Chapter 1

Introduction

This thesis is concerned with experiments on the automatic induction of German semantic verb classes. In other words, (a) the focus of the thesis is verbs, (b) I am interested in a semantic classification of the verbs, and (c) the induction of the classification is performed automatically. Why this interest in verbs? What is the idea and usage of a verb classification? Why is there a focus on the semantic properties of the verbs, and what does the term 'semantic' refer to? And, last but not least, why and how is the classification performed by automatic means? Within this introductory chapter of the thesis, I will address the above questions as a motivation and definition of my work.

Central Role of the Verb The verb is an especially relevant part of the sentence, since it is central to the structure and the meaning of the sentence: The verb determines the number and kind of the obligatory and facultative participants within the sentence, and the proposition of the sentence is defined by the structural and conceptual interaction between the verb and the sentence participants.

For example, consider the German verb *liegen* 'to lie'. From the semantic point of view, the verb describes a state which demands an entity that lies and a place where the entity lies, as obligatory participants in the sentence. From the syntactic point of view, the entity is realised as the subject of the sentence, and the place is realised as a locative adverbial. Example (1.1) satisfies the demands and provides (i) a subject for the verb, which is semantically selected as an entity which has the ability to lie: the cat, and (ii) a prepositional phrase for the verb, whose head is a locative preposition and subcategorises a place: the sofa.

(1.1) *Die Katze liegt auf dem Sofa.*
'The cat lies on the sofa.'

Given a verb, we intuitively realise the lexically specific demands on the verb usage, i.e. as speakers of a language we know which kinds of participants are compatible with the selectional preferences of a verb, and which are the possibilities to structurally encode the grammatical functions for the participants. Therefore, the verb tells us the core information about the sentence.

Lexical Verb Resources in Natural Language Processing Within the area of Natural Language Processing (NLP), computational applications depend on reliable language resources. As demonstrated in the above example, verbs play a central role with respect to the structure and the meaning of the sentence, so resources on verb information are especially valuable. But it is tedious and rather impossible to manually define the details of human language, particularly when it comes to semantic knowledge. Therefore, lexical semantic resources represent a bottleneck in NLP, and methods for the acquisition of large amounts of semantic knowledge with comparably little manual effort have gained importance. Within this thesis, I am concerned with the potential and limits of creating a semantic knowledge base by automatic means, semantic classes for German verbs.

Lexical Semantics and Conceptual Structure Which notion of lexical semantics and conceptual structure is relevant for my work? A verb is lexically defined by its meaning components, those aspects of meaning which are idiosyncratic for the verb. But even though the meaning components are specific for a verb, parts of the conceptual semantic structure which the verb evokes might overlap for a number of verbs. Compare Example (1.2) with Example (1.1). The German verb *sitzen* ‘to sit’ expresses a different state as *liegen* ‘to lie’; the verbs therefore define different lexical concepts. But it is possible to define a more general conceptual structure on which the verbs agree: Both verbs describe an entity and a location where the entity is situated. The verbs agree on this conceptual level, and the difference between the verbs is created by the lexical semantic content of the verbs, which in this case defines the specific way of being in the location. The agreement on the conceptual level is the basis for defining verb classes.

(1.2) *Die Katze sitzt auf dem Sofa.*
‘The cat sits on the sofa.’

Semantic Verb Classes Verb classes are an artificial construct of natural language which generalises over verbs. They represent a practical means to capture large amounts of verb knowledge without defining the idiosyncratic details for each verb. The class labels refer to the common properties of the verbs within the class, and the idiosyncratic lexical properties of the verbs are either added to the class description or left underspecified. On the one hand, verb classes reduce redundancy in verb descriptions, since they encode the common properties of verbs; on the other hand, verb classes can predict and refine properties of a verb that received insufficient empirical evidence, with reference to verbs in the same class.

Semantic verb classes are a sub-type of verb classes and generalise over verbs according to their semantic properties. The class definition is based on a conceptual structure which comprises a number of semantically similar verbs. Examples for the conceptual structures are *Position* verbs such as *liegen* ‘to lie’, *sitzen* ‘to sit’, *stehen* ‘to stand’, and *Manner of Motion with a Vehicle* verbs such as *fahren* ‘to drive’, *fliegen* ‘to fly’, *rudern* ‘to row’.

