AIMS

Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung

Lehrstuhl für Theoretische Computerlinguistik Universität Stuttgart

2003 • Vol. 9 • No. 2

Experiments on the Automatic Induction of German Semantic Verb Classes

Sabine Schulte im Walde

AIMS
Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung
Universität Stuttgart

ISSN 1434-0763

AIMS

In January 1994 we initiated a series of working papers: AIMS. AIMS stands for ’Arbeitpapiere des Instituts für Maschinelle Sprachverarbeitung’ (Working Papers of the Department of Natural Language Processing), at University of Stuttgart. So far every issue of AIMS is centered around a specific topic, either by presenting individual dissertations or by compiling articles and master theses on this topic written by members of the IMS and associates. The work has pre-publication status and might appear in other form elsewhere. We invite other departments and sites to exchange their reports with us. Otherwise our working papers are also available at the cost of 20 EUR per issue plus costs for shipping and handling. To initiate an exchange, or to obtain the working papers on a regular or issue-wise basis, contact:

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Azenbergstr. 12
70174 Stuttgart, Germany

aims@IMS.Uni-Stuttgart.DE
Tel.: +49 (0)711 121 1379
Fax: +49 (0)711 121 1366
www.ims.uni-stuttgart.de/phonetik/

Previous Issues of AIMS

Vol. 1 (1) 1994: Wigner Distribution in Speech Research. Master thesis by Wolfgang Wokurek (in German), papers by Grzegorz Dogil, Wolfgang Wokurek, and Krzysztof Marasek (in English), and a bibliography.

Vol. 2 (1) 1995: Sprachentstörung. Doctoral Dissertation, University of Vienna, 1994 by Wolfgang Wokurek (in German with abstract in English). Full title: Sprachentstörung unter Verwendung eines Lautklassendetektors (Speech enhancement using a sound class detector).

Vol. 2 (2) 1995: Word Stress. Master thesis by Stefan Rapp (in German) and papers mostly by Grzegorz Dogil, Michael Jessen, and Gabriele Scharf (in English).

Vol. 2 (3) 1995: Language and Speech Pathology. Master theses by Gabriele Scharf and by Jörg Mayer (in German) and papers by Hermann Ackermann, Ingo Hertrich, Jürgen Konczak, and Jörg Mayer (mostly in German).

Vol. 3 (1) 1997: Tense versus Lax Obstruents in German. Revised and expanded version of Ph.D. Dissertation, Cornell University 1996 by Michael Jessen (in English). Full title: Phonetics and phonology of the tense and lax obstruents in German.

Vol. 3 (2) 1997: Electroglottographic Description of Voice Quality. Habilitationsschrift, University of Stuttgart, by Krzysztof Marasek (in English).

Vol. 3 (3) 1997: Aphasie und Kernbereiche der Grammatiktheorie (Aphasia and core domains in the theory of grammar). Doctoral Dissertation, University of Stuttgart, by Annegret Bender (in German with abstract in English).

Vol. 3 (4) 1997: Intonation und Bedeutung (Intonation and meaning). Doctoral Dissertation, University of Stuttgart, by Jörg Mayer (in German with summary in English).

Vol. 3 (5) 1997: Koartikulation und glottale Transparenz (Coarticulation and glottal transparency). Doctoral Dissertation, University of Bielefeld, by Kerstin Vollmer (in German with abstract in English).

Vol. 3 (6) 1997: Der TFS-Repräsentationsformalismus und seine Anwendung in der maschinellen Sprachverarbeitung (The TFS Representation Formalism and its Application to Natural Language Processing). Doctoral Dissertation, University of Stuttgart, by Martin C. Emele (in German with summary in English).

Vol. 4 (1) 1998: Automatisierte Erstellung von Korpora für die Prosodieforschung (Automated generation of corpora for prosody research). Doctoral Dissertation, University of Stuttgart, by Stefan Rapp (in German with abstract in English).

Vol. 4 (2) 1998: Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese (Theory-based modelling of German intonation for speech synthesis). Doctoral Dissertation, University of Stuttgart, by Gregor Möhler (in German with abstract in English).

Vol. 4 (3) 1998: Inducing Lexicons with the EM Algorithm. Papers by M. Rooth, S. Riezler, D. Prescher, S. Schulte im Walde, G. Carroll, F. Beil (in English).

Vol. 5 (1) 1999: Probabilistic Constraint Logic Programming. Doctoral Dissertation, University of Stuttgart, by Stefan Riezler (in English).

Vol. 5 (2) 1999: Timing bei Dysarthrophonie (Timing in Dysarthria). Doctoral Dissertation, University of Stuttgart, by Gabriele Scharf (in German with abstract in English).

Vol. 5 (3) 1999: Semantikkonstruktion (Semantic Construction). Doctoral Dissertation, University of Stuttgart, by Michael Schiehlen (in German with abstract in English).

Vol. 6 (1) 2000: YAP: Parsing and Disambiguation with Feature-Based Grammars. Doctoral Dissertation, University of Stuttgart, by Helmut Schmid (in English).

Vol. 6 (2) 2000: Semantikbasierter Transfer - Maschinelle Übersetzung auf Mengen linguistischer Constraints. Doctoral Dissertation, University of Stuttgart, by Michael Dorna (in German).

Vol. 6 (3) 2000: The Contribution of Intonation to the Perception of Foreign Accent. Doctoral Dissertation, University of Stuttgart, by Matthias Jilka (in English).

Vol. 6 (4) 2000: Research Papers from the Phonetics Lab. Contributions by Kathrin Claßen, Grzegorz Dogil, Michael Jessen, Bernd Möbius, Karin Müller, Detlef Prescher and Wolfgang Wokurek (most papers in English).

Vol. 7 (1) 2001: Automatische Segmentierung, Verwaltung und Abfrage von Korpora gesprochener Sprache (Automated Segmentation, Administration and Retrieval of Spoken Language Corpora). Doctoral Dissertation, University of Stuttgart, by Marcus L. Fach (in German with summary in English).

Vol. 7 (2) 2001: Topics in the Clausal Syntax of German. Doctoral Dissertation, University of Stuttgart, by Judith Berman (in English).

Vol. 7 (3) 2001: Formal and Computational Aspects of Optimality-Theoretic Syntax. Doctoral Dissertation, University of Stuttgart, by Jonas Kuhn (in English).

Vol. 7 (4) 2001: German and Multilingual Speech Synthesis. Habilitationsschrift, University of Stuttgart, by Bernd Möbius (in English).

Vol. 8 (1) 2002: The Automatic Acquisition of Lexical Information from Portuguese Text Corpora with a Probabilistic Context-free Grammar. Doctoral Dissertation, University of Stuttgart, by Erika Fonseca de Lima (in English).

Vol. 8 (2) 2002: EM-basierte maschinelle Lernverfahren für natürliche Sprachen (EM–Based Maschine Learning in Natural Language Processing). Doctoral Dissertation, University of Stuttgart, by Detlef Prescher (in German with summary in English).

Vol. 8 (3) 2002: Probabilistic Syllable Modeling Using Unsupervised and Supervised Learning Methods. Doctoral Dissertation, University of Stuttgart, by Karin Müller (in English).

Vol. 8 (4) 2002: Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. Doctoral Dissertation, University of Stuttgart, by Wolfgang Lezius (in German with summary in English).

Vol. 9 (1) 2003: Implementing and Documenting Large-Scale Grammars – German LFG. Doctoral Dissertation, University of Stuttgart, by Stefanie Dipper (in English with summary in German).

Vol. 9 (2) 2003: Experiments on the Automatic Induction of German Semantic Verb Classes. Doctoral Dissertation, University of Stuttgart, by Sabine Schulte im Walde (in English with summary in German).

Experiments on the Automatic Induction of German Semantic Verb Classes

Von der Philosophisch-Historischen Fakultät der Universität Stuttgart
zur Erlangung der Würde eines Doktors der
Philosophie (Dr. phil.) genehmigte Abhandlung

Vorgelegt von

Sabine Schulte im Walde

aus Rheine

Hauptberichter:	Prof. Dr. h.c. Hans Kamp, PhD
Mitberichter:	Assistant Prof. Chris Brew, D.Phil.
Tag der mündlichen Prüfung:	18. Juni 2003

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

2003

D93

Acknowledgements

The work on this thesis was performed while I was a member of the DFG-funded PhD program 'Graduiertenkolleg' *Sprachliche Repräsentationen und ihre Interpretation*. I was supervised internally by Hans Kamp at the IMS, and externally by Chris Brew at the Ohio State University. To both of them as well as to Helmut Schmid for his permanent unofficial supervision I am grateful for their interest in my work, their motivation and feedback, and especially for the fact that their doors were always open for me. In addition, Stefan Evert took part in my work by repeatedly discussing mathematical questions and problems with me; and the evaluation of subcategorisation frames in Chapter 3 was performed in an enjoyable collaboration with *Bibliographisches Institut & F. A. Brockhaus AG*.

On various occasions I had the opportunity to receive valuable motivation and feedback from (computational) linguists working in related research. For this I wish to thank Ana Fernández, Mirella Lapata, Diana McCarthy, Paola Merlo, Patrick Saint-Dizier and Gloria Vázquez who followed up my invitations to Stuttgart; Eric Joanis, Anna Korhonen, Lillian Lee and Suzanne Stevenson who were willing to meet with me during my visits, and Uli Heid and John Nerbonne who were around in Stuttgart. Furthermore, I received many valuable comments at various conferences and workshops.

I am grateful to the professors at the IMS, Grzegorz Dogil, Hans Kamp and Christian Rohrer, as well as to the administrative staff, particularly Sabine Dieterle, Edgar Hoch and Sybille Laderer, who all took care that my environment was both working and comfortable. In addition, the 'Mensa-Runde' including Stefanie Dipper, Arne Fitschen, Martin Forst, Josef van Genabith, Effi Georgala, Jonas Kuhn, Wolfgang Lezius, Tina Säuberlich and Jasmin Saric made sure that I could forget about work for at least one hour per day, enjoying an entertaining meal in the university's cafeteria, possibly outside in the sun, but certainly not with a coffee after lunch. For a change in the daily routine, Inga Frese and Michael Klein joined me to explore the 'Mittagstisch' neighbourhood of the IMS.

Special thanks go to my office-mate Heike Zinsmeister with whom I enjoyed to discussing work matters as well as questions and problems concerning social issues. Furthermore, I am so lucky to have a loving family and many friends inside and outside the computational linguistics world who believe in me even if I stop doing so. My volleyball team provided for the physical and for yet another social aspect of my life. And, last but not least, most thanks go to Matthias who shared both the fun and the burden of my work with me.

Contents

List of Tables

List of Figures

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.

Chapter 2

German Verb Classes

This chapter describes the manual creation of German semantic verb classes: 168 partly ambiguous German verbs are classified into 43 semantic classes. The verb classes are described on the conceptual level and illustrated by corpus examples at the syntax-semantic interface. The purpose of the manual classification within this thesis is to evaluate the reliability and performance of the clustering experiments, both with respect to (i) the underlying relationship between verb meaning components and verb behaviour, and (ii) the usage of clustering methodologies for the automatic acquisition of a high-quality and large-scale lexical semantic resource. Besides using the verb classification as gold standard in the clustering evaluation within this thesis, its size is 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).

Section 2.1 introduces the general idea of verb classes and illustrates the German verb classes as well as related theoretical work on verb classes for various frameworks and languages. In Section 2.2, the German verb classes are exemplified in some detail, to illustrate the properties as well as the similarities and differences of verbs and verb classes, and to present a basis for discussions about the clustering experiments and outcomes. Section 2.3 describes applications of verb classes in Natural Language Processing, in order to show the potential of a large-scale verb classification.

2.1 Idea and Definition of Verb Classes

In Section 2.1.1, I start with an introduction to the basic idea of verb classes, before the German verb classification is presented in Section 2.1.2. Section 2.1.3 describes related work on the creation of verb classes for various frameworks and languages.

2.1.1 Idea of Verb Classes

A classification is a mapping which assigns objects to classes in such a way that objects in the same class are as similar as possible and objects in different classes are as dissimilar as possible. A classification is a means (i) to reduce redundancy in object descriptions by referring to the common properties of objects, and (ii) to predict and refine properties of an object that received insufficient empirical evidence, with reference to objects in the same class.

Verb classes are a natural language classification which generalises over verbs. A priori, the properties underlying the verb classes might be of various kinds, semantic, syntactic, morphological, aspectual, etc. As introduced in Chapter 1, this thesis concentrates on a semantic verb classification: I define verb classes which generalise over verbs according to their semantic properties. The verb class labels refer to the common semantic properties of the verbs in a class at a general conceptual level, and the idiosyncratic lexical semantic properties of the verbs are left underspecified.

My construction of the German verb classes is therefore based on semantic intuition. Verbs are assigned to classes according to their similarity of lexical and conceptual meaning, and each verb class is assigned a semantic class label. But as mentioned before, a long-standing linguistic hypothesis asserts a tight connection between the meaning components of a verb and its behaviour, cf. Pinker (1989); Levin (1993). Because of the meaning-behaviour relationship at the syntax-semantic interface, the verbs grouped in one class show a certain agreement in their behaviour. Even though the syntax-semantic relationship is not perfect, the verb behaviour can in turn be utilised in order to probe for class membership: lexical properties which are induced from corpus data and computational resources represent a basis for the automatic induction of verb classes.

2.1.2 Verb Classification

In this section I present the manual classification of the German verbs. Without any doubt, each reader will disagree on some aspect of the verb classes, whether because of the class label or because of the membership of a verb. I do neither claim completeness nor a general consensus of the verb classes, but I am aware of the subjective character of the definition. The evidence used in the class creation process was given by subjective conceptual knowledge, monolingual and bilingual dictionary entries and corpus search. The purpose of the manual classification within this thesis is to evaluate the reliability and performance of the clustering experiments in Chapter 5. In addition, the classification might be useful for further purposes in NLP applications, cf. Section 2.3.

I classified 168 verbs into 43 semantic verb classes. The class size is between 2 and 7, with an average of 3.9 verbs per class. Eight verbs are ambiguous with respect to class membership and marked by subscripts. The classes include both high and low frequency verbs, in order to exercise the clustering technology in both data-rich and data-poor situations: the corpus frequencies of the verbs range from 8 to 71,604. The class labels are given on two semantic levels; coarse labels such as *Manner of Motion* are sub-divided into finer labels, such as *Locomotion, Rotation, Rush, Vehicle, Flotation*. The fine labels are relevant for the clustering experiments, as the numbering indicates. I tried to balance the classification not to include any kind of bias, i.e. in the classification are no majorities of high frequent verbs, low frequent verbs, strongly ambiguous verbs, verbs from specific semantic areas, etc. Any bias in the classification could influence the evaluation of clustering methods.

As mentioned before, the classification is primarily based on semantic intuition, not on facts about the syntactic behaviour. As an extreme example, the *Support* class (23) contains the verb *unterstützen*, which syntactically requires a direct object, together with the three verbs *dienen, folgen, helfen* which dominantly subcategorise an indirect object.

German Semantic Verb Classes:

1. *Aspect*: anfangen, aufhören, beenden, beginnen, enden
2. *Propositional Attitude*: ahnen, denken, glauben, vermuten, wissen

- *Desire*
  3. *Wish*: erhoffen, wollen, wünschen
  4. *Need*: bedürfen, benötigen, brauchen

5. *Transfer of Possession (Obtaining)*: bekommen, erhalten, erlangen, kriegen

- *Transfer of Possession (Giving)*
  6. *Gift*: geben, leihen, schenken, spenden, stiften, vermachen, überschreiben
  7. *Supply*: bringen, liefern, schicken, vermitteln$_1$, zustellen
- *Manner of Motion*
  8. *Locomotion*: gehen, klettern, kriechen, laufen, rennen, schleichen, wandern
  9. *Rotation*: drehen, rotieren
  10. *Rush*: eilen, hasten
  11. *Vehicle*: fahren, fliegen, rudern, segeln
  12. *Flotation*: fließen, gleiten, treiben
- *Emotion*
  13. *Origin*: ärgern, freuen
  14. *Expression*: heulen$_1$, lachen$_1$, weinen
  15. *Objection*: ängstigen, ekeln, fürchten, scheuen

16. *Facial Expression*: gähnen, grinsen, lachen$_2$, lächeln, starren
17. *Perception*: empfinden, erfahren$_1$, fühlen, hören, riechen, sehen, wahrnehmen
18. *Manner of Articulation*: flüstern, rufen, schreien
19. *Moaning*: heulen$_2$, jammern, klagen, lamentieren

20. *Communication*: kommunizieren, korrespondieren, reden, sprechen, verhandeln

- *Statement*
    21. *Announcement*: ankündigen, bekanntgeben, eröffnen, verkünden
    22. *Constitution*: anordnen, bestimmen, festlegen
    23. *Promise*: versichern, versprechen, zusagen

24. *Observation*: bemerken, erkennen, erfahren$_2$, feststellen, realisieren, registrieren
25. *Description*: beschreiben, charakterisieren, darstellen$_1$, interpretieren
26. *Presentation*: darstellen$_2$, demonstrieren, präsentieren, veranschaulichen, vorführen
27. *Speculation*: grübeln, nachdenken, phantasieren, spekulieren
28. *Insistence*: beharren, bestehen$_1$, insistieren, pochen
29. *Teaching*: beibringen, lehren, unterrichten, vermitteln$_2$

- *Position*
    30. *Bring into Position*: legen, setzen, stellen
    31. *Be in Position*: liegen, sitzen, stehen

32. *Production*: bilden, erzeugen, herstellen, hervorbringen, produzieren
33. *Renovation*: dekorieren, erneuern, renovieren, reparieren
34. *Support*: dienen, folgen$_1$, helfen, unterstützen
35. *Quantum Change*: erhöhen, erniedrigen, senken, steigern, vergrößern, verkleinern
36. *Opening*: öffnen, schließen$_1$
37. *Existence*: bestehen$_2$, existieren, leben
38. *Consumption*: essen, konsumieren, lesen, saufen, trinken
39. *Elimination*: eliminieren, entfernen, exekutieren, töten, vernichten
40. *Basis*: basieren, beruhen, gründen, stützen
41. *Inference*: folgern, schließen$_2$
42. *Result*: ergeben, erwachsen, folgen$_2$, resultieren
43. *Weather*: blitzen, donnern, dämmern, nieseln, regnen, schneien

Regarding the overall task of automatically inducing the verb classes, the reader could ask why I deliberately complicate the clustering task. I could either (i) define the verb classes on a purely syntactic basis, since the meaning-behaviour mapping is not perfect, and syntactic properties are easier to induce from non-annotated corpus data than semantic properties, so a syntactic classification would be more promising for clustering. Or I could (ii) define larger classes of verbs, such that the distinction between the classes is not based on fine-grained verb properties which are difficult to induce from corpus data. In addition, I could (iii) disregard clustering complications such as verb ambiguity and low-frequency verbs.

But the overall goal is not to perform a perfect clustering on the given 168 verbs, but to develop a clustering methodology which is suitable for the demands of natural language and allows the automatic acquisition of a high-quality and large-scale lexical semantic resource. For that, we should learn as much as possible about the practical aspects of the theoretical relationship between verb behaviour and meaning components.

2.1.3 Related Work on Verb Classes

The current section describes related work on the construction of verb classes. It starts by describing a German valency dictionary (Helbig and Schenkel, 1969), which does not induce a classification of verbs, but does provide verb information on both the syntactic valency and the semantic distribution of German verbs. The next part describes criteria for an aspectual classification system for English verbs (Vendler, 1967; Dowty, 1979). The lexicographic database frameworks FrameNet (Baker *et al.*, 1998; Johnson *et al.*, 2002) and WordNet (Miller *et al.*, 1990; Fellbaum, 1998) have defined verb classes for English, and have since been transfered to additional languages. The final section presents multi-lingual semantic classifications for verbs, which refer to the syntax-semantic interface of verb properties and which are developed for Natural Language Processing, so they are most similar to the background of this work.

A Dictionary of Verb Valency and Distribution

Helbig and Schenkel (1969) do not induce a classification of verbs, but provide information on both the syntactic valency and the semantic distribution of German verbs, in the form of a dictionary. The purpose of the dictionary is German second language teaching. Therefore, the dictionary concentrates on the most common and the most difficult German verbs.

The dictionary contains approximately 500 verbs. The verb entries are ordered alphabetically. Each verb in the dictionary is described on three levels: Level (i) defines the number of complements for the verb, distinguishing obligatory and facultative complements on the basis of an elimination test. The number of obligatory complements is given without, and the number of facultative complements is given within brackets. Level (ii) gives qualitative syntactic information, by naming the kinds of complements the verb may take. Among the possible complement categories are noun phrases with case definition (Nom, Akk, Dat, Gen), prepositional phrases (Prep), adjectives (Adj) and adverbs (Adv), verbal infinitives (Inf) and participles (Part), and subordinated clauses (Sent). Level (iii) then associates the complements with semantic labels, providing selectional preferences on a coarse level, such as human (hum), or abstract (abstr). Ambiguous verbs are assigned multiple verb entries. Multiple verb senses are distinguished by paraphrases on level (i).

