleischware, Fleisch	‘meat’	37.52
höheres Lebewesen	‘higher creature’	34.51
Tier	‘animal’	26.18
Backware	‘pastry’	25.96
Gericht, Speise, Essen	‘food’	22.36
Grünzeug	‘vegetables’ (coll.)	20.78
Gewebetier	‘animal’	19.93
Artefakt, Werk	‘artefact’	19.61
Attribut, Eigenschaft, Merkmal	‘attribute’	17.73
Brot	‘bread’	17.00
Qualität, Beschaffenheit	‘quality’	16.96
Chordatier	‘animal’	14.93
Wirbeltier	‘vertebrate’	14.93
Gemüse	‘vegetables’	14.91
Pflanze, Gewächs	‘plant’	14.39
Nichts, Nichtsein	‘nothing’	14.35
Maßeinheit, Maß, Messeinheit	‘measurement’	13.70
Zeit	‘time’	11.98
Stoff, Substanz, Materie	‘substance’	11.88
Industriepflanze, Nutzpflanze	‘agricultural crop’	11.48
kognitives Objekt	‘cognitive object’	10.70
Zeitpunkt	‘point of time’	10.48
Fisch	‘fish’	9.94
Kuchen	‘cake’	8.96
nicht definite Raumeinheit	‘non-defined space unit’	8.66
Raumeinheit, Raummaß, Kubikmaß, Hohlmaß	‘space unit’	8.66
Menge	‘amount’	8.55
Struktur	‘structure’	8.55
Messgerät, Messinstrument	‘measure’	8.48
Uhrzeit, Zeit	‘time’	8.48
Uhr, Zeitmessinstrument, Zeitmesser	‘time measure’	8.48
Uhr	‘watch’	8.48
Mensch, Person, Persönlichkeit, Individuum	‘individual’	8.32
Wurstware, Wurst	‘meat’	7.70

Table 5.6: Selectional preference definition for *essen* in na as based on GermaNet nodes

To restrict the variety of noun concepts to a general level, I consider only the frequency distributions over the top GermaNet nodes. Since GermaNet had not been completed at the point of time I have used the hierarchy, I have manually added few hypernym definitions, such that the most commonly used branches realise the following 15 conceptual top levels. Most of them were already present; the additional links might be regarded as a refinement.

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

Since the 15 nodes exclude each other and the frequencies sum to the total joint verb-frame frequency, we can use the frequencies to define a probability distribution. Therefore, the 15 nodes define the selectional preferences for a verb-frame-slot combination. Tables 5.7 and 5.8 present examples of selectional preference definition with GermaNet top nodes. The relevant frame slot is underlined.

The last step towards the refined subcategorisation frame definition of German verbs needs to consider the question of how to include the selectional preferences into the frames. Two possibilities are listed below.

(a) Each argument slot in the subcategorisation frames is substituted by the verb-frame-slot combination refined by the selectional preference, e.g. instead of having a feature for the verb *beginnen* and the intransitive frame `n`, the joint frequency is distributed over `n_NP.Nom(Lebewesen)`, `n_NP.Nom(Sache)`, etc. An example is given in Table 5.9.

Remarks:

- The argument slots of frame types with several arguments are considered independently, e.g. `na` would be split into `na_NP.Nom(Lebewesen), na_NP.Nom(Sache)`, etc., and `na_NP.Akk(Lebewesen), na_NP.Akk(Sache)`, etc., but there is no direct connection between the `NP.Nom` role and the `NP.Akk` role.

- In the case of probability distributions, we either pick one (interesting) role per frame over which the joint value of verb and frame type is distributed, (e.g. `NP.Dat` in `nd`), to keep to the definition of a probability distribution, or we consider each role in the frame types, so the joint probability of verb and frame type is distributed several times, over each of the roles. By that, we have a richer preference information on the verb distribution, but the distribution is not a probability distribution per definitionem.

(b) The subcategorisation frames are substituted by the combinations of selectional preferences for the argument slots, e.g. the joint probability of a verb and `na` is distributed over `na(Lebewesen:Nahrung), na(Lebewesen:Sache), na(Sache:Nahrung)`, etc. An example is given in Table 5.10, for the most probable combinations (presentation cut-off: 0.001). The grammar only defines frequencies for the separate roles, but not for the combinations.

Remarks:

- The linguistic idea of a relationship between the different argument slots in a frame is represented in the feature combinations.
- The number of features explodes: for a frame type with one argument slot we face 15 features, for a frame type with two argument slots we face 15^2 features, for a frame type with three argument slots we face 15^3 features.
- The magnitudes of probabilities for the frame types differ strongly, as the frame probabilities are distributed over 15, 15^2 or 15^3 features.

To summarise, there is a slight linguistic bias towards version (b) which is closer in realising the relationship between different arguments in a frame, but a strong practical bias towards version (a) to prevent us from severe data sparseness. The favour for version (a) is confirmed by results by Schulte im Walde (2000a), and preliminary clustering results which showed the difficulty to encode the data in style (b). I therefore decided to encode the selectional preferences in style (a). As for the prepositional preferences, the coarse frame description can either be substituted by the refined information, or the refined information can be given in addition to the coarse definition. For the clustering experiments, I will apply both versions.

A final thought on selectional preferences is concerned with the choice of frame types to be refined with preference information. Are selectional preferences equally necessary and informative in all frame types? I empirically investigated which of the overall frame roles may be realised by different selectional preferences and are therefore relevant and informative for a selectional preference distinction. For example, the selectional preferences in ‘na’ strongly vary with respect to the subcategorising verb, but the selectional preferences in ‘ni’ mostly refer to agents and are therefore less interesting for refinement. The analysis is given in Appendix B; the results confirm the assumption that the degree of informativeness of selectional preferences in frame types differs according to their potential in distinguishing verb classes. Therefore, in parts of the clustering experiments, I will concentrate on a specific choice of frame-slot combinations to be refined by selectional preferences: `n, na, nd, nad, ns-dass`.

Verb	Frame	Synset	Freq	Prob
verfolgen 'to follow'	na	Situation	140.99	0.244
		Kognitives Objekt	109.89	0.191
		Zustand	81.35	0.141
		Sache	61.30	0.106
		Attribut	52.69	0.091
		Lebewesen	46.56	0.081
		Ort	45.95	0.080
		Struktur	14.25	0.025
		Kognitiver Prozess	11.77	0.020
		Zeit	4.58	0.008
		Besitz	2.86	0.005
		Substanz	2.08	0.004
		Nahrung	2.00	0.003
		Physis	0.50	0.001
essen 'to eat'	na	Nahrung	127.98	0.399
		Sache	66.49	0.207
		Lebewesen	50.06	0.156
		Attribut	17.73	0.055
		Zeit	11.98	0.037
		Substanz	11.88	0.037
		Kognitives Objekt	10.70	0.033
		Struktur	8.55	0.027
		Ort	4.91	0.015
		Zustand	4.26	0.013
		Situation	2.93	0.009
		Besitz	1.33	0.004
		Mittel	0.67	0.002
		Physis	0.67	0.002
		Kognitiver Prozess	0.58	0.002

Table 5.7: Selectional preference definition with GermaNet top nodes (1)

Verb	Frame	Synset	Freq	Prob
beginnen ‘to begin’	n	Situation	1,102.26	0.425
		Zustand	301.82	0.116
		Zeit	256.64	0.099
		Sache	222.13	0.086
		Kognitives Objekt	148.12	0.057
		Kognitiver Prozess	139.55	0.054
		Ort	107.68	0.041
		Attribut	101.47	0.039
		Struktur	87.08	0.034
		Lebewesen	81.34	0.031
		Besitz	36.77	0.014
		Physis	4.18	0.002
		Substanz	3.70	0.001
		Nahrung	3.29	0.001
nachdenken ‘to think’	np:Akk.über ‘about’	Situation	46.09	0.380
		Attribut	18.83	0.155
		Kognitives Objekt	12.57	0.104
		Zustand	11.10	0.092
		Besitz	6.16	0.051
		Sache	6.12	0.051
		Struktur	5.28	0.044
		Ort	5.12	0.042
		Lebewesen	3.90	0.032
		Zeit	3.34	0.028
		Kognitiver Prozess	2.05	0.017
		Physis	0.63	0.005

Table 5.8: Selectional preference definition with GermaNet top nodes (2)

Frame	Freq	Prob	Bin
na	1,026.07	0.644	1
na_NP.Akk(Situation)	140.99	0.157	1
na_NP.Akk(Kognitives Objekt)	109.89	0.123	1
na_NP.Akk(Zustand)	81.35	0.091	1
na_NP.Akk(Sache)	61.30	0.068	1
na_NP.Akk(Attribut)	52.69	0.059	1
na_NP.Akk(Lebewesen)	46.56	0.052	1
na_NP.Akk(Ort)	45.95	0.051	1
na_NP.Akk(Struktur)	14.25	0.016	1
na_NP.Akk(Kognitiver Prozess)	11.77	0.013	1
na_NP.Akk(Zeit)	4.58	0.005	0
na_NP.Akk(Besitz)	2.86	0.003	0
na_NP.Akk(Substanz)	2.08	0.002	0
na_NP.Akk(Nahrung)	2.00	0.002	0
na_NP.Akk(Physis)	0.50	0.001	0

Table 5.9: Frame+Pref distributions of *verfolgen* and frame type na

Frame	Prob	Bin
na	0.418	1
na(Lebewesen:Nahrung)	0.136	1
na(Lebewesen:Sache)	0.071	1
na(Lebewesen:Lebewesen)	0.053	1
na(Lebewesen:Attribut)	0.019	1
na(Lebewesen:Zeit)	0.013	1
na(Lebewesen:Substanz)	0.013	1
na(Lebewesen:KognitivesObjekt)	0.011	1
na(Lebewesen:Struktur)	0.009	0
na(Situation:Nahrung)	0.007	0
na(Sache:Nahrung)	0.006	0
na(KognitivesObjekt:Nahrung)	0.006	0
na(Struktur:Nahrung)	0.005	0
na(Lebewesen:Ort)	0.005	0
na(Lebewesen:Zustand)	0.005	0
na(Zeit:Nahrung)	0.004	0
na(Ort:Nahrung)	0.004	0
na(Situation:Sache)	0.003	0
na(Sache:Sache)	0.003	0
na(Lebewesen:Situation)	0.003	0
na(KognitivesObjekt:Sache)	0.003	0
na(Struktur:Sache)	0.003	0
na(Nahrung:Nahrung)	0.003	0
na(Situation:Lebewesen)	0.003	0
na(Attribut:Nahrung)	0.002	0
na(Sache:Lebewesen)	0.002	0
na(KognitivesObjekt:Lebewesen)	0.002	0
na(Struktur:Lebewesen)	0.002	0
na(Zeit:Sache)	0.002	0
na(Ort:Sache)	0.002	0
na(Zeit:Lebewesen)	0.001	0
na(Ort:Lebewesen)	0.001	0
na(Lebewesen:Besitz)	0.001	0
na(Nahrung:Sache)	0.001	0
na(Attribut:Sache)	0.001	0
na(Nahrung:Lebewesen)	0.001	0

Table 5.10: Combined Frame+Pref distributions of *essen* and frame type `na`

B) Strengthening

Assuming that the feature values of the verb description point into the desired linguistic direction but nevertheless include noise, the feature values are strengthened by squaring them, i.e. the joint frequency of each verb v and feature f_i is squared: $freq(v, f_i) = freq(v, f_i)^2$. The total verb frequency v_{freq} is adapted to the changed feature values, representing the sum of all verb feature values: $v_{freq} = \sum_i freq(v, f_i)$. The strengthened probability and binary values are based on the strengthened frequency distribution. There is no theoretical basis for the strengthening. The idea behind the manipulation was to find emphasise strong empirical evidence and ignore low frequency values.

C) Smoothing

In addition to the absolute verb descriptions described above, a simple smoothing technique is applied to the feature values. The smoothing is supposed to create more uniform distributions, especially with regard to adjusting zero values, but also to assimilate high and low frequency, probability and binary values. The smoothed distributions are particularly interesting for distributions with a large number of features, since they typically contain persuasive zero values on the one hand and severe outliers on the other hand.

Chen and Goodman (1998) present a concise overview of smoothing techniques, with specific regard towards language modelling. I decided to apply a simple smoothing algorithm which they refer to as *additive smoothing*, as a compromise between the wish to test the effect of smoothing on the verb data, and time and goal restrictions on not spending too much effort on this specific and secondary aspect.

The smoothing is performed simply by adding 0.5 to all verb features, i.e. the joint frequency of each verb v and feature f_i is changed by $freq(v, f_i) = freq(v, f_i) + 0.5$. The total verb frequency v_{freq} is adapted to the changed feature values, representing the sum of all verb feature values: $v_{freq} = \sum_i freq(v, f_i)$. The smoothed probability and binary values are based on the smoothed frequency distributions.

D) Noise

In order to discuss the usefulness and purity of the ‘linguistic’ properties in the verb distributions, the feature values in the verb descriptions are added noise. Each feature value in the verb description is assigned an additional random fraction of the verb frequency, such that the sum of all noise values equals the verb frequency. I.e. the sum of the former feature values f_i is the verb frequency $v_{freq} = \sum_i f_i$, each feature f_i is added random noise f_i^{noise}, such that the sum of the noise values equals the verb frequency: $v_{freq} = \sum_i f_i^{noise}$, so the total sum of the noisy feature values is twice the verb frequency: $2 * v_{freq} = \sum_i f_i + f_i^{noise}$. In this way, each verb feature is assigned a random value, with the random value related to the verb frequency.

5.1.3 Data Illustration

The previous section has described the feature choice for verb descriptions on three different levels. The current section is not necessary in order to understand the clustering experiments, but aims to supplement the verb distributions by various means of illustration, in order to provide the reader with an intuition on the clustering data, and to illustrate that the descriptions appear reliable with respect to their desired linguistic content. Section 5.1.3 provides a number of examples of verb distributions, followed by an illustration of the verb similarity in Section 5.1.3.

Illustration of Verb Distributions

In order to illustrate the definition of verb distributions, six verbs from different verb classes and with different defining properties have been chosen. For each of the verbs, the ten most frequent frame types are given with respect to the three levels of verb definition, both accompanied by the probability values. Each distribution level refines the previous level by substituting the respective information ('S'). On `frame+ppS+prefS`, the preferences are given for the argument roles as determined in Appendix B. Several slots within a frame type might be refined at the same time, so we do not have a probability distribution any longer.

The first column for *beginnen* defines `np` and `n` as the most probable frame types, followed by `ni` and `na` with probabilities in the next lower magnitude. Refining the prepositional phrase information shows that even by splitting the `np` probability over the different PP types, a number of prominent PPs are left, the time indicating *um*$_{Akk}$ and *nach*$_{Dat}$, *mit*$_{Dat}$ defining the begun event, *an*$_{Dat}$ as date and `in`$_{Dat}$ as place indicator. It is obvious that not all PPs are argument PPs, but the adjunct PPs also define a part of the typical verb behaviour. The refinement by selectional preferences illustrates that typical beginning roles are *Situation, Zustand, Zeit, Sache*. An indication of the verb alternation behaviour is given by `na_NP.Akk(Situation)` which refers to the same role for the direct object in a transitive situation as `n_NP.Nom(Situation)` in an intransitive situation.

As expected, *essen* as an object drop verb shows strong preferences for both an intransitive and transitive usage. The argument roles are strongly (i.e. catching a large part of the total verb-frame probability) determined by *Lebewesen* for both `n` and `na` and *Nahrung* for `na`. *fahren* chooses typical manner of motion frames (`n, np, na`) with the refining PPs being directional (*in*$_{Akk}$, *zu*$_{Dat}$, *nach*$_{Dat}$) or defining a means (*mit*$_{Dat}$, *in*$_{Dat}$, *auf*$_{Dat}$). The selectional preferences represent the desired alternation behaviour: the object drop case by *Lebewesen* in `n` and in `na`, and the inchoative/causative case by *Sache* in `n` and in `na`. An example for the former case is *Peter fährt* 'Peter drives' vs. *Peter fährt das Auto* 'Peter drives the car', an example for the latter case is *Das Auto fährt (langsam)* 'The car goes (slowly)' vs. *Peter fährt das Auto* 'Peter drives the car'.

An example of verb ambiguity is given by *dämmern* which –on the one hand– shows strong probabilities for `n` and `x` as typical for a weather verb, but –on the other hand– shows strong probabilities for `xd`, `nd` and subcategorising finite clauses which refer to its sense of understanding (e.g. *ihm*$_{Dat}$ *dämmert* ...). Similarly, *laufen* represents a manner of motion verb, which is indicated by strong preferences for `n, np, na`, with refining directional prepositions *in*$_{Dat}$, *auf*$_{Akk}$, *gegen*$_{Akk}$, but is also used within the existential collocational expression *es läuft* ‘it works’, as indicated by `x`.

The distributions for *glauben* show strong probabilities for finite clauses (referring to the ‘to think’ sense), and minor probabilities for `na` (ditto) and `n, np, nd, nad` (referring to the ‘to believe’ sense). The PP refinement in this case illustrates the restricted use of the specific preposition *an*$_{Akk}$, compared to the multi-fold categorial usage of directional/means/etc. PPs of e.g. manner of motion verbs. The main usage of selectional preferences is represented by *Lebewesen* for `ns-dass, na, nd` and `n` (object drop of `nd`).

Verb	Distribution					
	frame		frame+ppS		frame+ppS+prefS	
beginnen	np	0.428	n	0.278	np:Akk.um	0.161
	n	0.278	np:Akk.um	0.161	n_NP.Nom(Situation)	0.118
	ni	0.087	ni	0.087	ni	0.087
	na	0.071	np:Dat.mit	0.082	np:Dat.mit	0.082
	nd	0.036	na	0.071	np:Dat.an	0.056
	nap	0.032	np:Dat.an	0.056	np:Dat.in	0.055
	nad	0.019	np:Dat.in	0.055	n_NP.Nom(Zustand)	0.032
	nir	0.012	nd	0.036	n_NP.Nom(Zeit)	0.027
	ns-2	0.009	nad	0.019	n_NP.Nom(Sache)	0.024
	xp	0.005	np:Dat.nach	0.014	na_NP.Akk(Situation)	0.023
dämmern	n	0.195	n	0.195	xd	0.179
	xd	0.179	xd	0.179	nd_NP.Dat(Lebewesen)	0.103
	nd	0.132	nd	0.132	na_NP.Akk(Lebewesen)	0.080
	na	0.123	na	0.123	nd_NP.Nom(Sache)	0.066
	ns-dass	0.122	ns-dass	0.122	n_NP.Nom(KognitiverProzess)	0.061
	x	0.061	x	0.061	x	0.061
	nds-dass	0.046	nds-dass	0.046	ns-dass_NP.Nom(Zeit)	0.052
	ndp	0.035	ns-2	0.033	nds-dass	0.046
	ns-2	0.033	ndp:Dat.nach	0.015	na_NP.Akk(Sache)	0.043
	nas-dass	0.015	nas-dass	0.015	na_NP.Nom(Lebewesen)	0.041
essen	na	0.418	na	0.418	na_NP.Nom(Lebewesen)	0.329
	n	0.261	n	0.261	na_NP.Akk(Nahrung)	0.167
	nad	0.101	nad	0.101	na_NP.Akk(Sache)	0.087
	np	0.056	nd	0.053	n_NP.Nom(Lebewesen)	0.083
	nd	0.053	ns-2	0.018	na_NP.Akk(Lebewesen)	0.065
	nap	0.041	np:Dat.auf	0.017	n_NP.Nom(Nahrung)	0.056
	ns-2	0.018	ns-w	0.013	n_NP.Nom(Sache)	0.043
	ns-w	0.013	ni	0.012	nd_NP.Nom(Lebewesen)	0.038
	ni	0.012	np:Dat.mit	0.010	nd_NP.Dat(Nahrung)	0.023
	nas-2	0.007	np:Dat.in	0.009	na_NP.Akk(Attribut)	0.023

Table 5.11: Examples of most probable frame types (1)

Verb	Distribution					
	frame		frame+ppS		frame+ppS+prefS	
fahren	n	0.339	n	0.339	n_NP.Nom(Sache)	0.118
	np	0.285	na	0.193	n_NP.Nom(Lebewesen)	0.095
	na	0.193	np:Akk.in	0.054	na_NP.Nom(Lebewesen)	0.082
	nap	0.059	nad	0.042	na_NP.Akk(Sache)	0.063
	nad	0.042	np:Dat.zu	0.041	n_NP.Nom(Ort)	0.057
	nd	0.040	nd	0.040	np:Akk.in	0.054
	ni	0.010	np:Dat.nach	0.039	na_NP.Nom(Sache)	0.047
	ns-2	0.008	np:Dat.mit	0.034	np:Dat.zu	0.041
	ndp	0.008	np:Dat.in	0.032	np:Dat.nach	0.039
	ns-w	0.004	np:Dat.auf	0.018	np:Dat.mit	0.034
glauben	ns-dass	0.279	ns-dass	0.279	ns-2	0.274
	ns-2	0.274	ns-2	0.274	ns-dass_NP.Nom(Lebewesen)	0.217
	np	0.100	n	0.088	np:Akk.an	0.083
	n	0.088	np:Akk.an	0.083	na_NP.Akk(Sache)	0.065
	na	0.080	na	0.080	na_NP.Nom(Lebewesen)	0.062
	ni	0.050	ni	0.050	n_NP.Nom(Lebewesen)	0.060
	nd	0.034	nd	0.034	ni	0.050
	nad	0.023	nad	0.023	nd_NP.Nom(Lebewesen)	0.026
	nds-2	0.010	np:Dat.an	0.019	ns-dass_NP.Nom(Sache)	0.020
	nai	0.009	nds-2	0.010	np:Dat.an	0.019
laufen	n	0.382	n	0.382	n_NP.Nom(Situation)	0.118
	np	0.324	na	0.103	n_NP.Nom(Sache)	0.097
	na	0.103	np:Dat.in	0.060	np:Dat.in	0.060
	nap	0.041	nd	0.036	n_NP.Nom(Zustand)	0.037
	nd	0.036	np:Akk.auf	0.029	np:Akk.auf	0.029
	nad	0.026	np:Dat.auf	0.029	np:Dat.auf	0.029
	x	0.026	nad	0.026	n_NP.Nom(Attribut)	0.028
	ns-2	0.018	x	0.026	na_NP.Akk(Zeit)	0.027
	ndp	0.011	np:Dat.seit	0.022	x	0.026
	xa	0.010	np:Akk.gegen	0.020	na_NP.Nom(Sache)	0.025

Table 5.12: Examples of most probable frame types (2)

Illustration of Verb Similarity

The similarity between the different verbs is illustrated in three ways: Table 5.13 lists the five closest verbs for the above sample verbs, according to the similarity measures *cosine* and *skew divergence*, for each of the three verb description levels. The examples show that the neighbour relationship varies with the verb description and the similarity measure. Strongly related verb pairs such as *essen/trinken* or *fahren/fliegen* are invariant with respect to the used parameters, i.e. *trinken* is indicated as the closest verb of *essen* in each of the six columns. Verb pairs whose similarity relies on a similar usage of prepositional phrases (such as *beginnen/enden*) are recognised as close neighbours when refining the frame information by PPs. Few verbs in the sample need the refinement by selectional preferences in order to be recognised as similar, e.g. *essen/saufen*, in some cases the refined information seems to confuse the previous information level; for example, *anfangen* and *aufhören* are recognised as near neighbours of *beginnen* on basis of `frame+ppS`, but not on basis of `frame+ppS+prefS`. Concerning ambiguity, *dämmern* defines as nearest neighbours those verbs which agree in the subcategorisation of `nd`, such as *helfen* and *bedürfen* (incorrect choices), but the weather sense is not represented in the nearest neighbour set. For *laufen*, both nearest neighbours in the manner of motion sense (such as *fahren, fliegen*) and in the existence sense (such as *existieren, bestehen*) are realised.