But how can we obtain a semantic classification of verbs, avoiding a tedious manual definition of the verbs and the classes? A semantic classification demands a definition of semantic properties, but it is difficult to automatically induce semantic features from available resources, both with respect to lexical semantics and conceptual structure. Therefore, the construction of semantic classes typically benefits from a long-standing linguistic hypothesis which asserts a tight connection between the lexical meaning of a verb and its behaviour: To a certain extent, the lexical meaning of a verb determines its behaviour, particularly with respect to the choice of its arguments, cf. Levin (1993, page 1). We can utilise this meaning-behaviour relationship in that we induce a verb classification on basis of verb features describing verb behaviour (which are easier to obtain automatically than semantic features) and expect the resulting behaviour-classification to agree with a semantic classification to a certain extent.

However, it is still an open discussion (i) which exactly are the semantic features that define the verb classes, (ii) which exactly are the features that define the verb behaviour, and (iii) to what extent the meaning-behaviour relationship holds. Concerning (i), the semantic features within this thesis refer to conceptual class labels. Related work by Levin (1993) provides similar class labels, but she varies the semantic and syntactic content of the labels; related work in *FrameNet* (Baker *et al.*, 1998; Johnson *et al.*, 2002) explicitly refers to the conceptual idea of verb classes. The exact level of conceptual structure for the German verbs needs to be discussed within the experiments in this thesis.

Concerning (ii), a widely used approach to define verb behaviour is captured by the *diathesis alternation* of verbs, see for example Levin (1993); Dorr and Jones (1996); Lapata (1999); Schulte im Walde (2000a); Merlo and Stevenson (2001); McCarthy (2001); Joanis (2002). Alternations are alternative constructions at the syntax-semantic interface which express the same or a similar conceptual idea of a verb. In Example (1.3), the most common alternations for the *Manner of Motion with a Vehicle* verb *fahren* 'to drive' are illustrated. The participants in the conceptual structure are a driver, a vehicle, a driven person or thing, and a direction. Even if a certain participant is not realised within an alternation, its contribution might be implicitly defined by the verb. In (a), the vehicle is expressed as subject in a transitive verb construction, with a prepositional phrase indicating the direction of the movement. The driver is not expressed overtly, but we know that there is a driver. In (b), the driver is expressed as subject in a transitive verb construction, again with a prepositional phrase indicating the direction of the movement. The vehicle is not expressed overtly, but we know that there is a vehicle for the drive. In (c), the driver is expressed as subject in a transitive verb construction, with an accusative noun phrase indicating the vehicle. We know that there is a path for the movement, but it is not explicitly described. And in (d), the driver is expressed as subject in a ditransitive verb construction, with an accusative noun phrase indicating a driven person, and a prepositional phrase indicating the direction of the movement. Again, the vehicle is not expressed overtly, but we know that there is a vehicle for the drive.

(1.3) (a) *Der Wagen fährt in die Innenstadt.*
'The car drives to the city centre.'

(b) *Die Frau fährt nach Hause.*
'The woman drives home.'

(c) *Der Filius fährt einen blauen Ferrari.*
'The son drives a blue Ferrari.'

(d) *Der Junge fährt seinen Vater zum Zug.*
'The boy drives his father to the train.'

Assuming that the verb behaviour can be captured by the diathesis alternation of the verb, which are the relevant syntactic and semantic properties one would have to obtain for a verb description? The syntactic structures are relevant for the argument functions of the participants, the prepositions are relevant to distinguish e.g. directions from locations, and the selectional preferences of the conceptual entities are relevant, since they determine the participant roles. Therefore, I will choose exactly these three feature levels to describe the verbs by their behaviour.

Concerning (iii), the meaning-behaviour relationship is far from being perfect: It is not the case that verbs within the same semantic class behave the same, and it is not the case that verbs which behave the same are within the same semantic class. Consider the most specific conceptual level of semantic classes, a classification with classes of verb synonyms.[1] But even the verb behaviour of synonyms does not overlap perfectly, since e.g. selectional preferences of synonyms vary. For example, the German verbs *bekommen* and *erhalten* 'to get, to receive' are synonymous, but they cannot be exchanged in all contexts, cf. *einen Schnupfen bekommen* 'to catch a cold' vs. **einen Schnupfen erhalten*. Vice versa, consider the example that the two verbs *töten* 'to kill' and *unterrichten* 'to teach' behave similarly with respect to their subcategorisation properties, including a coarse level of selectional preference, such as a group or a person performing an action towards another person or group. They are similar on a very general conceptual level, so one might expect verbs with such similar behaviour to belong to the same semantic class on a more specific level of conceptual structure, but this is not the case. In conclusion, the meaning-behaviour relationship is valid to a certain extent, and it is an interesting task by itself to find the optimal level of overlap. Even though the relationship is not perfect, it supports the automatic induction of a semantic verb classification.