Example entry for *achten* ‘to respect’ and ‘to care for, to pay attention’:

I.	achten$_2$ (V$_1$ = *hochschätzen* ‘to respect’)	
II.	achten → Nom, Akk	
III.	Nom →	1. Hum : *Die Schüler* achten den Lehrer. ‘*The pupils* respect the teacher.’
		2. Abstr : *Die Universität* achtet den Forscher. ‘*The university* respects the researcher.’
	Akk →	1. Hum : Wir achten *den Lehrer*. ‘We respect *the teacher*.’
		2. Abstr : Wir achten *die Regierung*. ‘We respect *the government*.’
		3. Abstr : Wir achten *seine Meinung*. ‘We respect *his opinion*’
I.	achten$_2$ (V$_2$ = *aufpassen* ‘to take care, to pay attention’)	
II.	achten → Nom, Prep/Sent$_{dass}$/Inf	
III.	Nom →	1. Hum : *Das Kind* achtet auf die Worte des Vaters. ‘*The child* pays attention to the father’s words.’
		2. Abstr : *Die Schule* achtet auf Pünktlichkeit. ‘*The school* takes care of punctuality.’
	Prep = auf,	
	Prep$_{Akk}$ →	no selectional restrictions : Er achtet *auf das Kind, auf den Hund, auf den Betrieb, auf das Pfeifen, auf die Schwierigkeiten*. ‘He pays attention to *the child, the dog, the company, the whistle, the problems*.’
	Sent →	Act : Der Lehrer achtet darauf, *daß niemand abschreibt*. ‘The teacher pays attention *that nobody copies*.’
	Inf →	Act : Der Lehrer achtet darauf, *verständlich zu sprechen*. ‘The teacher pays attention *to speak clearly*.’

The verb information Helbig and Schenkel provide is very similar to my verb description to follow, in that they describe and illustrate the possible verb arguments on both the syntactic and semantic level. The dictionary has therefore been used as part of evaluating the statistical subcategorisation information, cf. Chapter 3.

Aspectual Verb Classification

Vendler (1967) suggests a classification for English verbs on basis of their aspectual properties. He proposes the four categories *activities, accomplishments, achievements, states* and demarcates the categories by aspectual verb properties according to their restrictions on time adverbials, tenses, and logical entailment. For example, activities but not states can be used in the progressive and as imperatives, as illustrated in Examples (2.1) and (2.2).

(2.1) John is running home.
*John is knowing the answer.

(2.2) Run home!
*Know the answer!

Dowty (1979) adjusts the targets of Vendler's classification to verb phrases instead of verbs. He adopts Vendler's verb classes and provides an aspect calculus for the classification. The calculus is based on a lexical decompositional analysis of the verb classes, which takes into account a more refined version of restrictions on aspect markers compared to Vendler: Dowty uses (i) a distinction between statives and non-statives based on verb properties such as the tense behaviour (e.g. progressive usage, the habitual interpretation of simple present tense), the combination with adverbs such as *deliberately*, the appearance in pseudo-cleft constructions, (ii) restrictions on the form and entailments of time adverbials, and (iii) the interpretation of verb phrases with indefinite plural or mass-noun direct objects. The decompositional definition of the classes refers to logical operators such as CAUSE, BECOME and DO. For example, simple statives are defined by $\pi(\alpha_1, ..., \alpha_n)$, simple activities are defined by $DO\ (\alpha_1, [\pi(\alpha_1, ..., \alpha_n)])$, inchoation of activity is defined by $BECOME\ [DO\ (\alpha_1, [\pi_n(\alpha_1, ..., \alpha_n)])]$.

Neither Vendler nor Dowty provide an explicit classification; rather, they suggest those properties of verbs which are relevant for a classification as based on verb aspect. Their ideas appear in various studies of the lexical semantic properties of verbs, e.g. Jackendoff's lexical conceptual structure (1983; 1990), Pustejovsky's event structure (1991; 1995), Levin's and Rappaport Hovav's lexical templates (1998), and Engelberg's lexical event structure (2000b; 2000a).

FrameNet

The Berkeley *FrameNet* project (Baker *et al.*, 1998; Johnson *et al.*, 2002) produces frame-semantic descriptions for English verbs, nouns, and adjectives. The descriptions are based on Fillmore's scenes-and-frames semantics (Fillmore, 1977, 1982), where a scene describes the conceptual embedding of a word, and the usage of the word is captured by the range of syntactic and semantic combinatory possibilities (frame variants). The aim of the project is to document the range of combinatory valences of each word in each of its senses. The descriptions build a computational lexicographic database resource, which contains basic conceptual structures, participating elements, and annotated example sentences. Recent work in Saarbrücken is in the early stages of a German version of FrameNet (Erk *et al.*, 2003), in order to semantically annotate the German *TIGER* corpus (Brants *et al.*, 2002).

The conceptual structures in the database posit a hierarchical classification structure on the English words. A subframe in the database can inherit elements and semantics from its parent. An example class is the *transportation* frame, with one subframe *driving*. The role names in the frames are local to the particular conceptual structure.

Frame: *Transportation*

Frame-Elements(Mover, Means, Path)
Scene(Mover move along Path by Means)

Frame: *Driving*

Inherit(Transportation)
Frame-Elements(Driver(=Mover), Vehicle(=Means), Rider(=Mover), Cargo(=Mover))
Scenes(Driver starts Vehicle, Driver controls Vehicle, Driver stops Vehicle)

Grammatical functions describe the ways in which the participants of a conceptual structure may be realised. The functions are illustrated by annotated corpus sentences from the British National Corpus. The following examples are taken from (Baker *et al.*, 1998).

Frame Elements	Example
Driver	[$_D$ Kate] **drove** [$_P$ home] in a stupor.
Driver, Path	And that was why [$_D$ I] **drove** [$_P$ eastwards along Lake Geneva].
Driver, Rider, Path	Now [$_D$ Van Cheele] was **driving** [$_R$ his guest] [$_P$ back to the station].

A lexical entry in the database is defined by its belonging to a conceptual class (such as *to drive* being a member of the transportation frame), the inherited frame elements, and a list of valency patterns for the respective entry, according to the corpus data.

The detailed description of my German verb classes (Section 2.2) relies on FrameNet's classification ideas. As in FrameNet, the semantic content of the German verb classes is characterised by the conceptual embedding of the verbs, participant roles are defined, and corpus examples underline the classification.

WordNet and EuroWordNet

WordNet is a lexical semantic ontology developed at the University of Princeton (Miller *et al.*, 1990; Fellbaum, 1998). The ontology organises English nouns, verbs, adjectives and adverbs into synonymous conceptual classes, which are connected by lexical and conceptual relations such as hyponymy, hypernymy, meronymy, etc. The hypernym-hyponym relation imposes a hierarchical structure on the ontology. Words with several senses are assigned to multiple classes. The lexical database is inspired by psycholinguistic research on human lexical memory.

The WordNet verb classes are described by (a) the verbs in the classes themselves, possibly (b) a comment on the class concept, (c) syntactic frame types, and possibly (d) example sentences. Within WordNet 1.5, a total of 14,734 English verbs with 25,768 verb senses are classified into 11,363 verb concepts. The following example illustrates the nine senses of the verb *to drive*, accompanied by the respective synonyms, the comments, frame types and super-ordinated classes, i.e. hypernym concepts indicated by '$\Rightarrow$'. WordNet is known for defining senses on a rather fine-grained basis. It provides little syntactic information; the frame types are given an intuitive notation.

1. **Concept:** *drive, operate a vehicle*

   Comment: drive a car or bus, for example

   Frame: somebody ___s something

   Hierarchy: drive, operate a vehicle
   ⇒ operate, control
   ⇒ manipulate
   ⇒ handle, palm
   ⇒ touch

2. **Concept:** *drive, motor*

   Comment: travel in a vehicle

   Frames: somebody ___s, somebody ___s PP

   Hierarchy: drive, motor
   ⇒ travel, go, move, locomote

3. **Concept:** *drive*

   Comment: cause someone or something to move by driving

   Frames: somebody ___s somebody PP, somebody ___s something PP

   Hierarchy: drive
   ⇒ move, displace, make move

   Example: *She drove me to school every day.*

4. **Concept:** *force, run, drive, ram*

   Comment: physical or metaphorical

   Frame: somebody ___s something PP

   Hierarchy: force, run, drive, ram
   ⇒ thrust
   ⇒ push, force
   ⇒ move, displace, make move

   Example: *She rammed her mind into focus.*

5. **Concept:** *drive*

   Frame: somebody ___s somebody to INFINITIVE

   Hierarchy: drive
   ⇒ induce, stimulate, cause, have, get, make

   Example: *He drove me to divorce my wife.*

6. **Concept:** *drive*

   Comment: to compel or force or urge relentlessly or exert coercive pressure on

   Frames: somebody ___s somebody, something ___s somebody

   Hierarchy: drive
   ⇒ coerce, pressure, force
   ⇒ compel, oblige, obligate
   ⇒ induce, stimulate, cause, have, get, make

   Example: *She is driven by her passion.*

7. **Concept:** *tug, labor, labour, push, drive*

   Comment: exert oneself

   Frame: somebody ___s to INFINITIVE

   Hierarchy: tug, labor, labour, push, drive
   ⇒ fight, struggle
   ⇒ try, seek, attempt, essay, assay
   ⇒ act, move, take measures, take steps, take a step, take action, perform an action, do something

   Example: *She tugged for years to make a decent living.*

8. **Concept:** *pull, drive*

   Comment: of a car

   Frame: something is ___ing PP

   Hierarchy: pull, drive
   ⇒ travel, go, move, locomote

   Example: *The van pulled up.*

9. **Concept:** *drive, ride*

   Comment: have certain properties when driven

   Frame: something ___s Adjective/Noun

   Example: *My new truck drives well.*

The idea of WordNet has been transfered to other languages than English, most extensively within the framework of *EuroWordNet* (Vossen, 1999). Multi-lingual wordnets are created for the languages Dutch, Italian, Spanish, French, German, Czech and Estonian. They are linked to the English WordNet version 1.5. The wordnets use a shared top-level ontology and are connected by an inter-lingual index. For German, the University of Tübingen is developing the German version of WordNet, *GermaNet* (Hamp and Feldweg, 1997; Kunze, 2000); the GermaNet version from October 2001 contains 6,904 verb concepts. Compared to my verb classification, the verb sense distinctions in GermaNet are more fine-grained, since they are built on a concept of synonymy.

Multi-Lingual Semantic Verb Classes

The following paragraphs present semantic verb classifications which refer to the syntax-semantic interface of verb properties. Differently to the FrameNet and WordNet approaches, the verb classes are not included in a larger framework.

German Verb Classes: A Classification for Language Learners Schumacher (1986) defines a classification of approximately 1,000 German verbs and verbal expressions, according to the verbs' syntactic and semantic characteristics. Like the valency dictionary in Helbig and Schenkel (1969), the purpose of the classification is German second language teaching. Schumacher distinguishes seven 'semantic fields' and divides them into sub-fields. The verbs in the fields are grouped from the general to the specific, and from the simple to the complex.

1. Allgemeine Existenz 'general existence'
2. Spezielle Existenz 'specific existence'
3. Differenz 'difference'
4. Relation und geistiges Handeln 'relation and mental act'
5. Handlungsspielraum 'scope of acting'
6. Sprachlicher Ausdruck 'language expression'
7. Vitale Bedürfnisse 'vital needs'

The syntactic environment of the verbs is defined by obligatory and facultative complements. Among the possible complements are nominative (NomE), accusative (AkkE), genitive (GenE) and dative complements (DatE), and prepositional (PrepE), adverbial (AdvE), predicative (PredE) and verbative complements (VerbE). The distinction between obligatory and facultative complements is based on elimination tests; facultative complements are marked by brackets. The semantic environment of the verbs is described by conceptual paraphrases of the syntactic usage, idiosyncratic selectional descriptions of complements, example sentences and idiomatic expressions. A morphological component is included by listing word formations based on the respective verb entry.

An example verb entry is given by *anfangen* 'to begin, to start', which is assigned to the semantic field of specific existence (sub-field: existential situation). As in FrameNet and my verb classification to follow in more detail, Schumacher formulates a conceptual description for the verb classes, which provides the semantic basis of the classes. The concepts are then illustrated by semantic and syntactic properties. The absolute number of classes, the relative number of classes compared to the number of verbs, and the thematic variety of classes and verbs is more restricted than in my classification. Given the larger number of verbs per class, the classes comprise less semantic content than my classes, and the syntactic behaviour of the verbs is more variable within a semantic class.

Lemma:	*anfangen*
Principal forms:	fängt an – fing an – hat angefangen
Verbal complex:	anfangen
Construction:	NomE (AdvE)
Structural example:	Der x_{NomE} fängt (an dem y_{AdvE}) an. ‘x_{NomE} begins (with y_{AdvE}).’
Paraphrase:	x existiert an y, und alle Punkte, an denen x außerdem existiert, sind y nachgeordnet. ‘x exists at y, and all points of times when x exists in addition are following y.’
Selectional description:	NomE x: Individuum, das immer an Punkten existiert ‘individual that exists at specific points of times’ AdvE y: Parameterpunkt um_{Akk}, an_{Dat}, bei_{Dat}, in_{Dat}, mit_{Dat}, etc. ‘parameter point expressed by ...’
Remark:	meist gilt: y ist Teil von x. *Die Symphonie fängt mit dem Paukenschlag an.* usually valid: y is part of x. *The symphony starts with the bass drum.*
Passive construction:	no passive possible
Example sentences:	*Wo fängt der Weltraum überhaupt an?* *Die Sache fängt um 16 Uhr an.*
Idiomatic expression:	*Das fängt ja gut/schön an!*
Word formation:	*der Anfang, anfänglich, anfangs*
Other meanings:	• *anfangen* as modal verb • *etwas/jemand fängt an zu* + infinitive • *anfangen mit etwas*

German Verb Classes: A Verb Process Classification Ballmer and Brennenstuhl (1986) classify approximately 13,000 common German verbs according to their meaning. The classification is based on process models, which in turn comprise classes of verbs designating the phases of a process, i.e. an initial situation, a transition from initial to end situation, an end situation, preconditions, results, and consequences. Verb meaning is defined as the participation of the verb within a process, and the verb classes within a model realise the model’s process with all its sequentially and alternatively proceeding sub-processes. For example, the life model refers to the existence of living organisms and contains 13 verbs, each representing a process part of the model and therefore constituting an distinct verb class.

The syntax of the verbs within a class is given by a single argument construction which characterises the syntactic embedding of the respective verb with respect to the class meaning. The syntactic categories come with rough selectional descriptions, such as *etwas* ‘something’, *jemand* ‘somebody’, *pflanzlich* ‘plant’, or *Vorgang* ‘process’, and a valency number for the number of arguments. The verbs within a class might differ in their syntactic behaviour.

Lebensmodell ‘life model’:

gezeugt werden	‘to be conceived’
im Embryo-Stadium sein	‘to be an embryo’
geboren werden	‘to be born’
Kind sein	‘to be child’
aufwachsen	‘to grow up’
pubertieren	‘to be pubescent’
sich einleben	‘to settle in’
sich paaren	‘to mate with someone’
Nachkommen haben	‘to have descendants’
leben	‘to live’
altern	‘to age’
sterben	‘to die’
tot sein	‘to be dead’

There is a total of 44 models. A class example within the *Lebensmodell* ‘life model’ is given by the class *aufwachsen* ‘to grow up’, with some example verbs.

aufwachsen ‘to grow up’

aufwachsen ‘to grow up’	jemand 1
werden ‘to become’	jemand 1
ranken ‘to etwine’	pflanzlich 1
wuchern ‘to proliferate’	pflanzlich 1
sich entwickeln ‘to develop’	jemand 1

The classification by Ballmer and Brennenstuhl is close to my classification in that it is basically organised by verb meaning, but the meaning refers to a process identification instead of a conceptual definition. In addition, in contrast to my verb class approach there is less emphasis on the variety of verb behaviour and the syntax-semantic interface. Nevertheless, it might be interesting to use their classification as gold standard in comparison to my classification.

English Verb Classes Levin (1993) has established a detailed manual classification for 3,104 English verbs, based on the assumption that

> *[...] the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning. Thus verb behavior can be used effectively to probe for linguistically relevant pertinent aspects of verb meaning.*
>
> Levin (1993, page 1)

The classification concentrates on the idea of verb alternations: In a first part, Levin defines a range of 79 possible verb alternations in English, such as the *unspecified object alternation* in Example (2.3) where the verb alternates between a transitive usage with a direct object and an intransitive usage omitting the object; the subject role is identical in both syntactic environments. Levin's definition of alternations refers to the syntactic embedding of the verbs and their obligatory and facultative complements, as well as to informal semantic role descriptions for the participating frame complements. The explicitness of the alternation descriptions varies from case to case. Levin lists all English verbs participating in each alternation.

(2.3) Mike ate the cake.
Mike ate.

In a second part, Levin defines a hierarchical system of semantic verb classes with 49 major verb classes and a total of 191 sub-classes. Each class is accompanied by the verbs in the class and the alternations which the verbs of the class typically undergo; exceptions as well as negative evidence are marked. The verb behaviour thereby probes for meaning aspects. Multiple senses of verbs are represented by multiple class assignment. Even though Levin explicitly states in the beginning of her book that the semantic class definitions are based on the alternations, exceptions to the alternation-based classification show that the syntax-semantic interface is imperfect.

An example of the Levin classes is the class of *give* verbs, a sub class of *verbs of change of possession*:

<u>Class</u>: *Give*

<u>Class members</u>:
feed, give, lease, lend, loan, pass, pay, peddle, refund, render, rent, repay, sell, serve, trade

<u>Class Properties</u>:

1. They lent a bicycle to me.
2. *They lent a bicycle near me / behind me.
3. Dative Alternation:
   They lent a bicycle to me.
   They lent me a bicycle.
4. *Fulfilling Alternation:
   They lent a bicycle to me.
   *They lent me with a bicycle.
5. *Causative Alternations:
   They lent a bicycle to me.
   *A bicycle lent (to me).

<u>Class Comments</u>: These verbs of change of possession display the dative alternation, though there may be some differences of opinion concerning whether some of the verbs actually are found in the double object construction. Although the prepositional phrase is optional with some of these verbs, when it does appear, it must be headed by the preposition *to*.

My definition of German verb classes is very close to Levin's approach. Both classifications are primarily based on semantic intuition, and the semantic classes are underlined by the verb behaviour and possible embeddings. Concerning conceptual information and selectional preferences, Levin comments on them instead of providing systematic labels. Levin's classification has been widely used in Natural Language Processing applications, as Section 2.3 will illustrate.

French Verb Classes Saint-Dizier (1996; 1998) defines contexts for 1,673 frequent French verb senses, which are used to classify the verbs into equivalence classes. The contexts capture the alternation behaviour of the verbs, in that they contain a set of syntactic frames which are subcategorised by the verbs, prepositional phrase information, and thematic roles. The scope of a context is the proposition. The range of contexts for French has been derived from the English alternations by Levin (1993), from French syntactic descriptions, corpus data and linguistic intuition. Saint-Dizier defines about 200 different contexts in Prolog notation. Verbs with exact overlap in the context definition are assigned to a common class. The method results in a total of 953 verb classes, with a large number of singletons and an average of 2 verbs per class.

Saint-Dizier describes the context information as mainly syntactic; in addition, the contexts contain prepositional information and thematic role definitions. He evaluates the resulting classification against the WordNet verb classes, by calculating the largest intersection between his verb classes and the WordNet classes. Saint-Dizier shows that the greater the number of contexts associated with a context verb class, the more homogeneous is the class with respect to WordNet.

The French verb class information is similar to my German verb classes in that the context definition contains the variety of syntactic frames, prepositional information, and thematic roles. But Saint-Dizier does not directly relate the syntactic to the semantic information, and the thematic roles are taken from a restricted set, without idiosyncratic relation to the different verbs and frame types. In addition, since the verb classes require exact overlap of the context information, he obtains a large number of singletons, compared to e.g. the Levin classes and my classification, which allow idiosyncratic irregularities and therefore result in larger verb classes on average.

Spanish Verb Classes Fernández, Martí, Vázquez, and Castellón (1999) and Vázquez, Fernández, Castellón, and Martí (2000) present a verb classification which shares the basic assumptions with Levin (1993): Verb meaning and verb behaviour are related, and semantic information can be inferred from the syntactic verb behaviour.

Vázquez *et al.* (2000) classify Spanish, Catalan and English verbs in parallel. Approximately 1,000 verbs from each language are assigned to verb classes. The characteristics of the classes are defined by common (i) meaning components, (ii) event structure, and (iii) diathesis alternations: (i) The meaning components are based on ideas by Talmy (1985) who addresses the systematic relations in language between lexical meaning and surface express. Vázquez *et al.* distinguish the meaning components carried by a lexical item and the meaning components which are added syntagmatically by the item's complements. (ii) The event structure distinguishes only states and events, and is oriented towards Pustejovsky (1995), decomposing an event into sub-events and specifying the temporal relations between the sub-events and the participants. (iii) Concerning diathesis alternations, Vázquez *et al.* deliberately keep to nine very general alternation types such as causative and passive, compared to the large number of fine-grained alternations by Levin (1993).

According to the verb properties, the verbs are assigned to two major semantic classes, *change of state* and *transfer*, with approximately 600 and 400 verb members for each language, respectively. Applying a refining mechanism by more specific alternation combinations, the verbs are grouped into 14 sub-classes. The verb classes show different degrees of homogeneity. A rough description of the *change of state* class is given, as taken from Fernández *et al.* (1999).

