Chapter 6

Conclusion

This thesis has performed experiments on the automatic induction of German semantic verb classes. The verb is central to the structure and the meaning of a sentence, and therefore lexical verb resources play an important role in supporting computational applications in Natural Language Processing. But especially semantic lexical resources represent a bottleneck in NLP, and methods for the acquisition of large amounts of knowledge with comparably little manual effort have gained importance. In this context, I have investigated the potential and the limits of an automatic acquisition of semantic classes for German verbs. A good methodology will support NLP applications such as word sense disambiguation, machine translation, and information retrieval.

Sometimes it is something of a black art when applying multivariate clustering to high-dimensional natural language data, since we do not necessarily find out about the relevance of data types or the interpretation of the data by the clustering algorithm. But the data and the clustering techniques should be based on the linguistic background of the task. Therefore, I have focused on the sub-goals of the clustering task: I have empirically investigated the definition and the practical usage of the relationship between verb meaning and verb behaviour, i.e. (i) which exactly are the semantic features that define verb classes, (ii) which exactly are the features that define verb behaviour, and (iii) can we use the meaning-behaviour relationship of verbs to induce verb classes, and to what extent does the meaning-behaviour relationship hold? In addition, I have investigated the relationship between clustering idea, clustering parameters and clustering result, in order to develop a clustering methodology which is suitable for the demands of natural language. The clustering outcome cannot be a perfect semantic verb classification, since (i) the meaning-behaviour relationship on which we rely for the clustering is not perfect, and (ii) the clustering method is not perfect for the ambiguous verb data. But only if we understand the potential and the limits of the sub-goals, we can develop a methodology which can be applied to large-scale data.

6.1 Contributions of this Thesis

The contribution of my work comprises three parts. Each of the parts may be used independently.

6.1.1 A Small-Scale German Verb Classification

I manually defined 43 German semantic verb classes containing 168 partly ambiguous German verbs. The construction of the German verb classes is primarily based on semantic intuition: Verbs are assigned to classes according to similarity of lexical and conceptual meaning, and each verb class is assigned a conceptual class label. Because of the meaning-behaviour relationship at the syntax-semantic interface, the verbs grouped in one class show a certain agreement in their behaviour.

The class size is between 2 and 7, with an average of 3.9 verbs per class. Eight verbs are ambiguous with respect to class membership and marked by subscripts. The classes include both high and low frequency verbs: the corpus frequencies of the verbs range from 8 to 71,604. The class labels are given on two conceptual levels; coarse labels such as *Manner of Motion* are sub-divided into finer labels, such as *Locomotion, Rotation, Rush, Vehicle, Flotation*.

The class description is closely related to Fillmore's scenes-and-frames semantics (Fillmore, 1977, 1982), which is computationally utilised in FrameNet (Baker *et al.*, 1998; Johnson *et al.*, 2002). Each verb class is given a conceptual scene description which captures the common meaning components of the verbs. Annotated corpus examples illustrate the idiosyncratic combinations of verb meaning and conceptual constructions, to capture the variants of verb senses. The frame-semantic class definition contains a prose scene description, predominant frame participant and modification roles, and frame variants describing the scene. The frame roles have been developed on basis of a large German newspaper corpus from the 1990s. They capture the scene description by idiosyncratic participant names and demarcate major and minor roles. Since a scene might be activated by various frame embeddings, I have listed the predominant frame variants as found in the corpus, marked with participating roles, and at least one example sentence of each verb utilising the respective frame. The frame variants with their roles marked represent the alternation potential of the verbs, by connecting the different syntactic embeddings to identical role definitions.

Within this thesis, the purpose of the manual classification was to evaluate the reliability and performance of the clustering experiments. But the size of the gold standard is also sufficient for usage in NLP applications, cf. analogical examples for English such as Lapata (1999); Lapata and Brew (1999); Schulte im Walde (2000a); Merlo and Stevenson (2001). In addition, the description details are a valuable empirical resource for lexicographic purposes, cf. recent work in Saarbrücken which is in the early stages of a German version of FrameNet (Erk *et al.*, 2003) and semantically annotates the German TIGER corpus (Brants *et al.*, 2002).