Clustering Methodology Assuming that we are provided with a feature description for verb behaviour, how can we obtain a semantic verb classification? I suggest a clustering algorithm which uses the syntactico-semantic descriptions of the verbs as empirical verb properties and learns to induce a semantic classification from this input data. The clustering of the German verbs is performed by the k-Means algorithm, a standard unsupervised clustering technique as proposed by Forgy (1965). With k-Means, initial verb clusters are iteratively re-organised by assigning each verb to its closest cluster and re-calculating cluster centroids until no further changes take place. Applying the k-Means algorithm assumes that (i) verbs are represented by distributional vectors. I follow the hypothesis that 'each language can be described in terms of a

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

distributional structure, i.e. in terms of the occurrence of parts relative to other parts', cf. Harris (1968), and define distributional vectors as verb description. And (ii) verbs which are closer to each other in a mathematically defined way are also more similar to each other in a linguistic way.

k-Means includes various cluster parameters: The number of clusters is not known beforehand, so the clustering experiments investigate this parameter. Related to this parameter is the level of conceptual structure: the more verb clusters are found, the more specific the conceptual level, and vice versa. The clustering input may be varied according to how much pre-processing we invest. k-Means is sensitive to the input, and the resulting cluster shape should match the idea of verb classes. I therefore experiment with random and pre-processed cluster input to investigate the impact of the input on the output. In addition, we can find various notions of defining the similarity between distributional vectors. But which does best fit the idea of verb similarity? The potential and the restrictions of the natural language clustering approach are developed with reference to a small-scale German verb classification and discussed and tested on the acquisition of a large-scale German verb classification.

Verb Class Usage What is the usage of the verb classes in Natural Language Processing applications? From a practical point of view, verb classes represent a lexical resource for NLP applications. On the one hand, verb classes reduce redundancy in verb descriptions, since they encode the common properties of verbs: a verb classification is a useful means for linguistic research, since it describes the verb properties and regularities at the syntax-semantic interface. On the other hand, verb classes can predict and refine properties of a verb that received insufficient empirical evidence, with reference to verbs in the same class: under this aspect, a verb classification is especially useful for the pervasive problem of data sparseness in NLP, where little or no knowledge is provided for rare events. Previous work at the syntax-semantic interface has proven the usefulness of verb classes: particularly the English verb classification by Levin (1993) has been used for NLP applications such as word sense disambiguation (Dorr and Jones, 1996), machine translation (Dorr, 1997), document classification (Klavans and Kan, 1998), and subcategorisation acquisition (Korhonen, 2002b).

Automatic Induction of German Semantic Verb Classes: Task Definition I summarise the thesis issues in an overall task definition. This thesis is concerned with experiments on the automatic induction of German semantic verb classes. To my knowledge, no German verb classification is available for NLP applications. Such a classification would therefore provide a principled basis for filling a gap in available lexical knowledge. However, the preceding discussion has shown that a classification of verbs is an interesting goal, but there are more tasks on the way which have not been addressed. The overall idea of inducing verb classes is therefore split into the following sub-goals.

Firstly, I perform an empirical investigation of the practical usage of the relationship between verb behaviour and meaning components. As said before, it is still an open discussion (i) which exactly are the semantic features that define verb classes, (ii) which exactly are the features that define verb behaviour, and (iii) to what extent the meaning-behaviour relationship holds. This thesis will investigate the relationship between verb features, where the semantic features refer to various levels of conceptual structure, and the syntactic features refer to various levels of verb alternation behaviour. In addition, I will investigate the practical usage of the theoretical hypothesis, i.e. is there a benefit in the clustering if we improve the syntax-semantic interface?