Class: *Change of State*

Meaning components: entity/initiator

Event structure:
e_1: event
e_2: resulting state
$e_1 < e_2$

Diathesis alternations:

1. Anticausative:
   The heat has melted the ice.
   The ice has melted.

2. Aspectual opposition:
   $[event:]$ María dances the tango.
   $[state:]$ María dances the tango (very well).

The classification is very similar in its ideas to Levin (1993), with the exception that it takes into account various syntactic, semantic and aspectual properties and deliberately concentrates on a more general verb grouping.

Japanese Verb Classes Oishi and Matsumoto (1997) perform a two-dimensional semantic classification of Japanese verbs. In the thematic dimension, 858 verbs with more than 50 examples within the EDR Japanese Cooccurrence Dictionary are classified according to their case marking particle patterns. For example, the Japanese translations of *be lost, become clear, get fewer, decrease, rise* are assigned to one thematic class since they agree in the case particle combinations $< ga >$ and $< ga, ni >$. In the aspectual dimension, the Japanese adverbs are classified according to the aspectual class of verbs they modify; for example, continuance adverbs are adverbs which can modify both a state and a process (with reference to Pustejovsky, 1991). The verbs are classified according to the combination of adverb classes they allow for modification; for example, the verbs *be settled, decrease, become larger, occur* are assigned to one aspectual class since they agree in allowing transition and quantity adverbs for modification. Combining the thematic and aspectual verb classifications results in a semantic classification of the Japanese verbs, which is illustrated by Lexical Conceptual Structures (Jackendoff, 1990).

2.2 Class Properties

After the brief introduction of the German verb classes and a description of related work on verb class creation in various frameworks and languages, my manual verb classes are exemplified in some detail, to illustrate the properties as well as the similarities and differences between the verbs and verb classes, and to present a basis for discussions about the clustering experiments and outcomes. As said before, the class description is closely related to Fillmore's scenes-and-frames semantics, as computationally utilised in FrameNet: 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.

Each verb class from Section 2.1.2 is repeated, accompanied by the verbs, a frame-semantic description, and annotated corpus examples. The frame-semantic class definition contains a prose scene description, predominant frame participant and modification roles, and frame variants describing the scene. The frame roles have been developed on the basis of a large German newspaper corpus from the 1990s. I do not claim objective agreement on the role and name definitions, but rather try to 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 list 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. Verbs allowing the respective frame variant are marked by '+', verbs allowing the frame variant only in company of an additional adverbial modifier are marked by '$+_{adv}$', and verbs not allowing the frame variant are marked by '$\neg$'. In the case of ambiguities, frame variants are only given for the senses of the verbs with respect to the class label.

The number of examples does not correspond to the empirical realisation of the verb-frame variants, but is restricted by space requirements. In addition, the example sentences are abbreviated such that they fit into one line. Relevant clause participants are kept in the sentence, and if space allows some context is displayed as well. Only the relevant verbs of the class and its participants are indicated, context information is left unmarked. Discontinuous roles obtain an identical label, e.g. *[$_{Role2}$ Das], denkt [$_{Role1}$ er], [$_{Role2}$ ist eine Falle]*.

The frame variants with their roles marked represent the alternation potential of the verbs, by connecting the different syntactic embeddings to identical role definitions. For example, the causative-inchoative alternation assumes the syntactic embeddings $\mathbf{n}_X\mathbf{a}_Y$ and $\mathbf{n}_Y$, indicating that the alternating verbs are realised by a transitive frame type (containing a nominative NP 'n' with role *X* and an accusative NP 'a' with role *Y*) and the corresponding intransitive frame type (with a nominative NP ‘n’ only, indicating the same role *Y* as for the transitive accusative). Passivisation of a verb-frame combination is indicated by [P].

The frame definitions will be introduced in detail in Chapter 3, which presents the potential of the statistical grammar in frame information induction. In order to understand the frame variants in the following class description, the reader is referred to Appendix A, where all frame variants and illustrative examples are listed. In case of verbs subcategorising for a specific PP, detailed prepositional phrase information is indicated by case and preposition, such as ‘Dat.mit’, ‘Akk.für’.

In addition to the scene-and-frame descriptions, I assign each class comparable class labels from Levin’s English and Schumacher’s German verb classifications (cf. Section 2.1.3), if such exist. The labels are given as a guide. If verb classes do not have a counterpart in Levin or Schumacher, it means that they fail to classify the respective verbs in the respective sense.

Last but not least, in contrast to my own example sentences in this thesis, the corpus examples are given in ‘the old German spelling notation’, since the newspaper corpus dates from the times before the spelling reform. Therefore, the reader might notice spelling inconsistencies between my own and the corpus examples. I do not give translations of the corpus examples, since on the one hand this would go beyond the scope of this work, and on the other hand, the relevant information is illustrated by the combination of class label and role marking.

Class 1: *Aspect*

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

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

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

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

Levin class: 55.1 *(Aspectual Verbs → Begin Verbs)*

Schumacher class: 2.1 *(Verben der speziellen Existenz → Verben der Existenzsituierung)*
beenden is not classified.

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

Class 2: *Propositional Attitude*

Verbs: *ahnen, denken, glauben, vermuten, wissen*

Scene: [$_B$ Somebody] has a propositional attitude with respect to [$_T$ something]. There might be a projection of the attitude onto [$_P$ somebody].

Frame Roles: B(eliever), T(heme)

Modification Roles: P(atient)

Levin classes: 29.4/5 *(Verbs with Predicative Complements → Declare Verbs / Conjecture Verbs)*

Schumacher class: 4.11 *(Verben der Relation und des geistigen Handelns → Verben der Aufmerksamkeit)*
Only *denken* is classified.

Frame	Participating Verbs & Corpus Examples
n$_B$	+ denken, glauben / ¬ ahnen, vermuten, wissen
	... [$_B$ der] langfristig **denkt** und investiert ... [$_B$ Er] hat zuviel gesehen, um zu **glauben**.
n$_B$ **a**$_T$	+ ahnen, vermuten, wissen / ¬ denken, glauben
	... als **ahne** [$_B$ er] [$_T$ sein späteres Schicksal] ... [$_B$ Boeing] **vermutet** hingegen [$_T$ nur eine Nachfrage von wenigen hundert Stück]. [$_T$ Was] **vermuten** [$_B$ Sie]? [$_B$ Hölscher] **wußte** [$_T$ das]. [$_T$ Den Rest] **wissen** [$_B$ die Götter].
n$_B$ **a**$_T$ [P]	+ vermuten / ¬ ahnen, denken, glauben, wissen
	... Werkstätten, [$_T$ die] in Speyer, Konstanz und Esslingen selbst **vermutet** werden ... Der Coup, hinter dem [$_T$ Suharto] selbst **vermutet** wurde ...
n$_B$ **p**$_T$:Akk.an	+ denken, glauben / ¬ ahnen, vermuten, wissen
	[$_B$ Doyé] **denkt** vor allem [$_T$ an Strickwaren] ... Ob [$_B$ sie] [$_T$ an den Frieden] **glauben**? [$_B$ Choegyal] **glaubt** [$_T$ an die Reinkarnation].
n$_B$ **p**$_T$:Dat.in	+ denken / ¬ ahnen, glauben, vermuten, wissen
	... [$_B$ der] ausschließlich [$_T$ in Dollar] **denkt**.
n$_B$ **p**$_T$:Akk.um	+ wissen / ¬ ahnen, denken, glauben, vermuten
	[$_B$ Jeder] ... **wisse** [$_T$ um die Schwierigkeiten des Lebenshaltungskostenindexes].
n$_B$ **s-2**$_T$	+ denken, glauben, vermuten, wissen / ¬ ahnen
	[$_T$ Da sind wir nun], **dachte** [$_B$ Berence] ... Weil [$_B$ sie] **glauben**, [$_T$ in der Leichtigkeit liege eine Oberflächlichkeit] ... [$_B$ Das Unternehmen] **vermutete**, [$_T$ ein Materialfehler ... habe zu dem Unglück geführt]. ... [$_B$ ich] **weiß**, [$_T$ am Ende des Weges könnte die Nummer eins stehen].
n$_B$ **s-dass**$_T$	+ ahnen, denken, glauben, vermuten, wissen
	[$_B$ Niemand] **ahnte**, [$_T$ daß man damit den Kern der Sache getroffen hatte]. [$_B$ Ich] **denke**, [$_T$ daß die Zeit reif war] ... [$_B$ Wir] **glauben** nicht, [$_T$ daß sich der Versuch lohnen würde] ... [$_B$ Die Polizei] **vermutet**, [$_T$ daß die Täter Separatisten des Bodo-Volks sind] ... [$_B$ Wir] **wissen** jetzt, [$_T$ daß das Wetter ... ein irreversibler Prozeß ist] ...
n$_B$ **s-w**$_T$	+ ahnen, denken, glauben, wissen / ¬ vermuten
	[$_B$ Wer es nicht kennt], **ahnt**, [$_T$ wo er umkehren muß]. ... wenn [$_B$ man] daran **denke**, [$_T$ welche Befürchtungen mit ... verbunden worden seien]. [$_B$ Dankert] **weiß**, [$_T$ wovon er spricht]. [$_B$ Ich] **weiß** nie, [$_T$ wann sie kommen].
n$_B$ **s-ob**$_T$	+ ahnen, wissen / ¬ denken, glauben, vermuten
	... ohne auch nur zu **ahnen**, [$_T$ ob dessen wichtigste Raumkante ... definiert werden wird]. [$_B$ Ich] **weiß** nicht, [$_T$ ob es die Nerven waren] ...
n$_B$ **i**$_T$	+ glauben / +$_{daran}$ denken / ¬ ahnen, vermuten, wissen
	... wenn [$_B$ die Mailänder] daran **dächten**, [$_T$ ihr Theater wieder abzubauen]. [$_B$ Die Sozialpolitik] **glaubte** [$_T$ die Regeln der Marktwirtschaft ignorieren zu können]. [$_T$ Das] **wußte** [$_B$ die polnische Führung] [$_T$ auszunutzen].
n$_B$ **d**$_R$	+ glauben / ¬ ahnen, denken, vermuten, wissen
	[$_R$ Wem] wollt [$_B$ ihr] **glauben**? **Glaubt** [$_B$ man] [$_R$ der Genesis] ...
n$_B$ **a**$_T$ **d**$_R$	+ glauben / ¬ ahnen, denken, vermuten, wissen
	Im Dunkeln **glaubst** [$_B$ Du] [$_R$ mir] [$_T$ alles].

Class 3: *Desire* → *Wish*

Verbs: *erhoffen, wollen, wünschen*

Scene: [$_E$ Somebody] has [$_W$ a wish]. There might be [$_R$ an explicit receiver] for the wish.

Frame Roles: E(xperiencer), W(ish)

Modification Roles: R(eceiver)

Levin class: 32.1 *(Desire Verbs → Want Verbs)*
Levin classifies both sub-classes of desire (my classes 3 and 4) into the same class.

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_E$ **a**$_W$	+ wollen, wünschen / ¬ erhoffen
	Dies ist der Beginn des Endes, [$_W$ das] [$_E$ wir alle] **wünschen**.
	[$_E$ Sie] **wünschen** [$_W$ ein Beispiel]?
	... [$_E$ Urlauber] **wollen** [$_W$ neue Erlebnisdimensionen] ...
n$_E$ **r**$_R$ **a**$_W$	+ erhoffen, wünschen / ¬ wollen
	... denn [$_E$ sie] **erhofft** [$_R$ sich] [$_W$ Auskünfte über Zlatas Verbleiben].
	[$_E$ Man] **erhofft** [$_R$ sich] [$_W$ neuen Schwung, frischen Wind, neue Dynamik].
	[$_E$ João Havelange] **wünscht** [$_R$ sich] ... [$_W$ ein Länderspiel der Nationalteams] ...
	[$_E$ Man] **wünscht** [$_R$ sich] langfristig [$_S$ ein eigenes Ensemble und einen Chor] ...
n$_E$ **a**$_W$ **d**$_R$	+ wünschen / ¬ erhoffen, wollen
	[$_E$ Boutros-Ghali] **wünschte** vor allem [$_R$ seinem Amtsnachfolger ...] [$_W$ viel Erfolg].
	Dann **wünsche** [$_E$ ich] [$_R$ seinen Bemühungen] [$_W$ Erfolg].
n$_E$ **s-dass**$_W$	+ wollen, wünschen / ¬ erhoffen
	Lebed sagte, [$_E$ er] **wünsche**, [$_W$ daß der Präsident gesund werden möge].
	[$_E$ Ich] **wünschte**, [$_W$ daß die ... in dieser Beziehung einander vertrauen].
	... daß [$_E$ die ... Kreise Rußlands] nicht **wollen**, [$_W$ daß ...]
	... [$_E$ er] **wolle** nicht, [$_W$ daß der Solidaritätszuschlag ... gesenkt werde].
n$_E$ **r**$_R$ **s-dass**$_W$	+ erhoffen, wünschen / ¬ wollen
	[$_E$ Man] **erhofft** [$_R$ sich], [$_W$ daß mancher Musikfreund die Gelegenheit ... nutzt].
	[$_E$ Wir] **wünschen** [$_R$ uns], [$_W$ daß jeder Mitgliedstaat einen Aktionsplan entwickelt] ...
n$_E$ **d**$_R$ **s-dass**$_W$	+ wünschen / ¬ erhoffen, wollen
	[$_E$ Ich] **wünsche** [$_R$ denen], [$_W$ daß sie in der Bundesliga bleiben].
n$_E$ **s-2**$_W$	+ wünschen / ¬ erhoffen, wollen
	[$_E$ Man] **wünschte**, [$_W$ man könnte es als erste Geste der Vernunft werten] ...
	[$_E$ Ich] **wünschte**, [$_W$ seine Eminenz fände den Mut] ...
n$_E$ **r**$_R$ **s-2**$_W$	+ wünschen / ¬ erhoffen, wollen
	... und [$_E$ man] mag [$_R$ sich] **wünschen**, [$_W$ sie hätte ... noch häufiger bebildert].
n$_E$ **i**$_W$	+ wollen, wünschen / ¬ erhoffen
	[$_E$ Die Arbeitgeber] **wollen** [$_W$ die Sozialsysteme auf Kernrisiken beschränken] ...
	[$_E$ Wir] **wollen** alle [$_W$ in Frieden und Versöhnung leben].
	[$_E$ Milosevic] **will** [$_W$ Zeit gewinnen].
	... [$_E$ der] **wünschte**, [$_W$ ein Bürger zu sein] ...
	[$_E$ Ich] **wünschte** nicht, [$_W$ Euch irrezuführen].

Class 4: *Desire* → *Need*

Verbs: *bedürfen, benötigen, brauchen*

Scene: [$_E$ Somebody] needs [$_N$ something]. The need might be defined with respect to [$_B$ a benefactive].

Frame Roles: E(xperiencer), N(eed)

Modification Roles: B(enefactive)

Levin class: 32.1 *(Desire Verbs → Want Verbs)*
Levin classifies both sub-classes of desire (my classes 3 and 4) into the same class.

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_E$ **a**$_N$	+ benötigen, brauchen / ¬ bedürfen
	... [$_E$ sie] **benötigen** stets [$_N$ zahlreiche Helfer]. [$_E$ Der transplantierte Patient] **benötigt** [$_N$ eine geschützte Umgebung] ... ... daß [$_E$ die Gemeinschaft] [$_N$ institutionelle Reformen] **benötige** ... [$_E$ Wir] **brauchen** [$_N$ mehr Flexibilität] ... Inzwischen **brauchen** [$_E$ hunderttausend Einwohner] [$_N$ mehr Platz]. Und in Norddeich **braucht** [$_E$ man] noch [$_N$ ein Taxi], um zum Bahnhof zu kommen.
n$_E$ **a**$_N$ [P]	+ benötigen, brauchen / ¬ bedürfen
	... [$_N$ die] [$_B$ für einen Golf der aktuellen Baureihe] **benötigt** wird. [$_N$ Mitarbeiter der Supermärkte] werden **benötigt** ... ... in welcher Höhe [$_N$ Mittel] **benötigt** werden. [$_B$ Für sie] werden [$_N$ Basiskurse] **benötigt**. [$_N$ Das Werk] würde nicht **gebraucht** bei erhöhtem Wirkungsgrad ... [$_N$ Ein globales Engagement Amerikas] wird heute mehr denn je **gebraucht**. ... obwohl [$_N$ die Absolventen] dringend **gebraucht** werden? [$_N$ Berkowitz und Revivo] werden bei ihren Klubs **gebraucht**.
n$_E$ **g**[1]$_N$	+ bedürfen / ¬ benötigen, brauchen
	[$_E$ Das deutsche Aktiengesetz] **bedarf** [$_N$ einer grundlegenden Reform] ... [$_E$ Er] **bedarf** [$_N$ des Beistandes des ganzen Hauses] ... [$_E$ Jede Marke] **bedürfe** [$_N$ intensiver Pflege].
n$_E$ **i**$_N$	+ brauchen / ¬ bedürfen, benötigen
	[$_E$ Wir] **brauchen** [$_N$ das Haus im Falle des totalen Chaos nicht mehr zu verlassen]. ... daß [$_E$ die Mailänder] [$_N$ nur zu fragen] **brauchten** ... [$_E$ Wer weniger als 1,37 Meter groß ist], **braucht** [$_N$ sich noch nicht mal zu bücken].

[1]Genitive NPs are not coded in the grammar.

Class 5: *Transfer of Possession (Obtaining)*

Verbs: *bekommen, erhalten, erlangen, kriegen*

Ambiguity: *erhalten* has a sense of 'to conserve'.

Scene: [$_R$ Somebody] obtains [$_T$ something]. There might be [$_S$ a source] or [$_C$ a reason] for the transfer of possession.

Frame Roles: R(eceiver), T(heme)

Modification Roles: S(ource), C(ause)

Levin class: 13.5.2 *(Verbs of Change of Possession → Verbs of Obtaining → Obtain Verbs)*

Schumacher class: 7.1(7) *(Verben der vitalen Bedürfnisse → Verben des Besitzes und Besitzwechsels, Kausative Handlungsverben)*
Schumacher class 7.1(7) subsumes *Transfer of Possession (Obtaining)* and *Transfer of Possession (Giving) → Gift* (my classes 5 and 6).

Frame	Participating Verbs & Corpus Examples
n$_R$ **a**$_T$	+ bekommen, erhalten, erlangen, kriegen
	[$_R$ Mancher Politiker] wird dann [$_T$ Angst vor der eigenen Courage] **bekommen**.
	... [$_R$ der] mit Sicherheit [$_T$ Verstärkung] **bekommen** wird ...
	[$_R$ Wir] **bekommen** zwar zunehmend [$_T$ mehr und interessantere Stellenangebote] ...
	[$_R$ Helmut Kohl] hat [$_T$ einen Brief] [$_S$ aus Hollywood] **bekommen** ...
	[$_R$ Horst Vetter] hatte ... [$_S$ von einem Gericht in Hongkong] [$_T$ recht] **bekommen**.
	... [$_T$ die] [$_R$ er] [$_C$ für Medikamente] **bekommen** haben soll.
	... falls [$_R$ er] [$_T$ das Geld] nicht **erhalten** sollte.
	[$_R$ Sie] **erhalten** zwar [$_T$ eine gleichwertige Berufsausbildung] ...
	[$_C$ Dafür] **erhalten** [$_R$ sie] [$_T$ eine Rundumversorgung].
	[$_R$ Sie] **erhalten** [$_T$ Aktien der BIZ] [$_S$ aus einer neuen Tranche].
	Um [$_T$ den Eindruck des Bunten] zu **erlangen** ...
	So konnte [$_R$ Kroatien] 1991 [$_T$ die Unabhängigkeit] **erlangen**.
	... [$_S$ von der Schaubühne] [$_T$ eine Auskunft] darüber zu **erlangen** ...
	[$_R$ Einer] muß [$_T$ die Prügel] ja **kriegen** ...
	[$_R$ Die Unis] wollen Autonomie und **kriegen** [$_T$ sie].
	Wenn die Punker kommen, **kriegen** [$_R$ sie] [$_T$ was auf die Finger] ...
	Sie glauben doch nicht, daß [$_R$ Sie] [$_S$ von uns] [$_T$ etwas] **kriegen** ...
	... [$_R$ die] normalerweise [$_T$ keinen Pfennig] [$_S$ von einer Bank] **kriegen** ...
	... [$_T$ die] [$_R$ wir] hier [$_C$ für den Verkauf unserer Häuser] **kriegen** ...
n$_R$ **a**$_T$ [P]	+ erlangen / ¬ bekommen, erhalten, kriegen
	... mit denen [$_T$ die Aufmerksamkeit der Wähler] **erlangt** werden soll ...
	... daß hier [$_T$ kein richtiges Verstehen] **erlangt** wird ...
	... es sei [$_T$ ein Geldbetrag in sechsstelliger Höhe] zu Unrecht **erlangt** worden.
	[$_T$ 3000 Unterschriften] seien bislang **erlangt** worden.
n$_R$ **i**$_T$	+ bekommen / ¬ erhalten, erlangen, kriegen
	[$_R$ Die Leser] **bekommen** [$_T$ das an allen Ecken und Enden zu spüren].
	[$_T$ Die Folgen] **bekommen** [$_R$ die Mitglieder] meist erst dann [$_T$ zu spüren] ...
	Das erste Opfer, [$_T$ das] [$_R$ wir] [$_T$ zu sehen] **bekommen**, ist Madame Groneille.

Class 6: *Transfer of Possession (Giving) → Gift*

Verbs: *geben, leihen, schenken, spenden, stiften, vermachen, überschreiben*

Ambiguity: *überschreiben* has a sense of 'to overwrite'. Depending on the frame, *leihen* has a sense of either giving or obtaining. In a reflexive use (frame: `nar`), *schenken* has a sense of 'to refuse'.