6.1.2 A Statistical Grammar Model for German

I developed a German statistical grammar model which provides empirical lexical information, specialising on but not restricted to the subcategorisation behaviour of verbs. Within the thesis, the model serves as source for the German verb description at the syntax-semantic interface which is used in the clustering experiments. But in general, the empirical data are valuable for various kinds of lexicographic work.

For example, Schulte im Walde (2003a) presents examples of lexical data which are available in the statistical grammar model. The paper describes a database of collocations for German verbs and nouns. Concerning verbs, the database concentrates on subcategorisation properties and verb-noun collocations with regard to their specific subcategorisation relation (i.e. the representation of selectional preferences); concerning nouns, the database contains adjectival and genitive nominal modifiers, as well as their verbal subcategorisation. As a special case of noun-noun collocations, a list of 23,227 German proper name tuples is induced. All collocation types are combined by a perl script which can be queried by a lexicographic user in order to filter relevant co-occurrence information on a specific lexical item. The database is ready to be used for lexicographic research and exploitation.

Schulte im Walde (2002b) describes the induction of a subcategorisation lexicon from the grammar model. The trained version of the lexicalised probabilistic grammar serves as source for the computational acquisition of subcategorisation frames for lexical verb entries. A simple methodology is developed to utilise the frequency distributions in the statistical grammar model. The subcategorisation lexicon contains 14,229 verbs with a frequency between 1 and 255,676 (according to the training corpus). Each lexical verb entry defines the verb lemma, the frequency, and a list of those subcategorisation frames which are considered to be lexicon-relevant. The frame definition is variable with respect to the inclusion of prepositional phrase refinement. Schulte im Walde (2002a) performs an evaluation of the subcategorisation data against manual dictionary entries and shows that the lexical entries hold a potential for adding to and improving manual verb definitions. The evaluation results justify the utilisation of the subcategorisation frames as a valuable component for supporting NLP-tasks.

In addition to the verb subcategorisation data in the grammar model, there is empirical lexical information on all structural definitions in the base grammar. For example, Zinsmeister and Heid (2003b) utilise the same statistical grammar framework (with a slightly different base grammar) and present an approach for German collocations with collocation triples: Five different formation types of adjectives, nouns and verbs are extracted from the most probable parses of German newspaper sentences. The collocation candidates are determined automatically and then manually investigated for lexicographic use. Zinsmeister and Heid (2003a) use the statistical grammar model to determine and extract predicatively used adverbs. Other sources for lexical information refer to e.g. adverbial usage, tense relationship between matrix and sub-ordinated clauses, and so on.

6.1.3 A Clustering Methodology for NLP Semantic Verb Classes

As main concern of this thesis, I have developed a clustering methodology which can be applied to the automatic induction of semantic verb classes. Key issues of the clustering methodology refer to linguistic aspects on the one hand, and to technical aspects on the other hand. In the following paragraphs, I will describe both the linguistic and the technical insights into the cluster analysis.

Linguistic Aspects I have empirically investigated the definition and the practical usage of the relationship between verb meaning and verb behaviour, i.e. (i) which exactly are the semantic features that define verb classes, (ii) which exactly are the features that define verb behaviour, and (iii) can we use the meaning-behaviour relationship of verbs to induce verb classes, and to what extent does the meaning-behaviour relationship hold?

The linguistic investigation referred to the following definitions. The semantic properties of the verbs were captured by the conceptual labels of the semantic verb classes. As a subjective manual resource, the classes refered to different levels of conceptual description. The behaviour of the verbs was described by distributions over properties at the syntax-semantic interface. Assuming that the verb behaviour can be captured by the diathesis alternation of the verb, I empirically defined syntactic subcategorisation frames, prepositional information and selectional preferences as verb properties. The meaning-behaviour relationship referred to the agreement of the behavioural and conceptual properties on the verb classes.