Table 5.14 is supposed to represent especially strong similarities between pairs of verbs: The table defines two verbs as a pair of respective nearest neighbours if each is the other's most similar verb, according to the skew divergence. Comparing the verb pair lists with the possible list of verb pairs as defined by the manual verb classification, recall decreases with refining the frame distributions, but precision increases. Later in the clustering experiments, we will see that the symmetrically nearest neighbour verbs pervasively appear within the same verb clusters.

Table 5.15 compares the similarities between verbs in the same semantic class with similarities between verbs in different semantic classes. The verbs are described on different frame levels, and the similarity in the whole table is based on the skew divergence. The first rows concerning *beginnen* until the horizontal line present the distances between *beginnen* and the four other *Aspect* verbs *anfangen, aufhören, beenden, enden*. The following rows present the distances between *beginnen* and the 10 most similar verbs which are not in the *Aspect* class. For example, the second column based on `frame+ppS` tells us that the similarity between *beginnen* and *enden* is strong (because of a small distance), the similarity to *anfangen* and *aufhören* is strong, but not distinguishing the common class membership (because there are more similar verbs which are not in the same semantic class), and the similarity to *beenden* is weak, compared to the verbs which are not in the same semantic class.

The first rows concerning *fahren* present the distances between *fahren* and the three other verbs in the *Manner of Motion* sub-class *Vehicle*. The following rows present the distances to all other *Manner of Motion* verbs, and the last lines present the distances between *fahren* and the 10 most similar verbs which are not in the *Manner of Motion* class. For example, the second column based on `frame+ppS` shows that *fliegen* is by far the most similar verb to *fahren*, and *laufen* and *wandern* (among others) are more similar to *fahren* than the other verbs from the same *Means* sub-class. But many verbs from other classes are more similar to *fahren* than several *Manner of Motion* verbs. The table demonstrates that it is not necessarily the case that the verbs in the same class are those which are most similar. The coherence of the verbs in the same classes varies according to the verb distributions, which corresponds to the examples of closest verbs in Table 5.13.

Verb	Closest Neighbours					
	frame		frame+ppS		frame+ppS+prefS	
	cos	skew	cos	skew	cos	skew
beginnen	sprechen	liegen	enden	enden	enden	enden
	resultieren	bestehen	anfangen	anfangen	laufen	liegen
	segeln	leben	kommunizieren	leben	segeln	laufen
	verhandeln	sprechen	rudern	rudern	liegen	stehen
	liegen	verhandeln	aufhören	verhandeln	bestehen	bestehen
dämmern	helfen	bedürfen	saufen	bedürfen	helfen	helfen
	saufen	gehen	helfen	feststellen	bedürfen	gehen
	lamentieren	feststellen	rufen	glauben	rufen	rufen
	riechen	glauben	fliegen	bemerken	nieseln	flüstern
	rufen	helfen	folgen	lamentieren	unterrichten	kriechen
essen	trinken	trinken	trinken	trinken	trinken	trinken
	lesen	spenden	lesen	produzieren	saufen	fahren
	spenden	produzieren	schließen	lesen	rufen	rufen
	entfernen	lesen	entfernen	hören	lesen	produzieren
	hören	rufen	spenden	spenden	produzieren	lesen
fahren	fliegen	fliegen	fliegen	fliegen	fliegen	fliegen
	laufen	demonstrieren	saufen	laufen	wandern	wandern
	demonstrieren	laufen	laufen	fließen	segeln	laufen
	fließen	sprechen	rufen	rufen	rotieren	verhandeln
	reden	verhandeln	hasten	wandern	starren	stehen
glauben	folgern	denken	versichern	denken	versichern	denken
	versichern	folgern	vermuten	versichern	folgern	versichern
	denken	versichern	folgern	vermuten	denken	fürchten
	vermuten	fürchten	denken	folgern	fürchten	folgern
	fürchten	vermuten	fürchten	fürchten	jammern	klagen
laufen	fließen	fließen	heulen	fliegen	segeln	stehen
	reden	fliegen	donnern	fahren	enden	liegen
	leben	leben	existieren	fließen	stehen	fahren
	wandern	sprechen	blitzen	existieren	existieren	bestehen
	starren	fahren	hasten	leben	liegen	existieren

Table 5.13: Examples of closest verbs

Distribution					
frame		frame+ppS		frame+ppS+prefS	
ahnen	– wissen	ahnen	– wissen	anfangen	– aufhören
anfangen	– aufhören	anfangen	– aufhören	basieren	– beruhen
bekommen	– brauchen	basieren	– beruhen	beginnen	– enden
bemerken	– feststellen	beginnen	– enden	bekommen	– erhalten
benötigen	– erhalten	bekanntgeben	– erkennen	bemerken	– feststellen
beruhen	– resultieren	bekommen	– erhalten	bringen	– treiben
beschreiben	– realisieren	bemerken	– feststellen	denken	– glauben
bestimmen	– kriegen	beschreiben	– charakterisieren	dienen	– folgen
bringen	– schicken	bestimmen	– kriegen	erfahren	– hören
darstellen	– senken	bringen	– schicken	erhöhen	– steigern
dienen	– folgen	darstellen	– senken	essen	– trinken
eilen	– gleiten	denken	– glauben	fahren	– fliegen
entfernen	– lesen	dienen	– folgen	freuen	– ärgern
erhöhen	– stützen	eröffnen	– gründen	gründen	– sehen
erzeugen	– vernichten	essen	– trinken	lächeln	– schreien
essen	– trinken	existieren	– leben	präsentieren	– stellen
fahren	– fliegen	fahren	– fliegen	reden	– sprechen
fließen	– leben	freuen	– ärgern	regnen	– schneien
freuen	– fühlen	jammern	– klagen	rennen	– starren
gehen	– riechen	leihen	– wünschen	schenken	– vermachen
gähnen	– lamentieren	liegen	– sitzen	schließen	– öffnen
jammern	– klagen	lächeln	– schreien	sitzen	– stehen
kommunizieren	– nachdenken	nachdenken	– spekulieren	versprechen	– zusagen
kriechen	– rennen	produzieren	– vermitteln		
lachen	– schreien	präsentieren	– stellen		
leihen	– wünschen	reden	– sprechen		
liegen	– stehen	regnen	– schneien		
produzieren	– unterrichten	schenken	– vermachen		
präsentieren	– stellen	steigern	– vergrößern		
regnen	– schneien	unterstützen	– vernichten		
schenken	– vermachen	versprechen	– zusagen		
sprechen	– verhandeln	vorführen	– zustellen		
versprechen	– zusagen				

Table 5.14: Examples of nearest neighbour verb pairs

Verb	Verb Distances					
	frame		frame+ppS		frame+ppS+prefS	
beginnen	anfangen	0.329	anfangen	0.525	anfangen	1.144
	aufhören	0.600	aufhören	0.703	aufhören	1.475
	beenden	1.279	beenden	1.349	beenden	2.184
	enden	0.171	enden	0.421	enden	0.572
	liegen	0.113	leben	0.580	liegen	0.772
	bestehen	0.121	rudern	0.581	laufen	0.811
	leben	0.122	verhandeln	0.583	stehen	0.830
	sprechen	0.126	fahren	0.592	bestehen	0.862
	verhandeln	0.127	fliegen	0.663	verhandeln	0.911
	segeln	0.129	schreien	0.664	klettern	0.927
	stehen	0.135	bestehen	0.665	leben	0.928
	resultieren	0.144	demonstrieren	0.669	sitzen	0.945
	sitzen	0.157	kommunizieren	0.671	fahren	1.051
	rudern	0.158	laufen	0.677	sprechen	1.060
fahren	fliegen	0.030	fliegen	0.123	fliegen	0.323
	rudern	0.356	rudern	0.807	rudern	1.376
	segeln	0.205	segeln	0.502	segeln	0.643
	drehen	0.811	drehen	0.975	drehen	1.611
	eilen	0.223	eilen	0.497	eilen	0.822
	fließen	0.097	fließen	0.288	fließen	0.816
	gehen	0.382	gehen	0.519	gehen	0.700
	gleiten	0.265	gleiten	0.741	gleiten	0.999
	hasten	0.349	hasten	0.612	hasten	1.240
	klettern	0.103	klettern	0.501	klettern	0.688
	kriechen	0.158	kriechen	0.499	kriechen	0.945
	laufen	0.078	laufen	0.249	laufen	0.533
	rennen	0.224	rennen	0.437	rennen	0.768
	rotieren	0.341	rotieren	0.878	rotieren	0.991
	schleichen	0.517	schleichen	0.747	schleichen	1.407
	treiben	0.613	treiben	0.705	treiben	1.265
	wandern	0.126	wandern	0.363	wandern	0.501
	demonstrieren	0.074	rufen	0.332	verhandeln	0.575
	sprechen	0.086	schreien	0.383	stehen	0.579
	verhandeln	0.096	essen	0.405	leben	0.588
	erwachsen	0.123	leben	0.443	sprechen	0.647
	reden	0.126	verhandeln	0.462	rufen	0.737
	leben	0.132	demonstrieren	0.469	demonstrieren	0.759
	donnern	0.135	enden	0.485	sitzen	0.765
	enden	0.163	donnern	0.487	reden	0.782
	rufen	0.168	trinken	0.503	starren	0.787
	beginnen	0.172	sprechen	0.510	liegen	0.816

Table 5.15: Examples distances between verbs in same or different classes

5.1.4 Summary

This section has provided the necessary data background for the clustering experiments. I once more presented the gold standard verb classes (the full set and a reduced set of the classes), accompanied by their empirical properties. A choice of features to describe the verbs has been given, referring to three levels of verb description: purely syntactic frame types, prepositional phrase information, and selectional preferences. I pointed to difficulties in encoding the verb features both in general and with respect to the linguistic task. Variations of the verb attributes will be discussed separately in Section 5.4, which optimises the setup of the clustering experiments.

Finally, I illustrated the verb similarity by various means, in order to provide the reader with an intuition on the clustering data. It is important to notice that the basic verb descriptions appear reliable with respect to their desired linguistic content. The definition includes the desired features and some noise, and the possible effects of verb ambiguity. Verb similarity is represented as expected, i.e. verbs from the same semantic class are assigned a strong degree of similarity, and verbs from different semantic classes are assigned weak degrees of similarity, including some noise with respect to an intuitive definition of similarity. The question now is whether and how the clustering algorithm is able to benefit from the linguistic properties and to abstract from the noise in the distributions. This question is addressed in the following sections.

5.2 Verb Class Experiments

This section brings together the clustering concept, the clustering data and the clustering techniques, and presents the clustering experiments as performed by k-Means. Section 5.2.1 reminds the reader of the clustering methodology and its parameters, Section 5.2.2 introduces the baseline as well as the upper bound of the experiments, and Section 5.2.3 finally lists and describes the clustering results.

5.2.1 Clustering Methodology

The clustering methodology describes the application of k-Means to the clustering task: The verbs are associated with distributional vectors over frame types and assigned to starting clusters, the k-Means algorithm 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. As Chapter 4 has illustrated, this simple description of the clustering methodology contains several parameters which need to be varied, since it is not clear which setup results in the optimal cluster analysis. The following paragraphs summarise the variation of the experiment setup.

Number of Verbs and Verb Classes The experiments partly refer to the reduced set of 57 verbs (in 14 manual classes), since this concise set facilitates the interpretation of the various clustering setups. But most experiments are also applied to the full set of 168 verbs (in 43 manual classes).

Frame Distribution The representation of the verbs is realised by vectors which describe the verbs by distributions over their features. The German verbs are described on three levels at the syntax-semantic interface, purely syntactic frame types, prepositional phrase information, and selectional preferences. Each level refers to frequencies, probabilities, and binaries, with their original, strengthened, smoothed or noisy values.

Input Cluster The starting clusters for a k-Means cluster analysis are generated either randomly or by a pre-processing cluster analysis. For random cluster input the verbs are randomly assigned to a cluster, with cluster numbers between 1 and the number of manual classes. An optimisation of the number of clusters is ignored in this section, but Section 5.4 will come back to this issue. For pre-processing clusters, agglomerative hierarchical analyses are performed, referring to all amalgamation methods as introduced in Chapter 4: single-linkage, complete-linkage, centroid distance, average distance, and Ward's method.

Similarity Measure The experiments vary the similarity measures which determine the similarity of verbs and clusters, cf. Chapter 4.

5.2.2 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 self-created desideratum. In case of clustering the larger set of verbs, the manual classification is adapted before calculating the upper bound, by deleting more than one sense of the verbs, i.e. each verb should only belong to one class, since k-Means as a hard clustering algorithm cannot model ambiguity.

Table 5.16 lists the baseline and upper bound values for the clustering experiments. All evaluation measures are cited except for sum-of-squared-error and silhouette, which depend on the similarity measure.

5.2.3 Experiment Results

Following, several tables present the results of the diverse clustering experiments. Each table concentrates on one parameter of the clustering process; the final table then focuses on per-

Evaluation	Baseline	Optimum	Baseline	Optimum
	57 verbs (unambiguous)		168 verbs (ambiguous)	
PairR	6.96	100	2.14	91.96
PairP	5.89	100	2.03	100
PairF	6.37	100	2.08	95.81
ClassR	14.42	100	4.92	93.98
ClassP	14.31	100	5.18	100
ClassF	14.36	100	5.05	96.90
APP	0.017	0.291	0.005	0.277
MI	0.234	0.493	0.302	0.494
Rand	0.877	1	0.956	0.998
Rand$_{adj}$	-0.002	1	-0.004	0.909
B-k	0.064	1	0.020	0.911

Table 5.16: k-Means experiment baseline and upper bound

forming a cluster analysis with the ‘best’ parameter set, in order to illustrate the linguistically interesting parameter concerning the feature choice for the verbs. To facilitate the understanding of the tables without spending to much time on reading them, the main statements of the tables are summarised. As said before, the applied evaluation measures are the adjusted pair-wise precision APP, the f-score of pair-wise P/R $PairF$, and the adjusted Rand index $Rand_{adj}$ (shorthand: R$_a$).

Tables 5.17 to 5.20 illustrate the effect of the frame distributions on the clustering result. All distributions are tested on both verb sets, described by the features `frame` (only) and frames refined by PPs (`frame+pp`), with various inputs, and the cosine as similarity measure (since it works on all kinds of distributions). To summarise the results, (i) the original distributions (‘orig’) are more useful than their strengthened variants (‘mani’), except for the case of producing binary distributions. The latter might be explained by a more demanding dividing line between binaries 0 and 1, when based on strengthened conditions. (ii) Smoothing of the feature values (‘smooth’) does help the clustering in two cases: in case of probabilities the more objects and features are present in the clustering process, the more does smoothing support the analysis, which is exactly the effect I desired; in case of frequencies, the less objects and features are present in the clustering process, the more does smoothing support the analysis, i.e. for large number of features the smoothing of frequencies does not help the clustering. (iii) Adding noise to the verb features (‘noise’) has a similar, but less severe effect on the clustering results than smoothing the distributions. This insight is surprising, since I have expected the noisy distributions to perform more poorly then the original or smoothed distributions. The effect might be due to the fact that (a) the original distributions obtained from the unsupervised trained grammar model need to be considered noisy, too, and (b) the range of the additional noise is limited to the respective verb frequency. So the resulting distributions are on the one hand ‘noisier than before’, but on the other hand smoothed, since zero values are added some verb frequency proportion and the difference between high and low frequency feature values is assimilated. (iv) There is no preference for either probabilities or frequencies. Interestingly, one is favoured compared to the other with respect to the chosen clustering parameter combination. Including smoothing, however, the probability distributions are favoured in clustering. Further experiments will therefore concentrate on probabilities.

Tables 5.21 to 5.24 illustrate the usage of different similarity measures. As before, the experiments are performed on both verb sets and the two feature sets `frame` and `frame+pp`, with various inputs. The similarity measures are applied to the relevant verb distributions, probabilities if possible, binaries otherwise. The tables point out that there is no unique best performing similarity measure in the clustering processes. Especially with few features, it might be either cosine, L1, Euclidean distance, information radius, or skew divergence which achieve the comparably best cluster analysis; the τ coefficient and the binary measures provide less reliable results, compared to the former similarity measures. On larger feature sets, the Kullback-Leibler variants information radius and (mainly:) skew divergence tend to outperform all other similarity measures. In further experiments, I will therefore concentrate on using the latter two measures.

Tables 5.25 to 5.28 compare the effect of varying the input clusters for the k-Means algorithm. The experiments are performed on both verb sets and the two feature sets `frame` and `frame+pp`, on basis of probability distributions, with the two similarity measures information radius and skew divergence. For random and hierarchical input, I cite both the evaluation scores for the k-Means input cluster analysis (i.e. the output clustering from the random assignment or the pre-processing hierarchical analysis), and for the k-Means result. The following insights are based on the input analysis:

1. The *manual* column in the tables refers to a cluster analysis where the input clusters to k-Means are the manual classification, i.e. the gold standard. An optimal cluster analysis would realise the ‘perfect’ clustering and not perform any re-organising iteration on the clusters. In the experiments, k-Means does perform iterations, so the clustering result is sub-optimal. Since the input is the desired result, we can regard the clustering output as a kind of upper bound as defined by the data, i.e. in a given parameter space the clustering could not be better with the respective feature description of the verbs. Comparing the minimal pairs of clustering experiments only distinguished by the feature description, the clustering result should (and actually ‘is’) therefore be better with an enlarged feature set, as I hope to improve the verb description by the feature description. For illustration purposes, the following list shows a manual clustering result for the reduced verb set, based on the coarse frame descriptions only. Verbs not correctly belonging to the same class (according to the gold standard) are marked by different subscripts.

   - anfangen aufhören
   - ahnen glauben vermuten wissen
   - beenden$_1$ bekommen$_2$ erhalten$_2$ erlangen$_2$ konsumieren$_3$ kriegen$_2$
   - bringen$_1$ eröffnen$_2$ liefern$_1$ schicken$_1$ vermitteln$_1$ zustellen$_1$
   - beginnen$_1$ blitzen$_2$ donnern$_2$ enden$_1$ fahren$_3$ fliegen$_3$ rudern$_3$
   - freuen ärgern

- ankündigen bekanntgeben verkünden
- beschreiben$_1$ charakterisieren$_1$ darstellen$_1$ interpretieren$_1$ unterstützen$_2$
- beharren$_1$ bestehen$_1$ denken$_2$ insistieren$_1$ pochen$_1$
- liegen$_1$ segeln$_2$ sitzen$_1$ stehen$_1$
- dienen folgen helfen
- lesen$_1$ schließen$_2$ öffnen$_2$
- essen saufen trinken
- dämmern nieseln regnen schneien

In comparison, a second list presents the verb classes resulting from the same experiment setup, except for using the verb descriptions enriched by prepositional phrase information. Obviously, the cluster analysis with the additional information introduces similar but less errors into the manual classification, so the verb description data is more appropriate for the classification.