Scene: [$_S$ A possessor as source of a transfer] gives [$_G$ something] as a gift. There might be [$_R$ a receiver] or [$_P$ a purpose] for the transfer of possession.

Frame Roles: S(ource), G(ift)

Modification Roles: R(eceiver), P(urpose)

Levin class: 13.1 *(Verbs of Change of Possession → Give Verbs)*

Schumacher class: 7.1(7) *(Verben der vitalen Bedürfnisse → Verben des Besitzes und Besitzwechsels, Kausative Handlungsverben)*
Schumacher class 7.1(7) subsumes *Transfer of Possession (Obtaining)* and *Transfer of Possession (Giving) → Gift* (my classes 5 and 6).

Frame	Participating Verbs & Corpus Examples
n$_S$ **a**$_G$	+ geben, schenken, spenden, stiften, vermachen, überschreiben / ¬ leihen
	Dann kann [$_S$ man] [$_G$ Vollgas] **geben** ...
	[$_S$ Vier Meisterschaften] **geben** [$_G$ Zeugnis vom Erfolg der Ära Lee].
	... [$_S$ sie selbst] **schenkt** [$_G$ Genuß].
	[$_S$ Ein amerikanisches Ehepaar] hatte [$_G$ das benötigte Geld] **gespendet**.
	[$_S$ Es] **spendet** [$_G$ das Vergnügen] ...
	Dabei **spendet** [$_S$ die negativ geladene Kathode] [$_G$ die Elektronen] ...
	Und [$_S$ sein Trainer Nevio Scala] **spendete** [$_G$ Beifall] für so viel Zuversicht.
	[$_G$ Raketenmotor und Sonde] **stiftet** [$_S$ die Jugend-Initiative].
	... [$_S$ der] überall [$_G$ nur Unfrieden] **stiften** will.
	Tatsächlich hat [$_S$ diese These] jedoch [$_G$ nur Verwirrung] **gestiftet**.
	... [$_S$ sie] **vermachen** [$_G$ ihre Gemälde- und Fotosammlungen, Häuser, Geld].
	Beispielsweise könne [$_S$ man] [$_G$ sein Haus] **überschreiben** und darin weiterhin leben.
n$_S$ **a**$_G$ [P]	+ geben, schenken, spenden, stiften / ¬ leihen, vermachen, überschreiben
	... wie [$_G$ Interviews vor Ort] auf gepflegten Villengrundstücken **gegeben** wurden ...
	... [$_G$ die] vor vierzig und mehr Jahren als "Museumsspende" **gegeben** wurden ...
	[$_G$ Seiffener Reifenfiguren] werden wieder **geschenkt**.
	[$_G$ Futter] wurde **gespendet**.
	... [$_G$ die] [$_S$ von einer Firma oder Bank] **gespendet** werden.
	[$_G$ Zwei Backöfen für 125000 Mark] wurden [$_S$ von einer Lebensmittelkette] **gespendet**.
	... [$_G$ die] [$_S$ von Tänzerinnen] **gestiftet** worden seien ...
	[$_G$ Der Preis] war [$_P$ zum fünfzigsten Jubiläum der Filmfestpiele] **gestiftet** worden.

Frame	Participating Verbs & Corpus Examples
$\mathbf{n}_S$ $\mathbf{a}_G$ $\mathbf{d}_R$	+ geben, leihen, schenken, spenden, stiften, vermachen, überschreiben
	[$_S$ Sie] **geben** [$_R$ dem Winterhimmel] [$_G$ das prachtvolle Aussehen].
	Möge [$_S$ Gott] [$_R$ uns] [$_G$ die Kraft] **geben** [$_P$ für die Arbeit] ...
	Aber mitunter **leihen** [$_S$ Leute wie der Arbeiter im Overall] [$_R$ ihnen] [$_G$ die Zunge].
	[$_S$ Wir] **leihen** [$_R$ ihnen] [$_G$ Geld] und [$_S$ sie] [$_R$ uns].
	[$_S$ Er] **schenkte** [$_R$ ihm] [$_G$ eine Mineraliensammlung] ...
	Aber [$_S$ die Fans] **schenken** [$_R$ dir] gleich [$_G$ die Fahrkarte für den nächsten Bus] ...
	Damit **schenkte** [$_S$ er] [$_R$ meiner Freundin] [$_G$ ein zweites Leben].
	[$_S$ Er] **schenkt** dabei [$_R$ seinen Lesern] [$_G$ nichts] ...
	... will [$_S$ der Verein] nun [$_G$ 200000 Mark] [$_R$ der Caritas] **spenden** ...
	[$_G$ Trost] **spendet** [$_R$ den Marktteilnehmern] hierzulande [$_S$ nur die Tatsache] ...
	... [$_S$ sie] **stifte** [$_R$ der Allgemeinheit] [$_G$ einen erheblichen externen Nutzen].
	[$_S$ Sie] **stiftete** [$_R$ den Händlern und Kursmaklern] [$_G$ zwei Marzipantorten] ...
	... [$_G$ das] [$_S$ Picasso] [$_R$ Madrid] **vermacht** hatte.
	In seinem Testament **vermachte** [$_S$ er] [$_R$ Mapplethorpe] [$_G$ drei Viertel] ...
	[$_G$ Seine Eingeweide] **vermachte** [$_S$ er] [$_R$ der Pathologie] ...
	... [$_S$ Joao] ... [$_R$ seiner Exfrau] [$_G$ eine Hypothek] ... **überschreiben** wird.
	... indem [$_S$ sie] [$_R$ ihrer Frau] [$_G$ das Unternehmen] **überschreiben**.
$\mathbf{n}_S$ $\mathbf{a}_G$ $\mathbf{d}_R$ [P]	+ geben, leihen, schenken, spenden, stiften, vermachen, überschreiben
	Dabei solle jedoch [$_R$ der kooperativen Gesamtschule] [$_G$ Vorrang] ... **gegeben** werden.
	[$_R$ Den Hochschulen] soll [$_G$ mehr Einfluß ...] **gegeben** werden.
	[$_R$ Dem besiegten Kriegsgegner] ... wurden [$_G$ drei Milliarden Dollar] ... **geliehen**.
	... [$_R$ dem] [$_G$ die Statue von Mussolini] **geschenkt** worden war.
	... [$_S$ von der Welthungerhilfe] bekommen [$_R$ die Bauern] [$_G$ die Pflüge] nicht **geschenkt**.
	[$_G$ Frische Sträuße] ... sollten [$_R$ Krankenhäusern und Altenheimen] **gespendet** werden.
	... wie [$_G$ er] [$_R$ ihm] [$_S$ von der Fraktion] zuvor nie **gespendet** worden war.
	... [$_G$ die] jetzt [$_R$ den Städtischen Kunstsammlungen] **gestiftet** wurden.
	... [$_G$ das] [$_R$ dem Rotterdamer ... Museum] [$_S$ von Vitale Bloch] **vermacht** wurde ...
	... wonach [$_G$ die Gebeine Dunstans] ... [$_R$ dem Kloster Glastonbury] **vermacht** wurden ...
	... [$_G$ das] [$_R$ ihm] 1986 [$_S$ von einem Verwandten] ... **überschrieben** wurde ...
$\mathbf{n}_S$ $\mathbf{a}_G$ $\mathbf{p}_R$:Akk.an	+ geben, leihen, schenken, spenden, stiften, vermachen, überschreiben
	[$_S$ Ford] **gibt** [$_G$ keinen Anschlußauftrag] [$_R$ an den Karosseriebauer Karmann].
	[$_S$ Diese] wiederum **gibt** [$_G$ eine Empfehlung] [$_R$ an den Beirat der IG Bau] ...
	[$_S$ Der Fonds] **leiht** [$_G$ seine Gelder] nicht direkt [$_R$ an die Endkreditnehmer] ...
	... **spenden** [$_S$ sie] [$_G$ Geld] [$_R$ an gemeinnützige Organisationen] ...
	... [$_S$ diese Gruppen] **stifteten** [$_G$ Unfrieden] [$_R$ an der Basis] ...
	[$_S$ Beide] **stifteten** [$_G$ weitere Bilder] [$_R$ an französische Museen].
	[$_S$ Die preußischen Könige] ... **vermachten** [$_G$ diese Erbschaft] [$_R$ an das Kaiserreich].
$\mathbf{n}_S$ $\mathbf{r}_G$	$+_{adv}$ geben / ¬ leihen, schenken, spenden, stiften, vermachen, überschreiben
	[$_S$ Die Mullahs] **geben** [$_G$ sich] fortschrittlich.
	[$_S$ Die Institute] **geben** [$_G$ sich] in ihrem Bericht zuversichtlich ...
	[$_S$ Linke wie rechte Politiker] **geben** [$_G$ sich] gleichermaßen entsetzt.
$\mathbf{n}_S$ $\mathbf{d}_R$ $\mathbf{i}_G$	+ geben / ¬ leihen, schenken, spenden, stiften, vermachen, überschreiben
	[$_S$ Das] sollte [$_R$ allen jenen] [$_G$ zu denken] **geben** ...
	... so will [$_R$ uns] [$_S$ der Film] [$_G$ zu verstehen] **geben** ...
	[$_S$ Dies] wird [$_R$ den Verantwortlichen] sicher [$_G$ zu denken] **geben** ...
$\mathbf{x}$ $\mathbf{a}_G$	+ geben / ¬ leihen, schenken, spenden, stiften, vermachen, überschreiben
	... es könne auch [$_G$ etwas Schnee] **geben** ...
	An der Pazifikküste von Big Sur werde es [$_G$ Erdrutsche] **geben** ...
	Es könne [$_G$ keinen Anlaß zur Hysterie] **geben** ...

Class 7: *Transfer of Possession (Giving) → Supply*

Verbs: *bringen, liefern, schicken, vermitteln*$_1$*, zustellen*

Ambiguity: *vermitteln* has a sense of teaching, which refers to a supply on the mental level, cf. class 29; in addition, *vermitteln* has a sense of 'to mediate'.

Scene: [$_S$ A possessor as source of a transfer] gives [$_G$ something] as a supply. There might be [$_R$ a receiver] of the transfer of possession. The supply might enforce a specific kind of [$_C$ circumstance] (with respect to location, situation, etc.).

Frame Roles: S(ource), G(iven)

Modification Roles: R(eceiver), C(ircumstance)

Levin class: 11.1/3 and 13.4.1 *(Verbs of Sending and Carrying → Send Verbs / Bring and Take; Verbs of Change of Possession → Verbs of Providing → Verbs of Fulfilling)*

Schumacher class: 6.2 *(Verben des sprachlichen Ausdrucks → Verben des Übermittelns)*
Only four German supply verbs (different verbs than my verbs in class 7) are classified by the respective semantic class in Schumacher.

Frame	Participating Verbs & Corpus Examples
n$_S$ **a**$_G$	+ bringen, liefern, schicken, vermitteln, zustellen
	Kann [$_S$ die Wissenschaft] [$_G$ Lösungen für diese Gefahren] **bringen** ?
	[$_S$ Neue Teleskope] **bringen** automatisch [$_G$ neue Entdeckungen].
	... [$_S$ welche] [$_G$ die Wand] [$_C$ zum Einsturz] **brachten**:
	[$_S$ Die Amerikaner] setzen alles daran, [$_G$ Geister] [$_C$ auf die Bühne] zu **bringen** ...
	... [$_G$ Gewalt] [$_C$ unter Kontrolle] zu **bringen** ...
	Damit könnte [$_S$ er] [$_G$ eine Lawine] [$_C$ ins Rollen] **bringen**.
	[$_S$ Willibald von Gluck] **lieferte** [$_G$ Kompositionen].
	[$_S$ Kaiser] **schickt** [$_G$ die Leser] [$_C$ auf eine Besichtigungstour] ...
	Dann **schickt** [$_S$ er] [$_G$ die Kandidaten] [$_C$ in eine Dunkelkammer] ...
	Zu Weihnachten **schickte** [$_S$ der Brauer] [$_G$ einen Kasten mit Bier].
	[$_S$ Sixt Travel] ... **vermittele** über lokale Partner [$_G$ Mietwagen am Zielort] ...
	[$_S$ Die Sozialämter] sind gehalten, [$_G$ Jobs] zu **vermitteln**.
	... [$_S$ der] ... [$_G$ das Ergebnis] einmal in der Woche über E-mail **zustellt**.
	[$_G$ Diesen Zugang zur Geschichte] ... wird [$_S$ keine Fachhistorie] ... **zustellen**.
n$_S$ **a**$_G$ [P]	+ bringen, liefern, schicken, vermitteln, zustellen
	[$_G$ Kaffee und Scotch] werden **gebracht**.
	[$_G$ Es] hat das Recht, [$_C$ aus der Gefahrenzone "Familie"] **gebracht** zu werden ...
	Damit sollen [$_G$ die Fischerboote] ... [$_C$ unter Kontrolle] **gebracht** werden.
	Soll [$_G$ der Garten] [$_C$ in Form] **gebracht** werden ...
	... [$_G$ das] heutzutage [$_S$ vom Fernsehen] **geliefert** wird.
	[$_G$ Die Motoren] dazu sollen [$_S$ von Ford und Peugeot] **geliefert** werden.
	... [$_G$ die] vor allem [$_R$ an Pharma- und Kosmetikindustrie] **geliefert** werden ...
	... als [$_G$ Aragon] [$_C$ an die Front] **geschickt** wird ...
	... wie [$_G$ sie] [$_S$ vom Kunden] **geschickt** werden.
	Vermittelt werden aber auch [$_G$ Grabsteine, Foto auf der Urne, güldene Inschrift].
	... obwohl [$_G$ etwa fünfzig der 178 offerierten Werke] nicht **vermittelt** werden konnten.
	[$_G$ Die Klageschrift] sei ordnungsgemäß **zugestellt** worden ...
n$_S$ **a**$_G$ **d**$_R$	+ bringen, liefern, schicken, vermitteln, zustellen
	[$_S$ Waffengeschäfte] **bringen** [$_R$ nur noch den Händlern] [$_G$ das große Geld].
	... [$_G$ die] [$_R$ ihnen] [$_S$ der Sozialismus] **gebracht** hat ...
	[$_S$ Wir] **liefern** [$_R$ den jeweils Verantwortlichen] [$_G$ nur die Munition] dazu.
	[$_S$ Sie] **lieferte** [$_R$ ihr] [$_G$ neue Impulse] ...
	[$_S$ Der Eindringling] **schickte** [$_R$ ihr] [$_G$ den Nachbarn] zur Hilfe ...
	[$_R$ Dem Geliebten] **schickt** [$_S$ sie] [$_G$ zartbittere Zettel] ...
	Jedes Jahr **schickt** [$_S$ er] [$_R$ ihr] [$_G$ Geld] ...
	... wenn [$_S$ sie] [$_R$ ihrem Kunden] [$_G$ eine ... Versicherungsleistung] **vermittelt** haben.
	[$_S$ Sie] kann [$_R$ der Tochterbank] [$_G$ Kunden] ... **vermitteln** ...
	... [$_G$ die] [$_S$ sie] [$_R$ dem Premierminister] **zustellen** will ...
	[$_R$ Der ... Bischofskongregation] wird [$_S$ Schönborn] [$_G$ eine ... Dokumentation] **zustellen**.
n$_S$ **a**$_G$ **d**$_R$ [P]	+ bringen, liefern, schicken, vermitteln, zustellen
	... [$_G$ ihre Republik] sei [$_R$ Europa] zum Opfer **gebracht** worden ...
	Da wird einerseits [$_R$ dem Auge] [$_G$ Stoff] **geliefert** ...
	... [$_G$ die] nach dem Just-in-time-Prinzip [$_R$ dem Kunden] **geliefert** werden.
	... [$_G$ das] [$_R$ ihm] [$_C$ nach Hause] **geschickt** wird.
	... [$_G$ das] [$_R$ ihm] [$_C$ nach Athen] **geschickt** wurde.
	[$_G$ Was] [$_R$ dem Konsumenten] nachhaltig **vermittelt** wurde ...
	Oder [$_G$ sie] können [$_R$ dem gebeutelten Bürger] nicht mehr **vermittelt** werden.
	[$_G$ Der Bericht] werde [$_R$ den OSZE-Delegationen] **zugestellt** ...
	... [$_G$ sie] werden [$_R$ mir] gleich dreifach per Post **zugestellt**.

Class 8: *Manner of Motion → Locomotion*

Verbs: *gehen, klettern, kriechen, laufen, rennen, schleichen, wandern*

Ambiguity: *gehen* is used in several idiomatic expressions, e.g. *es geht um etwas* 'to deal with' and *es geht jemandem gut/schlecht/etc.* 'to feel good/bad/etc.'

Scene: [$_M$ Somebody or something able to move in some (abstract) way] moves agentively, or [$_M$ something] moves by external causation. The motion might be described by [$_K$ the specific kind] of motion. Typically, [$_P$ one ore more paths or directions] for the motion are specified.

Frame Roles: M(oving entity)

Modification Roles: P(ath), K(ind)

Levin class: 51.3.2 *(Verbs of Assuming a Position → Manner of Motion Verbs → Run Verbs)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_M$	+ gehen, klettern, kriechen, laufen, rennen, schleichen, wandern
	Entweder [$_M$ sie] **gehen** [$_K$ zu Fuß] ...
	So **geht** [$_M$ das Leben] [$_P$ dahin] ...
	Dann **ging** [$_M$ er] [$_P$ nach Paris] ...
	Trotzdem blieb er öffentlich bei seiner Behauptung, [$_M$ kein Wort] **gehe** [$_P$ verloren].
	Selbstverständlich **gehen** [$_M$ Sie] [$_P$ zum Notar] ...
	[$_M$ Ein Mann] **geht** morgens [$_P$ ins Bad] ...
	Im Sommer **kletterten** [$_M$ sie], im Winter fuhren sie Ski.
	Im Gegenzug **kletterten** [$_M$ die Zinsen].
	[$_M$ Die Benzinpreise] **kletterten** [$_P$ auf das Doppelte].
	Nachts **klettern** [$_M$ sie] dann [$_P$ über den Zaun] und holen sich die Beute.
	[$_M$ Er] muß **kriechen** und **klettern**.
	[$_M$ Voss] **kriecht** [$_K$ auf allen vieren].
	In dieser Zeit **kriechen** [$_M$ die Klapperschlangen] ... [$_P$ unter die Steine].
	[$_M$ Die Kälte] **kriecht** [$_P$ durch alle Poren].
	[$_M$ Hunderte Autos] **krochen** [$_K$ im Schrittempo] [$_P$ durch die Straßen] ...
	[$_M$ Wer] **kriecht** zuerst [$_P$ ins Sterbebett] ...?
	Bei günstigeren Temperaturen wäre [$_M$ ich] schneller **gelaufen**.
	[$_M$ Die Testwochen] **laufen** [$_P$ vom 10. Januar] [$_P$ bis zum 10. März 1997] ...
	So **läuft** [$_M$ ihr Wunsch] [$_P$ ins Leere].
	[$_M$ Eine Träne] **läuft** ihr [$_P$ über das faltige Gesicht].
	[$_M$ Ich] bin **gerannt**.
	[$_M$ Freeman] **rennt** für Australien ...
	Damals **rannten** [$_M$ die Spanier] [$_P$ gegen Mauern] ...
	In Arabien **rennen** [$_M$ Kamele gegen Pferde] [$_M$ durch die Wüste] ...
	[$_M$ Menschen] **rannten** in Panik [$_P$ auf die Straßen] ...
	[$_M$ Er] **schleicht** [$_P$ zur Bühne] ...
	Wortlos **schlichen** [$_M$ die Stars] am Ende [$_P$ von dannen] ...
	[$_M$ Hopkins] **schleicht** eher gebeugt [$_P$ durchs Leben].
	Statt [$_P$ in den ersten Stock] **schleicht** [$_M$ er] [$_P$ in den Computerraum] ...
	[$_M$ Der hager-spitzbärtige Nikolai Tscherkassow] **schleicht** [$_P$ durch düstere Gewölbe] ...
	Langsam **wandert** [$_M$ der Friedhof] ...
	[$_M$ Die Wellen] sind über Nacht nicht **gewandert**.
	[$_M$ Wir] **wanderten** [$_P$ über weite Bergmatten] ...
	So **wandert** [$_M$ die Diskette] auf wundersamem Wege unbemerkt [$_P$ in den Müll] ...
	[$_M$ Ein Stück Kreidefelsen] **wandert** [$_P$ ins Handschuhfach des Autos] ...
	Mit Schneeschuhen kann [$_M$ man] in diesem Winter [$_P$ durch die Rhön] **wandern**.
n$_M$ **a**$_P$	+ gehen, laufen, rennen, wandern / ¬ klettern, kriechen, schleichen
	[$_M$ Die Planverfasser] **gehen** [$_P$ einen anderen Weg] ...
	[$_P$ Keinen Schritt] hätte [$_M$ ich] **laufen** können.
	[$_M$ Der neue Tarifvertrag] soll [$_P$ zwölf Monate] **laufen**.
	[$_M$ Er] spielt lieber Tennis und **läuft** gern [$_P$ Strecken bis zu 10 Kilometer].
	... **läuft** [$_M$ sie] [$_P$ Gefahr], selbst jene zu verlieren.
	... denn genaugenommen war [$_M$ Kipketer] ... [$_P$ neuen Weltrekord] **gerannt**.
	... [$_M$ der] [$_P$ 100 Meter] **rennt** ...
	... [$_M$ sie] **wandern** mitunter [$_P$ dreihundert Kilometer] flußaufwärts ...
n$_M$ **r**	+$_{adv}$ schleichen / ¬ gehen, klettern, kriechen, laufen, rennen, wandern
	... **schleicht** sich dem Jungen [$_M$ die Sehnsucht nach der Mutter] [$_P$ ins Herz].
	Als Elfjähriger **schlich** sich [$_M$ "Petit Julien"] [$_P$ auf den Dreimaster "Coralie"].
	... [$_M$ man] **schleicht** sich [$_P$ zur Wasserflasche] ...