I have illustrated the verb descriptions and the realisation of verb similarity as defined by common similarity measures on the verb vectors. Of course, there is noise in the verb descriptions, but it is important to notice that the basic verb descriptions appear reliable with respect to their desired linguistic content. The reliability was once more confirmed by an evaluation of the subcategorisation frames against manual dictionary definitions.

The fact that there were at all verbs which were clustered semantically on basis of their behavioural 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. A series of post-hoc experiments which analysed the influence of specific frames and frame groups on the coherence of the verb classes illustrated the tight connection between the behaviour of the verbs and the verb meaning components.

Low frequent verbs have been determined as problem in the clustering experiments. Their distributions are noisier than those for more frequent verbs, so they typically constitute noisy clusters. The effect was stronger in a large-scale clustering, because the number of low frequent events represents a substantial proportion of all verbs.

The ambiguity of verbs cannot be modelled by the hard clustering algorithm k-Means. Ambiguous verbs were typically assigned either (i) to one of the correct clusters, or (ii) to a cluster whose verbs have distributions which are similar to the ambiguous distribution, or (iii) to a singleton cluster.

The interpretation of the clusterings unexpectedly pointed to meaning components of verbs which had not been discovered by the manual classification before. In the analysis, 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. In a similar way, the clustering interpretation exhibited semantically related verb classes, manually separated verb classes whose verbs were 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.

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

The experiments presented evidence for a linguistically defined limit on the usefulness of the verb features, which is driven by the dividing line between the common and idiosyncratic features of the verbs in a verb class. Recall the underlying idea of verb classes, that the meaning components of verbs to a certain extent determine their behaviour. This does not mean that all properties of all verbs in a common class are similar and we could extend and refine the feature description endlessly. The meaning of verbs comprises both (a) properties which are general for the respective verb classes, and (b) idiosyncratic properties which distinguish the verbs from each other. As long as we define the verbs by those properties which represent the common parts of the verb classes, a clustering can succeed. But by step-wise refining the verb description and including lexical idiosyncrasy, the emphasis of the common properties vanishes. From the theoretical point of view, the distinction between common and idiosyncratic features is obvious, but from the practical point of view there is no unique perfect choice and encoding of the verb features. The feature choice depends on the specific properties of the desired verb classes, but even if classes are perfectly defined on a common conceptual level, the relevant level of behavioural properties of the verb classes might differ.

For a large-scale classification of verbs, we need to specify a combination of linguistic verb features as basis for the clustering. Which combination do we choose? Both the theoretical assumption of encoding features of verb alternation as verb behaviour and the practical realisation by encoding syntactic frame types, prepositional phrases and selectional preferences have proven successful. In addition, I determined a (rather linguistically than technically based) choice of selectional preferences which represents a useful compromise for the conceptual needs of the verb classes. Therefore, this choice of features utilises the meaning-behaviour relationship best.

Technical Aspects I have investigated the relationship between clustering idea, clustering parameters and clustering result, in order to develop a clustering methodology which is suitable for the demands of natural language.

Concerning the clustering algorithm, I have decided to use the k-Means algorithm for the clustering, because it is a standard clustering technique with well-known properties. The parametric design of Gaussian structures realises the idea that objects should belong to a cluster if they are very similar to the centroid as the average description of the cluster, and that an increasing distance refers to a decrease in cluster membership. As a hard clustering algorithm, k-Means cannot model verb ambiguity. But starting clustering experiments with a hard clustering algorithm is an easier task than applying a soft clustering algorithm, especially with respect to a linguistic investigation of the experiment settings and results.