- anfangen aufhören beginnen
- ahnen denken glauben vermuten wissen
- beenden$_1$ bekommen$_2$ erhalten$_2$ erlangen$_2$ kriegen$_2$
- bringen liefern schicken vermitteln zustellen
- donnern$_1$ enden$_2$ fahren$_2$ fliegen$_2$ rudern$_2$
- freuen ärgern
- ankündigen bekanntgeben eröffnen verkünden
- beschreiben$_1$ charakterisieren$_1$ darstellen$_1$ interpretieren$_1$ unterstützen$_2$
- beharren bestehen insistieren pochen
- blitzen$_1$ liegen$_2$ segeln$_3$ sitzen$_2$ stehen$_2$
- dienen folgen helfen
- schließen öffnen
- essen konsumieren lesen saufen trinken
- dämmern nieseln regnen schneien

2. For *random* clustering input to k-Means, the tables present both the best and the average clustering results. The best results are coupled with the evaluation of their input clusters, i.e. the random clusterings. As the tables show, the input clusters are given 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. The following list shows an example random clustering input, with those verbs actually belonging to the same class marked in bold font.

   - konsumieren kriegen vermuten
   - anfangen
   - ahnen bekanntgeben bestehen **fahren fliegen** liefern nieseln pochen
   - aufhören **bekommen erhalten** essen insistieren regnen segeln vermitteln

- beginnen freuen interpretieren
- rudern saufen schneien ärgern
- eröffnen folgen glauben
- zustellen
- charakterisieren dämmern stehen
- blitzen verkünden wissen
- beschreiben **dienen** donnern schließen **unterstützen**
- beenden darstellen **liegen sitzen**
- ankündigen denken enden lesen schicken öffnen
- beharren bringen erlangen helfen trinken

k-Means is able to cope with the high degree of perturbation: the resulting clusters are comparable with those based on pre-processed hierarchical clustering. The competitiveness decreases with both an increasing number of verbs and features. Experiments based on a considerably enlarged set of verbs (not presented here) show that k-Means fails on a meaningful re-organisation of the random cluster input.

The average values of the random input experiments are clearly below the best ones, but still comparable to a part of the pre-processed clustering results, especially when based on a small feature set.

3. Cluster analyses 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, cf. Chapter 4; the effect is observable in the analysis, which typically contains one very large cluster and many clusters with few verbs, mostly singletons. The following list of clusters represents a typical result of this method. It is based on the reduced verb set with coarse frame description, similarity measure: skew divergence.

   - ahnen$_2$ wissen$_2$
   - anfangen$_1$ aufhören$_1$ beginnen$_1$ beharren$_9$ bestehen$_9$ blitzen$_{14}$ denken$_2$ donnern$_{14}$ enden$_1$ fahren$_5$ fliegen$_5$ liegen$_{10}$ pochen$_9$ rudern$_5$ saufen$_{13}$ segeln$_5$ sitzen$_{10}$ stehen$_{10}$
   - ankündigen$_7$ beenden$_1$ bekanntgeben$_7$ bekommen$_3$ beschreiben$_8$ bringen$_4$ charakterisieren$_8$ darstellen$_8$ erhalten$_3$ erlangen$_3$ eröffnen$_7$ essen$_{13}$ interpretieren$_8$ konsumieren$_{13}$ kriegen$_3$ lesen$_{13}$ liefern$_4$ schicken$_4$ schließen$_{12}$ trinken$_{13}$ unterstützen$_{11}$ verkünden$_7$ vermitteln$_4$ öffnen$_{12}$
   - dienen$_{11}$ folgen$_{11}$
   - dämmern$_{14}$
   - freuen$_6$
   - glauben$_2$
   - helfen$_{11}$
   - insistieren$_9$

- nieseln$_{14}$
- regnen$_{14}$ schneien$_{14}$
- vermuten$_{2}$
- zustellen$_{4}$
- ärgern$_{6}$

k-Means obviously cannot compensate for the strong bias in cluster sizes (and their respective centroids); the re-organisation improves the clusterings, but the result is still worse than for any other input.

4. With *average distance* and *centroid distance* amalgamation, both the clusterings and the evaluation results are less extreme than single-linkage, since the chaining effect is smoothed. The hierarchical clusterings contain few large and many small clusters, but with less verbs in the larger clusters and fewer singletons. The overall results are better than for single-linkage, but hardly improved by k-Means.

5. Hierarchical clusters based on *complete-linkage* amalgamation are more compact, theory-conform, and result in closer relation to the gold standard than the previous methods. The hierarchical input is hardly improved by k-Means, in some cases the k-Means output is worse than its hierarchical input.

6. *Ward's method* seems to work best on hierarchical clusters and k-Means input. The cluster sizes are more balanced, corresponding to compact cluster shapes, as the following example illustrates which is based on the same methodology as for single-linkage above.

   - ahnen$_{2}$ wissen$_{2}$
   - anfangen$_{1}$ aufhören$_{1}$ rudern$_{5}$
   - ankündigen$_{7}$ beenden$_{1}$ bekanntgeben$_{7}$ bekommen$_{3}$ beschreiben$_{8}$ bringen$_{4}$ charakterisieren$_{8}$ darstellen$_{8}$ erhalten$_{3}$ erlangen$_{3}$ eröffnen$_{7}$ interpretieren$_{8}$ konsumieren$_{13}$ kriegen$_{3}$ liefern$_{4}$ schicken$_{4}$ unterstützen$_{11}$ vermitteln$_{4}$
   - beginnen$_{1}$ beharren$_{9}$ bestehen$_{9}$ liegen$_{10}$ pochen$_{9}$ segeln$_{5}$ sitzen$_{10}$ stehen$_{10}$
   - blitzen$_{14}$ donnern$_{14}$ enden$_{1}$ fahren$_{5}$ fliegen$_{5}$
   - denken$_{2}$ glauben$_{2}$
   - dienen$_{11}$ folgen$_{11}$ helfen$_{11}$
   - dämmern$_{14}$
   - essen$_{13}$ lesen$_{13}$ schließen$_{12}$ trinken$_{13}$ öffnen$_{12}$
   - freuen$_{6}$ ärgern$_{6}$
   - insistieren$_{9}$ saufen$_{13}$
   - nieseln$_{14}$ regnen$_{14}$ schneien$_{14}$
   - verkünden$_{7}$ vermuten$_{2}$
   - zustellen$_{4}$

> As for complete-linkage, k-Means hardly improves the clusterings, in some cases the k-Means output is worse than its hierarchical input. A cluster analysis based on Ward's hierarchical clusters is performing best of all applied methods, when compared to the gold standard, especially with an increasing number of verbs (and features). The similarity of Ward's clusters (and similarly: complete-linkage clusters) and k-Means is not by chance, since both methods aim to optimise the same issue, the sum of distances between the verbs and their respective cluster centroids.

To summarise the overall insights for my needs, utilising a hierarchical clustering based on Ward's method as input to k-Means is the most stable solution. Since Ward's method is the most time-consuming, random input (and its best output) might be used as long as we concentrate on few verbs and few features, and hierarchical clustering with complete-linkage might be used, since its clustering hypothesis and performance is similar to Ward's, but it is less time consuming. When applying Ward's or complete-linkage clustering, k-Means is not expected to improve the result significantly.

The last part of the experiments applies the algorithmic insights from the previous experiments to a linguistic variation of parameters. The verbs are described by probability distributions on different levels of linguistic information (frames, prepositional phrases, selectional preferences). Similarities are measured by the skew divergence. A pre-processing hierarchical cluster analysis is performed by complete-linkage and Ward's method, and k-Means is applied to re-organise the clusters. Tables 5.29 and 5.30 present the results, with frames only (frame), substitutional and additional prepositional phrase information (ppS/ppA), and substitutional and additional selectional preferences (prefS/prefA), either on specified frame slots (n, na, nd, nad, ns-dass for prefS, and n, na, nd, nad, ns-dass for prefA), 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 cited in the relevant column. Smoothing is omitted in the experiments; it does improve the results, but for comparing the feature choice the original probabilities are more suitable.

The tables demonstrate that already a purely syntactic verb description allows a verb clustering clearly above the baseline. Refining the coarse subcategorisation frames by prepositional phrases considerably improves the verb clustering results, with no obvious difference concerning the distinction between substitutional and additional PP definition. Unfortunately, there is no consistent effect of adding the selectional preferences to the verb description. With the reduced set of verbs, I have expected the results to decrease when adding selectional preferences, since the increasing number of features per object represents a problem to the cluster analysis. For the full set of 168 verbs, a careful choice of selectional preference roles does improve the clustering results compared to the coarse syntactic frame information frame. But compared to frame+pp, in some cases the refining selectional information does help the clustering, in others it does not. In the case of adding role information on all NP (and all PP) slots, the problem might be caused by sparse data; but specifying only a linguistically chosen subset of argument slots does not increase the number of features considerably, compared to frame+pp, so I assume additional linguistic reasons directly relevant for the clustering outcome.

Eval	Input	Distribution											
		prob				freq				bin			
		orig	mani	smooth	noise	orig	mani	smooth	noise	orig	mani	smooth	noise
APP	Random	0.139	0.140	0.142	0.153	0.130	0.098	0.134	0.140	0.085	0.106	0.075	0.040
	H-Comp	0.072	0.069	0.072	0.096	0.071	0.067	0.079	0.094	0.061	0.077	0.049	0.010
	H-Ward	0.102	0.083	0.102	0.103	0.103	0.068	0.102	0.100	0.065	0.110	0.072	0.005
PairF	Random	31.80	25.21	31.69	32.96	33.47	30.26	36.19	31.63	28.97	32.91	24.17	11.54
	H-Comp	22.78	21.08	22.78	26.67	21.23	20.62	21.86	27.24	18.25	26.61	14.81	3.96
	H-Ward	29.17	21.97	27.10	27.30	29.73	20.80	30.24	27.59	26.13	28.57	20.39	3.81
R_a	Random	0.259	0.181	0.258	0.274	0.287	0.244	0.317	0.268	0.239	0.277	0.186	0.054
	H-Comp	0.153	0.134	0.153	0.200	0.136	0.127	0.142	0.205	0.115	0.208	0.077	-0.025
	H-Ward	0.230	0.145	0.205	0.207	0.235	0.130	0.241	0.209	0.207	0.233	0.149	-0.029

Table 5.17: Comparing distributions (frame only, reduced verb set)

Eval	Input	Distribution											
		prob				freq				bin			
		orig	mani	smooth	noise	orig	mani	smooth	noise	orig	mani	smooth	noise
APP	Random	0.148	0.144	0.152	0.126	0.128	0.106	0.139	0.089	0.099	0.102	0.100	0.062
	H-Comp	0.100	0.074	0.104	0.090	0.100	0.074	0.097	0.090	0.100	0.107	0.090	0.057
	H-Ward	0.119	0.069	0.128	0.109	0.115	0.068	0.116	0.133	0.108	0.113	0.115	0.110
PairF	Random	36.23	28.97	38.69	29.83	32.41	30.91	34.96	26.40	27.72	31.96	31.92	14.91
	H-Comp	23.28	22.31	23.61	22.63	23.28	22.31	23.13	22.63	21.83	32.33	22.69	17.24
	H-Ward	29.93	21.98	30.77	26.99	28.94	22.22	30.93	31.68	27.32	30.90	29.67	26.47
R_a	Random	0.310	0.219	0.332	0.230	0.265	0.245	0.326	0.198	0.229	0.270	0.271	0.085
	H-Comp	0.154	0.140	0.156	0.146	0.154	0.140	0.151	0.146	0.160	0.267	0.167	0.110
	H-Ward	0.238	0.138	0.246	0.202	0.225	0.139	0.249	0.256	0.224	0.256	0.248	0.215

Table 5.18: Comparing distributions (frame+pp, reduced verb set)

Eval	Input	Distribution											
		prob				freq				bin			
		orig	mani	smooth	noise	orig	mani	smooth	noise	orig	mani	smooth	noise
APP	Random	0.060	0.060	0.062	0.057	0.054	0.047	0.052	0.044	0.030	0.039	0.036	0.015
	H-Comp	0.041	0.024	0.042	0.039	0.041	0.026	0.040	0.030	0.017	0.027	0.022	0.013
	H-Ward	0.038	0.031	0.039	0.044	0.041	0.033	0.037	0.033	0.024	0.035	0.023	0.015
PairF	Random	12.67	12.04	12.72	12.87	14.06	13.62	14.14	12.92	12.19	11.42	11.29	6.03
	H-Comp	11.31	9.91	11.27	10.23	12.59	10.21	11.27	10.75	8.16	8.83	9.13	3.22
	H-Ward	11.40	11.21	11.70	12.36	11.56	11.25	11.37	11.24	8.40	9.10	8.72	3.99
R$_a$	Random	0.090	0.077	0.090	0.092	0.102	0.098	0.102	0.089	0.089	0.075	0.081	0.034
	H-Comp	0.074	0.057	0.074	0.064	0.087	0.059	0.074	0.068	0.050	0.052	0.061	0.007
	H-Ward	0.079	0.071	0.081	0.087	0.080	0.070	0.076	0.064	0.057	0.057	0.060	0.015

Table 5.19: Comparing distributions (frame only, full verb set)

Eval	Input	Distribution											
		prob				freq				bin			
		orig	mani	smooth	noise	orig	mani	smooth	noise	orig	mani	smooth	noise
APP	Random	0.074	0.067	0.073	0.066	0.053	0.038	0.053	0.056	0.038	0.045	0.036	0.041
	H-Comp	0.042	0.029	0.040	0.042	0.039	0.031	0.040	0.044	0.034	0.035	0.028	0.031
	H-Ward	0.046	0.018	0.056	0.051	0.048	0.031	0.043	0.048	0.047	0.045	0.042	0.038
PairF	Random	14.98	12.04	15.37	15.09	14.82	14.15	15.07	14.72	13.25	13.62	12.67	13.98
	H-Comp	10.67	9.27	10.77	10.39	10.61	9.10	10.41	10.86	12.91	12.02	11.59	10.76
	H-Ward	10.57	9.84	13.71	13.27	11.65	9.24	9.98	10.95	14.04	13.25	12.91	10.71
R$_a$	Random	0.104	0.075	0.113	0.107	0.107	0.097	0.109	0.101	0.102	0.102	0.096	0.110
	H-Comp	0.064	0.047	0.065	0.061	0.061	0.045	0.069	0.063	0.096	0.083	0.084	0.076
	H-Ward	0.065	0.052	0.096	0.090	0.075	0.047	0.056	0.068	0.112	0.101	0.100	0.079

Table 5.20: Comparing distributions (frame+pp, full verb set)

Eval	Input	Similarity Measure									
		prob-orig						bin-orig			
		Cos	L1	Eucl	IRad	Skew	τ	Match	Dice	Jaccard	Overlap
APP	Random	0.139	0.141	0.139	0.145	0.150	0.093	0.119	0.095	0.095	-
	H-Comp	0.072	0.095	0.103	0.087	0.091	0.079	0.051	0.046	0.046	0.068
	H-Ward	0.102	0.105	0.117	0.101	0.102	0.077	0.058	0.077	0.081	0.020
PairF	Random	31.80	36.51	33.58	36.36	37.45	30.55	28.57	31.39	31.39	-
	H-Comp	22.78	27.08	30.23	23.50	22.89	27.07	18.33	16.38	16.38	15.24
	H-Ward	29.17	27.65	31.82	27.30	27.65	23.63	23.81	25.12	26.47	13.74
R_a	Random	0.259	0.314	0.280	0.310	0.327	0.246	0.223	0.263	0.263	-
	H-Comp	0.153	0.203	0.239	0.160	0.154	0.210	0.118	0.090	0.090	0.066
	H-Ward	0.230	0.211	0.262	0.207	0.211	0.177	0.171	0.200	0.215	0.040

Table 5.21: Comparing similarity measures (frame only, reduced verb set)

Eval	Input	Similarity Measure									
		prob-orig						bin-orig			
		Cos	L1	Eucl	IRad	Skew	τ	Match	Dice	Jaccard	Overlap
APP	Random	0.148	0.167	0.155	0.171	0.147	0.073	0.036	0.036	0.036	0.036
	H-Comp	0.100	0.112	0.102	0.123	0.126	0.103	0.084	0.090	0.090	0.089
	H-Ward	0.119	0.130	0.095	0.160	0.167	0.147	0.079	0.121	0.098	0.055
PairF	Random	36.23	39.84	36.24	38.49	41.63	30.77	10.28	10.28	10.28	10.28
	H-Comp	23.28	24.02	28.37	30.62	33.78	28.24	17.31	24.49	24.49	27.52
	H-Ward	29.93	29.90	27.31	34.81	40.75	44.67	25.18	34.69	27.27	13.19
R_a	Random	0.310	0.350	0.307	0.334	0.370	0.255	0.041	0.041	0.041	0.041
	H-Comp	0.154	0.165	0.222	0.244	0.279	0.224	0.098	0.185	0.185	0.223
	H-Ward	0.238	0.236	0.215	0.293	0.358	0.410	0.188	0.304	0.225	0.029

Table 5.22: Comparing similarity measures (frame+pp, reduced verb set)

Eval	Input	Similarity Measure									
		prob-orig						bin-orig			
		Cos	L1	Eucl	IRad	Skew	τ	Match	Dice	Jaccard	Overlap
APP	Random	0.060	0.064	0.057	0.057	0.054	0.044	-	0.035	0.035	-
	H-Comp	0.041	0.030	0.036	0.033	0.032	0.036	0.028	0.014	0.014	0.012
	H-Ward	0.038	0.040	0.039	0.039	0.041	0.031	0.028	0.012	0.013	0.019
PairF	Random	12.67	13.11	13.85	14.19	14.13	13.51	-	11.11	11.11	-
	H-Comp	11.31	10.01	11.39	10.16	11.00	14.41	6.69	7.89	7.89	5.25
	H-Ward	11.40	13.65	12.88	13.07	12.64	10.34	7.73	7.88	7.68	5.31
R_a	Random	0.090	0.094	0.101	0.101	0.105	0.103	-	0.076	0.076	-
	H-Comp	0.074	0.059	0.075	0.065	0.072	0.113	0.025	0.045	0.045	0.007
	H-Ward	0.079	0.099	0.093	0.097	0.094	0.074	0.037	0.048	0.047	0.008

Table 5.23: Comparing similarity measures (frame only, full verb set)

Eval	Input	Similarity Measure									
		prob-orig						bin-orig			
		Cos	L1	Eucl	IRad	Skew	τ	Match	Dice	Jaccard	Overlap
APP	Random	0.074	0.066	0.073	0.061	0.063		-	0.044	0.044	-
	H-Comp	0.042	0.052	0.054	0.053	0.057	0.048	0.000	0.000	0.000	0.000
	H-Ward	0.046	0.051	0.045	0.066	0.068	0.060	0.030	0.038	0.036	0.026
PairF	Random	14.91	15.20	16.10	16.15	18.01	13.62	-	13.91	13.91	-
	H-Comp	10.67	12.73	12.27	14.44	13.81	16.62	4.84	4.84	4.84	4.84
	H-Ward	10.57	15.51	13.11	17.49	19.30	22.44	10.99	13.33	11.42	5.84
R_a	Random	0.104	0.109	0.123	0.118	0.142		-	0.107	0.107	-
	H-Comp	0.064	0.087	0.083	0.105	0.102	0.133	0.001	0.001	0.001	0.001
	H-Ward	0.065	0.116	0.092	0.142	0.158	0.192	0.076	0.104	0.088	0.013

Table 5.24: Comparing similarity measures (frame+pp, full verb set)

Eval	Distance	k-Means cluster initialisation		
		Manual	Random	
			best	avg
APP	IRad	0.181	0.022 → 0.145	0.108
	Skew	0.199	0.022 → 0.150	0.107
PairF	IRad	52.52	7.73 → 36.36	28.21
	Skew	60.30	2.00 → 37.45	28.65
R_a	IRad	0.490	-0.003 → 0.310	0.215
	Skew	0.577	-0.045 → 0.327	0.222

Eval	Distance	k-Means cluster initialisation				
		Hierarchical				
		single	complete	average	centroid	ward
APP	IRad	0.043 → 0.043	0.085 → 0.087	0.079 → 0.079	0.073 → 0.073	0.101 → 0.101
	Skew	0.043 → 0.043	0.091 → 0.091	0.068 → 0.068	0.062 → 0.062	0.102 → 0.102
PairF	IRad	20.08 → 20.08	21.61 → 23.50	21.46 → 21.46	21.49 → 21.49	27.30 → 27.30
	Skew	20.08 → 20.08	22.89 → 22.89	21.30 → 21.30	21.61 → 21.61	27.65 → 27.65
R_a	IRad	0.114 → 0.114	0.137 → 0.160	0.133 → 0.133	0.131 → 0.131	0.207 → 0.207
	Skew	0.114 → 0.114	0.154 → 0.154	0.130 → 0.130	0.133 → 0.133	0.211 → 0.211