Class 9: *Manner of Motion → Rotation*

Verbs: *drehen, rotieren*

Ambiguity: *drehen* has a sense similar to the verbs in class 8, but it does not specify the kind of motion. In addition, *drehen* has obtained a sense of ‘to shoot a film’.

Scene: [$_M$ Something] moves around its own axis. The motion might have an external [$_C$ cause].

Frame Role: M(over)

Modification Role: C(ause)

Levin class: 51.3.1 *(Verbs of Assuming a Position → Manner of Motion Verbs → Roll Verbs → Motion around an Axis)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_M$	+ rotieren / ¬ drehen
	Endlos **rotierten** [$_M$ die Dreiklangsbrechungen] ...
	[$_M$ Odinshühnchen] ... **rotieren** dabei einmal um die eigene Achse.
	In der sogenannten Paul-Fallen **rotiert** [$_M$ ein elektromagnetisches Feld].
	...dabei **rotieren** [$_M$ die jeweils aufeinanderfolgenden Walzen] immer schneller.
	Während [$_M$ die Blaulichter umstehender Polizeiwagen] **rotierten** ...
n$_C$ **a**$_M$	+ drehen / ¬ rotieren
	... [$_C$ der Notar] **drehte** [$_M$ seinen Schlüssel].
	... wie [$_C$ man] [$_M$ ihn] auch **dreht** und wendet.
	[$_C$ Er] ... **dreht** [$_M$ sie] vielmehr hoch ins Zwielicht ...
n$_C$ **a**$_M$ [P]	+ drehen / ¬ rotieren
	... [$_M$ der Sitz] kann **gedreht** werden ...
	... daß [$_M$ das Bild] ... entsprechend dem gedrehten Gehäuse **gedreht** wird ...
	... [$_M$ die] [$_C$ von Pilgern und Mönchen] ... **gedreht** werden sollten ...
n$_M$ **r**	+ drehen / ¬ rotieren
	[$_M$ Ein Mühlrad] **dreht** sich zweckfrei um die eigene Achse ...
	... an denen sich [$_M$ Windrädchen] **drehen**.
	... [$_M$ die Koalitionsparteien] **drehen** sich munter im Kreis.
	[$_M$ Deren Beine] hielten inne und **drehten** sich ...
	Nachdem sich plötzlich [$_M$ der Wind] **drehte** ...
	Schon **dreht** sich [$_M$ das Kandidatenkarussell] ...

Class 10: ***Manner of Motion → Rush***

Verbs: *eilen*, *hasten*

Scene: [$_M$ Somebody or something] moves in a rush. The movement is typically specified with respect to some [$_P$ path].
In a closely related, but rather idiomatic sense, *eilen* specifies an [$_U$ urgent matter], a necessity to rush, often with respect to [$_E$ an experiencer].

Frame Roles: M(over) or U(rgent matter)

Modification Roles: P(ath), E(xperiencer)

Levin class: 53.2 *(Verbs of Lingering and Rushing → Verbs of Rushing)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_U$	+ eilen / ¬ hasten
	[$_U$ Das Referendum] **eilt**.
	[$_U$ Die Sache] **eilt** nicht!
	[$_U$ Eine Entscheidung] **eile**, betonte ... auch der Rechtsanwalt der Gruppe ...
n$_M$	+$_{adv}$ eilen, hasten
	... als [$_M$ sie] [$_P$ ins Gericht] **eilt** ...
	... daß [$_M$ Trainer Murdoch] [$_P$ aufs Eis] **eilte** ...
	... [$_M$ die Tagesordnung] **eilte** ohne jede auffällige Hast [$_P$ zum nächsten Punkt].
	[$_M$ Der deutsche Aktienmarkt] **eilt** [$_P$ von Rekord zu Rekord].
	[$_M$ Ein junger Musiker] **eilt** ... aus den Armen der Geliebten [$_P$ nach Hause].
	[$_M$ Die großen Banken der Welt] **eilen** [$_P$ Südkorea zu Hilfe].
	... [$_M$ sie] **hasten** [$_P$ von Bedeutsamkeit zu Bedeutsamkeit] ...
	... und [$_M$ ich] **hastete** [$_P$ hinterher].
	[$_P$ Quer durch deren Zimmer] **hasten** [$_M$ die Männer] ...
	... **hasten** [$_M$ Spaziergänger] [$_P$ über die Straßen] ...
x d$_E$	+ eilen / ¬ hasten
	[$_E$ Den vier Mächten] **eilte** es.
	[$_U$ Mit einer Einigung] **eilt** es [$_E$ beiden Regierungsparteien] nicht.
	[$_E$ Den Amerikanern] **eilt** es ja nicht, wo doch Europa sich so vordrängelte ...

Class 11: ***Manner of Motion → Vehicle***

Verbs: *fahren, fliegen, rudern, segeln*

Scene: There is movement with respect to [$_M$ a specific vehicle]. The external cause for movement might be specified with respect to [$_I$ an initiator]. Next to a reference to a vehicle, it is possible to either refer to [$_M$ an animal] performing the respective kind of motion, or to abstract the manner of motion with respect to [$_T$ things] which can be brought into the respective kind of motion by external force. Typically, [$_P$ an origin or a path] defines the direction of the movement. In addition, there might be [$_E$ a driven person or thing] within the moving vehicle.

Frame Roles: M(over) or (Moved) T(hing), I(nitiator)

Modification Roles: P(ath), E(xperiencer)

Levin class: 51.4.2 *(Verbs of Assuming a Position → Verbs of Motion using a Vehicle → Verbs that are not Vehicle Names)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_M$	+$_{adv}$ fahren, fliegen, segeln / ¬ rudern
	Auf kurvenreichen Strecken **fährt** [$_M$ er] dank Frontantrieb völlig untückisch.
	[$_M$ Busse und Bahnen] **fahren** pünktlich ...
	Sagen Sie, [$_M$ welcher Bus] **fährt** [$_P$ in die Innenstadt]?
	[$_M$ Dutzende Kranken- und Feuerwehrwagen] **fuhren** [$_P$ zum Ort der Explosion].
	Anschließend **flog** [$_M$ der Hubschrauber] [$_P$ in ein Fußballstadion im Süden der Hauptstadt].
	Noch **fliegt** [$_M$ der Airbus] mit schwerem Ballast.
	Wie ein Pfeil **flog** [$_M$ das leichte Boot] [$_P$ über die silberne Bahn] ...
	[$_M$ Diese Art] **fliegt** hoch und hat einen kräftigen Flügelschlag.
	[$_M$ Sie] sollte im Auftrage des Zaren [$_P$ um die Welt] **segeln**.
	Annas Aal malt, [$_M$ Peters Esel] **segelt**, Iris' Iltis ißt.
	[$_M$ Möwen] **segeln** krächzend im Aufwind des Strombetts.
	1995 **segelte** [$_M$ der Siebzehn-Meter-Katamaran] ... [$_P$ durch den Südpazifik].
	... als [$_M$ Schiffe] nach Kompaß und Lot [$_P$ über die Weltmeere] **segelten** ...
n$_T$	+ fliegen, segeln / ¬ fahren, rudern
	[$_T$ Patronenhülsen] **fliegen** [$_P$ durch die Gegend] ...
	... [$_T$ die Erde] **flog** [$_P$ zu allen Seiten].
	... wenn in Sofia nicht [$_T$ Steine und Gummikugeln] **fliegen**.
	[$_T$ Haarbüschel] **flogen**.
	Dort aber **fliegen** [$_T$ die Fäuste] mit einer Vehemenz ...
	Es **segeln** [$_T$ Sterne] [$_P$ durch die Luft] ...
	... [$_P$ von den Rängen] **segelten** [$_T$ Dutzende von Sitzkissen] [$_P$ aufs Eis].
	Doch dann **segelte** [$_T$ der Ball] nach Tretschoks Schuß [$_P$ ins Tor] ...
n$_I$	+ rudern, segeln / +$_{adv}$ fahren, fliegen
	[$_I$ Wer] **fährt** schon zum Skifahren [$_P$ ins nördliche Griechenland]?
	Zunächst **fuhr** [$_I$ man] [$_P$ durch den Wald].
	Danach **fuhr** [$_I$ der Mann] [$_P$ nach Karlsruhe].
	... Gesellschaft, [$_I$ die] [$_P$ von vielen deutschen Flughäfen] **fliegt** ...

	... **flog** [$_I$ Luschkow] am Freitag [$_P$ nach Sewastopol] ...
	... [$_I$ ich] **fliege** nicht [$_P$ in die Luft].
	Bei der WM 1995 **flog** [$_I$ der Läufer] [$_P$ aus der Nationalmannschaft] ...
	20 Prozent der Kunden, [$_I$ die] [$_P$ von Kassel] aus **fliegen** ...
	[$_I$ Die Lufthansa] **fliegt** [$_P$ nach Almaty (Kasachstan)] ...
	[$_I$ Kavaliere] **rudern** mit Zylindern auf den Seen ...
	Seit zwei Jahren **rudert** [$_I$ Maréchal] [$_P$ durch die Texte Claudels, Becketts und nun Préverts].
	Einmal **ruderte** [$_I$ er] [$_P$ über die Bucht] [$_P$ nach San Francisco] ...
	[$_I$ Ich] **rudere** seit dem zehnten Lebensjahr ...
	Bislang hat der Vater von vier Söhnen, [$_I$ der] in seiner Freizeit gern **segelt** ...
	... aber [$_I$ sie] **segelten** alle [$_P$ in unterschiedliche Richtungen] ...
n$_I$ **a**$_M$	+ fahren, fliegen, rudern / ¬ segeln
	[$_I$ Der Filius] **fährt** inzwischen [$_M$ einen blauen Ferrari] ...
	Weil [$_I$ man] ... sowieso [$_M$ keine großen Maschinen] **fahren** darf ...
	[$_I$ Der Favorit] **fuhr** [$_M$ Kowalit].
	[$_I$ Schroeders Teamkollege Uli Schwenk] **fliegt** [$_M$ eine Maschine] ...
	[$_I$ Eine neue Besatzung] **flog** [$_M$ die Maschine] [$_P$ nach Budapest] weiter.
	... wie [$_I$ man] [$_M$ einen Vierer] **rudern** muß.
	... [$_I$ der] [$_M$ das Boot] **rudern** sollte ...
n$_I$ **a**$_M$ [P]	+ fahren, fliegen, rudern / ¬ segeln
	... und daß [$_M$ neue Autos] mehr **gefahren** werden als alte ...
	... **gefahren** wird [$_M$ ein alter Ford Granada].
	Daß [$_M$ ein dafür konstruierter Ski] [$_I$ von jedermann] **gefahren** werden kann ...
	... daß [$_M$ die Maschine] ... mit Höchstgeschwindigkeit **geflogen** werden könne.
	... Ruderboote ... [$_M$ die] dann [$_P$ zu ... Zahlungsstellen] **gerudert** werden ...
	Sie hätte das lecke Boot, [$_M$ das] zwei Stunden **gerudert** wurde ...
n$_I$ **a**$_E$	$+_{adv}$ fahren, fliegen, rudern / ¬ segeln
	Und Chauffeure. [$_I$ Sie] **fahren** [$_E$ die Kinder] [$_P$ durch Eis und Schnee] ...
	[$_I$ 14 Hubschrauber] **flogen** [$_E$ Sandsäcke].
	Den Bootsleuten, [$_I$ die] [$_E$ sie] [$_P$ über den Sone-Fluß] **gerudert** hatten ...
n$_I$ **a**$_E$ [P]	$+_{adv}$ fahren, fliegen, rudern / ¬ segeln
	[$_E$ Georges Mathieu] war ... [$_P$ ins Atelier von Otto Piene] **gefahren** worden.
	... Dinge, [$_E$ die] im Wunderkarren [$_P$ von einem Zirkus] [$_P$ zum andern] **gefahren** werden.
	... [$_E$ Anhänger Garangs] [$_P$ von Uganda aus] [$_P$ an die Front] ... **geflogen** worden sind.
	[$_M$ In einem Hubschrauber] seien ebenfalls [$_E$ mehrere Ärzte] [$_P$ nach Vlora] **geflogen** worden.
	[$_E$ Die Verletzten] wurden [$_P$ in Krankenhäuser] [$_P$ nach Winnipeg] **geflogen**.
	... Pirogge, ... [$_E$ die] [$_I$ von Haida-Indianern] [$_P$ von Rouen] [$_P$ bis ...] **gerudert** wurde.
n$_I$ **a**$_P$	+ fahren, fliegen, rudern, segeln
	[$_I$ Sie] sollte häufiger [$_P$ Slalom] **fahren** ...
	Ich weiß, daß [$_I$ er] [$_P$ eine schnelle Runde] **fahren** kann ...
	... [$_P$ 170 Meilen je Stunde] zu **fahren**.
	... [$_I$ sie] **fahre** [$_P$ die harte Linie].
	... **fliegt** [$_I$ sie] bisweilen [$_P$ einige hundert Meter weit] ...
	[$_I$ Der Pilot] habe [$_P$ eine Schleife] **geflogen** ...
	[$_I$ Sie] sollte im Skiff [$_P$ eine Runde auf dem Küchensee] **rudern**.
	... und **segelt** nebenbei [$_P$ die Weltmeisterschaft im Flying Dutchman].
	So [$_P$ ein Rennen] **segelt** [$_I$ man] nur einmal im Leben.
n$_I$ **a**$_P$ [P]	+ fahren, rudern / ¬ fliegen
	... zweite Riesenslalom, [$_P$ der] am Samstag **gefahren** werden sollte ...
	Bis dahin ... werde am Staatstheater Kassel "[$_P$ ein harter Sparkurs] **gefahren**".
	... die längste Strecke, [$_P$ die] ... problemlos **gerudert** werden konnte.

Class 12: *Manner of Motion* $\rightarrow$ *Flotation*

Verbs: *fließen, gleiten, treiben*

Ambiguity: *treiben* has obtained additional senses such as ‘to do’, and ‘to have sex’.

Scene: [$_M$ Something] moves by floating. A possible means of flotation is defined by a moving surface. Typically, [$_P$ an origin or a path] defines the direction of the movement. In causative cases of flotation, [$_C$ an external cause] might be given.

Frame Roles: M(over)

Modification Roles: P(ath), C(ause)

Levin class: 51.3.1 *(Verbs of Assuming a Position $\rightarrow$ Manner of Motion Verbs $\rightarrow$ Roll Verbs)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_M$	+ fließen, gleiten / +$_{adv}$ treiben
	Es **fließt** weiterhin [$_M$ ein elektrischer Strom].
	[$_M$ Die Gewinne] ... **fließen** [$_P$ auf Konten im Ausland].
	[$_M$ Die Gelder] **fließen** vor allem [$_P$ an die Ortskrankenkassen] ...
	Sie erwärmen den Kältestrom, [$_M$ der] [$_P$ durch seine Filme] **fließt**.
	... daß [$_M$ Tränen] **fließen** ...
	[$_M$ Der Wein] **fließt** in Strömen.
	[$_M$ Die Wupper] **fließt** [$_P$ durch Bochum] ...
	Vielmehr **gleitet** [$_M$ die indische Platte] sehr flach [$_P$ unter die eurasische Platte].
	Instinktsicher **gleiten** [$_M$ die Ströme].
	Die Hände geben auf, [$_M$ der Blick] **gleitet** [$_P$ über das Meer].
	[$_M$ Das Boot] **glitt** lautlos [$_P$ durch das eisige Wasser].
	[$_M$ Der Zug] **gleitet** [$_P$ an Hell’s Gate vorbei] ...
	Zu nahe **gleiten** [$_M$ Kunst und Design] [$_P$ in dekorative Bahnen].
	Nur gelegentlich **glitt** [$_M$ ein leichtes Lächeln] [$_P$ über sein Gesicht].
	Aber **gleiten** [$_M$ nur die Augen]?
	... [$_M$ der Bug] **trieb** später [$_P$ auf ein Riff vor der Küste].
	Jetzt **treibt** [$_M$ diese Flaschenpost] [$_P$ von Frankreich] [$_P$ nach Deutschland].
	Zwischen Mainz und Koblenz **trieben** [$_M$ einige tausend Tonnen tote Fische].
n$_C$ **a**$_M$	+ treiben / ¬ fließen, gleiten
	[$_C$ Hunger und Kälte] **treiben** [$_M$ sie] [$_P$ auf die Straße].
	[$_C$ Steuer] **trieb** [$_M$ die Ursachenforschung] [$_P$ noch ein wenig weiter].
	[$_M$ Ihn] **treibt** [$_C$ der Ehrgeiz] ...
	[$_C$ Schlechtes Führungsverhalten] **treibt** [$_M$ Fehlzeiten] [$_P$ in die Höhe].
	Jeden Frühsommer **treibt** [$_C$ er] noch [$_M$ seine Kühe] ...
	[$_C$ Der Küstenwind] **treibt** [$_M$ die weiß leuchtenden Dünen] ... [$_P$ vorwärts] ...
n$_C$ **a**$_M$ [P]	+ treiben / ¬ fließen, gleiten
	Wenn [$_M$ das Volk] aber [$_C$ von Armut und Verzweiflung] **getrieben** werde ...
	... in dem [$_M$ Konflikte] gerne [$_P$ auf die Spitze] **getrieben** werden.
	Sehen [$_M$ Sie] es als Privileg, [$_C$ von meist gut bemittelten Leuten] **getrieben** zu werden?
	... dürften [$_M$ Heimarbeiter] nicht [$_P$ in die Scheinselbständigkeit] **getrieben** werden.

Class 13: ***Emotion*** $\rightarrow$ ***Origin***

Verbs: *ärgern, freuen*

Scene: [$_E$ Somebody] experiences an inner emotion. Typically, [$_C$ the cause] of the emotion is expressed.

Frame Roles: E(xperiencer)

Modification Roles: C(ause)

Levin class: 31.1 *(Verbs of Psychological State* $\rightarrow$ *Amuse Verbs)*
Levin's class of amuse verbs comprises my classes 13 and 15.

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_C$ **a**$_E$	+ ärgern, freuen
	[$_C$ Boris Becker] wird [$_E$ den Deutschen Tennis-Bund] **ärgern** ...
	[$_C$ Was] [$_E$ uns] jetzt **ärgert**, ist der Gleichmut in Bonn.
	[$_C$ Die Berichterstattung in vielen Medien] **ärgert** [$_E$ ihn] da schon ein wenig.
	Denn nicht nur [$_C$ zu schwache Feldstärken] **ärgern** [$_E$ Mobilfunker].
	... [$_E$ ihren bekanntesten Gemeindebürger] zu **ärgern** ...
	[$_C$ Vermeidbare Niederlagen] **ärgern** [$_E$ mich] ...
	[$_C$ Der Abschied vom Leistungsfähigkeitsprinzip] mag [$_E$ die selbsternannte Partei] ... **freuen**.
	[$_C$ Das] **freut** möglicherweise [$_E$ die bayerischen Brauer] ...
	... [$_C$ das] hätte [$_E$ Casanovas Bruder] sehr **gefreut** ...
n$_C$ **a**$_E$ [P]	+ ärgern / ¬ freuen
	... nachdem [$_E$ sie] über viele Jahre hinweg [$_C$ von den Amateuren] **geärgert** worden waren ...
	... wie [$_E$ Scientology] [$_C$ mit Nadelstichen] **geärgert** werden kann.
	... wie [$_E$ eine Türkin] **geärgert** wird ...
	... [$_E$ die] **geärgert** worden war [$_C$ von ihrer ehemaligen Schülerin] ...
n$_E$ **r**	+ ärgern, freuen
	Richtig **ärgern** kann [$_E$ er] sich [$_C$ darüber] nicht.
	... im Vorjahr haben [$_E$ wir] uns sehr **geärgert** ...
	Andererseits **ärgerte** [$_E$ sie] sich [$_C$ über die Entscheidung der Trainer].
	Wenn zwei sich treffen, **ärgert** sich [$_E$ der Dritte]:
	Nebenbei **ärgert** [$_E$ er] sich [$_C$ über die Werbegrafiken] ...
	[$_E$ Ich] **ärgere** mich [$_C$ über die Scheintoleranz] ...
	[$_E$ Viele] **ärgern** sich derzeit [$_C$ über die Anflüge amerikanischer Arroganz].
	[$_E$ Hongkong] **freut** sich [$_C$ auf 1997].
	[$_E$ Der Kanzler und die Sternsinger] **freuen** sich [$_C$ auf Spenden].
	Mögen sich viele Leser finden, [$_E$ die] sich mit ihm **freuen** können.
	[$_E$ Beide Seiten] **freuen** sich [$_C$ über die gute Zusammenarbeit].
	... diesmal **freuen** [$_E$ wir] uns wieder [$_C$ über unseren zweiten Platz].
	Ja, [$_E$ ich] **freute** mich und war sogar ein wenig stolz darauf ...
	[$_E$ Unser Schatzmeister] wird sich nicht **freuen**.
	[$_E$ Kein Börsianer] **freut** sich [$_C$ über die hohe Arbeitslosigkeit].
n$_E$ **r i**$_C$	+ ärgern, freuen
	[$_E$ Wer es sieht], **ärgert** sich, [$_C$ nicht selbst die Idee gehabt zu haben]:
	[$_E$ Ich] **freue** mich riesig, [$_C$ eine gesunde Tochter zu haben].
	[$_E$ Man] **freue** sich darauf, [$_C$ Veba partnerschaftlich dabei zu unterstützen] ...
	Heute **freue** [$_E$ ich] mich, [$_C$ mich mit ihm zu treffen].
n$_E$ **r s-dass**$_C$	+ ärgern, freuen
	[$_E$ Ricco Groß] **ärgert** sich nur, [$_C$ daß seine Siegform den Olympiatermin verpaßt hat].
	[$_E$ Er] **freute** sich, [$_C$ daß sein ehemaliger Stabsunteroffizier ... da war].
	[$_E$ Die Wagenbewohner] selbst **freuen** sich, [$_C$ daß sie ihrem Hobby ... nachgehen können].
a$_E$ **s-dass**$_C$[2]	+ ärgern, freuen
	[$_E$ Den Sozialtechnokraten] **ärgert** es, [$_C$ daß die Bürger sich nicht ... anvertrauen] ...
	Weil es [$_E$ sie] **ärgerte**, [$_C$ daß Kinderliteratur plötzlich nur noch realistisch sein sollte] ...
	Zudem **ärgerte** [$_E$ ihn], [$_C$ daß die Polen unverändert Zweifel säten ...
	Es **freut** [$_E$ mich] ganz besonders, [$_C$ daß Silvio Meißner zwei Tore gemacht hat].
	[$_C$ Daß der Kiki wieder mal ein Tor gemacht hat], **freut** [$_E$ mich] besonders.