The experiments confirmed that the clustering input plays an important role. k-Means needs similarly-sized clusters in order to achieve a linguistically meaningful classification. Perturbation in the clusters is corrected for a small set of verbs and features, but fatal for extensive classifications. The linguistically most successful input clusters are therefore based on hierarchical clustering with complete linkage or Ward's method, since their clusters are comparably balanced in size and correspond to compact cluster shapes. The hierarchical clusterings actually reach similar clustering outputs than k-Means, 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. The similarity measure used in the clustering experiments was of secondary importance, since the differences in clustering with varying the similarity measure are negligible. For larger object and feature sets, Kullback-Leibler variants show a tendency to outperform other measures, confirming language-based results on distributional similarity by Lee (2001). Both frequencies and probabilities represent a useful basis for the verb distributions. A simple smoothing of the distributions supported the clustering, but to be sure of the effect one would need to experiment with solid smoothing methods. The number of clusters only plays a role concerning the magnitude of numbers. Inducing fine-grained clusters as given in the manual classification seems an ambitious intention, because the feature distinction for the classes is fine-grained, too. Inducing coarse clusters provides a coarse classification which is object to less noise and easier for manual correction.

Clustering Methodology In conclusion, I have presented a clustering methodology for German verbs whose results agree with the manual classification in many respects. 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, but the slight gain in clustering success will not be worth it; in addition, I would risk overfitting of the parameters. Without any doubt the cluster analysis needs manual correction and completion, but represents a plausible basis.

A large-scale experiment confirmed the potential of the clustering methodology. Based on the linguistic and practical insights, the large-scale cluster analysis resulted 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.

Compared to related work on clustering, my work is the first approach on automatic verb classification (i) where more than 100 verbs are clustered, (ii) which does not define a threshold on verb frequency, (iii) which evaluates the clustering result against fine-grained verb classes, (iv) which does not rely on restricted verb-argument structures, and (v) with a manual gold standard verb classification for evaluation purposes. In addition, the approach is the first one to cluster German verbs.

6.2 Directions for Future Research

There are various directions for future research, referring to different aspects of the thesis. The main ideas are illustrated in the following paragraphs.

Extension of Verb Classification The manual definition of the German semantic verb classes might be extended in order to include a larger number and a larger variety of verb classes. An extended classification would be useful as gold standard for further clustering experiments, and more general as manual resource in NLP applications. As a different idea, one might want to use the large-scale manual process classification by Ballmer and Brennenstuhl (1986) for comparison reasons.

Extension and Variation of Feature Description Possible features to describe German verbs might include any kind of information which helps classify the verbs in a semantically appropriate way. Within this thesis, I have concentrated on defining the verb features with respect to the alternation behaviour. Other features which are relevant to describe the behaviour of verbs are e.g. their auxiliary selection and adverbial combinations.

Variations of the existing feature description are especially relevant for the choice of selectional preferences. The experiment results demonstrated that the 15 conceptual GermaNet top levels are not sufficient for all verbs. For example, the verbs *töten* and *unterrichten* need a finer version of selectional preferences to be distinguished. It might be worth either to find a more appropriate level of selectional preferences in WordNet, or to apply a more sophisticated approach on selectional preferences such as the MDL principle (Li and Abe, 1998), in order to determine a more flexible choice of selectional preferences.

Clustering and Classification Techniques With respect to a large-scale classification of verbs, it might be interesting to run a classification technique on the verb data. The classification would presuppose more data manually labelled with classes, in order to train a classifier. But the resulting classifier might abstract better than k-Means over the different requirements of the verb classes with respect to the feature description.

As an extension of the existing clustering, I might apply a soft clustering algorithm to the German verbs. The soft clustering enables us to assign verbs to multiple clusters and therefore address the phenomenon of verb ambiguity. The clustering outcomes should be even more useful to discover new verb meaning components and semantically related classes, compared to the hard clustering technique.

NLP Application for Semantic Classes The verb clusters as resulting from the cluster analysis might be used within an NLP application, in order to prove the usefulness of the clusters. For example, replacing verbs in a language model by the respective verb classes might improve a language model's robustness and accuracy, since the class information provides more stable syntactic and semantic information than the individual verbs.