Table 5.25: Comparing clustering initialisations (frame only, reduced verb set)

Eval	Distance	k-Means cluster initialisation		
		Manual	Random	
			best	avg
APP	IRad	0.248	0.033 → 0.171	0.110
	Skew	0.248	0.020 → 0.147	0.097
PairF	IRad	81.25	6.03 → 38.49	29.50
	Skew	81.25	7.73 → 41.63	28.52
R_a	IRad	0.801	-0.002 → 0.334	0.232
	Skew	0.801	0.014 → 0.370	0.224

Eval	Distance	k-Means cluster initialisation				
		Hierarchical				
		single	complete	average	centroid	ward
APP	IRad	0.092 → 0.101	0.123 → 0.123	0.123 → 0.123	0.081 → 0.081	0.160 → 0.160
	Skew	0.092 → 0.101	0.126 → 0.126	0.118 → 0.118	0.081 → 0.081	0.167 → 0.167
PairF	IRad	19.06 → 25.23	30.62 → 30.62	26.34 → 26.34	23.73 → 23.73	34.81 → 34.81
	Skew	19.06 → 25.23	33.78 → 33.78	25.85 → 25.85	23.73 → 23.73	40.75 → 40.75
R_a	IRad	0.097 → 0.175	0.244 → 0.244	0.189 → 0.189	0.156 → 0.156	0.293 → 0.293
	Skew	0.097 → 0.175	0.279 → 0.279	0.183 → 0.183	0.156 → 0.156	0.358 → 0.358

Table 5.26: Comparing clustering initialisations (frame+pp, reduced verb set)

Eval	Distance	k-Means cluster initialisation		
		Manual	Random	
			best	avg
APP	IRad	0.066	0.004 → 0.057	0.041
	Skew	0.074	0.004 → 0.054	0.040
PairF	IRad	18.56	2.16 → 14.19	11.78
	Skew	20.00	1.90 → 14.13	12.17
R_a	IRad	0.150	-0.004 → 0.101	0.078
	Skew	0.165	-0.005 → 0.105	0.083

Eval	Distance	k-Means cluster initialisation				
		Hierarchical				
		single	complete	average	centroid	ward
APP	IRad	0.016 → 0.028	0.031 → 0.033	0.030 → 0.031	0.019 → 0.025	0.039 → 0.039
	Skew	0.012 → 0.026	0.032 → 0.032	0.034 → 0.033	0.027 → 0.027	0.040 → 0.041
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
R_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

Table 5.27: Comparing clustering initialisations (frame only, full verb set)

Eval	Distance	k-Means cluster initialisation		
		Manual	Random	
			best	avg
APP	IRad	0.160	0.007 → 0.061	0.045
	Skew	0.171	0.004 → 0.063	0.042
PairF	IRad	40.23	1.34 → 16.15	13.37
	Skew	47.28	2.41 → 18.01	14.07
R_a	IRad	0.358	0.001 → 0.118	0.093
	Skew	0.429	-0.002 → 0.142	0.102

Eval	Distance	k-Means cluster initialisation				
		Hierarchical				
		single	complete	average	centroid	ward
APP	IRad	0.012 → 0.031	0.054 → 0.053	0.043 → 0.042	0.030 → 0.037	0.066 → 0.066
	Skew	0.014 → 0.026	0.058 → 0.057	0.046 → 0.046	0.022 → 0.029	0.068 → 0.068
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
R_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

Table 5.28: Comparing clustering initialisations (frame+pp, full verb set)

Eval	Input	Verb Description						
					specified		all	
		frame	ppS	ppA	ppS+prefS	ppA+prefA	ppA+prefA_NP	ppA+prefA_NP-PP
		[38]	[178]	[183]	[253]	[288]	[906]	[2,726]
APP	H-Comp	0.091	0.126	0.153	0.116	0.130	0.111	0.097
	H-Ward	0.102	0.167	0.145	0.136	0.150	0.145	0.138
PairF	H-Comp	22.89	33.78	37.40	30.90	29.86	35.57	28.27
	H-Ward	27.65	40.75	34.35	32.71	35.79	31.94	32.39
R$_a$	H-Comp	0.154	0.279	0.322	0.281	0.231	0.304	0.221
	H-Ward	0.211	0.358	0.289	0.271	0.302	0.260	0.265

Table 5.29: Comparing feature descriptions on reduced verb set

Eval	Input	Verb Description						
					specified		all	
		frame	ppS	ppA	ppS+prefS	ppA+prefA	ppA+prefA_NP	ppA+prefA_NP-PP
		[38]	[178]	[183]	[253]	[288]	[906]	[2,726]
APP	H-Comp	0.032	0.057	0.060	0.048	0.050	0.045	0.050
	H-Ward	0.041	0.068	0.067	0.069	0.064	0.066	0.067
PairF	H-Comp	11.00	13.81	18.34	16.25	19.03	17.72	14.02
	H-Ward	12.64	19.30	18.81	20.73	22.19	19.29	21.11
R$_a$	H-Comp	0.072	0.102	0.145	0.123	0.147	0.139	0.106
	H-Ward	0.094	0.158	0.151	0.168	0.182	0.158	0.176

Table 5.30: Comparing feature descriptions on full verb set

5.2.4 Summary

This section has presented the k-Means clustering setups, experiments and results. The experiments were based on various parameter settings concerning the verb distributions, the clustering input, and the similarity measures. The experiment results show that frequencies and probabilities are both useful for describing the verbs, either in their original form or as a smoothed version. As input clusters, hierarchical clusters based on complete-linkage or even more on Ward's amalgamation method, are most compatible with the k-Means algorithm. In fact, k-Means does not improve the results considerably, which is due to the similarity of the clustering methods with respect to the common clustering criterion of optimising the sum of distances between verbs and cluster centroids. Random input clusters are only useful for small sets of objects. Using the gold standard classes as input to the clustering process, the (non-desired) changes performed by k-Means point to deficiencies in the verb description, with respect to the desired classification; refining the verb description is reflected by less deficiencies in the clustering and therefore underlines the linguistic improvement of the description. With regard to similarity measures in clustering, there is no unique best performing method, but on larger feature sets the Kullback-Leibler variants information radius and even more skew divergence tend to be the most stable solutions.

The various choices of verb features illustrate that already a purely syntactic verb description allows a verb clustering clearly above the baseline. Refining the syntactic features by prepositional phrase information considerably improves the clustering results, but there is no consistent effect when adding the selectional preferences to the verb description. I assume that not only sparse data is responsible for the latter negligible improvement in clustering, but more importantly that linguistic reasons are directly relevant for the clustering outcome. The following clustering interpretation in Section 5.3 will investigate the correlations in more detail.

5.3 Experiment Interpretation

The clustering setup, proceeding and results provide a basis for a linguistic investigation concerning the German verbs, their empirical characteristics, syntactic properties and semantic classification. The interpretation is started by an analysis of the experiment outcomes in Section 5.3.1. In Section 5.3.2, a series of post-hoc cluster analyses explores the influence of specific frames and frame groups on the coherence of the verb classes.

5.3.1 Interpretation of Experiment Outcome

The first part of interpreting the cluster outcomes considers example clusterings for the various levels of feature definition. For each of the levels, a clustering is presented and described, with reference to the underlying feature values determining the respective clustering, and the semantic content of the verbs and verb classes.

The cluster analysis which is based on the coarse syntactic verb descriptions refers to the reduced set of verbs, providing an easy understanding of the clustering phenomena. The analysis is accompanied by its clustering pendant based on the refined version of verb descriptions where the prepositional phrase information substitutes the coarse p-frames. The more extensive verb descriptions containing selectional preferences are investigated for the full set of verbs, with references to the clustering pendants with restricted feature sets. All cluster analyses have been performed by k-Means with hierarchical clustering input (Ward's method) on probability distributions, with the similarity measure being skew divergence.

Coarse Syntactic Definition of Subcategorisation The following list of verbs represents the clustering output based on the coarse syntactic verb descriptions. The ordering of the clusters is irrelevant. The verbs in the clusters are sorted alphabetically; only for large clusters a visually easier ordering is given.

(1) ahnen$_{2}$ wissen$_{2}$

(2) denken$_{2}$ glauben$_{2}$

(3) anfangen$_{1}$ aufhören$_{1}$ rudern$_{5}$

(4) blitzen$_{14}$ donnern$_{14}$ enden$_{1}$ fahren$_{5}$ fliegen$_{5}$

(5) beginnen$_{1}$ beharren$_{9}$ bestehen$_{9}$ liegen$_{10}$ pochen$_{9}$ segeln$_{5}$ sitzen$_{10}$ stehen$_{10}$

(6) insistieren$_{9}$ saufen$_{13}$

(7) beschreiben$_{8}$ charakterisieren$_{8}$ darstellen$_{8}$ interpretieren$_{8}$
bekommen$_{3}$ erhalten$_{3}$ erlangen$_{3}$ kriegen$_{3}$
bringen$_{4}$ liefern$_{4}$ schicken$_{4}$ vermitteln$_{4}$
ankündigen$_{7}$ bekanntgeben$_{7}$ eröffnen$_{7}$
beenden$_{1}$
konsumieren$_{13}$
unterstützen$_{11}$

(8) zustellen$_{4}$

(9) dienen$_{11}$ folgen$_{11}$ helfen$_{11}$

(10) essen$_{13}$ lesen$_{13}$ schließen$_{12}$ trinken$_{13}$ öffnen$_{12}$

(11) freuen$_{6}$ ärgern$_{6}$

(12) verkünden$_{7}$ vermuten$_{2}$

(13) nieseln$_{14}$ regnen$_{14}$ schneien$_{14}$

(14) dämmern$_{14}$

Clusters (1) and (2) are sub-classes of the semantic verb class *Propositional Attitude*. The verbs agree in their syntactic subcategorisation of a direct object (`na`) and finite clauses (`ns-2, ns-dass`); *glauben* and *denken* are assigned to a different cluster, because they also appear as intransitives, and show especially strong probabilities for `ns-2`.

Cluster (3) contains the two *Aspect* verbs *anfangen* and *aufhören*, polluted by the verb *rudern* 'to row'. All *Aspect* verbs show a 50% preference for an intransitive usage, and a minor 20% preference for the subcategorisation of non-finite clauses. By mistake, the infrequent verb *rudern* (corpus frequency 49) shows a similar preference for `ni` in its frame distribution and therefore appears within the same cluster as the *Aspect* verbs. The frame confusion has been caused by parsing mistakes for the infrequent verb; `ni` is not among the frames possibly subcategorised by *rudern*.

Cluster (4) is formed by verbs from the semantic *Weather*, *Aspect* and *Manner of Motion* classes. All verbs show high probabilities for an intransitive usage (for the weather verbs, this is a learning confusion with the expletive, based on tag ambiguity) and for subcategorising a prepositional phrase. The *Manner of Motion* verbs additionally have a large probability for an transitive usage, and are therefore often assigned to a separate class, in other cluster analyses. As we will see below, adding information about the specific prepositional head used in the `np` frame helps to distinguish the verbs, since *Weather* verbs typically appear with a locative (adjunct), *Aspect* verbs with the specific preposition mit_{Dat}, and *Manner of Motion* verbs with directional prepositions.

Cluster (5) comprises three *Insistence* verbs (*bestehen*, *beharren*, *pochen*), all three *Position* verbs (*liegen*, *sitzen*, *stehen*), the *Aspect* verb *beginnen* and the *Manner of Motion* verb *segeln*. All verbs show strong preferences for (i) an intransitive usage (incorrect for the *Insistence* verbs), and (ii) subcategorising a prepositional phrase. Similarly to cluster (4), the verbs are distinguishable when adding prepositional head information: *beginnen* uses mit_{Dat}, *segeln* directional prepositions, the *Insistence* verbs auf_{Dat}, and the *Position* verbs locative prepositions.

A syntactic overlap in frame usage clusters the verbs *insistieren* and *saufen* into cluster (6): a strong preference for an intransitive usage, or transitively with a direct object, a subcategorised PP, or a finite clause (verb second). These statements in the frame distribution are partly correct, but contain severe noise; the noise might –once again– refer to the fact that both verbs are rather low frequent (corpus frequencies 36 and 80, respectively).

The 18 verbs in cluster (7) –I ordered them according to their semantic affinity, one class per line– comprise the complete verb classes *Description* and *Obtaining*, the verb classes *Supply* and *Announcement* with only one verb missing, plus three singletons. The verbs agree in an approximately 50% probability for the subcategorisation of a direct accusative object, and a substantial probability for an additional prepositional phrase (nap). Most of the verbs have additional frames with respect to their verb classes (e.g. *Supply* verbs subcategorise a ditransitive frame), but those seem to be ignored with the weight of agreeing material.

The singleton cluster (8) is defined by the *Supply* verb *zustellen*, which distinguishes itself from the other verbs in its class by a comparably strong preference for the ditransitive.

Cluster (9) correctly clusters three of the four *Support* verbs, based on their common strong preference for subcategorising an indirect dative object. The only missing verb is *unterstützen* –as expected– which needs an accusative object.

Cluster (10) comprises *Consumption* and *Opening* verbs, which is a frequent coincidence in many cluster analyses. The commonsense of the verbs is an approximately 20% probability of intransitive and 40% probability of transitive frames. Unfortunately, the *Opening* verbs do not show a distinguishable strong preference for their reflexive usage, as hoped.

Cluster (11) finds the two *Emotion* verbs with their characteristic reflexive usage (possibly with a PP adjunct), and minor probabilities for na and finite clauses (correct).

The two verbs in cluster (12) agree in a syntactic frame mixture which prevents them from clustering with their desired class: about 10% n (parsing noise), 30% na, possibly with a PP adjunct (another 20%, rather noisy), and about 20% for finite clauses.

Cluster (13) perfectly comprises *Weather* verbs, agreeing in their characteristic expletive behaviour. *dämmern* in cluster (14) is not contained in (13), because of its ambiguous usage, which models –next to its weather sense– a sense of understanding by various possible syntactic frames.

Syntactico-Semantic Definition of Subcategorisation with Prepositional Preferences The preceding clustering result and interpretation clearly demonstrate the potential for an improved cluster analysis, especially with respect to prepositional head refinements. The following list of verbs is a clustering result based on a frame description with PP refinement.

(1) ahnen$_2$ vermuten$_2$ wissen$_2$

(2) denken$_2$ glauben$_2$

(3) anfangen$_1$ aufhören$_1$ beginnen$_1$ enden$_1$ rudern$_5$

(4) beharren$_9$ insistieren$_9$ pochen$_9$

(5) liegen$_{10}$ sitzen$_{10}$ stehen$_{10}$

(6) donnern$_{14}$ fahren$_5$ fliegen$_5$

(7) bestehen$_9$ blitzen$_{14}$ segeln$_5$

(8) beschreiben$_8$ charakterisieren$_8$ darstellen$_8$ interpretieren$_8$
bekommen$_3$ erhalten$_3$ erlangen$_3$ kriegen$_3$
ankündigen$_7$ bekanntgeben$_7$ eröffnen$_7$
liefern$_4$ vermitteln$_4$
beenden$_1$
unterstützen$_{11}$

(9) bringen$_4$ schicken$_4$ zustellen$_4$

(10) dienen$_{11}$ folgen$_{11}$ helfen$_{11}$

(11) essen$_{13}$ konsumieren$_{13}$ lesen$_{13}$ saufen$_{13}$ schließen$_{12}$ trinken$_{13}$ verkünden$_7$ öffnen$_{12}$

(12) freuen$_6$ ärgern$_6$

(13) nieseln$_{14}$ regnen$_{14}$ schneien$_{14}$

(14) dämmern$_{14}$

Clusters (1) and (2) together constitute the complete set of *Propositional Attitude* verbs. Again, the verbs are split over two classes because *glauben* and *denken* show especially strong probabilities for `ns-2`.

Cluster (3) now contains all *Aspect* verbs except for *beenden*. The verbs were formerly split over three clusters, but based on their common usage of prepositional phrases headed by *mit*$_{Dat}$ as well as time prepositions they form a more coherent class.

Clusters (4) and (5) successfully comprise and distinguish the *Insistence* and *Position* verbs formerly thrown together in one cluster, now distinguished by their relevant prepositions, *auf*$_{Dat}$ and locative prepositions, respectively. Similarly, the *Manner of Motion* verbs *fahren* and *fliegen* are distinguished by cluster (6) on basis of their directional prepositions, e.g. *durch*$_{Akk}$, *nach*$_{Dat}$, *zu*$_{Dat}$. *donnern* is assigned to the same cluster because of its possible motion sense referring to the sound emission, as in *Ein roter Polo donnert durch die schwarze Nacht* ‘A red polo rumbles through the black night’.

Cluster (7) represents an incoherent collection of three verbs which share a preference for an intransitive usage, but in addition only agree in using several possible prepositional phrase adjuncts. There is neither a close syntactic nor semantic relation.

Cluster (8) has changed by separating part of the *Supply* verbs into cluster (9), which now represents a correct semantic sub-class, and separating *konsumieren* correctly into the *Consumption* cluster (11). The remaining verbs are still characterised by their common subcategorisation of transitive direct objects.

Clusters (10) and (12)-(14) have not been expected to change, since they are distinguished by frames without distinctive prepositional phrases. They are identical to the previous cluster analysis. Cluster (11) has been improved and now comprises all *Consumption* verbs. As before, the verbs are mixed with *Opening* verbs, plus additionally *verkünden*. The verbs agree in their non-prepositional behaviour, as explained before.

Conclusion I Clearly, refining the syntactic verb information by prepositional phrases is helpful for the semantic clustering. This is the case because on the one hand, more structural information is provided concerning the usage of the verbs, and on the other hand the prepositions contain semantic content themselves, distinguishing e.g. locative and directional verb complementation. The detailed prepositional phrase information is not only useful in the clustering of verbs where the PPs are obligatory, but also in the clustering of verbs with optional PP arguments. For example, the *Consumption* verbs as well as the *Supply* verbs are clustered sufficiently, not because of obligatory PPs, but because of their similar usage of PP adjuncts (and, certainly, their non-usage of PP arguments, compared to other verbs).

This notion of PP knowledge in the verb description is confirmed by an experiment: eliminating all PP information from the verb descriptions (not only the delicate PP information, but also PP argument information in the coarse frames) produces obvious deficiencies in most of the semantic classes, among them *Weather* and *Support*, whose verbs do not require PPs as arguments.

Clusters such as (8) and (11) confirm the idea that selectional preferences should help distinguishing verbs from different classes. The verbs have similar strong preferences for a common frame (in this case: `na`), which is more specified for their semantics by additional selectional preferences. I assume that additional selectional preference information is too subtle for the reduced set of verbs, so I proceed the clustering investigation on the larger set of verbs.

Syntactico-Semantic Definition of Subcategorisation with Prepositional and Selectional Preferences Following a cluster analysis is presented which is based on the same clustering setup as above, the features being the frame plus additional prepositional phrase information and additional selectional preference information on specified frame-slots. The cluster analysis is described and compared with its pendants based on less verb information.