[2] This frame is not coded in the grammar.

Class 14: *Emotion → Expression*

Verbs: *heulen*$_1$, *lachen*$_1$, *weinen*

Ambiguity: *heulen* is classified both as an overt non-verbal expression of emotion (this class) and the specific kind of moaning (class 19). *lachen* is classified both as an expression of motion which includes sound emission (this class) and as facial expression without sound emission (class 16).

Scene: [$_E$ Somebody] expresses a non-verbal emotion by a facial expression and overt sounds specific for the emotion. [$_C$ The cause] for the emotion might be mentioned; in addition, the emotion might be described by [$_I$ illustrative means]. The verbal events are often transfered to [$_E$ things] with sound emission similar to the original emotional sound.

Frame Roles: E(mitter)

Modification Roles: C(ause), I(llustration)

Levin class: 40.2 *(Verbs Involving the Body → Verbs of Nonverbal Expression)*
Levin does not distinguish the non-verbal expression of emotion with regard to emitting a sound or not (compared to the distinction between my classes 14 and 16).

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_E$	+ heulen, lachen, weinen
	In Melbourne **heulten** ... in diesem Jahr wieder ganz offiziell [$_E$ die Motoren]. ... könnte [$_E$ man] [$_C$ vor Wut] **heulen**. [$_E$ Er] verliert und **heult** erbarmungslos in die Kameras. [$_E$ Eine Sirene] **heult**. Doch dann **heult** und orgelt auch schon [$_E$ die Musik von Jerzy Satanowski]. Schließlich **heulte** [$_E$ nur Oksana Bajul] [$_C$ vor Glück]. Da hat [$_E$ er] am Telefon **geheult** [$_C$ vor Erleichterung]. [$_E$ Laura] **lacht** und begrüßt Nummer 94: [$_E$ Er] schaut listig und **lacht** ... Am Telefon **lacht** [$_E$ er] bitter [$_C$ über die Stadt] ... Wenn [$_E$ man] schon [$_C$ über Banalitäten] **lachen** soll ... ... [$_E$ ich] würde [$_C$ um dich] **weinen**! [$_E$ Mutti] sitzt oben und **weint**. [$_E$ Sie] **weint** [$_C$ über den Verlust des kindlichen Geheimnisstandes] ... ... [$_E$ die Gottesmutter] **weine** [$_C$ aus Enttäuschung über die Menschen]. [$_E$ Sie] **weinen** [$_C$ über die Laufmasche in der Strumpfhose] ... [$_E$ Ich] könnte **weinen** [$_C$ vor Glück].
n$_E$ **a**$_I$	+ heulen / ¬ lachen, weinen
	[$_E$ Sie] **heulten** hemmungslos [$_I$ Rotz und Wasser] ... [$_E$ Eine Hündin] **heult** [$_I$ sieben Laute] ...
n$_E$ **r**	+$_{adv}$ lachen / ¬ heulen, weinen
	[$_E$ Die] **lachten** sich kaputt. Aber [$_E$ ich] **lach** mich tot.

Class 15: *Emotion* → *Objection*

Verbs: *ängstigen, ekeln, fürchten, scheuen*

Scene: [$_E$ Somebody or a group of people] feels an objection against [$_C$ something].

Frame Roles: E(xperiencer)

Modification Roles: C(ause)

Levin class: 31.2 *(Verbs of Psychological State → Admire Verbs, Negative Verbs)*
Levin's class of amuse verbs comprises my classes 13 and 15.

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
$\mathbf{n}_E$	+ scheuen / $+_{adv}$ fürchten / ¬ ängstigen, ekeln
	Fischer, [$_E$ die] [$_C$ um ihren Fang von Muscheln, Seetang und Krabben] **fürchten** ... Zwar müssen [$_E$ Wintersportler] nicht [$_C$ um den Schnee] **fürchten** ... Als ein Hund bellte, **scheute** [$_E$ das Tier].
$\mathbf{n}_C\ \mathbf{a}_E$	+ ängstigen, ekeln / ¬ fürchten, scheuen
	[$_E$ Ihn] **ängstigte** [$_C$ das fremd gewordene Land] ... Nur [$_E$ Fremde] **ängstigt** [$_C$ die Szenerie]. [$_C$ Gestank, Schmutz und Unrat] **ekeln** [$_E$ sie]. ... [$_E$ den] [$_C$ er] aus dem Land **ekeln** half ... Jetzt ist er ja Maler, und es gibt Tage, da **ekeln** [$_E$ ihn] [$_C$ die Worte].
$\mathbf{n}_E\ \mathbf{a}_C$	+ fürchten, scheuen / ¬ ängstigen, ekeln
	[$_E$ Die Wirtschaft] **fürchtet** [$_C$ den Machtwechsel in Hongkong] kaum. In Madrid **fürchtete** [$_E$ man] ... [$_C$ eine unnötige Konkurrenz um potentielle Anleger]. [$_E$ Sie] scheinen [$_C$ ihn] mehr zu **fürchten** als seine Vorgängerin ... [$_E$ Er] **fürchte** [$_C$ Rußland] nicht. [$_E$ Die Sowjets] **scheuten** [$_C$ ein Abenteuer]. [$_E$ Viele Arbeitgeber] **scheuten** [$_C$ das Risiko] ... [$_C$ Den offenen Streit] **scheuen** jedoch [$_E$ beide] aus Klugheit. Nicht zuletzt **scheuen** [$_E$ die Gegner] [$_C$ die Kosten] ...
$\mathbf{n}_E\ \mathbf{i}_C$	+ fürchten, scheuen / ¬ ängstigen, ekeln
	Aber [$_E$ sie] **fürchteten**, [$_C$ majorisiert zu werden] ... Offensichtlich **fürchten** [$_E$ sie], [$_C$ ins Hintertreffen zu geraten] ... Leider **scheuen** heutzutage [$_E$ nur noch wenige], [$_C$ sich ... zu blamieren].
$\mathbf{n}_E\ \mathbf{r}$	+ ekeln, fürchten / $+_{adv}$ ängstigen, scheuen
	Sein Nein würde ihn isolieren, [$_E$ er] **ängstigt** sich [$_C$ vor der Einsamkeit]. ... daß [$_E$ die junge Frau an seiner Seite] sich halb zu Tode **ängstigte**. [$_E$ Goethe] **ängstigte** sich [$_C$ vor Berlin und den Berlinern] ... ... [$_E$ ich] **ekelte** mich [$_C$ vor mir selbst]. Erst paart das Tier sich, und dann **ekelt** sich [$_E$ der Mensch]. ... [$_E$ die] sich [$_C$ vor Pornographie] **ekeln** ... [$_E$ Sie] **fürchten** sich [$_C$ weder vor Gesetzen noch vor Strafen]. Sie ist wie eine Rasierklinge, und [$_E$ ich] **fürchte** mich [$_C$ vor ihren Schnitten]. [$_E$ Sie] **fürchten** sich, ziehen sich zurück. [$_E$ Ihr klarer, lyrischer Sopran] **scheute** sich fast [$_C$ vor den Gewaltsamkeiten der Rolle] ... Deshalb **scheut** [$_E$ er] sich auch nicht [$_C$ vor der schrecklichen Vorstellung] ...
$\mathbf{n}_E\ \mathbf{r}\ \mathbf{i}_C$	+ ängstigen, fürchten, scheuen / ¬ ekeln
	... [$_E$ die] sich **ängstigen**, [$_C$ nicht zum Volk zu gehören]. [$_E$ Die Leute] **fürchten** sich auch nicht mehr, [$_C$ etwas in der Straße zu verkaufen]. [$_E$ Menschen] ... **fürchteten** sich, [$_C$ mit ihrer Meinung an die Öffentlichkeit zu gehen]. Auf allen Organisationsebenen **scheut** [$_E$ man] sich, [$_C$ Risiken zu übernehmen]. [$_E$ Lehrer] **scheuen** sich, [$_C$ auffällige Kinder zu melden] ...
$\mathbf{n}_E\ \mathbf{s\text{-}2}_C$	+ fürchten / ¬ ängstigen, ekeln, scheuen
	... [$_E$ ich] **fürchte**, [$_C$ Sie mißverstehen da etwas]. [$_E$ Montenegro] **fürchtet**, [$_C$ es werde ... mitbezahlen müssen]. ... [$_E$ ich] **fürchte**, [$_C$ in dieser wilden Dekade ist ein Geschlecht herangewachsen] ...
$\mathbf{n}_E\ \mathbf{s\text{-}dass}_C$	+ fürchten / ¬ ängstigen, ekeln, scheuen
	[$_E$ Sie] **fürchteten**, [$_C$ daß das Beispiel Schule machen ... könnte]. [$_E$ Ich] **fürchte**, [$_C$ daß der Fußball kaputtgeht]. ... sobald [$_E$ sie] **fürchten**, [$_C$ daß ihre Wohlfahrt leidet] ...

Class 16: *Facial Expression*

Verbs: *gähnen, grinsen, lachen$_2$, lächeln, starren*

Ambiguity: *lachen* is classified both as an expression of motion which includes sound emission (class 14) and as facial expression without sound emission (this class).

Scene: [$_E$ Somebody] expresses a specific emotion on the face, without sound emission. The characteristics of the looks can be transfered to [$_E$ similar looking things].

Frame Role: E(xperiencer)

Levin class: 40.2 *(Verbs Involving the Body → Verbs of Nonverbal Expression)*
Levin does not distinguish the non-verbal expression of emotion with regard to emitting a sound or not, compared to the distinction between classes 14 and 16.

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_E$	+ gähnen, grinsen, lachen, lächeln / +$_{adv}$ starren
	[$_E$ Man] schaut lange hin, wird schon müde und **gähnt** ...
	Da baumelt er nun; unter ihm **gähnt** [$_E$ der Abgrund].
	Wo [$_E$ die Leere] **gähnt**, will er Autos parken ...
	[$_E$ Die Journalisten und Einkäufer] **gähnten** etwas häufiger während der Defilees ...
	[$_E$ Keiner] **grinst** feister ...
	[$_E$ Sensenmänner] **grinsen** in der Werbung, Skelette tanzen Tango ...
	Guck mal, [$_E$ der] **grinst** schon wieder so dämlich.
	[$_E$ Unser Lehrer] **lacht** nie ...
	[$_E$ Sophie] **lacht** jetzt.
	[$_E$ Er] strahlte, **lachte**, umarmte alle Freunde und Förderer ...
	[$_E$ Sie] **lächelt** verlegen und wischt sich die Hände an ihrer Schürze ab.
	[$_E$ Man] durfte **lächeln** über Anna ...
	[$_E$ Mocsai] hebt nur die Schultern und **lächelt**.
	Zuerst sah er mich an, dann **starrte** [$_E$ er] auf das Bild von van Gogh.
	[$_E$ Alk] **starrt** eine Weile gebannt auf die Wohnungstür ...
	Die Türen schließen, [$_E$ die Fahrgäste] **starren** aus dem Fenster.
	Schaut man in die Zukunft, **starrt** [$_E$ man] in ein schwarzes Loch.

Class 17: *Perception*

Verbs: *empfinden, erfahren*$_1$*, fühlen, hören, riechen, sehen, wahrnehmen*

Ambiguity: Most verbs of this class show some ambiguity: *erfahren* has a sense of observation (class 24); the class 24 is closely related in the sense that somebody perceives something, but this class 17 refers to the specific body senses, class 24 to realising an observation. *hören* has a meaning of 'to obey'. *wahrnehmen* has a sense of 'to execute'.

Scene: [$_E$ Somebody] perceives [$_T$ information] with respect to a specific sense. The information might be specified with respect to the individual [$_P$ perception].

Frame Roles: E(xperiencer), T(heme)

Modification Roles: P(erception)

Levin class: 30.1 / 43.3 *(Verbs of Perception → See Verbs / Verbs of Emission → Verbs of Smell Emission)*
Levin's verbs of smell emission contain a subset of the see verbs, which show the specific alternation of things emitting a smell which on the other hand can be perceived.

Schumacher class: 4.9 *(Verben der Relation und des geistigen Handelns → Verben der Evaluation)*
Only *sehen* is classified.

Frame	Participating Verbs & Corpus Examples
n$_T$	+$_{adv}$ riechen / ¬ empfinden, erfahren, fühlen, hören, sehen, wahrnehmen
	[$_T$ Die] **riechen** schon [$_P$ etwas komisch] ...
	[$_T$ Die Luft] **riecht** förmlich [$_P$ nach Urlaub].
	Noch ist [$_T$ er] weiß und **riecht** [$_P$ nach Milch] ...
	[$_T$ Er] **riecht** überhaupt nicht!
	[$_P$ Rot] **roch** [$_T$ die Nacht], [$_P$ nach Farbe].
n$_E$ **a**$_T$ **i**$_P$	+ hören, sehen / ¬ empfinden, erfahren, fühlen, riechen, wahrnehmen
	... **hörten** [$_E$ seine Fans] [$_T$ ihn] nun [$_P$ im Radio zur Gitarre singen] ...
	[$_E$ Man] **hört** [$_T$ das Zuckerrohr] [$_P$ im Wind rauschen] ...
	Nachmittags ... **hört** [$_E$ man] [$_T$ sie] [$_P$ in die kreolische Stille donnern].
	[$_E$ Man] **hört** [$_T$ sie] [$_P$ grummeln] ...
	Auf dem Bildschirm ... **sieht** [$_E$ er] [$_T$ seine Tasche] [$_P$ vorbeiziehen].
	Sieht [$_E$ man] [$_T$ ihn] [$_P$ dirigieren] ...
x p$_T$:Dat.nach	+ riechen / ¬ empfinden, erfahren, fühlen, hören, sehen, wahrnehmen
	Es **riecht** [$_T$ nach brennenden Holzscheiten im Herd] ...
	Aber es **riecht** überhaupt nicht [$_T$ nach Schwein].
	Es **riecht** [$_T$ nach Öl, Metall und Schmierfett] ...

Frame	Participating Verbs & Corpus Examples
$\mathbf{n}_E$ $\mathbf{a}_T$	+ empfinden, erfahren, fühlen, hören, riechen, sehen, wahrnehmen
	Daß gerade [$_E$ die konservativen Eliten] ... [$_T$ den tiefsten Groll] **empfinden** ...
	[$_T$ Was] **empfindet** [$_E$ jemand, der mitmachte]?
	[$_E$ Ich] **empfand** [$_T$ die Größe ihrer Einsamkeit] ...
	[$_E$ Phoenix] dürfte in Deutschland [$_T$ ähnlichen Gegenwind] **erfahren** ...
	... wenn [$_E$ man] [$_T$ internationales Geistesleben] **erfahren** wollte.
	Auch [$_E$ die Nachgeborenen] **erfuhren** ... [$_T$ die "Gnade des Schmerzes"] ...
	... ohne dafür [$_T$ den Puls] **fühlen** zu müssen.
	[$_E$ Man] konnte [$_T$ den Schmerz] **fühlen** ...
	... [$_E$ er] **fühle** [$_T$ tiefe Trauer über Dengs Tod] ...
	[$_E$ Man] **hört** [$_T$ islamische Gebetsmusik und amerikanische Pop-Hits].
	[$_E$ Ein Fußgänger] **hörte** [$_T$ die Hilfeschreie der davontreibenden Männer] ...
	In den deutschen Großstädten **hört** [$_E$ man] [$_T$ Russisch] jeden Tag auf der Straße ...
	... weil [$_E$ sie] [$_T$ Gas] im Haus **gerochen** hatten.
	Riechen [$_E$ Sie] [$_T$ den Gestank der Sprache]?
	[$_E$ Ich] **rieche** schon [$_T$ den Duft der deutschen Linden] ...
	[$_E$ Wir] **sehen** [$_T$ die Welt] durch ihre Augen ...
	Immer wieder **sieht** [$_E$ man] auch [$_T$ Pilger] ...
	[$_T$ Neue Märkte] **sieht** [$_E$ man] im europäischen Ausland, vor allem im Osten.
	[$_E$ Sie] haben zunächst [$_T$ nur das Wasser der Spree] **wahrgenommen** ...
	... um ganz automatisch [$_T$ den vollen Umfang] ... **wahrzunehmen**.
	Trotz Freud-Lektüre ... sträuben [$_E$ wir] uns ... [$_T$ Abgründe] **wahrzunehmen**.
$\mathbf{n}_E$ $\mathbf{a}_T$ [P]	+ erfahren, fühlen, sehen, wahrnehmen / ¬ empfinden, hören, riechen
	[$_T$ Das Schwesteruniversum] ... kann ... nur [$_E$ von Frauen] ... **erfahren** werden.
	... [$_T$ die] jetzt [$_P$ als ganze und in ganzer Vielfalt] erstmals **erfahren** werden kann.
	Stündlich wird den Motoren [$_T$ der Puls] **gefühlt** ...
	Dabei dürfe [$_T$ der Mensch] nicht nur [$_P$ als Faktor Arbeit] ... **gesehen** werden.
	... [$_T$ es] stößt an die Decke ..., kann also [$_E$ vom Betrachter] ... **gesehen** werden.
	... daß ... [$_T$ diese Funktionen] [$_P$ auf andere Weise] **wahrgenommen** werden können.
	Daß [$_T$ er] nicht nur subjektiv **wahrgenommen** wird ...
$\mathbf{n}_E$ $\mathbf{a}_T$ $\mathbf{p}_P$:Akk.vgl	+ empfinden, erfahren, fühlen, sehen, wahrnehmen / ¬ hören, riechen
	[$_E$ Sie] **empfinden** [$_T$ das] [$_P$ als Öko-Kolonialismus].
	[$_T$ Dies] **empfinde** [$_E$ man] [$_P$ als besonders störend] ...
	[$_T$ Die Tour] **empfindet** [$_E$ Sanquer] [$_P$ als schöne Belohnung].
	[$_E$ Er] **fühlte** [$_T$ sich] [$_P$ als Preuße] ...
	... [$_E$ der] [$_T$ sich] von 1933 bis zu seinem Tod ... stets [$_P$ als Emigrant] **fühlte** ...
	[$_E$ Er] **sah** [$_T$ sich] schon nach der Landtagswahl [$_P$ als ... Ministerpräsidenten] ...
	[$_E$ Monsengwo] hat keine politischen Ambitionen, **sieht** [$_T$ sich] [$_P$ als Seelsorger] ...
	Dieser Blick, [$_E$ der] [$_T$ politische Bilder] unweigerlich [$_P$ als Pop-art] **wahrnimmt** ...
	... weil [$_E$ er] imstande war, [$_T$ die Deutschen] ... [$_P$ als Individuen] **wahrzunehmen**.
$\mathbf{n}_E$ $\mathbf{a}_T$ $\mathbf{p}_P$:Akk.vgl [P]	+ empfinden, erfahren, sehen, wahrnehmen / ¬ fühlen, hören, riechen
	Zum einen, weil [$_T$ hartnäckige Fragen] ... [$_P$ als störend] **empfunden** wurden;
	... [$_T$ jeder andere Star] ... wäre [$_P$ als Tabubruch] **empfunden** worden ...
	... in dem [$_T$ das Gegenwärtige] [$_P$ als das je schon Gewesene] **erfahren** wird.
	[$_T$ Die russische Revolution] ist [$_P$ als Naturgeschehen] **erfahren** worden ...
	... daß [$_T$ die Energiewirtschaft] ... nicht [$_P$ als Gesamtbild] **gesehen** wird.
	[$_T$ Dies] sollte nicht [$_P$ als letzte Runde der Erweiterung] **gesehen** werden ...
	... daß [$_T$ Umweltrisiken] [$_P$ als Gesundheitsgefahren] **wahrgenommen** werden.
	... wurden [$_T$ Umweltbelastungen] ... kaum [$_P$ als Probleme] **wahrgenommen**.
	... daß [$_T$ sie] [$_E$ von der Fachwelt] ... [$_P$ als Signum] **wahrgenommen** wurden.