Secondly, I aim to develop a clustering methodology which is suitable for the demands of natural language. As described above, I apply the hard clustering technique k-Means to the German verb data. I decided to use the k-Means algorithm for the clustering, because it is a standard clustering technique with well-known properties. The reader will learn that there are other clustering and classification techniques which might fit better to some aspects of the verb class task, e.g. with respect to verb ambiguity. But k-Means is a good starting point, because it is easy to implement the algorithm and vary the clustering parameters, and the relationship between parameters and clustering result is easy to follow and interpret.

Finally, I bring together the insights into the meaning-behaviour relationship and the experience with clustering, in order to investigate the automatic acquisition of German semantic verb classes. As obvious from the discussions, the clustering outcome will not 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 it should be clear by now that the goal of this thesis is not necessarily to obtain the optimal clustering result, but to understand what is happening. Only in this way we can develop a methodology which abstracts from the given, small-scale data and can be applied to a large-scale application.

Contributions of this Thesis The contribution of my work comprises three parts. Each of the parts may be used independently from the others, for various purposes in NLP.

1. <u>A small-scale German verb classification</u>

   I manually define 43 German semantic verb classes containing 168 partly ambiguous German verbs. The verb classes are described on the conceptual level and illustrated by corpus examples at the syntax-semantic interface. Within this thesis, the purpose of this manual classification is 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).

2. A statistical grammar model for German

   I describe the implementation and training of a German lexicalised probabilistic context-free grammar. The statistical grammar model provides empirical lexical information, specialising on but not restricted to the subcategorisation behaviour of verbs. The empirical data are useful for any kind of lexicographic work. For example, Schulte im Walde (2003a) presents the range of lexical data which are available in the statistical grammar model, concentrating on verb and noun collocations. And Schulte im Walde (2002b) describes the induction of a subcategorisation lexicon from the grammar model, with Schulte im Walde (2002a) referring to the evaluation of the subcategorisation data against manual dictionary entries.

3. A clustering methodology for NLP semantic verb classes

   I present clustering experiments which empirically analyse and utilise the assumption of a syntax-semantic relationship between verb meaning and verb behaviour. Based on the experimental results, I define the relevant aspects of a clustering methodology which can be applied to automatically induce a semantic classification for German verbs. The variation of the clustering parameters illustrates both the potential and limit of (i) the relationship between verb meaning components and their behaviour, and (ii) the utilisation of the clustering approach for a large-scale semantic verb classification as lexical NLP resource.

Overview of Chapters The chapters are organised as follows.

Chapter 2 describes the manual definition of the small-scale German semantic verb classes. As said above, the purpose of the manual classification within this thesis is to evaluate the reliability and performance of the clustering experiments. The chapter introduces the general idea of verb classes and presents related work on verb class definition in various frameworks and languages. The German classification is described in detail, to illustrate the syntactic, lexical semantic and conceptual properties of the verbs and verb classes, and to present a basis for discussions about the clustering experiments and outcomes. The final part of the chapter refers to the usage of verb classes in Natural Language Processing applications, in order to show the potential of a verb classification.

Chapter 3 describes the German statistical grammar model. The model serves as source for the German verb description at the syntax-semantic interface, which is used within the clustering experiments. The chapter introduces the theoretical background of lexicalised probabilistic context-free grammars and describes the German grammar development and implementation, the grammar training and the resulting statistical grammar model. The empirical lexical information in the grammar model is illustrated, and the core part of the verb information, the subcategorisation frames, are evaluated against manual dictionary definitions.

Chapter 4 provides an overview of clustering algorithms and evaluation methods which are relevant for the natural language task of clustering verbs into semantic classes. The chapter introduces clustering theory and relates the theoretical assumptions to the induction of verb classes. A range of possible evaluation methods are described, and relevant measures for a verb classification are determined.

Chapter 5 presents the clustering experiments which investigate the automatic induction of semantic classes for German verbs. The clustering data are described by introducing the German verbs and the gold standard verb classes from an empirical point of view, and by illustrating the verb data and feature choice. The clustering setup, process and results are presented, followed by a detailed interpretation and a discussion of possibilities to optimise the experiment setup and performance. The preferred clustering methodology is applied to a large-scale experiment on 883 German verbs. The chapter closes with related work on clustering experiments.

Chapter 6 discusses the contributions of the thesis and suggests directions for future research. The main focus of the contributions interprets the clustering data, the clustering experiments, and the clustering results with respect to the empirical relationship between verb meaning and verb behaviour, the development of a methodology for natural language clustering, and the acquisition of semantic verb classes.