(1) ahnen$_{2}$ fürchten$_{15}$ vermuten$_{2}$ wissen$_{2}$

(2) anfangen$_{1}$ aufhören$_{1}$ rudern$_{11}$

(3) ankündigen$_{21}$ anordnen$_{22}$ bekanntgeben$_{21}$ empfinden$_{17}$ erkennen$_{24}$ interpretieren$_{25}$ scheuen$_{15}$ sehen$_{17}$

(4) basieren$_{40}$ beharren$_{28}$ beruhen$_{40}$ pochen$_{28}$

(5) bedürfen$_{4}$ dienen$_{34}$ folgen$_{34/41}$ helfen$_{34}$

(6) beenden$_{1}$ beschreiben$_{25}$ charakterisieren$_{25}$ eröffnen$_{21}$ realisieren$_{24}$ registrieren$_{24}$ unterstützen$_{34}$ veranschaulichen$_{26}$ wahrnehmen$_{17}$

(7) beginnen$_{1}$ bestehen$_{28/37}$ enden$_{1}$ existieren$_{37}$ laufen$_{8}$ liegen$_{31}$ sitzen$_{31}$ stehen$_{31}$

(8) beibringen$_{29}$ leihen$_{6}$ schenken$_{6}$ vermachen$_{6}$

(9) bekommen$_{5}$ benötigen$_{4}$ brauchen$_{4}$ erhalten$_{5}$ erneuern$_{33}$ gründen$_{40}$ herstellen$_{32}$ kriegen$_{5}$ schicken$_{7}$

(10) bemerken$_{24}$ erfahren$_{17/24}$ feststellen$_{24}$ hören$_{17}$ lesen$_{38}$ rufen$_{18}$ verkünden$_{21}$

(11) bestimmen$_{22}$ bringen$_{7}$ darstellen$_{25/26}$ erlangen$_{5}$ erzeugen$_{32}$ hervorbringen$_{32}$ liefern$_{7}$ produzieren$_{32}$ stiften$_{6}$ treiben$_{12}$ vermitteln$_{7/29}$ vernichten$_{39}$

(12) bilden$_{32}$ erhöhen$_{35}$ festlegen$_{22}$ senken$_{35}$ steigern$_{35}$ vergrößern$_{35}$ verkleinern$_{35}$

(13) erniedrigen$_{35}$

(14) geben$_{6}$

(15) denken$_{2}$ folgern$_{41}$ glauben$_{2}$ versichern$_{23}$

(16) demonstrieren$_{26}$ lehren$_{29}$

(17) blitzen$_{43}$ insistieren$_{28}$ rotieren$_{9}$

(18) donnern$_{43}$ hasten$_{10}$ heulen$_{14/19}$

(19) eilen$_{10}$ gleiten$_{12}$ kriechen$_{8}$ rennen$_{8}$ starren$_{16}$

(20) fahren$_{11}$ fliegen$_{11}$ fließen$_{12}$ klettern$_{8}$ segeln$_{11}$ wandern$_{8}$

(21) drehen$_{9}$ ergeben$_{42}$ stützen$_{40}$

(22) eliminieren$_{39}$ exekutieren$_{39}$

(23) töten$_{39}$ unterrichten$_{29}$

(24) entfernen$_{39}$ legen$_{30}$ präsentieren$_{26}$ schließen$_{36/40}$ setzen$_{30}$ stellen$_{30}$ öffnen$_{36}$

(25) erhoffen$_{3}$ wünschen$_{3}$

(26) erwachsen$_{41}$ resultieren$_{41}$

(27) essen$_{38}$ konsumieren$_{38}$ spenden$_{6}$ trinken$_{38}$

(28) flüstern$_{18}$ schleichen$_{8}$

(29) gehen$_{8}$ riechen$_{17}$

(30) freuen$_{13}$ fühlen$_{17}$ ärgern$_{13}$

(31) ängstigen$_{15}$

(32) ekeln$_{15}$

(33) grinsen$_{16}$ grübeln$_{27}$ jammern$_{19}$ klagen$_{19}$ lachen$_{14/16}$ lächeln$_{16}$ schreien$_{18}$ weinen$_{14}$

(34) gähnen$_{16}$ lamentieren$_{19}$

(35) kommunizieren$_{20}$ leben$_{37}$ nachdenken$_{27}$ reden$_{20}$ spekulieren$_{27}$ sprechen$_{20}$ verhandeln$_{20}$

(36) korrespondieren$_{20}$

(37) phantasieren$_{27}$ saufen$_{38}$

(38) renovieren$_{33}$ reparieren$_{33}$

(39) dekorieren$_{33}$

(40) versprechen$_{23}$ wollen$_{3}$ zusagen$_{23}$

(41) vorführen$_{26}$ zustellen$_{7}$ überschreiben$_{6}$

(42) nieseln$_{43}$ regnen$_{43}$ schneien$_{43}$

(43) dämmern$_{43}$

Cluster (1) contains three *Propositional Attitude* verbs, together with the *Emotion* verb *fürchten*. Semantically, *fürchten* does fit into the class, because it also expresses a propositional attitude, with an additional emotional denotation. Syntactically, the common cluster is based on similar preferences for the frames `na, ns-dass`. In addition, the role preferences on `na` are similar for a living entity as subject and a thing or situation as direct object.
The *Propositional Attitude* verbs *denken* and *glauben* are split into a separate cluster (15); as said before, the splitting is caused by stronger preferences for `ns-2`. The verbs are clustered together with the *Inference* verb *folgern* and the *Promise* verb *versichern*, which share the frame preferences –including the selectional preferences, mainly living entities as subjects. Semantically, *folgern* does have a meaning of thinking, which it shares with *denken* and *glauben*, *versichern* shares a sense of saying with the two verbs.
The respective clusters are identical with only PP refinement on the frames, i.e. the refinement by selectional preferences is not crucial for the cluster formation.

In cluster (2), we find the two *Aspect* verbs *anfangen* and *aufhören* together with *rudern*, based on the common `ni` frame. The two verbs are in different clusters than *beginnen* and *enden* – cluster (7), because the former have stronger preferences for an intransitive usage and relevant selectional preferences (mainly: situation), and the latter have stronger preferences for subcategorising a PP (head: mit_{Dat}). The split is the same when basing the clustering on the less refined verb descriptions.

Cluster (3) contains verbs from the verb classes *Announcement* (*ankündigen, bekanntgeben*) and *Constitution* (*anordnen*), all sub-classes of *Statement*, together with two *Perception* verbs and three single verbs from other classes, with *erkennen* having a similar meaning to the *Perception* verb *sehen*. The verbs in this cluster agree in a strong subcategorisation for a direct accusative object, including the specified selectional preferences in the frame, a living entity as subject and a situation or thing as object. In addition, they subcategorise `nap` with an obligatory PP for the transitive frame. The meaning differences of the verbs are subtle, such that only selectional preferences on a fine-grained level could capture them. The coarse definitions I use, help the clustering (which is better than without the role descriptions) but do not represent the necessary details for distinguishing the verbs.

The verbs in cluster (4) are mainly affected by the common strong syntactic preference for subcategorising a PP with head *auf* and dative case for the former three verbs, accusative case for *pochen*. In addition, the verbs show a similar strong preference for intransitive usage, which is only justified for *beharren* and *pochen*. Semantically, the verbs are from two different classes, but related in their meaning. In a cluster analysis without selectional preferences, *pochen* is not found as belonging to this cluster, so obviously the additional preferences help to overcome the PP frame difference (referring to the preposition case).

Cluster (5) is syntactically based on the subcategorisation of an indirect dative object, correctly constituting the *Support* verb class, incorrectly including the *Desire* verb *bedürfen*. The latter verb is special in its subcategorisation, demanding a genitive object, which is not coded in the grammar. Therefore, the most similar noun phrase, the dative NP, determines the verb behaviour. As said before, the *Support* class can be instantiated by the coarse verb description, without selectional preference information.

Similarly to cluster (3), cluster (6) contains verbs from various semantic classes, mainly *Observation*, *Perception*, *Description*, which obviously share the semantic meaning components of watching and realising. The class constitution is determined by main preferences for `na` (with a living entity as subject and a situation/thing as direct object) and `nap`, also similar to cluster (3). The union of clusters (3) and (6) would constitute the majority of the semantic classes above, so the selectional preferences create an unnecessary cut between the classes.

Cluster (7) contains verbs of *Aspect*, *Existence* and *Position*. Admittedly, they are also close in their semantics, with a common sense of existence. The additional verb *laufen* fits into the cluster with its sense of ‘working’. Syntactically, the verbs are similar in their intransitive usage and subcategorisation of PPs. The prepositional semantics is captured by diverse locative heads, such as in_{Dat}, auf_{Dat}, an_{Dat}. The ambiguity of the latter preposition referring to a point of time causes the union of the *Aspect* with the other verbs. For this cluster, the selectional preferences not necessarily constitute an improvement. With only the PP refinement, the *Position* verbs are correctly classified in a pure cluster.

Cluster (8) is constituted by *Gift* verbs and a *Teaching* verb, which share a strong syntactic ditransitive behaviour. The selectional preferences (particularly on the accusative slot in the verb description, but also on other frame slots) are similar and could only be distinguished by subtle roles, which are not realised in the verb description. But considering the fact that *beibringen* also means giving something (→ knowledge) to somebody, the cluster is considerably clean. Cluster analyses based on less verb information group a part of these verbs together, but succeed in a smaller cluster only.

In cluster (9), we find a semantically interesting and coherent group of *Need* and *Obtaining*, *Production* and *Renovation* verbs. They can be summarised by a common sense of need and achievement (by different means). *gründen* is in class *Basis*, but except for its meaning of ‘to be based on' it also has a meaning of 'to found'. *schicken* does only belong to this class union through the *Giving–Obtaining* relation. Syntactically, all verbs agree in a strong preference for a direct accusative object. The selectional preferences for the subject are restricted to living entities, but variable for the object. Many verbs also specify a purpose in the frame, by `nap` with *für*$_{Akk}$ or *zu*$_{Dat}$.

Cluster (10) basically contains verbs of *Observation*, *Perception* and *Announcement*, with some noise. I have already stated the similarity of *Observation* and *Perception* verbs. Since *feststellen* has both a meaning of observation and of announcing, the related *Announcement* verb *verkünden* is clustered here as well. *rufen*, in addition, has a sense of manner of announcement, so it fits to *verkünden*. The verbs do not show strong overlap in frame usage, but agree to some degree in `n` and `na`, mainly with a living entity as subject, and the subcategorisation of finite clauses of diverse kinds. Without selectional preferences (with and without PP refinement), the cluster actually contains less noise.

The core of cluster (11) is determined by verbs of the related classes *Production* (*erzeugen, hervorbringen, produzieren*) and *Giving* (*bringen, liefern, stiften, vermitteln*), with diverse noise. The verbs in the cluster agree in a strong preference for a direct object. The selectional preferences seem variable, both for the subject and the object. In addition to `na`, there are minor preferences for `nap` (mainly with *in*$_{Dat}$) and `nad`. The respective cluster contains more noise without the selectional preference information.

Cluster (12) contains most verbs of *Quantum Change*, together with one verb of *Production* and *Constitution* each. The semantics of the cluster is therefore rather pure. The verbs in the cluster also typically subcategorise a direct accusative object, but the frame alternates with a reflexive usage, `nr` and `npr` with mostly *auf*$_{Akk}$ and *um*$_{Akk}$. The selectional preferences help to distinguish this cluster: in a cluster analysis based on `frame+pp` the number of correct verbs is smaller and the noise larger. The verbs often demand a thing or situation as subject, and various objects such as attribute, cognitive object, state, structure or thing as object. The only missing change of quantum verb *erniedrigen* is split into a singleton cluster (13), probably because it is not as frequently used as reflexive. Without selectional preferences, the change of quantum verbs are not found together with the same degree of purity.

geben also represents an own cluster (14). Syntactically, this is caused by being the only verb with a strong preference for `xa`. From the meaning point of view, this specific frame represents an idiomatic expression, only possible with *geben*. The respective frame usage overlaps the *Giving* sense of the verb.

demonstrieren (*Presentation* class) and *lehren* (*Teaching* class) in (16) are a typical pair in the cluster analyses, semantically similar in the sense of showing somebody something. Syntactically, the commonality is based on similar probabilities for the frames `na, n, nad, np`.

The three verbs in cluster (17) have nothing in common concerning their meaning. Their clustering is based on a similar strong preference for an intransitive usage, which is accidentally confused with the expletive in the case of the *Weather* verb *blitzen*.

Cluster (18) seems equally confused on the first sight, but the three verbs *donnern* (*Weather*), *hasten* (*Rush*) and *heulen* (*Emotion/Moaning*) can all express a manner of motion, in the first and third case based on the respective sound of emission. This meaning is expressed by a strong preference for an intransitive (with selectional preferences demanding a thing) as well as subcategorising a prepositional phrase, often headed by $durch_{Akk}$.

The verbs in clusters (19) and (20) represent almost pure sub-classes of *Manner of Motion* verbs. All verbs alternate between a purely intransitive usage and subcategorising a PP, with diverse directional heads, e.g. $nach_{Dat}$, zu_{Dat}, in_{Akk}. It is not clear to me why the verbs are split into two clusters in exactly this way. The *Manner of Motion* verbs are not much dependent on the selectional preference information. The PP description seems sufficient to distinguish them.

As in cluster (17), the three verbs in cluster (21) have nothing in common concerning their meaning. In this case, their clustering is based on a strong syntactic preference for `npr`, but already the syntactic realisation and the semantic contributions of the prepositions are clearly different.

Cluster (22) is a small but perfect sub-class of *Elimination*. Both verbs in the cluster have strong syntactic preferences for `na`, with strong selectional preferences for living entities in both the subject and the object slot. The selectional preferences are responsible for the successful clustering, without them the verbs are split into different clusters. The verbs in cluster (23) are very similar in their behaviour to those in cluster (22), and *töten* is actually an *Elimination* verb, but *unterrichten* is a *Teaching* verb. The selectional behaviour of all verbs is very similar, though, and could not be distinguished in an obvious way.

Cluster (24) mainly contains verbs of *Bring into Position* and *Opening* which is essentially nothing else than a special case of bringing something into a certain position. The verbs agree in strong preferences for `na` and `nap`, with basically the verbs of *Bring into Position* demanding auf_{Akk} and in_{Akk}, the verbs of *Opening* demanding instrumental prepositions such as mit_{Dat}. The selectional preferences appear important for this cluster, without them the verbs are split over several clusters.

Clusters (25) and (26) are pure sub-classes of *Wish* and *Result*, respectively. Both clusters are characterised by special syntactic behaviour, the former by `nar` and the latter by `np` with aus_{Dat}. For cluster (26), the coarse syntactic behaviour is distinctive enough to cluster the respective verbs, without further preference information.

Cluster (27) mainly contains *Consumption* verbs, except for *spenden*, rather an opposite being of class *Gift*. As expected, the *Consumption* verbs alternate between `n` with a living entity realisation and `na` with the same as subject and food as object. For *konsumieren*, the selectional preferences for the objects are more variable. The selectional preferences are essential for the formation of the cluster.

Clusters (28) and (29) confuse verbs from different classes because of partly similar syntactic behaviour. *flüstern* and *schleichen* agree in a similar preference for the intransitive, specifically with a living entity; the other frame probabilities differ from each other. *gehen* and *riechen* are probably clustered together because of an overlap in the *riechen*-specific frame `xp`. *gehen* is an ambiguous verb with many frame realisations, among them `xp`.

Clusters (30) to (32) contain verbs of *Emotion*, with the exception of *fühlen* which has not been classified as *Emotion* verb (but should). The three verbs in cluster (30) agree in strong preferences for `nr` and `npr` with the preposition mainly being *über*$_{Akk}$. Differences in the selectional preferences in `na` (thing as subject, living entity as object for *ärgern* and *freuen*, the opposite for *fühlen*) are overlapped by the strong reflexive characteristics, so the cluster is formed in the same way without the selectional preferences. *ekeln* and *ängstigen* use a different preposition to express the cause of the emotion, *vor*$_{Akk}$.

The verbs in cluster (33) are from the semantic classes *Facial Expression*, *Moaning*, *Speculation*, *Manner of Articulation*, *Emotion*, but all refer to the expression of emotion, by face or by voice. The commonality is realised by a strong preference for intransitive usage (almost exclusively with a living entity), a verb second finite clause, and a prepositional phrase, often *über*$_{Akk}$. The two verbs in cluster (34) should also belong to (33), but do not appear with as strong preferences as the previous verbs.

Except for *leben*, all verbs in clusters (35) and (36) express communication. The verbs belong to the semantic classes *Communication* and *Speculation* and preferably use `n` with strong preferences for living entities, and `np` with *mit*$_{Dat}$ in case of communication, and *über*$_{Akk}$ in case of speculation. The coarse syntactic environment of the verbs is almost sufficient to distinguish them from other semantic classes; without further information, most of the verbs are clustered correctly on basis of the coarse frames only. With PP information, the cluster output is rather cleaner than with the selectional preferences in addition.

phantasieren and *saufen* represent an incoherent cluster (37). There is no obvious overlap except for an intransitive usage (with living entity). Both verbs are low frequent verbs (corpus frequencies of 26 and 80, respectively).

Clusters (38) and (39) both contain verbs of *Renovation*, unfortunately *dekorieren* is split from the other two. Frame overlap appears in `na`, with typical selectional preferences on the direct object being thing and place. Differently to the other two verbs, *dekorieren* has an additional meaning of adding some decoration when ‘renovating’ it and therefore subcategorises a PP with *mit*$_{Dat}$.

Cluster (40) contains verbs of *Promise* and *Wish*. Obviously, there is some close semantic relation between the verbs. The verbs agree in an alternation behaviour on `na` with typically a living entity as subject and a situation as object, `nad` and subcategorising finite (verb second) and non-finite clauses.

Cluster (41) comprises two *Giving* verbs and the *Presentation* verb *vorführen*. The three verbs agree in a strong preference for the ditransitive, plus a strong preference for `na`. There is no typical selectional preferences on the relevant frames.

Clusters (42) and (43) are identical to the smaller clusterings above. The common expletive frame preferences are so strong that no further information destroys their effect.

Conclusion II The description and interpretation of the clustering results gives insight into the relationship between verb properties and clustering outcome. Following, I first summarise minor issues, before a more extensive discussion concerning the relevance of the feature choice takes place.

- The fact that there are verbs which are clustered semantically on basis of their corpus-based and knowledge-based empirical properties, indicates (i) a **relationship between the meaning components of the verbs and their behaviour**, and (ii) that the clustering algorithm is able to benefit from the linguistic descriptions and to abstract from the noise in the distributions. The relationship between verb properties and semantic clusters is investigated in more detail in the following Section 5.3.2.

- The **verb properties** determining the cluster membership are (i) **observable** in the verb distributions. But with an increasing number of features, the intuitive judgement about strength and proportions of the feature values is growing more complicated. In addition, (ii) the description of verb properties by automatic means is as **expected**, i.e. capturing the features in a way we have expected. But some feature values determining the cluster membership are due to parsing noise, especially with respect to the intransitive frame type n.

- The **low frequency** verbs are noisier than verbs with larger frequencies and constitute noisy clusters. The cluster description pointed to example verbs with total corpus frequencies below 50.

- The interpretation of the clusterings unexpectedly points to meaning components of verbs which have not been discovered by the manual classification before. Example verbs are *fürchten* expressing a propositional attitude which includes its more basic sense of an *Emotion* verb, and *laufen* expressing not only a *Manner of Motion* but also a kind of existence when used in the sense of operation. The **discovering effect** should be larger with an increasing number of verbs, since the manual judgement is more difficult, and also with a soft clustering technique, where multiple cluster assignment is enabled.

- In a similar way, the clustering interpretation exhibits **semantically related verb classes**: verb classes which are separated in the manual classification, but semantically merged in a common cluster. For example, *Perception* and *Observation* verbs are related in that all the verbs express an observation, with the *Perception* verbs additionally referring to a physical ability, such as hearing.

- Related to the preceding issue, the **verb classes** as defined in Chapter 2 are demonstrated as **detailed** and **subtle**. Compared to a more general classification which would appropriately merge several classes, the clustering confirms that I have defined a difficult task with subtle classes. I was aware of this fact but preferred a fine classification, since it allows insight into more verb and class properties. But in this way, verbs which are similar in meaning are often clustered wrongly with respect to the gold standard.

The description and interpretation of the extended clustering illustrates that the definition of selectional preferences once more improves the clustering results. But the improvement is not as persuasive as in the first step, when refining the purely syntactic verb descriptions by prepositional information. Why is that? The effect could be due to (i) noisy or (ii) sparse data, but the example distributions in Tables 5.11 and 5.12 demonstrate that –even if noisy– the basic verb descriptions appear reliable with respect to their desired linguistic content, and Tables 5.29 and 5.30 illustrate that even with adding little information (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) the effect exists.

Why do we encounter an unpredictability concerning the encoding and effect of verb features, especially with respect to selectional preferences? The clustering has presented evidence for a linguistically defined limit on the usefulness of the verb features, which is driven by the **idiosyncratic properties of the verbs**. Compare the following representative parts of the cluster analysis.

(i) The weather verbs in cluster (42) strongly agree in their syntactic expression and do not need feature refinements for a successful class constitution. *dämmern* in cluster (43) is ambiguous between a weather verb and expressing a sense of understanding; this ambiguity is idiosyncratically expressed by the syntactic features, so *dämmern* is never clustered together with the other weather verbs.

Summarising, the syntactic features are sufficient for some verb classes to distinguish them from others, and any refining information does not change the classes.

(ii) *Manner of Motion*, *Existence*, *Position* and *Aspect* verbs are similar in their syntactic frame usage and therefore merged together on the purely syntactic level, but adding PP information distinguishes the respective verb classes: *Manner of Motion* verbs primarily demand directional PPs, *Aspect* verbs are distinguished by patient mit_D and time and location prepositions, and *Existence* and *Position* verbs are distinguished by locative prepositions, with *Position* verbs showing more PP variation. The PP information is essential for successfully distinguishing these verb classes, and the coherence is partly destroyed by adding selectional preferences: *Manner of Motion* verbs (from the sub-classes 8-12) are captured well by clusters (19) and (20), since they inhibit strong common alternations, but cluster (7) merges the *Existence*, *Position* and *Aspect* verbs, since verb-idiosyncratic demands on selectional roles destroy the PP-based class demarcation. Admittedly, the verbs in cluster (7) are close in their semantics, with a common sense of (bringing into vs. being in) existence. Schumacher (1986) actually classifies most of the verbs into one existence class. *laufen* fits into the cluster with its sense of ‘to function’.

Summarising, (i) some verb classes are not distinguished by purely syntactic information, but need PPs. In addition, (ii) correct semantic verb classes might be destroyed by refining the features, since the respective verbs do not agree with each other and differ from verbs in other classes strongly enough.

(iii) Cluster (12) contains most verbs of *Quantum Change*, together with one verb of *Production* and *Constitution* each. The semantics of the cluster is therefore rather pure. The verbs in the cluster typically subcategorise a direct object, alternating with a reflexive usage, 'nr' and 'npr' with mostly auf_{Akk} and um_{Akk}. The selectional preferences help to distinguish this cluster: the verbs agree in demanding a thing or situation as subject, and various objects such as attribute, cognitive object, state, structure or thing as object. Without selectional preferences, the change of quantum verbs are not found together with the same degree of purity.

Summarising, some verb classes need not only syntactic information and PPs, but selectional preferences to be distinguished from other classes.