Frame	Participating Verbs & Corpus Examples
n$_E$ **s-dass**$_T$	+ fühlen, hören, riechen, sehen, wahrnehmen / ¬ empfinden, erfahren
	... [$_E$ wir] **fühlen**, [$_T$ daß wir mit aller Cultur zu weit gegangen sind] ... [$_E$ Ich] **fühle**, [$_T$ daß ich am Ende der Straße bin] ... Immer **hören** [$_E$ wir] von den Touristen, [$_T$ daß sie nun nach Alice Springs fahren] ... Als [$_E$ wir] **hörten**, [$_T$ daß wir einen Austausch mit Schweriner Schülern haben würden] ... Wenn [$_E$ Sie] **hören**, [$_T$ daß jemand Beamter oder Beamtin ist] ... ... wenn [$_E$ er] **riecht**, [$_T$ daß sein Napf zum Abkühlen auf dem Fensterbrett steht] ... [$_E$ Sie] **sehen**, [$_T$ daß Dubai nur 600000 Einwohner hat] ... [$_E$ Man] sollte klar **sehen**, [$_T$ daß der Änderungsbedarf ... zunimmt]. ... wird [$_E$ man] im Rückblick **sehen**, [$_T$ daß er mit den Billionen nicht gekauft worden ist]. [$_E$ Die internationalen Märkte] werden **wahrnehmen**, [$_T$ daß es ... Währung gibt] ... ... haben dann ... [$_E$ Hochschulen in Europa] **wahrgenommen**, [$_T$ daß die Initiative] ...
n$_E$ **s-w**$_T$	+ erfahren, fühlen, hören, riechen, sehen, wahrnehmen / ¬ empfinden
	[$_E$ Man] **erfährt**, [$_T$ wie es ist, mit einem Chopper-Motorrad über Bergstraßen zu fahren] ... Viele waren entsetzt, als [$_E$ sie] **erfuhren**, [$_T$ wo sie ihr Austauschjahr verbringen würden]. [$_E$ Er] **fühlt**, [$_T$ wie nach und nach der Atem ruhiger wird] ... Vielleicht **hört** [$_E$ er] sogar, [$_T$ wie sie oben nach ihm rufen] ... [$_E$ Man] muß **hören**, [$_T$ wie Geisler solche Sätze spricht]. [$_E$ Wer] ... **hörte**, [$_T$ welche Erwartungen sie auf das Reisegeschäft ... richten] ... Und wenn der Wind dann von Westen kam, hat [$_E$ man] **gerochen**, [$_T$ was passiert ist]. Der VW-Chef Piëch ... sei einer, [$_E$ der] förmlich **rieche**, [$_T$ wie man ein Produkt ... plaziert]. ... [$_E$ man] **roch**, [$_T$ wo einer Arbeit gehabt hatte] ... [$_E$ Niemand] **sieht**, [$_T$ wie er zum zehntenmal einen Satz streicht und leise verzweifelt] ... Als [$_E$ ich] hochblickte und **sah**, [$_T$ wo ich war] ... ... konnten [$_E$ wir] zum ersten Mal richtig **wahrnehmen**, [$_T$ welche Probleme] ...
n$_E$ **s-ob**$_T$	+ erfahren, fühlen, hören, riechen, sehen / ¬ empfinden, wahrnehmen
	Dabei würde [$_E$ man] gerne **erfahren**, [$_T$ ob denn das Unsichtbare auch gesehen wird]. [$_E$ Ich] muß regelmäßig **fühlen**, [$_T$ ob sie noch da ist]. Mal **sehen** und **hören**, [$_T$ ob sie auch so dröhnend hinlangen] ... Dann braucht [$_E$ man] nur noch zu spielen und zu **hören**, [$_T$ ob ... durchgekocht ist] ... ... müßten [$_E$ die Schaltermitarbeiter] "**riechen**", [$_T$ ob dieses Geld] ...
n$_E$ **s-dass**$_T$ [P]	+ sehen, wahrnehmen / ¬ empfinden, erfahren, fühlen, hören, riechen
	Bei der Diskussion ... wird nicht **gesehen**, [$_T$ daß die Staatsverschuldung ... ist] ... ... wurde **gesehen**, [$_T$ daß das Zielfahrzeug ... auf das Gelände der Tankstelle auffuhr]. ... werde kaum noch **wahrgenommen**, [$_T$ daß ... Friedhöfe geschändet worden seien] ... Es wird auch nicht **wahrgenommen**, [$_T$ daß das alte Links-rechts-Schema zerbricht] ...

Class 18: *Manner of Articulation*

Verbs: *flüstern, rufen, schreien*

Scene: [$_A$ Somebody or a source of sound] articulates [$_T$ something], with a specific manner of articulation. The utterance might express [$_G$ a target of the articulation]. The context might mention [$_L$ a listener].

Frame Roles: A(rticulator), T(heme)

Modification Roles: L(istener), (tar)G(et)

Levin class: 37.3 *(Verbs of Communication → Verbs of Manner of Speaking)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_A$	+ flüstern, rufen, schreien
	... **flüstert** [$_A$ er] beschwörend ...
	[$_A$ Er] **flüstert** undeutlich ...
	... daß [$_A$ man] etwa **flüstern** muß ...
	... senken [$_A$ die meisten Leute] ... die Stimme, **flüstern** fast.
	Schließlich **rief** [$_A$ es] aus der Menge ...
	... als [$_A$ sein Großvater] **rief** ...
	[$_A$ Der Berg] **ruft** nicht mehr, er kommt:
	[$_A$ Er] mußte **schreien** ...
	[$_A$ Die kleineren Kinder] **schrien**.
	Als [$_A$ der Junge] in der Nacht **schrie** ...
	[$_A$ Hunderte Anhänger] **schrien**, als die Kolonne vorbeikam.
n$_A$ [P]	+ flüstern, schreien / ¬ rufen
	... wie **geflüstert** wurde ...
	Im Büchertempel wird nur **geflüstert** ...
	In "Cop Land" wird **geschrien**, geprügelt und geschossen ...
	Es wurde geschimpft und **geschrien**.
n$_A$ **a**$_T$	+ flüstern, rufen / ¬ schreien
	Wissend **flüsterten** [$_A$ diejenigen] ... nach dem Rundgang [$_T$ dies und das].
	[$_A$ Sie] **flüstert** und gurgelt [$_T$ ihre Botschaften] ...
	... [$_A$ einer] **flüstert** [$_T$ den Namen, der in Großbuchstaben in den Stein gemeißelt ist] ...
	[$_A$ Der Bauer aus dem Hochschwarzwald] **rief** unermüdlich [$_T$ seine Botschaften] ...
	... doch [$_A$ die Stadt] **ruft** [$_T$ immer neue Namen] durch die Korridore.
	[$_A$ Er] **rief** [$_T$ "Disziplin" und "Gewissen"] ...
n$_A$ **a**$_T$ [P]	+ flüstern, rufen / ¬ schreien
	[$_T$ Was] lange **geflüstert** wurde, wird jetzt lauter gesagt ...
	In den Nachkriegsjahrzehnten wurde [$_T$ dieser Spruch] nur **geflüstert** ...
	... [$_T$ was] zwischen den Fachsimpeleien **gerufen** wird:

Frame	Participating Verbs & Corpus Examples
n$_A$ **a**$_G$	+ rufen / ¬ flüstern, schreien
	[$_A$ Ein 33 Jahre alter Mann] ... **rief** [$_G$ den Notarzt]. [$_A$ Der Küstenschutz] hat [$_G$ zwei Zerstörer der Marine] zur Hilfe **gerufen**.
n$_A$ **a**$_G$ [P]	+ rufen / ¬ flüstern, schreien
	[$_G$ Die Rettungs- und Notarztwagen] seien in 475 Fällen **gerufen** worden ... Dabei kann [$_G$ Rußland] - von den arabischen Ländern zu Hilfe **gerufen** ...
n$_A$ **a**$_T$ **d**$_L$	+$_{adv}$ flüstern, schreien / ¬ rufen
	[$_A$ Das erste Kind] **flüstert** [$_L$ einem zweiten] schnell [$_T$ ein Wort] ins Ohr. [$_A$ Herr Everding] **flüstert** [$_L$ Herrn BMW] [$_T$ das Wort Geldnot] ins Ohr. ... wie [$_L$ mir] [$_A$ ein Geheimdienstler] **geflüstert** hatte ... [$_A$ Er] möchte [$_L$ dem Kapitalismus] ... [$_T$ die Wahrheit] ins Gesicht **schreien**.
n$_A$ **p**$_T$:Akk.über	+ flüstern / ¬ rufen, schreien
	... [$_T$ über] die [$_A$ man] nur **flüstert**.
n$_A$ **p**$_T$:Akk.über [P]	+ flüstern / ¬ rufen, schreien
	Und [$_T$ über das so lange gepriesene Vorbild] ... werde ... **geflüstert**.
n$_A$ **p**$_G$:Dat.nach	+ rufen, schreien / ¬ flüstern
	Verzweifelt **rief** [$_A$ er] [$_G$ nach Krediten und Lebensmitteln]. ... Gutachter - [$_G$ nach denen] [$_A$ der Autor] **ruft** ... Oder nach der Philosophie **rufen**. [$_A$ Sie] **rufen** nicht [$_G$ nach staatlicher Hilfe] ... [$_A$ Der Ottocilindri] **schreit** [$_G$ nach Liebe] ... ... [$_A$ alle Gäste] **schreien** [$_G$ nach mir] ...
n$_A$ **p**$_G$:Dat.nach [P]	+ rufen, schreien / ¬ flüstern
	... daß [$_G$ nach Auslieferung von Erich Honecker] **gerufen** wird ... Wenn [$_A$ von einzelnen Unternehmen] [$_G$ nach einer Ökosteuer] **gerufen** wird ... Wenn Lear stirbt, wird wieder [$_G$ nach der Krankenschwester] **gerufen**. ... weil nun in der DDR [$_G$ nach der deutschen Einheit] **geschrien** wird. ... wo so laut wie nirgens [$_G$ nach dem Retter aus dem Westen] **geschrien** wurde ...
n$_A$ **s-2**$_T$	+ flüstern, rufen, schreien
	[$_T$ Mehr vertrage ihr Magen nicht], **flüstert** [$_A$ sie] wie zur Entschuldigung. [$_T$ Ich mag dieses Lied sehr], **flüstert** [$_A$ die wohlklingende Stimme zur Rechten]. [$_T$ Man sieht keine Menschen], **flüstert** [$_A$ der Film] unter den Hammerschlägen ... [$_T$ Mein Herz zerspringt vor Freude], **ruft** [$_A$ ein aufgebrachter Zeitungsverkäufer]. [$_T$ Wir sind Bayern], **rief** [$_A$ er] triumphierend ... [$_A$ Sein Promoter Don King] **schrie** verärgert: [$_T$ Auf geht's]. [$_T$ Alle sollen ihn duzen], **schreit** [$_A$ Norbert] ...
n$_A$ **s-dass**$_T$	+ rufen, schreien / ¬ flüstern
	... [$_A$ der] ... laut **ruft**, [$_T$ daß der Kaiser keine Kleider trägt]. ... [$_A$ die] **rufen**, [$_T$ daß sie pflegebedürftig sind] ... [$_A$ Die Frau] ... **schrie**, [$_T$ daß ich für ihren Tod Verantwortung trüge]. Da spricht ihn [$_A$ jemand] von hinten an, **schreit**, [$_T$ daß er allen Grund ... habe] ...
n$_A$ **d**$_L$ **s-dass**$_T$	+ flüstern / ¬ rufen, schreien
	Hat [$_L$ ihm] [$_A$ kein Berater] **geflüstert**, [$_T$ daß er ... ... halten muß] ...? Da sollte [$_A$ man] [$_L$ sich] besser **flüstern**, [$_T$ daß es sich ... um eine ... Ente handelt]. ... hatte [$_A$ man] bereits [$_L$ allen editors] ins Ohr **geflüstert**, [$_T$ daß ... würde].

Class 19: *Moaning*

Verbs: *heulen*$_2$*, jammern, klagen, lamentieren*

Ambiguity: *heulen* is classified both as an overt general non-verbal expression of emotion (class 14) and the specific kind of moaning (this class). In legal language, *klagen* has obtained a sense of 'to sue'.

Scene: [$_A$ Somebody] moans about [$_T$ something]. [$_C$ The cause] for moaning might be mentioned. And [$_L$ a listener] to the moaner might be defined.

Frame Roles: A(rticulator), T(heme)

Modification Roles: C(ause), L(istener)

Levin class: 31.3 *(Verbs of Psychological State → Marvel Verbs)*
The moaning verbs are a part of Levin's marvel verbs (closest to those subcategorising the prepositions *for, over*).

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_A$	+ jammern, klagen, lamentieren / ¬ heulen
	[$_A$ Wer] heute **jammert**, statt zu handeln ...
	Draußen vor der Tür **jammert** [$_A$ das Volk] ...
	Doch [$_A$ Ralph H. Neal, der Direktor der Schule], **klagt** nicht.
	Lerne **klagen**, ohne zu leiden.
	Einer, [$_A$ der] nicht **lamentiert** und grübelt, sondern zupackt.
	[$_A$ Der Handel] sollte hier nicht **lamentieren** ...
n$_A$ [P]	+ heulen, jammern, klagen, lamentieren
	Bei mir wird **geheult** oder geschimpft.
	In Deutschland werde zu viel **gejammert**.
	Es durfte - ja sollte - **geklagt** werden:
	... daß in Mönchengladbach nicht nur **lamentiert** werde ...
n$_A$ **p**$_C$:Akk.über	+ heulen, jammern, klagen, lamentieren
	... [$_A$ der] zu Hause weint und **heult** [$_C$ über der schlechten Zeiten] ...
	... daß [$_A$ wir alle] [$_C$ über den Sender] **heulen** ...
	In der Ausbildungsstätte **jammern** [$_A$ die Bürokraten] [$_C$ über die Lage] ...
	... doch in Vancouver selbst **jammert** [$_A$ niemand] [$_C$ über den Wetterbericht] ...
	[$_A$ DGB-Beschäftigte] **klagen** [$_C$ über Doppelzüngigkeit].
	[$_A$ Der Vorstandsvorsitzende] **klagt** jedoch [$_C$ über die billigen Importe] ...
	[$_A$ Man] braucht jedoch nicht [$_C$ über den Zustand der Curry-Wurst] zu **lamentieren**.
	[$_A$ Er] **lamentierte** [$_C$ über den Fußballehrer] ...
n$_A$ **p**$_C$:Akk.über [P]	+ jammern, klagen, lamentieren / ¬ heulen
	... überall wird mit Recht [$_C$ über sterbende Bäume] **gejammert**.
	Warum wird eigentlich jetzt [$_C$ über die Kosten der Einheit] so **gejammert**?
	Viel wurde **geklagt** [$_C$ über Filme wie The Moderns oder Henry and June].
	Während ... viel [$_C$ über die Menschenrechtsverhältnisse in Ost] **geklagt** wird ...
	... daß [$_C$ über die Ungerechtigkeit] ... **lamentiert** wird ...
	Es soll ... wieder [$_C$ über die ewigen Preissteigerungen] **lamentiert** werden.
n$_A$ **s-2**$_T$	+ jammern, klagen, lamentieren / ¬ heulen
	[$_T$ Niemand will mich], **jammerte** [$_A$ er] ...
	[$_T$ Das darf nicht sein], **jammerte** [$_A$ ein ... Teenager] immer wieder still vor sich hin ...
	[$_T$ Die Leute haben kein Geld], **klagt** [$_A$ eine Verkäuferin].
	[$_A$ Die Senatoren Helms und Burton] **klagten**, [$_T$ Clinton habe sich selbst entwaffnet].
	[$_T$ Er mag mich nicht], **klagte** [$_A$ Erhard] schon 1950.
	Später **lamentierte** [$_A$ er]: [$_T$ Zur Zeit sind wir harmlos].
	[$_T$ Dardenne], **lamentierte** [$_A$ er], [$_T$ sei der Aufgabe nicht gewachsen gewesen].
n$_A$ **s-dass**$_T$	+ heulen, jammern, klagen, lamentieren
	[$_A$ Ich] könnt' so **heulen**, [$_T$ daß ich so'n Mauerblümchen bin] ...
	[$_A$ Trainer] **jammern**, [$_T$ daß ihnen ... nur eine Notelf zur Verfügung stünde] ...
	... wenn [$_A$ alle Welt] heute **jammert**, [$_T$ daß es ... an Risikokapital mangelt].
	[$_A$ Die Geschäftsleute] **klagen**, [$_T$ daß die Parkgebühren ... bei 4 DM ... liegen] ...
	[$_A$ Die Zeitschrift "Ogonjok"] **klagte**, [$_T$ daß Politiker ... Zeit für Jagden finden] ...
	[$_A$ Sein Kollege Ewald Lienen] **lamentierte**, [$_T$ daß sein Team ...].
	[$_A$ Ich] will gar nicht **lamentieren**, [$_T$ daß Euer Reichtum ... erworben wurde] ...
n$_A$ **s-w**$_T$	+ jammern, klagen, lamentieren / ¬ heulen
	[$_A$ R.] **jammert**, [$_T$ wie schwierig es doch in Hamburg sei, eine Wohnung zu finden].
	Endlich konnten [$_A$ sie] [$_L$ jemandem] **klagen**, [$_T$ wie schlimm es .... steht] ...
	... zu **lamentieren**, [$_T$ wie sehr er sich um den Pokaltriumph gebracht sähe] ...

Class 20: *Communication*

<u>Verbs</u>: *kommunizieren, korrespondieren, reden, sprechen, verhandeln*

<u>Ambiguity</u>: *korrespondieren* has a sense of 'to correlate'.

<u>Scene</u>: Two or more parties communicate with each other. The parties might be expressed [$_{C_{1 \ldots n}}$ individually] or as [$_C$ a group summarising the individual parties]. The communication deals with [$_T$ a theme] and might take place over [$_M$ a means].

<u>Frame Roles</u>: $C_{(1 \ldots n)}$(ommunicator), T(heme)

<u>Modification Roles</u>: M(eans)

<u>Levin class</u>: 36.1 / 37.5 *(Verbs of Social Interaction → Correspond Verbs / Verbs of Communication → Talk Verbs)*

<u>Schumacher class</u>: 6.3 *(Verben des sprachlichen Ausdrucks → Verben des Diskutierens)*
The verbs *reden, sprechen, verhandeln* are mentioned, but not explicitly discussed in the classification.

Frame	Participating Verbs & Corpus Examples
n$_{C_1}$	+ kommunizieren, reden / +$_{adv}$ korrespondieren, sprechen, verhandeln
	... daß [$_{C_1}$ das Bewußtsein] nicht **kommunizieren** kann ...
	[$_{C_1}$ Therese] **kommunizierte** täglich ...
	[$_{C_1}$ Er] **korrespondierte** in mehreren Sprachen ...
	Auch [$_{C_1}$ sie] **reden** viel.
	[$_{C_1}$ Man] redet und **redet** und denkt und denkt.
	[$_{C_1}$ Heinz] ... **spricht** [$_T$ über Beziehungen zwischen Kybernetik und Erkenntnistheorie] ...
	Eule ist kein Typ, [$_{C_1}$ der] viel **spricht** ...
	In der Sache hat [$_{C_1}$ er] hart **verhandelt**.
	[$_{C_1}$ Ich] werde niemals [$_T$ über die Schließung eines Reaktors] **verhandeln** ...
n$_C$	+ kommunizieren, verhandeln / +$_{adv}$ korrespondieren, reden, sprechen
	[$_C$ Diese] können wiederum untereinander **kommunizieren** ...
	Beim Überspielen **kommunizieren** [$_C$ die zwei Komponenten] [$_M$ über das Bus-System] ...
	[$_C$ Wir] haben die ganze Zeit [$_M$ über Radio] **kommuniziert** ...
	[$_C$ Die Händler] **kommunizieren** untereinander [$_M$ über Computer und Telefon].
	Stets **korrespondierten** in seinen Arbeiten [$_C$ funktionelle Graphik und ... Fotoarbeiten] ...
	... und [$_C$ etwa 65 Prozent] **korrespondieren** regelmäßig [$_M$ mit E-mail].
	[$_C$ Metalltarifparteien] wollen [$_T$ über Vorruhestand] **reden**.
	[$_C$ Die Menschen] **reden** [$_T$ über ganz simple Dinge] ...
	[$_C$ CDU und SPD] **sprechen** [$_T$ über Energiekonsens] ...
	[$_C$ Bern und Warschau] **sprechen** [$_T$ über Vermögen von Nazi-Opfern] ...
	Vier Wochen **verhandelten** [$_C$ die Emmissäre] ... [$_T$ über die Stadt].
	[$_C$ Man] **verhandelt** noch.
	Gegenwärtig **verhandelt** ... [$_C$ eine gemischte Kommission] [$_T$ über die Durchsetzung] ...
	[$_C$ Vertreter beider Länder] **verhandeln** seit Januar [$_T$ über einen neuen Kredit].
	Seit 1995 **verhandeln** [$_C$ die EU und Südafrika] [$_T$ über ein Handelsabkommen] ...