(iv) There are verbs such as *töten* and *unterrichten* in cluster (23), whose properties are similar on each level of description, so a common cluster is established, but the verbs only have coarse common meaning components. Such verbs would need a finer version of selectional preferences to be distinguished.

Summarising, some verb classes cannot be distinguished by the verb features I provide, but would need finer features.

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

5.3.2 Feature Manipulation and Class Coherence

In order to directly illustrate the tight connection between the lexical meaning components of the verbs and their behaviour, this section performs a series of post-hoc cluster analyses to explore the influence of specific frames and frame groups on the coherence of the verb classes. For example, what is the difference in the clustering result (on the same starting clusters) if we deleted all frame types containing an expletive *es* (frame types including x)? Once again, the experiments are proceeded on the reduced set of verbs, in order to facilitate the interpretation of the feature variation.

The reference clustering for the experiments is the cluster analysis performed by k-Means with hierarchical clustering input (Ward's method) on probability distributions, with the similarity measure being skew divergence. The feature set contains PPs substituting the coarse syntactic `p`-frames. The cluster analysis is repeated here.

(1) ahnen$_2$ vermuten$_2$ wissen$_2$

(2) denken$_2$ glauben$_2$

(3) anfangen$_1$ aufhören$_1$ beginnen$_1$ enden$_1$ rudern$_5$

(4) beharren$_9$ insistieren$_9$ pochen$_9$

(5) liegen$_{10}$ sitzen$_{10}$ stehen$_{10}$

(6) donnern$_{14}$ fahren$_5$ fliegen$_5$

(7) bestehen$_9$ blitzen$_{14}$ segeln$_5$

(8) beschreiben$_8$ charakterisieren$_8$ darstellen$_8$ interpretieren$_8$
bekommen$_3$ erhalten$_3$ erlangen$_3$ kriegen$_3$
ankündigen$_7$ bekanntgeben$_7$ eröffnen$_7$
liefern$_4$ vermitteln$_4$
beenden$_1$
unterstützen$_{11}$

(9) bringen$_4$ schicken$_4$ zustellen$_4$

(10) dienen$_{11}$ folgen$_{11}$ helfen$_{11}$

(11) essen$_{13}$ konsumieren$_{13}$ lesen$_{13}$ saufen$_{13}$ schließen$_{12}$ trinken$_{13}$ verkünden$_7$ öffnen$_{12}$

(12) freuen$_6$ ärgern$_6$

(13) nieseln$_{14}$ regnen$_{14}$ schneien$_{14}$

(14) dämmern$_{14}$

By deleting a frame group from the verb description and then repeating the cluster analysis under the same conditions, a minimal pair of cluster analyses is created where the difference in clustering is supposedly the effect of the deleted frame group. To give an example, if the dative frames `nd, ndp` are taken from the verb description, most of the clusters in the clustering result are the same. But the coherence of the *Support* verbs in cluster (10) is destroyed: the verbs are split and distributed over other clusters, according to the remaining verb features. For example, *helfen* is assigned to the same cluster as two *Aspect* verbs, because of their common subcategorisation of non-finite clauses. Following the changed clusters are given, with the moved verbs underlined. (Of course, there are also changes with respect to other verbs, but those are ignored here.)

(3) anfangen$_1$ aufhören$_1$ helfen$_{10}$

(6) bestehen$_9$ donnern$_{14}$ fahren$_5$ fliegen$_5$ folgen$_{10}$

(7) blitzen$_{14}$ dienen$_{10}$

Deleting all finite clause frame types from the verb description causes mainly the verbs of *Propositional Attitude* to be split into other clusters. *denken* and *glauben* still remain in a common cluster because of their similarity as intransitives and subcategorising the specific PP with prepositional head *an*$_{Akk}$.

(2) $\underline{\text{denken}_{2}}$ $\underline{\text{glauben}_{2}}$

(8) $\underline{\text{ahnen}_{2}}$ $\underline{\text{vermuten}_{2}}$
beschreiben$_8$ charakterisieren$_8$ darstellen$_8$ interpretieren$_8$
bekommen$_3$ erhalten$_3$ erlangen$_3$ kriegen$_3$
ankündigen$_7$ bekanntgeben$_7$ eröffnen$_7$
liefern$_4$ vermitteln$_4$
beenden$_1$
unterstützen$_{11}$

(11) essen$_{13}$ konsumieren$_{13}$ lesen$_{13}$ schließen$_{12}$ trinken$_{13}$ verkünden$_7$ $\underline{\text{wissen}_{2}}$ öffnen$_{12}$

Without specifying features for the expletive, particularly the *Weather* verbs *nieseln, regnen, schneien* which formerly formed a coherent verb class are split over different clusters.

(3) anfangen$_1$ aufhören$_1$ $\underline{\text{nieseln}_{14}}$

(6) donnern$_{14}$ fahren$_5$ fliegen$_5$ $\underline{\text{regnen}_{14}}$

(14) dämmern$_{14}$ saufen$_{13}$ $\underline{\text{schneien}_{14}}$

Equivalent experiments were performed for each frame or frame group in the syntactic verb descriptions. The experiments illustrate the tight connection between the syntactic behaviour of the verbs and their meaning components, since a deleting of syntactic features is directly related to the coherence of the respective semantic classes.

5.3.3 Summary

This section has illustrated a tight connection between the induced verb behaviour and the constitution of the semantic verb classes. Additional profit from the clustering than expected concerns the detection of verb meaning components and the detection of relations between semantic classes. A number of low frequency verbs have presented themselves as difficult for clustering, since the verb descriptions are unreliable.

I demonstrated that the usefulness of verb features is limited by the specific properties of the desired verb classes, i.e. verb features referring to the common properties of verbs within a semantic class support the clustering, but verb features referring to the idiosyncratic properties of the verbs in a semantic class do not provide additional support for the clustering, but rather destroy coherent clusters. Since the properties of verbs in a common class depend on the semantic class, and the semantic classes exhibit properties on different levels of verb description, there is no unique perfect choice and encoding of the verb features.

5.4 Optimisation Criteria

This section discusses various ways to optimise the cluster analysis of the German verbs. The purpose of the section is to anticipate the reader's potential suggestions and objections concerning my choice of parameter setting, and to demonstrate that I applied a reasonable selection of parameters. Section 5.4.1 once more discusses the issue of feature variation: what other combinations of features have been tried or could be tried? Section 5.4.2 approaches the issue of feature choice from a practical point of view, applying a simple optimisation algorithm. In Section 5.4.3, the optimal number of clusters is discussed and varied. In Section 5.4.4, the problem of verb ambiguity is raised, and possibilities to handle the problem are illustrated.

5.4.1 Feature Variation

Now that the reader has gained an overview of what kind of features are used in the clustering experiments and what kind of effect they have on the cluster analysis of the German verbs, possible variations and extensions of the feature description are illustrated. I formerly described the feature choice and implementation on three levels. The following paragraphs pick up the distinction and discuss alternatives. Other features than the existing ones at the syntax-semantic interface are not mentioned in this section.

Coarse Syntactic Definition of Subcategorisation The verb description on the coarse level distinguishes 38 frame types. On this level, there is little room to vary the verb information. Possibilities for variation demand an extension or a change in the grammar and re-training, but are ignored because (i) on the one hand they are not considered as relevant, because the 38 frames cover the vast majority of the verb structures, and (ii) on the other hand they are not learned sufficiently, since further frames are rather infrequent or difficult to learn. To give some examples, rare frame types such as `naa` which are subcategorised by few verbs (e.g. *kosten*) could be coded in the grammar, but their few occurrences do rather confuse the learning of the different frame types than help distinguish them: e.g. the confusion of dative and accusative case in the grammar is strengthened when adding `naa` in addition to `nad`. In addition, subcategorised adjectives were coded in a previous grammar version, but they turned out unreliable and were therefore abandoned from the grammar.

To summarise, there is little potential in varying the coarse verb description. In addition, the main phenomena (according to German standard grammar, cf. Helbig and Buscha, 1998) are covered, sufficiently learned and successfully applied to clustering, so concentrating on marginal phenomena should provide little help to improve the cluster analysis.

Syntactico-Semantic Definition of Subcategorisation with Prepositional Preferences Various possibilities to include the prepositional phrases into the verb descriptions have already been discussed. Further variations of the PP information affect the amount of PP information refining the syntactic frames: (i) On the one hand, standard German grammar books such as Helbig and Buscha (1998) define a more restricted set of prepositional phrases than ours, since they distinguish categorise PPs with respect to their usage as arguments and adjuncts, and only argument PPs are relevant. (ii) In contrast, ignoring the constraint of 'reasonable corpus appearance' laid on the PP information increases the number and kinds of PPs in the frame, up to between 40 (on xp) and 140 (on np).

The clustering experiments on both the reduced and the full set of verbs are repeated, in order to compare the results based on the selected PP information in the previous experiments with both (i) the more restricted and (ii) the more generous inclusion of PPs. The experiments are performed on probability distributions, with the PP information either substituting or adding to the coarse frame types. As input, I choose hierarchical clusters, based on complete-linkage and Ward's method, similarity measure being the skew divergence. The results in Tables 5.31 and 5.32 demonstrate that in all PP experiments the cluster quality outperforms the clustering without PP information. But the differences in cluster quality vary depending on the input, the distribution and the evaluation measure, and there is no unique best performing PP distribution. Concluding, the PP varying experiments confirm the importance of prepositional phrase refinements in the syntactic frames; it appears that for larger sets of verbs the more detailed information becomes more relevant, but the exact effect of the PP information depends on the various experiment parameters.

Eval	Input	Distribution						
		frame	frame+ppS			frame+ppA		
			arg	chosen	all	arg	chosen	all
APP	H-Comp	0.091	0.125	0.126	0.122	0.126	0.153	0.160
	H-Ward	0.102	0.163	0.167	0.160	0.140	0.145	0.145
PairF	H-Comp	22.89	34.15	33.78	26.34	31.88	37.40	42.57
	H-Ward	27.65	38.31	40.75	34.81	33.46	34.35	34.35
R_a	H-Comp	0.154	0.284	0.279	0.189	0.256	0.322	0.380
	H-Ward	0.211	0.332	0.358	0.293	0.280	0.289	0.289

Table 5.31: Comparing the amount of PP information (reduced verb set)

Eval	Input	Distribution						
		frame	frame+ppS			frame+ppA		
			arg	chosen	all	arg	chosen	all
APP	H-Comp	0.032	0.064	0.057	0.062	0.057	0.060	0.055
	H-Ward	0.041	0.069	0.068	0.069	0.062	0.067	0.071
PairF	H-Comp	11.00	15.48	13.81	16.20	15.83	18.34	18.32
	H-Ward	12.64	19.71	19.30	18.08	18.53	18.81	19.65
R_a	H-Comp	0.072	0.119	0.102	0.122	0.119	0.145	0.146
	H-Ward	0.094	0.163	0.158	0.148	0.151	0.151	0.160

Table 5.32: Comparing the amount of PP information (full verb set)

Syntactico-Semantic Definition of Subcategorisation with Prepositional and Selectional Preferences The definition of selectional preferences leaves most room for variation.

Role Choice: The first issue to be discussed concerns the specificity of the role definition. I mentioned the potential of the grammar model to define selectional preferences on a fine-grained level, the word level. Obviously, with this amount of features in the verb description I would run into a severe sparse data problem, so I have not tried this variation. In contrast, I performed experiments which define a more generalised description of selectional preferences than 15 concepts, by merging the frequencies of the 15 top level nodes in GermaNet to only 2 (Lebewesen, Objekt) or 3 (Lebewesen, Sache, Abstraktum). The more general definition should suit the linguistic demarcation of the verb classes, but merging the frequencies resulted in noisy distributions and destroyed the coherence in the cluster analyses.

Role Integration. The way of integrating the selectional preferences into the verb description opens another source for variation. Remember the discussion whether to refine either single slots in the frame types, or slot-combinations. In order to repeat the main points of the discussion with respect to an optimisation of the verb features, the former solution is the more practical one, since the selectional preferences in the grammar are encoded separately on the frame slots, and the number of features remains within a reasonable magnitude; the latter solution is the more linguistic one, trying to capture the idea of alternations, but there is no ground for the combination in the grammar, and the number of features is unacceptable. I therefore based the experiments on the encoding of selectional preferences for single slots of the frames. Because of the sparse data problem, I have ignored the combination of argument slots.

Slot Choice: In order to choose the most informative frame roles in a linguistic way, I have provided a quantitative corpus analysis in Appendix B. Tables 5.33 and 5.34 present the clustering results when varying the slots in a more practical way, by considering only single slots for selectional preference refinements, or small combinations of argument slots. The variations are supposed to provide insight into the contribution of slots and slot combinations to the clustering. The experiments are performed on probability distributions, with PP and selectional preference information given in addition to the syntactic frame types. As input, I choose hierarchical clusters, based on complete-linkage and Ward's method, similarity measure being the skew divergence.

Table 5.33 shows that refining only a single slot (the underlined slot in the respective frame type) in addition to the `frame+pp` definitions results in no or little improvement. There is no frame-slot type which consistently improves the results, but the success depends on the parameter instantiation. Obviously, the results do not match linguistic intuition. For example, we would expect the arguments in the two highly frequent intransitive `n` and transitive `na` to provide valuable information with respect to their selectional preferences, but only those in `na` do improve `frame+pp`. On the other hand, `ni` which is not expected to provide variable definitions of selectional preferences for the nominative slot, does work better than `n`.

Eval	Input		Selectional Preferences for frame+ppA+prefA					
		frame+ppA	$\underline{\text{n}}$	$\underline{\text{n}}\text{a}$	$\text{n}\underline{\text{a}}$	$\underline{\text{n}}\text{ad}$	$\text{n}\underline{\text{a}}\text{d}$	$\text{na}\underline{\text{d}}$
APP	H-Comp	0.060	0.046	0.059	0.053	0.047	0.055	0.053
	H-Ward	0.067	0.068	0.059	0.071	0.064	0.065	0.068
PairF	H-Comp	18.34	15.42	16.97	19.92	16.46	18.99	18.45
	H-Ward	18.81	16.22	21.15	20.19	17.82	15.13	19.48
R_a	H-Comp	0.145	0.117	0.134	0.159	0.133	0.154	0.150
	H-Ward	0.151	0.125	0.176	0.164	0.144	0.115	0.161

Eval	Input		Selectional Preferences for frame+ppA+prefA						
		frame+ppA	$\underline{\text{n}}\text{d}$	$\text{n}\underline{\text{d}}$	$\underline{\text{n}}\text{p}$	$\underline{\text{n}}\text{i}$	$\underline{\text{n}}\text{r}$	$\underline{\text{n}}\text{s-2}$	$\underline{\text{n}}\text{s-dass}$
APP	H-Comp	0.060	0.060	0.057	0.061	0.058	0.061	0.058	0.056
	H-Ward	0.067	0.063	0.069	0.055	0.069	0.061	0.061	0.069
PairF	H-Comp	18.34	20.65	18.75	17.40	17.68	19.46	17.64	17.16
	H-Ward	18.81	18.88	17.92	16.77	18.26	17.22	15.55	19.29
R_a	H-Comp	0.145	0.168	0.153	0.139	0.140	0.160	0.136	0.135
	H-Ward	0.151	0.152	0.143	0.133	0.148	0.136	0.121	0.156

Table 5.33: Comparing selectional preference slot definitions on full verb set

In Table 5.34, few slots are combined to define selectional preference information, e.g. $\underline{\text{n}}/\underline{\text{na}}$ means that the nominative slot in 'n', and both the nominative and accusative slot in 'na' are refined by selectional preferences. It is clear that the clustering effect does not represent a sum of its parts, e.g. both the information in $\underline{\texttt{n}}\texttt{a}$ and in $\texttt{n}\underline{\texttt{a}}$ improve Ward's clustering based on `frame+ppA` (cf. Table 5.33), but it is not the case that $\underline{\texttt{na}}$ improves the clustering, too. As in Table 5.33, there is no combination of selectional preference frame definitions which consistently improves the results. The specific combination of selectional preferences as determined pre-experimental actually achieves the overall best results, better than using 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 5.30.

Role Means: Last but not least, I could use a different means for selectional role representation than GermaNet. But since the ontological idea of WordNet has been widely and successfully used and I do not have any comparable source at hand, I have to exclude this variation.

The various experiments on feature variation illustrate (i) that selectional preference information on single slots does not result in a strong impact on the clustering, but enlarging the information to several linguistically relevant slots shows small improvements, (ii) that there is no unique optimal encoding of the features, but the optimum depends on the respective clustering parameters, (iii) the linguistic intuition and the algorithmic clustering results do not necessarily align, and (iv) that the way I chose to define and implement the features was near-optimal, i.e. there is no feature variation which definitely outperforms the former results.

Eval	Input		Selectional Preferences				
			ppA+prefA				
		frame+ppA	n	na	n/na	nad	n/na/nad
APP	H-Comp	0.060	0.046	0.059	0.052	0.054	0.059
	H-Ward	0.067	0.068	0.060	0.071	0.055	0.067
PairF	H-Comp	18.34	15.42	14.58	18.03	13.36	15.69
	H-Ward	18.81	16.22	17.82	17.00	13.36	16.05
R_a	H-Comp	0.145	0.117	0.099	0.137	0.091	0.114
	H-Ward	0.151	0.125	0.137	0.128	0.088	0.118

Eval	Input		Selectional Preferences			
			ppA+prefA			
		frame+ppA	nd	n/na/nd	n/na/nad/nd	np/ni/nr/ns-2/ns-dass
APP	H-Comp	0.060	0.060	0.058	0.055	0.061
	H-Ward	0.067	0.064	0.058	0.072	0.064
PairF	H-Comp	18.34	18.77	14.31	18.44	16.99
	H-Ward	18.81	18.48	16.48	20.21	16.73
R_a	H-Comp	0.145	0.149	0.100	0.136	0.135
	H-Ward	0.151	0.150	0.124	0.161	0.131

Table 5.34: Comparing selectional preference frame definitions on full verb set

5.4.2 Feature Selection

It is necessary to find a compromise between the time spending on the search for the optimal feature set and the gain in cluster quality performed on basis of the features. I believe that (i) there is no global optimal feature set for the clustering task, since the evaluation of clusterings depends on the number and kinds of verbs, the desired cluster number, the available features, etc. And (ii) the optimal set of features for a given setting is still a compromise between the linguistic and practical demands on the cluster analysis, but never optimal in both the linguistic and the practical sense.

Nevertheless, I aim to prove that at least a simple algorithm for feature selection does not choose a linguistically desired set. Two greedy algorithms are implemented which perform feature selection in the following ways: (i) *Bottom-Up:* The search for a feature selection starts with no features, i.e. an empty feature set. In a first step, a cluster analysis is performed with each of the features, and the feature which induces the cluster analysis with the best result is chosen into the feature set. In a second step, each of the remaining features is tried in addition to the singleton feature set, a cluster analysis is performed, and the feature which induces the cluster analysis with the best result is added to the feature set. In this way, a feature is added to the feature set as long as there is an improvement in clustering. If the cluster analysis does not improve any more by adding any of the remaining features to the feature set, the search is halted. (ii) *Top-Down:* The search for a feature selection starts with all features in the feature set. In a first step, a cluster analysis is performed for each of the features deleted from the feature set, and the (abandoned) feature which induces the cluster analysis with the best result is deleted from the feature set. In this way, features are deleted from the feature set as long as there is an improvement in clustering. If the cluster analysis does not improve any more by deleting any of the remaining features from the feature set, the search is halted.

The above idea was developed by myself, but a literature search encounters similar ideas. For general discussions on the feature selection issue in machine learning, the reader is referred to e.g. Langley (1994) and Blum and Langley (1997) for general reviews on the problem, or John *et al.* (1994) and Kohavi and John (1998), as well as Koller and Sahami (1997) for more specific approaches. My approach is close to the *Wrapper Model* for feature selection, as introduced by John *et al.* (1994). Differently to pre-existing *Filter Models*, which perform a feature selection only on basis of the meaning and importance of the features, the wrapper model performs a greedy search through the space of feature combinations on basis of a task-oriented evaluation, i.e. the feature sets are evaluated with respect to the overall learning task. Differently to my approach, the wrapper model allows both deleting and adding a feature in each step of the search, independent of whether the search is performed bottom-up or top-down.

In order to demonstrate that there is no unique optimal feature set, I perform the bottom-up and top-down feature selection on both the reduced and the full set of verbs, with reference to the evaluation measures *APP* and $Rand_{adj}$. The feature description of the relevant verbs is based on the coarse syntactic frames, which facilitates the interpretation of the results. Table 5.35 illustrates that the feature sets are far away from uniformity. In fact, depending on the search direction, the feature set and the verb set, the resulting ‘optimal’ feature sets vary severely. The only tendency I carefully induce from the experiments concerns a slight preference for rare frame types (such as nar, ndp) compared to frequently used types (such as n, na), so in a purely practical sense they might be more informative.

Eval	Search	Verb Set	
		reduced	full
APP	bottom-up	nrs-dass nds-w	ns-ob
	top-down	n na nad ndp ni nai nir nr nar ndr npr ns-2 nas-2 nrs-2 ns-dass ns-w nas-w x xp	na nd nad np nap ndp npr ni nai ndi nir nr nar ndr ns-2 nas-2 nrs-2 ns-dass ns-w ns-ob nas-ob xa xd xp
R_a	bottom-up	nai ndi nar nas-2 nrs-2 ns-dass nas-dass nrs-dass ns-w nas-w nds-w x xs-dass	nr nar ns-dass nrs-dass ns-w x xd
	top-down	nd nad np ndp nai nir nr nar ndr ns-2 nas-2 ns-dass nas-dass x	na nd nad np nap npr ni nai nir nar ndr ns-2 nas-2 ns-dass nas-dass nds-dass nrs-dass ns-w nas-w nds-w nrs-w ns-ob nas-ob xa xd xp xr

Table 5.35: Comparing optimal feature sets

5.4.3 Optimising the Number of Clusters

It is not a goal within this thesis to optimise the number of clusters in the cluster analysis. I am not interested in the question whether e.g. 40, 42, 43, or 45 clusters represent the better 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 a different numbers of clusters. (i) I should make sure that the clustering methodology basically works the way we expect, i.e. 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. And the optimisation experiments have been used to detect biases of the evaluation measures concerning cluster sizes. (ii) I raise the question whether it makes sense to select a different magnitude of number of clusters as the goal of clustering, i.e. 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.

Figures 5.3 to 5.8 show the clustering results for series of cluster analyses performed by k-Means with hierarchical clustering input (Ward's method) on probability distributions, with the similarity measure being skew divergence. The feature description refers to the coarse syntactic frames with substituting prepositional phrases. Both for the reduced and the full set of verbs I vary the number of clusters from 1 to the number of verbs (57/168) and evaluate the clustering results by APP, $PairF$ and $Rand_{adj}$. Figures 5.3 and 5.6 illustrate that APP finds an optimal clustering result for a small number of clusters (12/17), whereas $PairF$ (Figures 5.4 and 5.7) and $Rand_{adj}$ (Figures 5.5 and 5.8) determine a range of numbers of clusters as optimal (13/71) or near-optimal (approx. 12-14/58-78). Loosely saying, with an evaluation based on $PairF$ or $Rand_{adj}$ I stay on the safe side, since the cluster analysis contains many small clusters and therefore provides a high precision, and with an evaluation based on APP I create larger clusters, with semantically more general content.

Following I list the 17 clusters on the full verb set, as taken from the APP-optimal hierarchical cluster analysis. The semantic content of the clusters can roughly be described (ignoring the noise) as (1) *Propositional Attitude*, (2) *Aspect*, (4) *Basis, Insistence*, (5) *Support*, (6) *Wish, Gift*, (7) *Existence, Position*, (9) *Supply*, (11) *Propositional Attitude/Thinking*, (12) *Manner of Motion*, (13) *Result*, (14) *Emotion, Facial Expression*, (15) *Emotion*, (16) *Moaning, Communication*, (17) *Weather*. Admittedly, clusters (8) and (10) contain too much noise to dare giving them a label, and cluster (3) comprises verbs from too many different areas to label it.

(1) ahnen bemerken erfahren feststellen fürchten verkünden vermuten wissen

(2) anfangen aufhören beginnen enden korrespondieren rudern

(3) ankündigen anordnen beenden bekanntgeben bekommen benötigen beschreiben bestimmen brauchen charakterisieren darstellen dekorieren eliminieren empfinden erhalten erkennen erlangen erneuern erzeugen eröffnen exekutieren gründen herstellen hervorbringen interpretieren konsumieren kriegen liefern produzieren realisieren registrieren renovieren reparieren scheuen sehen senken stiften töten unterrichten unterstützen veranschaulichen verkleinern vermitteln vernichten wahrnehmen

(4) basieren beharren beruhen klettern pochen starren

(5) bedürfen dienen dämmern folgen helfen

(6) beibringen erhoffen leihen schenken vermachen wünschen

(7) bestehen blitzen demonstrieren existieren leben liegen segeln sitzen stehen

(8) bilden drehen ekeln ergeben erhöhen festlegen präsentieren steigern stellen stützen vergrößern ängstigen

(9) bringen legen schicken setzen treiben vorführen zustellen überschreiben

(10) entfernen erniedrigen essen geben hören lehren lesen schließen spenden trinken versprechen wollen zusagen öffnen

(11) denken folgern glauben versichern

(12) donnern eilen fahren fliegen fließen gehen gleiten kriechen laufen rennen riechen rufen wandern

(13) erwachsen resultieren

(14) flüstern grinsen gähnen hasten heulen insistieren lachen lächeln phantasieren rotieren saufen schleichen schreien sprechen weinen

(15) freuen fühlen ärgern

(16) grübeln jammern klagen kommunizieren lamentieren nachdenken reden spekulieren verhandeln

(17) nieseln regnen schneien

The cluster analysis illustrates that a semantic classification with the number of clusters in a much smaller magnitude than I tried in previous experiments might be a real alternative. In this case, the semantic content of the clusters is a more general label with less noise, compared to the analyses with a more specific semantic content but more noise. In addition, the demarcation between class properties and idiosyncratic verb properties might be facilitated, since it takes place on a rather general level.

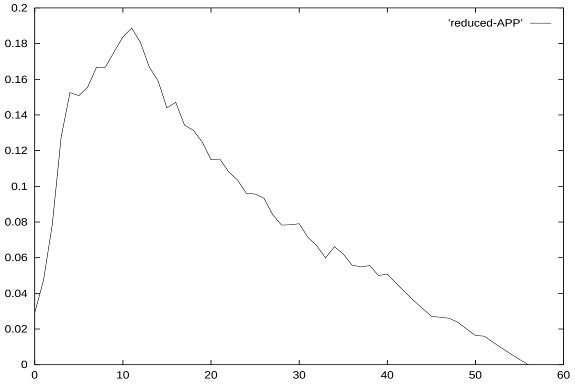

Figure 5.3: Varying the number of clusters on reduced verb set (evaluation: APP)

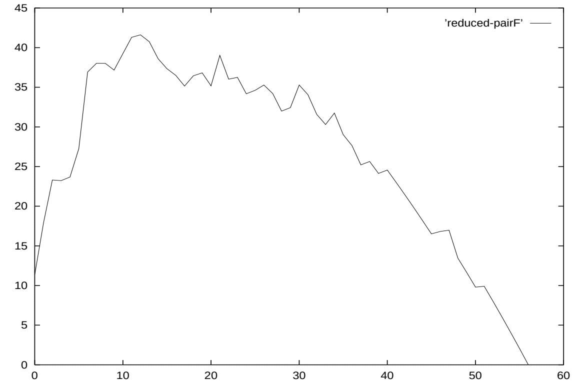

Figure 5.4: Varying the number of clusters on reduced verb set (evaluation: $PairF$)

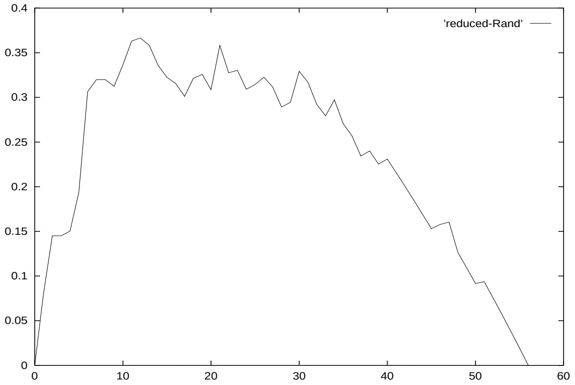

Figure 5.5: Varying the number of clusters on reduced verb set (evaluation: $Rand_{adj}$)

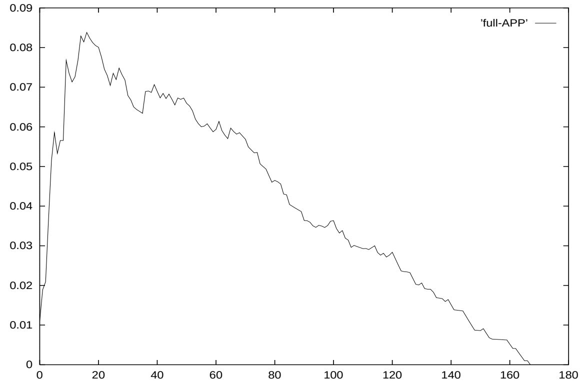

Figure 5.6: Varying the number of clusters on full verb set (evaluation: APP)

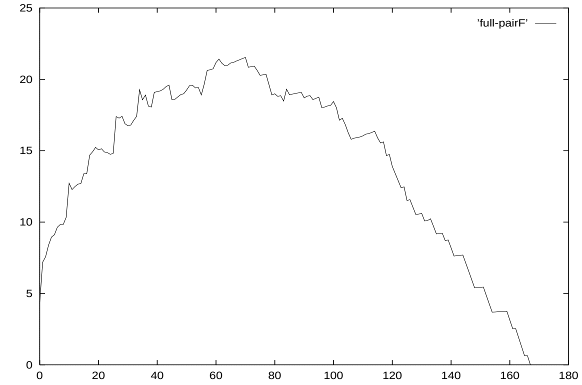

Figure 5.7: Varying the number of clusters on full verb set (evaluation: $PairF$)

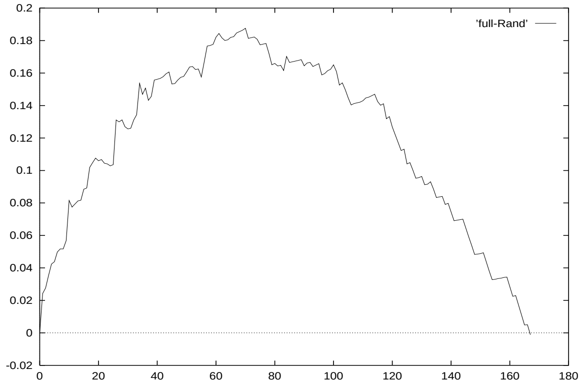

Figure 5.8: Varying the number of clusters on full verb set (evaluation: $Rand_{adj}$)

5.4.4 Verb Sense Disambiguation

As final part in the section of optimisation, I would like to discuss the problem of verb ambiguity in clustering, and possibilities to address the problem. Verb ambiguity is a pervasive phenomenon in natural language, so it should be taken into consideration in whatever natural language processing task. In this section, I do not try to solve the ambiguity problem in clustering, but discuss possibilities to cope with it.

In the clustering experiments, the German verbs are described by distributions over subcategorisation frames of pre-defined types. The distributional values for the different verb senses are hidden in the distributions, since the statistical grammar model does not distinguish verb senses and therefore the frequency information from the model is merged for the verb senses. For example, the verb *bestehen* has at least four different senses, each coupled with a preferred subcategorisation behaviour: (i) *bestehen* referring to *Insistence* subcategorises np with auf_{Dat}, (ii) *bestehen* referring to *Consistence* subcategorises np with aus_{Akk}, (iii) *bestehen* referring to *Existence/Survival* subcategorises n or np with in_{Akk}, and (iv) *bestehen* referring to *Passing* (e.g. an exam) subcategorises na. Considering only the coarse subcategorisation and PP information, each of the above frames has a comparably high frequency within the distributional verb description.

Using a hard clustering algorithm such as k-Means, in the best case the similarity measure realises the close similarities of *bestehen* with other verbs of (i) *Insistence*, (ii) *Consistence*, (iii) *Existence*, and (iv) *Passing*, but nevertheless *bestehen* is assigned to only one of the respective semantic classes, since the ambiguity cannot be modelled.

There is two general possibilities to model the verb ambiguity:

- The verb clustering is based on the existing verb descriptions, but a soft clustering algorithm is applied.
- The verb senses are disambiguated before they are given a descriptive distribution, i.e. a disambiguation method is defined which is able to state that there is $bestehen_1$ with a high frequency for np with auf_{Akk} but low frequencies for all other frame types, $bestehen_2$ with a high frequency for np with aus_{Akk} but low frequencies for all other frame types, etc. With the preceding verb sense disambiguation the clustering input would consider the different verb senses separately.

I do not go into further details here, since each of the issues deserves specific attention which is not subject of this chapter. Further work might deal with verb ambiguity in clustering experiments.

5.4.5 Summary

Summarising the above discussions on optimising the clustering of verbs, there is no unique combination of feature choice and clustering parameters which optimises the clustering outcome. The strategy of utilising subcategorisation frames, prepositional information and selectional preferences to define the verb features has proven successful, since the application at each level has generated a positive effect on the clustering. But the usefulness of the verb features is limited by the specific properties of the desired verb classes. In addition, subtle distinctions in the feature choice do not show a consistent effect on the clustering, and the results not necessarily align with linguistic intuition. These insights agree with the definition of overfitting, that applying an 'optimal' combination of feature choice and clustering parameters (as measured on a specific clustering setting) to a different set of verbs does not necessarily result in the desired optimal clustering.

The purposes of this section have therefore been fulfilled: (i) On the one hand, the optimisation criteria were a means to demonstrate the range of possibilities to set the different clustering parameters. If I had not illustrated the potential of the parameters, each reader would have different questions and suggestions concerning why I did not try this or that. The optimisation discussion should prevent me from such complaints. (ii) On the other hand, the discussions were a means to show that I did not arbitrarily set the parameters, but tried to find an at least near-optimal compromise between linguistic and practical demands. There is always a way to reach a better result if I went on trying more and more combinations of parameters, but the slight gain in clustering success will not be worth it; on the contrary, I would risk overfitting of the parameters.

5.5 Large-Scale Clustering Experiment

So far, all clustering experiments have been performed on a small-scale, preliminary set of manually chosen 168 German verbs. But a goal of this thesis is to develop a clustering methodology with respect to an automatic acquisition of a high-quality and large-scale German verb classification. I therefore apply the insights (i) on the theoretical relationship between verb meaning and verb behaviour and (ii) on the clustering parameters to a considerably larger amount of verb data.

- Verb Data:

  I extracted all German verbs from the statistical grammar model which appeared with an empirical frequency between 500 and 10,000 in the training corpus. This selection results in a total of 809 verbs, including 94 verbs from the preliminary set of 168 verbs. I added the remaining verbs of the preliminary set (because of evaluation reasons, see below), resulting in a total selection of 883 German verbs. The list of verbs and verb frequencies is given in Appendix C.

- Feature Choice:

  The feature description of the German verbs refers to the probability distribution over the coarse syntactic frame types, which are added prepositional phrase information on the 30 chosen PPs and selectional preferences for the linguistically and practically most successful combination n, na, nd, nad, and ns-dass. As in previous clustering experiments, the features are step-wise refined.

- Clustering Parameters:

  k-Means is provided hierarchical clustering input (based on complete-linkage and 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.

- Evaluation:

  For the large-scale set of German verbs no manual classification is provided. (A manual classification would actually disagree with the idea that an automatic induction of verb classes prevents the computational linguist from the manual effort of constructing a classification from scratch.) But to provide an indication of the clustering success, I have made sure that the preliminary set of 168 verbs is included in the large-scale set. On the basis of the 168 manually chosen verbs an 'auxiliary' evaluation of the clustering result is performed: All clusters in the resulting large-scale cluster analysis which contain any of the manually chosen verbs are extracted, only the manually chosen verbs are kept in the clusters, and this partial cluster analysis is evaluated against the gold standard of 43 verb classes. The result is not expected to keep up with clustering experiments on only the preliminary verb set, since the clustering task on 883 verbs is much more difficult, but it provides an indication for comparing different cluster analyses with each other.

Tables 5.36 to 5.38 present the clustering results on the large-scale verb set, based on syntactic frame information in Table 5.36, with additional prepositional phrase information in Table 5.37 and additional selectional preferences in Table 5.38. As said before, the evaluation is performed on the manually chosen set of verbs. The results are therefore compared to the respective clustering results on the set of 168 verbs (a) in 43 clusters which is the gold standard number of classes, and (b) in 72 clusters of the hierarchical clustering input and 64 clusters of the k-Means clustering outcome, since these are the number of clusters over which the manually chosen verbs are distributed in the large-scale experiments.

The large-scale clustering results once more confirm the general insights (i) that the step-wise refinement of features improves the clustering, (ii) that Ward's method is usually the optimal choice for the hierarchical clustering, and (iii) that Ward's hierarchical clustering is seldom improved by the k-Means application. In addition, several large-scale cluster analyses keep up well with the comparable clustering results on the small-scale set of verbs, especially when compared to 72 and 64 clusters. This means that the distributional value of the verb descriptions has not vanished within a large set of verb vectors.

Eval	Input	Verb Description: frame		
		small-scale		large-scale
		43 clusters	72 → 64 clusters	72 clusters
APP	H-Comp	0.032 → 0.032	0.025 → 0.029	0.022 → 0.022
	H-Ward	0.040 → 0.041	0.029 → 0.035	0.029 → 0.031
PairF	H-Comp	11.50 → 11.00	11.71 → 12.21	9.86 → 9.36
	H-Ward	12.44 → 12.64	10.83 → 11.73	12.15 → 12.88
R_a	H-Comp	0.077 → 0.072	0.091 → 0.094	0.067 → 0.063
	H-Ward	0.092 → 0.094	0.084 → 0.091	0.094 → 0.102

Table 5.36: Large-scale clustering on frames

Eval	Input	Verb Description: frame+ppA		
		small-scale		large-scale
		43 clusters	72 → 64 clusters	72 clusters
APP	H-Comp	0.062 → 0.060	0.045 → 0.048	0.037 → 0.040
	H-Ward	0.068 → 0.067	0.044 → 0.055	0.045 → 0.048
PairF	H-Comp	18.87 → 18.34	20.78 → 20.10	13.96 → 16.33
	H-Ward	18.64 → 18.81	17.56 → 18.81	18.22 → 16.96
R_a	H-Comp	0.150 → 0.145	0.180 → 0.171	0.119 → 0.134
	H-Ward	0.148 → 0.151	0.149 → 0.161	0.152 → 0.142

Table 5.37: Large-scale clustering on frames and PPs

Eval	Input	Verb Description: frame+ppA+prefA on $\underline{\mathtt{n}}/\underline{\mathtt{na}}/\underline{\mathtt{nd}}/\mathtt{n}\underline{\mathtt{a}}\mathtt{d}/\underline{\mathtt{n}}\mathtt{s\text{-}dass}$		
		small-scale		large-scale
		43 clusters	72 → 64 clusters	72 clusters
APP	H-Comp	0.047 → 0.050	0.036 → 0.038	0.028 → 0.029
	H-Ward	0.064 → 0.064	0.050 → 0.058	0.040 → 0.030
PairF	H-Comp	19.28 → 19.03	20.69 → 18.21	14.50 → 11.43
	H-Ward	22.86 → 22.19	19.47 → 20.48	19.92 → 15.06
R_a	H-Comp	0.153 → 0.147	0.174 → 0.144	0.122 → 0.074
	H-Ward	0.190 → 0.182	0.165 → 0.174	0.170 → 0.115

Table 5.38: Large-scale clustering on frames, PPs and preferences

Following, I present example clusters from the optimal large-scale cluster analysis (according to the above evaluation): Ward's hierarchical cluster analysis based on subcategorisation frames, PPs and selectional preferences, without running k-Means on the hierarchical clustering. As a general characterisation of the cluster analysis, some clusters are extremely good with respect to the semantic overlap of the verbs, some clusters contain a number of similar verbs mixed with semantically different verbs, and for some clusters it is difficult to recognise a common semantic aspect of the verbs. For each kind of result I will present examples. The verbs which I think semantically similar are marked in bold font. Differently to previous examples where the manual verbs were not translated, but identified by the semantic class label, the following analysis gives translations of the verbs. I will only refer to the semantic content of the clusters and the verbs, but not to the verb distributions on the syntax-semantic interface, since the latter have been discussed before in detail.

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

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

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

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

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

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

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

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

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

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

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

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

(13) ***beziffern*** ‘to amount to’, ***schätzen*** ‘to estimate’, ***veranschlagen*** ‘to estimate’

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

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

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

Clusters (1) to (3) are example clusters where the verbs do not share meaning aspects. In the overall cluster analysis, the semantically incoherent clusters tend to be rather large, i.e. with more than 15-20 verb members.

Clusters (4) to (7) are example clusters where a part of the verbs show overlap in their meaning aspects, but the clusters also contain considerable noise. Cluster (4) mainly contains verbs of buying and selling, cluster (5) contains verbs of wishing, cluster (6) contains verbs of expressing a speech act concerning a specific event, and cluster (7) contains verbs of quantum change.

Clusters (8) to (16) are example clusters where most or all verbs show a strong similarity in their conceptual structures. Cluster (8) contains verbs expressing a propositional attitude; the underlined verbs in addition indicate an emotion. The only unmarked verb *weißen* also fits into the cluster, since it is a morphological lemma mistake changed with *wissen* which belongs to the verb class. The verbs in cluster (9) describe a scene where somebody or some situation makes something possible (in the positive or negative sense). Next to a lemmatising mistake (*angebieten* is not an infinitive, but a morphologically mistaken perfect participle of *anbieten*), the only exception verb is *veranstalten*. The verbs in cluster (10) are connected more loosely, all referring to a verbal discussion, with the underlined verbs in addition denoting a negative, complaining way of utterance. In cluster (11) all verbs refer to a basis, in cluster (12) the verbs describe the process from arresting to treating a suspect, and cluster (13) contains verbs of estimating an amount of money. In cluster (14), all verbs except for *entschuldigen* refer to an emotional state (with some origin for the emotion). The verbs in cluster (15) except for *profitieren* all indicate a thinking (with or without talking) about a certain matter. Finally in cluster (16), we can recognise the same weather verb cluster as in previously discussed small-scale cluster analyses; the three verbs also cluster together in a large-scale environment.

I have experimented with two variations in the clustering setup:

- For the selection of the verb data, I 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 basis of the evaluation and on 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.

- The number of target clusters was set to 300 instead of 100, i.e. 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 introspection 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 have often preserved their constitution. To present one example, the following cluster from the 100-cluster analysis

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

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

  (a) ***anzeigen*** ‘to announce’, ***aufklären*** ‘to clarify’

  (b) ***beeindrucken*** ‘to impress’, ***enttäuschen*** ‘to disappoint’, ***überraschen*** ‘to surprise’, ***überzeugen*** ‘to persuade’

  (c) ***befreien*** ‘to free’, ***beruhigen*** ‘to calm down’, ***retten*** ‘to save’, ***schützen*** ‘to protect’, ***stören*** ‘to 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: non-protective) influence of one person towards another.

Summarising, the large-scale clustering experiment results in a mixture of semantically diverse verb classes and semantically coherent verb classes. I have presented a number of semantically coherent classes which need little manual correction as a lexical resource. Semantically diverse verb classes and clustering mistakes need to be split into finer and more coherent clusters, or to be filtered from the classification.

5.6 Related Work

The following section presents related work on the clustering experiments. The description and comparison of the related work refers to (i) the automatic induction of class-relevant features, which illustrates approaches that obtain syntactic and semantic properties of verbs and confirm the relationship between the verb meaning and verb behaviour, and (ii) classification and clustering experiments on the automatic induction of classes for verbs, nouns, and adjectives. For the description of related work on the usage of verb classes the reader is referred to Chapter 2.

5.6.1 Automatic Induction of Class-Relevant Features

The verb information underlying my clustering experiments basically describes the syntactic definition of verb subcategorisation, syntactico-semantic prepositional refinement, and the semantic definition of selectional preferences for verb arguments. The sum of the verb information inherently defines the verb alternation behaviour, as a combination of syntactic frame alternation and selectional preferences. Related work on class-relevant features for verb description refers to a similar arrangement of verb properties. The following paragraphs therefore refer to the empirical acquisition of subcategorisation frames, selectional preferences, and diathesis alternation.

Subcategorisation Frames

The following approaches on extracting subcategorisation frames to describe verb usage especially illustrate the strong relation between verb meaning and verb behaviour, providing empirical syntactic evidence for semantic verb classes.

Lapata and Brew (1999) show that the syntactic frame definition of English verbs can be used to disambiguate the semantic class affiliation of verb usage. The joint probabilities of verb, frame and semantic class are estimated by frequency counts from the lemmatised version of the British National Corpus. The simple model achieves high precision and can be extended to incorporate other sources of information which influence the class selection process. The approach emphasises the strong relationship between syntactic and semantic verb features, and presents empirical evidence for the English verb class construction with regard to verb-frame combinations.

As described earlier as approach to word sense disambiguation, Dorr and Jones (1996) parse the example sentences in the Levin classes (Levin, 1993) and extract syntactic patterns for the English verbs, according to the syntactic structures they do and they do not allow. The approach distinguishes positive and negative examples by 1 and 0, respectively. For example, the parsing pattern for the positive sentence *Tony broke the vase to pieces* would be `1-[np,v,np,pp(to)]`. Dorr and Jones show that the syntactic patterns of the verbs closely correspond to their distinction in semantic class affiliation, and therefore validate the strong relation between the syntactic and the semantic information in the verb classes.

Selectional Preferences

Computational approaches to defining selectional preferences for predicate-argument structures refine syntactic predicate (mainly: verb) environments by semantic demands on their arguments. Typical applications of the preference information next to verb class constitution are word sense disambiguation, statistical parsing, and anaphora resolution.

Resnik (1993, 1997) defines selectional preference as the statistical association between a predicate and its argument within a syntactic relationship. The association value is the relative entropy between (a) the posterior probability of the argument appearing within the given relationship to a specific predicate and (b) the prior probability of the argument appearing within the given relationship to any predicate. The frequency counts underlying the probabilities for the nominal arguments are assigned to and propagated upwards the WordNet hierarchy, such that the hierarchical nodes represent the selectional preferences. For ambiguous nouns, the noun frequency count is split over all WordNet conceptual classes containing the respective noun. The probabilistic preference model of association values is used for word sense disambiguation.

Ribas (1994, 1995) performs variations on the basic technique as defined by Resnik (1993). Mainly, he varies the definition of the prior probability distribution (by using the probability of the argument without reference to the syntactic environment), the assignment of ambiguous nominal frequency counts to classes (by splitting the counts of ambiguous nouns over all leaf nodes containing the respective noun), and the statistical measure (by using the log-likelihood ratio and mutual information). The resulting models show an improvement in the word sense disambiguation task.

Abe and Li (1996) and Li and Abe (1998) also use WordNet to define selectional preferences. As in the above approaches, their algorithm is based on co-occurrence counts of predicates and arguments within a specific syntactic relationship. The selectional preferences for a predicate-argument structure are described by a cut in the WordNet hierarchy, a set of WordNet nodes; the cut is determined by the Minimum Description Length (MDL), a principle from information theory for data compression and statistical estimation. The best probability model for given data is that which requires the least code length in bits for the encoding of the model itself (model description length) and the given data observed through it (data description length). A model nearer the WordNet root is simpler but with poorer fit to the data, and a model nearer the WordNet leaves is more complex but with a better fit to the data. The MDL principle finds that model which minimises the sum of both description length values.

Wagner (2000) introduces modifications on the model by Abe and Li: (i) He ensures that the levels of noun senses and conception in the WordNet hierarchy are separated, by splitting hybrid nodes and introducing extra hyponyms, (ii) he maps the WordNet directed acyclic graph onto a tree structure, (iii) he introduces a threshold for the tree cut calculation, and (iv) most importantly, he introduces a weighting for the MDL principle which transforms the principle into a Bayesian learning algorithm. The modifications improve the overall performance on the selectional preference acquisition.

Abney and Light (1999) provide a stochastic generation model for selectional preferences of a predicate-argument relationship. Co-occurrence counts are extracted from the British National Corpus by Abney's parser Cass (Abney, 1997), and the co-occurrence probabilities are estimated by a Hidden Markov Model (HMM) for each predicate structure. The HMM is defined and trained on the WordNet hierarchy, with the initial state being the (artificial) root node of WordNet. Each HMM run is a path through the hierarchy from the root to a word sense, plus the word generated from the word sense. The algorithm does not work sufficiently; the main reason seems to be that the estimation method is inappropriate for the problem.

Clark and Weir (2000, 2002) utilise the WordNet hierarchy to determine a suitable noun class as the optimal level of generalisation for a predicate-argument relationship. They obtain frequency triples for a verb and a noun within a specific syntactic relationship from the British National Corpus, using the parser by Briscoe and Carroll (1997). Estimating the joint frequencies for a predicate-argument relationship and a specific WordNet class as by Resnik (1993), the generalisation procedure by Clark and Weir uses the statistical X^2 test to find the most suitable class: Bottom-up the WordNet hierarchy, each node in the hierarchy is checked whether the probability of the parent class is significantly different to that of the children classes. In that case, the search is stopped at the respective child node as the most suitable selectional preference representation.

Brockmann and Lapata (2003) compare the approaches to selectional preference definition as given by Resnik (1993), Li and Abe (1998) and Clark and Weir (2002), with respect to German verbs and their NP and PP complements. The models as well as a combination of the models are evaluated against human ratings, with the result that there is no method which overall performs best. The model combination is performed by multiple linear regression and obtains a better fit with the experimental data than the single methods.

Gamallo, Agustini, and Lopes (2001) define selectional preferences by 'co-specification': Two syntactically related words impose semantic selectional restrictions on each other. For each two words w_1 and w_2 within a syntactic relationship r, Gamallo *et al.* collect co-occurrence triples $< r, w_1 \uparrow, w_2 \downarrow >$, with ↑ indicating the head and ↓ indicating the complement of the respective syntactic relationship. The co-occurrence counts are based on 1.5 million words of the *Portuguese General Attorney Opinions (PGR)*, a domain-specific Portuguese corpus of case-law documents. The set of co-occurrence triples for a specific word as either head or complement represents the selectional preferences for that word. Gamallo *et al.* use the co-occurrence triples for a semantic clustering. Following Harris' distributional hypothesis (Harris, 1968), words occurring in similar syntactic contexts are semantically similar and clustered into the same semantic class. Gamallo *et al.* define an agglomerative hierarchical clustering algorithm which forms clusters according to the agreement in the co-occurrence contexts. The resulting clusters are evaluated manually, i.e. by linguistic intuition of the authors.

Most approaches to selectional preference acquisition utilise the existing semantic ontology WordNet, which provides a hierarchical system of noun concepts, basically relating nouns by lexical synonymy and hypernymy. As in my usage of selectional preference definition, the ontology is a convenient resource, since it provides nominal concepts on various levels of generality.

It is much more difficult and seems rather intuitive to define own conceptual classes, which in addition are difficult to evaluate, cf. Gamallo *et al.* (2001).

As in all above approaches, I utilise the frequency counts for predicate-argument structures to define selectional preferences. My approach for the preference definition is comparably simple, since it does not define a model over the complete hierarchy, but considers only the top-level nodes. In addition, the top-level choice guarantees a restricted number of preference concepts. As a disadvantage, the resulting model is less flexible on the choice of preference node level.

Diathesis Alternations

The recognition of diathesis alternations provides a direct source for the definition of verb classes, since alternations capture verb meaning to a large extent. But the general identification of alternations is complicated, since the syntactic environment of verbs is only partly sufficient, e.g. for the dative and benefactive alternations in English, cf. Lapata (1999). For many alternations, such as the distinction between unergative and unaccusative verbs (cf. McCarthy (2001) and the verb classification by Merlo and Stevenson, 2001), it is necessary to take the selectional preferences into account. The following approaches are more detailed than my verb descriptions, since they make explicit reference to which verbs undergo which alternations, whereas my verb descriptions only inherently include diathesis alternation.

Lapata (1999) presents a case study for the acquisition of diathesis alternations, by examining the extent to which the dative and benefactive alternation for English verbs (cf. Examples (5.1) and (5.2) as taken from the paper) are attested in the British National Corpus.

(5.1) John offers shares to his employees.
John offers his employees shares.

(5.2) Leave a note for her.
Leave her a note.

Lapata acquires the alternating verbs by extracting the alternation-related syntactic structures (the double object frame ‘V NP_1 NP_2’, and the prepositional frames ‘V NP_1 *to* NP_2’ and ‘V NP_1 *for* NP_2’) by a shallow parser from the part-of-speech-tagged BNC. The parser output is filtered by linguistic heuristics and statistical scores, and the result is compared to the respective Levin semantic classes (Levin, 1993). The alternating verbs agree to a large extent with Levin’s classification, add verbs to the classes, and support the classes by empirical evidence.

McCarthy (2001) presents an identification methodology for the participation of English verbs in diathesis alternations. In a first step, she uses the subcategorisation frame acquisition system by Briscoe and Carroll (1997) to extract frequency information on 161 subcategorisation frame types for verbs from the written part (90 million words) of the British National Corpus. The subcategorisation frame types are manually linked with the Levin alternations (1993), and thereby define the verbal alternation candidates. Following the acquisition of the syntactic information, the nominal fillers of the noun phrase and prepositional phrase arguments in the verb-frame tuples are used to define selectional preferences for the respective argument slots. For this step, McCarthy utilises the selectional preference acquisition approach of Minimum Description Length (MDL) by Li and Abe (1998). In the final step, McCarthy defines two methods to identify the participation of verbs in diathesis alternations: (i) The MDL principle compares the costs of encoding the tree cut models of selectional preferences for the relevant argument slots in the alternation frames. If the cost of combining the models is cheaper than the cost of the separate models, the verb is decided to undergo the respective alternation. (ii) The similarity-based method calculates the similarity of the two tree cut models with reference to the alternating argument slots for verb participants in diathesis alternations. A threshold decides the participation.

5.6.2 Automatic Induction of Classes

The following sections describe classification and clustering experiments on the automatic induction of classes for verbs, nouns, and adjectives. The classifications refer to different aspects of the respective parts of speech, e.g. the verb classes represent aspectual properties (Siegel and McKeown, 2000), syntactic categories (Merlo and Stevenson, 2001; Merlo *et al.*, 2002; Tsang *et al.*, 2002), and –most similar to my approach– semantic categories (Schulte im Walde, 2000a; Joanis, 2002). According to the classification type, different kinds of properties are used to describe the underlying class words, with a dominant number of approaches utilising frequency counts for verb-noun relationships.

Verb Classes

Siegel and McKeown (2000) use three supervised and one unsupervised machine learning algorithms to perform an automatic aspectual classification of English verbs. (i) For the supervised classification, 97,973 parsed sentences on medical discharge summaries are used to extract frequencies for verbs on 14 linguistic indicators, such as manner adverb, duration *in*-PP, past tense, perfect tense. Logistic regression, decision tree induction and genetic programming are applied to the verb data to distinguish states and events. Comparing the ability of the learning methods to combine the linguistic indicators is claimed difficult, since they rank differently depending on the classification task and evaluation criteria. Decision trees achieve an accuracy of 93.9%, as compared to the uninformed baseline of 83.8%. (ii) For the unsupervised clustering, 14,038 distinct verb-object pairs of varying frequencies are extracted from 75,289 parsed novel sentences. The verbs are clustered semantically by a non-hierarchical algorithm, which produces a partition of the set of verbs according to the similarities of the verbs with regard to their subcategorised direct object nouns, cf. Hatzivassiloglou and McKeown (1993): For each verb pair, the distances between the verbs is calculated by Kendall's τ coefficient (Kendall, 1993). A random partition of the set of verbs is improved by a hill-climbing method, which calculates the sum of distances in all clusters and step-wise improves the partition by moving that verb to that different cluster where the decrease in the sum of distances is largest. For a small set of 56 verbs whose frequency in the verb-object pairs is larger than 50, Siegel and McKeown claim on basis of an evaluation of 19 verbs that the clustering algorithm discriminates event verbs from stative verbs.

In former work on English, I clustered 153 verbs into 30 verb classes as taken from Levin (1993), using an unsupervised hierarchical clustering method (Schulte im Walde, 2000a). The verbs are described by distributions over subcategorisation frames as extracted from maximum probability parses of a robust statistical parser, and completed by assigning WordNet classes as selectional preferences to the frame arguments. Using Levin's verb classification as evaluation basis, 61% of the verbs are classified correctly into semantic classes. The clustering is most successful when utilising syntactic subcategorisation frames enriched with PP information; selectional preferences decrease the performance of the clustering approach. With reference to the paper, the detailed encoding and therefore sparse data make the clustering worse with than without the selectional preference information. The paper empirically investigates the proposition that verbs can be semantically classified according to their syntactic alternation behaviour concerning subcategorisation frames and their selectional preferences for the arguments within the frames.

Merlo and Stevenson (2001) present an automatic classification of three types of English intransitive verbs, based on argument structure crucially involving thematic relations. They select 60 verbs with 20 verbs from each verb class, comprising unergatives, unaccusatives and object-drop. The verbs in each verb class show similarities in their argument structure, in that they all may be used as transitives and intransitives, as Examples (5.3) to (5.5) as taken from the paper show. Therefore, the argument structure alone does not distinguish the classes. In order to distinguish the classes, the subcategorisation information needs to be refined by thematic relations.

(5.3) Unergative Verbs:
The horse raced past the barn.
The jockey raced the horse past the barn.

(5.4) Unaccusative Verbs:
The butter melted in the pan.
The cook melted the butter in the pan.

(5.5) Object-Drop Verbs:
The boy played.
The boy played soccer.

Merlo and Stevenson define verb features based on linguistic heuristics which describe the thematic relations between subject and object in transitive and intransitive verb usage. The features include heuristics for transitivity, causativity, animacy and syntactic features. For example, the degree of animacy of the subject argument roles is 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 is described by a 5-feature-vector, and the vector descriptions are fed into a decision tree algorithm. Compared to a baseline performance of 33.9%, the decision trees classify the verbs into the three classes with an accuracy of 69.8%. Further experiments show the different degrees of contribution of the different features within the classification.

Compared to my work, Merlo and Stevenson perform a simpler task and classify a smaller number of 60 verbs in only three classes. The features of the verbs are restricted to those which should capture the basic differences between the verb classes, agreeing on 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) presents an extension of their work which uses 802 verbs from 14 classes in Levin (1993). He defines an extensive feature space with 219 core features (such as part of speech, auxiliary frequency, syntactic categories, animacy as above) and 1,140 selectional preference features taken from WordNet. As in my approach, the selectional preferences do not improve the clustering.

The classification methodology from Merlo and Stevenson (2001) is transfered to multi-linguality, by Merlo, Stevenson, Tsang, and Allaria (2002) and Tsang, Stevenson, and Merlo (2002). Merlo *et al.* show that the classification paradigm is applicable in other languages 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 use the content of Chinese verb features to refine the English verb classification, explained in more detail by Tsang *et al.* (2002). The English verbs are 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 outperforms either set of monolingual features, yielding an accuracy of 83.5%.

Compared to the above approaches, my work is the first approach on automatic verb classification (i) where more than 100 verbs are clustered, and (ii) without a threshold on verb frequency, and (iii) with fine-grained verb classes, and (iv) without concentration on specific verb-argument structures, and (v) with a gold standard verb classification for evaluation purposes. In addition, the approach is the first one to cluster German verbs.

Noun and Adjective Classes

The clustering approaches for noun and adjective classification are basically similar to verb classification. The following approaches present three soft clustering algorithms for noun classes, and a hard clustering algorithm for adjective classes.

Hindle (1990) presents a semantic classification of English nouns. He parses a six million word sample of Associated Press news stories and extracts 4,789 verbs from 274,613 parsed clausal structures. For each verb in each clause, the deep subject and object noun are determined, resulting in a total of 26,742 head nouns. For each verb-noun pair with respect to a predicate-argument relation, the mutual information between verb and noun is calculated. The similarity of each two nouns is then based on their agreement in the predicate-argument structures, i.e. the more two nouns agree in their appearance as subjects or objects of the same verbs, the more similar they are. The similarity for each noun pair is calculated as the sum of subject and object similarities over all verb-noun pairs, where subject similarity is the minimal mutual information value of the two verb-noun pairs $< v, n_1 >$ and $< v, n_2 >$ with the nouns as subject of the verb, and object similarity is the minimal mutual information value of the two verb-noun pairs $< v, n_1 >$ and $< v, n_2 >$ with the nouns as object of the verb. For each noun, the ten most similar nouns are determined to define a noun class. For example, the ten most similar nouns for *boat* are *boat, ship, plane, bus, jet, vessel, truck, car, helicopter, ferry, man*.

Pereira, Tishby, and Lee (1993) describe a hierarchical soft clustering method which clusters words according to their distribution in particular syntactic contexts. They present an application of their method to nouns appearing as direct objects of verbs. The clustering result is a hierarchy of noun clusters, where each noun belongs to each cluster with a membership probability. The input data for the clustering process are frequencies of verb-noun pairs in the direct object relationship, as extracted from parsed sentences of the Associated Press news wire corpus. On basis of the conditional verb-noun probabilities, the similarity of the distributions is determined by the Kullback-Leibler divergence, cf. Section 4.1.3. The EM algorithm (Baum, 1972) is used to learn the hidden cluster membership probabilities, and deterministic annealing performs the divisive hierarchical clustering. The resulting class-based model can be utilised for estimating information for unseen events, cf. Dagan, Lee, and Pereira (1999).

Rooth, Riezler, Prescher, Carroll, and Beil (1999) produce soft semantic clusters for English which at the same time represent a classification on verbs as well as on nouns. They gather distributional data for verb-noun pairs in specific grammatical relations from the British National Corpus. The extraction is based on a lexicalised probabilistic context-free grammar (Carroll and Rooth, 1998) and contains the subject and object nouns for all intransitive and transitive verbs in the parses, a total of 608,850 verb-noun types. The conditioning of the verbs and the nouns on each other is made through hidden classes, and the joint probabilities of classes, verbs and nouns are trained by the EM algorithm. The resulting model defines conditional membership probabilities of 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 are utilised for the induction of a semantically annotated verb lexicon.

Hatzivassiloglou and McKeown (1993) present a semantic classification of adjectives which is based on a non-hierarchical clustering algorithm. In a first stage, they filter adjective-noun pairs for 21 frequent adjectives from a 8.2 million word corpus of stock market reports from the Associated Press news wire. The 3,073 distinct tuples represent the basis for calculating distances between each two adjectives by Kendall's τ coefficient (Kendall, 1993). A random partition of the set of adjectives is improved by a hill-climbing method, which calculates the sum of distances in all clusters and step-wise improves the partition by moving that adjective to that different cluster where the decrease in the sum of distances is largest. An evaluation of the resulting clusters is performed by pair-wise precision and recall, referring to the manual solutions of nine human judges. Their best result corresponds to a clustering with 9 clusters, with recall of 49.74%, precision of 46.38% and f-score of 48.00%.