Frame	Participating Verbs & Corpus Examples
n$_C$ [P]	+ kommunizieren, reden, sprechen, verhandeln / ¬ korrespondieren
	... daß wesentlich mehr **kommuniziert** wird ...
	... wo **kommuniziert** wird ...
	Allenfalls [$_T$ über das Gefahrenpotential solcher Techniken] könne **geredet** werden.
	Da wurde immer nur **geredet** und geredet, und nichts tat sich.
	[$_T$ Darüber] wird aber nicht **gesprochen** ...
	Bisher sei meist nur [$_T$ über den Schaden] **gesprochen** worden ...
	Nun wird [$_T$ über eine Verkürzung der Sperre] **verhandelt**.
	Möglicherweise wird an diesem Mittwoch **verhandelt**.
n$_C$ **a**$_T$	+ kommunizieren, verhandeln / ¬ korrespondieren, reden, sprechen
	... weil [$_C$ sie] [$_T$ das Konzept] nicht ausreichend **kommunizieren**.
	... auf elektronischem Wege [$_T$ Versicherungen] ... zu **kommunizieren**.
	... [$_C$ der] [$_T$ sie] annimmt und **verhandelt**.
	Wenn [$_C$ Israel] [$_T$ alles] neu **verhandeln** wolle ...
n$_C$ **a**$_T$ [P]	+ kommunizieren, verhandeln / ¬ korrespondieren, reden, sprechen
	... [$_T$ was] **kommuniziert** werden kann ...
	... [$_T$ die] ... auch nach außen **kommuniziert** waren ...
	... an dem [$_T$ eine bessere Zukunft] **verhandelt** werden sollte ...
	... weniger spektakuläre Fälle, [$_T$ die] hier **verhandelt** werden.
n$_{C_1}$ **p**$_{C_2}$:Dat.mit	+ kommunizieren, korrespondieren, reden, sprechen, verhandeln
	[$_{C_1}$ Der ... Wohnbau] **kommuniziert** ... [$_{C_2}$ mit den vier Außenflügeln].
	[$_{C_1}$ Sie] **kommunizieren** nur [$_M$ über das Vorzimmer] [$_{C_2}$ mit den Beschäftigten] ...
	Die [$_{C_1}$ Cut Box] **korrespondiert** [$_M$ über Datenleitungen] [$_{C_2}$ mit den Zuspielern] ...
	[$_{C_1}$ Wer] selbst [$_{C_2}$ mit Kunden] **korrespondiert** ...
	[$_{C_2}$ Mit Berlinern] kann [$_{C_1}$ man] [$_T$ über] alles **reden** ...
	[$_{C_1}$ Die Regisseurin] **redet** [$_{C_2}$ mit den Leuten] ...
	Darüber **spreche** [$_{C_1}$ ich] [$_{C_2}$ mit dem Schulleiter].
	[$_{C_1}$ Man] **spreche** [$_{C_2}$ mit der spanischen Gesellschaft Iberia] ...
	... daß [$_{C_1}$ wir] nie [$_{C_2}$ mit Frankfurt] [$_T$ wegen Gaudino] **verhandelt** haben.
	... [$_{C_1}$ er] habe [$_{C_2}$ mit Netanjahu] zu **verhandeln**.
	Derzeit **verhandeln** [$_{C_1}$ wir] [$_{C_2}$ mit der Landesregierung] [$_T$ über Zuschüsse] ...
n$_{C_1}$ **p**$_{C_2}$:Dat.mit [P]	+ reden, sprechen, verhandeln / ¬ kommunizieren, korrespondieren
	... daß [$_{C_2}$ mit den "Bullen"] nicht **geredet** werden durfte.
	... daß erst dann [$_{C_2}$ mit der IRA] **geredet** werden könne ...
	... solle ... [$_{C_2}$ mit den Tarifpartnern] [$_T$ über den Krankenstand] **gesprochen** werden ...
	... dann wird [$_{C_2}$ mit diesem Priester] **gesprochen** ...
	... soll [$_T$ wegen der Verpflichtung ...] [$_{C_2}$ mit den Nordbadenern] **verhandelt** werden.
	Als im Sommer ... [$_{C_2}$ mit Moskau] [$_T$ über den Frieden] **verhandelt** wurde ...
n$_C$ **p**:Dat.mit	+ kommunizieren, korrespondieren, reden, sprechen, verhandeln
	[$_C$ Die Völker] **kommunizierten** miteinander ...
	[$_C$ Die drei Musiker] **kommunizieren** in einer Intensität miteinander ...
	... [$_C$ die] miteinander zu **korrespondieren** scheinen.
	[$_C$ Sie] **reden** vielleicht - ohne es zu wissen - miteinander.
	Inzwischen hat [$_C$ man] miteinander **geredet** ...
	[$_C$ Man] habe in einer "freundlichen" Atmosphäre miteinander **gesprochen** ...
	[$_C$ Man] **spreche** täglich miteinander und treffe Entscheidungen gemeinsam.
	... auf der Grundlage ... weiter miteinander zu **verhandeln**.
	... hatten [$_C$ Fujimori und sein bolivianischer Kollege] miteinander **verhandelt** ...

Class 21: *Statement* → *Announcement*

Verbs: *ankündigen, bekanntgeben, eröffnen, verkünden*

Ambiguity: *eröffnen* has a sense of 'to open, to establish' (e.g. a shop), which is related with its sense in class 21, but does not refer to an act of speaking.

Scene: [$_A$ Somebody or a source of announcement] announces [$_T$ something]. There might be [$_L$ a receiver] of the announcement.

Frame Roles: A(rticulator), T(heme)

Modification Roles: L(istener)

Levin class: 37.7 *(Verbs of Communication → Say Verbs)*

Schumacher class: 6.1 *(Verben des sprachlichen Ausdrucks → Verben des Mitteilens)*
Only *bekanntgeben* is classified.

Frame	Participating Verbs & Corpus Examples
n$_A$ **a**$_T$	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	[$_A$ Der frühere Emir] hatte für den Fall seiner Rückkehr [$_T$ Reformen] **angekündigt** ...
	... [$_A$ Bonino] hat ... [$_T$ Vertragsverletzungsverfahren] ... **angekündigt** ...
	... [$_A$ die] [$_T$ McVeighs Verteidigung] **angekündigt** hat.
	Kürzlich hatte [$_A$ Jelzin] [$_T$ Unterstützung für das Verbot] **angekündigt** ...
	[$_A$ Mexiko] hat schon [$_T$ die Emission weiterer Papiere] **angekündigt** ...
	[$_A$ Brie] will [$_T$ seine eigene Entscheidung] Anfang kommender Woche **bekanntgeben** ...
	[$_A$ Er] sei froh, im Namen beider Führer [$_T$ den Abschluß] ... **bekanntgeben** zu können.
	... als [$_A$ die Rheinländer] [$_T$ ihre Expansionspläne] **bekanntgegeben** hatten ...
	[$_A$ Er] hatte im Dezember [$_T$ die Trennung von seiner Frau Mette] **bekanntgegeben**.
	[$_A$ Jaruzelski] hat bereits [$_T$ das Kriegsrecht] **verkündet** ...
	[$_A$ Ein neues liberales Programm] **verkündete** [$_T$ die therapeutischen Möglichkeiten] ...
	[$_A$ Es] desinformiert nicht und **verkündet** [$_T$ keine Heilslehren].
n$_A$ **a**$_T$ [P]	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	[$_A$ Mit dem Katalog] **angekündigt** wurden [$_T$ Passagierschiffe des Konstrukteurs]:
	... im Verteidigungsausschuß seien [$_T$ nicht konkrete Kürzungen] ... **angekündigt** worden.
	[$_T$ Erweiterungen] sind **angekündigt**.
	[$_T$ Sie] war [$_A$ von ihm] 1993 ... bereits für frühere Jahre **angekündigt** worden.
	[$_T$ Dieser Tausch] war [$_A$ von Vogts] **angekündigt** worden.
	[$_T$ Finanzielle Einzelheiten] wurden nicht **bekanntgegeben**.
	[$_T$ Die Sieger] wurden **bekanntgegeben** ...
	[$_T$ Dies] wurde am Freitag **bekanntgegeben**.
	[$_T$ Der Preis für das Aktienpaket] ist nicht **bekanntgegeben** worden.
	[$_T$ Der Name der Supermarktkette] werde zunächst nicht **bekanntgegeben**.
	... unter Berufung auf die Verfassung werden dann [$_T$ neue ... Maximen] **verkündet** ...
	... noch ehe [$_T$ alle Wahlergebnisse] **verkündet** waren.
	Doch nach einer Regelung, [$_T$ die] vor wenigen Jahren **verkündet** worden war ...
	[$_T$ Das Gesamtergebnis für 1996] soll erst im März **verkündet** werden.

Frame	Participating Verbs & Corpus Examples
n$_A$ **a**$_T$ **d**$_L$	+ ankündigen, bekanntgeben, eröffnen, verkünden
	... daß [$_A$ die Enthusiasten] [$_L$ der Zunft] [$_T$ Denkmaschinen] **ankündigten** ... [$_A$ Er] hat [$_L$ mir] [$_T$ das Tor] ja vor dem Spiel schon **angekündigt**. ... [$_T$ sie] [$_L$ allen ihren Mitgliedern] **bekanntgeben** und diese darauf verpflichten. [$_T$ Was] [$_A$ der] [$_L$ ihm] zu **eröffnen** hatte ... ... Botschaft, [$_T$ die] [$_L$ ihm] [$_A$ Vorstandsvertreter Gerd Markus] **eröffnet** hat ... ... und **verkündet** [$_L$ ihm] [$_T$ die Gesetze einer neuen Religion]. [$_A$ Er] muß hinaus, [$_L$ uns] [$_T$ den Blues] **verkünden**.
n$_A$ **a**$_T$ **d**$_L$ [P]	+ ankündigen, eröffnen, verkünden / ¬ bekanntgeben
	Soll [$_L$ Eltern] [$_T$ ein behindertes Kind] **angekündigt** werden ... [$_T$ Dies] war [$_L$ den Aktionären] auf der Hauptversammlung ... **angekündigt** worden. In einem Gespräch wurden [$_L$ ihm] [$_T$ seine "Wahlmöglichkeiten"] **eröffnet**: Gestern vormittag wurden [$_L$ ihm] [$_T$ zwei Haftbefehle] **verkündet**.
n$_A$ **i**$_T$	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	[$_A$ China] hatte **angekündigt**, [$_T$ ... insgesamt 240 neue Flugzeuge zu kaufen]. [$_A$ Levy] hatte **angekündigt**, [$_T$ sein Amt aufzugeben] ... Darum hat [$_A$ er] nun **bekanntgegeben**, [$_T$ seinen angestammten Platz ... zu verlassen]. [$_A$ Das Unternehmen] ... hat **bekanntgegeben**, [$_T$ einen Produktionsstandort] ... ... hatte bereits [$_A$ Belgien] **verkündet**, [$_T$ 1996 ein Fünftel ... verkauft zu haben] ... [$_A$ Piëch] hat oft **verkündet**, [$_T$ Volkswagen zum besten Automobilhersteller ... zu machen].
n$_A$ **d**$_L$ **i**$_T$	+ ankündigen, eröffnen / ¬ bekanntgeben, verkünden
	[$_A$ China] hat [$_L$ dem ... Olympischen Komitee] **angekündigt**, [$_T$ sich ... zu bewerben]. ... und **eröffnete** [$_L$ ihm], [$_T$ sich unter den Schutz ... stellen ... zu wollen] ...
n$_A$ **s-2**$_T$	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	Auch [$_A$ andere Häuser] haben **angekündigt**, [$_T$ sie würden ... auf Bachmann setzen]. [$_A$ Bayrou] hat **angekündigt**, [$_T$ er werde ... "Sammlungsbewegung" gründen]. ... [$_A$ die] ... **bekanntgab**, [$_T$ sie hätte sich ... ein besseres Geschäft ... erwartet]. [$_A$ Die Organisation] hat ... **bekanntgegeben**, [$_T$ die Arbeit ...] [$_A$ Ein erpreßter Vater] ... **verkündet**, [$_T$ er werde das geforderte Lösegeld nicht zahlen] ... [$_A$ Preuß] **verkündete** stolz, [$_T$ ein wichtiger Sponsor habe ... seinen Vertrag verlängert] ... [$_T$ Ich bleibe beim VfB], **verkündete** [$_A$ Elber]. ... **verkündet** [$_A$ er] grinsend, [$_T$ sein Name sei ja noch nicht einmal geeignet] ... [$_A$ Der Anstaltsleiter] **verkündet**, [$_T$ ihm seien die Hände gebunden].
n$_A$ **s-2**$_T$ [P]	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	Zugleich wurde **angekündigt**, [$_T$ das Land entziehe dem Unternehmen seine Unterstützung]. Es reiche nicht, jetzt **anzukündigen**, [$_T$ man wolle das Spektrum erweitern]. In Thüringen wurde werbewirksam **verkündet**, [$_T$ man werde Beobachtungen ... aufgreifen]. [$_T$ "Der DJ ist der neue Star"], wird auf dem Festival **verkündet** ...
n$_A$ **d**$_L$ **s-2**$_T$	+ ankündigen, eröffnen, verkünden / ¬ bekanntgeben
	Dagegen hatte [$_A$ Thyssen] [$_L$ seinen Mitarbeitern] **angekündigt**, [$_T$ man werde ...] ... [$_A$ die Vernehmungsbeamten] [$_L$ ihm] **ankündigten**, [$_T$ er werde Zeugen] ... [$_T$ Die Bauarbeiter haben ... abgerissen], **eröffnet** [$_L$ ihm] [$_A$ der freundliche Mitmieter]. ... [$_A$ Ann] **eröffnet** [$_L$ ihm], [$_T$ sie wolle vom Dealen auf Naturkosmetik umsatteln]. ... um [$_L$ ihm] dort zu **verkünden**: ... [$_T$ Klestil entschuldigt sich] ... [$_A$ Dieser] **verkündete** [$_L$ ihm], [$_T$ er werde doch freigelassen] ...
n$_A$ **d**$_L$ **s-2**$_T$ [P]	+ ankündigen, eröffnen, verkünden / ¬ bekanntgeben
	[$_L$ Ihm] selbst wurde [$_A$ vom Chefarzt Wolf] **eröffnet**, [$_T$ er werde ... gesund werden] ... ... [$_T$ Klaus Klenke werde Vox ... verlassen], wurde [$_L$ den Mitarbeitern] **verkündet** ...

Frame	Participating Verbs & Corpus Examples
n$_A$ **s-dass**$_T$	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	[$_A$ Paris] hatte **angekündigt**, [$_T$ daß es diese Frage ... neu entscheiden wolle. [$_A$ Die Antikenverwaltung] hat **angekündigt**, [$_T$ daß sie vorhat] ... Wenn [$_A$ Mullen] **bekanntgibt**, [$_T$ daß er das Geld nicht zahlen wird] ... So habe [$_A$ die Notenbank] **bekanntgegeben**, [$_T$ daß sie das Wachstum ...] [$_A$ Die Rebellen] haben bereits **bekanntgegeben**, [$_T$ daß sie ihren Kampf... fortsetzen] ... [$_A$ Blüm] **verkündet** stolz, [$_T$ daß die Partei ... aufzuweisen habe] ... [$_A$ Ein LKA-Beamter] **verkündete**, [$_T$ daß Chemikalien ... gefunden worden seien] ...
n$_A$ **s-dass**$_T$ [P]	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	Es wurde ... **angekündigt**, [$_T$ daß mit einem Umsatzrückgang zu rechnen sei]. [$_T$ Daß man sich ... wehren werde], wird [$_A$ im Abschnitt] ... **angekündigt**. ... auf denen **angekündigt** wurde, [$_T$ daß die SPD ... nacheifern werde]. Gleichzeitig wurde **bekanntgegeben**, [$_T$ daß ein ... Beerdigungskomitee ...] Unlängst wurde offiziell **bekanntgegeben**, [$_T$ daß man auf einen Fingerzeig ... warte]. Am Mittwoch wurde ... **verkündet**, [$_T$ daß die Arbeitenden ... zur Arbeit ausrückten]. ... wurde **verkündet**, [$_T$ daß ... der Architektenwettbewerb ausgelobt werde].
n$_A$ **d**$_L$ **s-dass**$_T$	+ ankündigen, eröffnen, verkünden / ¬ bekanntgeben
	Als [$_A$ ich] [$_L$ ihm] **ankündigte**, [$_T$ daß ich Urlaub brauchte] ... ... hätten [$_L$ ihm] [$_A$ gleich vier Bauern] **angekündigt**, [$_T$ daß sie resignierten] ... ... habe [$_L$ ihm] [$_A$ Drach] **eröffnet**, [$_T$ daß er jemanden entführen wolle]. ... als ... [$_A$ sein Bruder] [$_L$ ihm] **eröffnete**, [$_T$ daß er in den Westen gehen werde]. ... [$_A$ der] [$_L$ mir] ... **eröffnete**, [$_T$ daß ich eine bessere Stellung erhalten würde] ... ... wenn [$_A$ sie] ... [$_L$ dem ganzen Publikum] **verkündet**, [$_T$ daß Hermann ...] Morgen wird [$_A$ Heyme] [$_L$ der Presse] **verkünden**, [$_T$ daß ... stattfinden könne].
n$_A$ **d**$_L$ **s-dass**$_T$ [P]	+ ankündigen, eröffnen, verkünden / ¬ bekanntgeben
	[$_L$ Der Exzellenz] war **angekündigt** worden, [$_T$ daß der Bildhauer ...]. [$_L$ Den Rentnern] wird **angekündigt**, [$_T$ daß ihre Bezüge ... besteuert werden] ... [$_L$ Ausländischen KorrespondentInnen] wurde **eröffnet**, [$_T$ daß sie ...] Dort wurde [$_L$ ihr] **eröffnet**, [$_T$ daß sie ... vom Dienst suspendiert sei] ...
n$_A$ **s-w**$_T$	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	[$_A$ Man] dürfe nicht **ankündigen**, [$_T$ was man nachher nicht einlösen könne]. [$_A$ Kohl] solle **ankündigen**, [$_T$ wann er zurücktreten werde]. [$_A$ Ballesteros] will ... **bekanntgeben**, [$_T$ wer die beiden "Wildcards" erhält]. ... [$_A$ der] im Dorf gerade **bekanntgab**, [$_T$ wann die ... Trauben ...] ... hat [$_A$ er] doch gerade erst ... **verkündet**, [$_T$ wie klar ihm alles ist]. ... indem [$_A$ sie] bloß **verkünden**, [$_T$ wer wann wo zur Zeit im Geschäft ist].
n$_A$ **s-w**$_T$ [P]	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	Bis jetzt ist aber nur **angekündigt**, [$_T$ wann man ihn nicht tun werde]: [$_T$ Was weiter damit passiert], wurde nicht **bekanntgegeben**. ... wurde ... **bekanntgegeben**, [$_T$ wann das neue System eingeführt werden soll].
n$_A$ **d**$_L$ **s-w**$_T$	+ ankündigen, eröffnen, verkünden / ¬ bekanntgeben
	... [$_A$ dessen Teilnehmer] [$_L$ sich] ... **eröffnen**, [$_T$ was sie wirklich glauben].
n$_A$ **s-ob**$_T$	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	... hat ... **angekündigt**, [$_T$ ob und welche politische Rolle er in Nigeria anstrebt]. [$_A$ Die UUP] will ... **bekanntgeben**, [$_T$ ob sie ... teilnehmen wird]. ... wollte [$_A$ das Bundesverfassungsgericht] **verkünden**, [$_T$ ob die Reform ...].
n$_A$ **s-ob**$_T$ [P]	+ ankündigen, bekanntgeben, verkünden / ¬ eröffnen
	... [$_A$ es] werde auch nicht **bekanntgegeben**, [$_T$ ob eine Versicherung ... erteilt hat].

Class 22: *Statement* → *Constitution*

Verbs: *anordnen, bestimmen, festlegen*

Ambiguity: *anordnen* has a sense of 'to arrange, to align'.

Scene: [$_C$ Somebody or a source of constitution] constitutes [$_T$ something]. There might be [$_P$ a patient] as object of the constitution.

Frame Roles: C(onstitutor), T(heme)

Modification Roles: P(atient)

Levin class: -

Schumacher class: 5 *(Verben des Handlungsspielraums, Kausative Handlungsverben)*
Schumacher's class 5 semantically contains my classes 22 and 23. Only *anordnen* is classified, together with verbs of allowing and order.

Frame	Participating Verbs & Corpus Examples
n$_C$ **a**$_T$	+ anordnen, bestimmen, festlegen
	[$_C$ Evita] selbst soll kurz vor ihrem Tod [$_T$ eine postmortale Maniküre] **angeordnet** ... [$_C$ Das Gericht] kann [$_T$ die Rückgabe von Vermögenswerten] **anordnen** ... ... [$_C$ der Staatssicherheitsgerichtshof] ... hat [$_T$ seine Verhaftung] **angeordnet**. ... [$_C$ Toesca und Wölfflin] **bestimmten** [$_T$ den Weg des Studenten] ... ... ein Protein entdeckt worden, [$_C$ das] [$_T$ die Geschwindigkeit] **bestimmt** ... [$_C$ Der Abstand zwischen den Messern] **bestimmt** [$_T$ die Länge] ... ... [$_T$ die Mülltonnengröße] selbst zu **bestimmen**. [$_C$ Diese Scham] **bestimmt** [$_T$ das Selbstgefühl des Jungen] ... Bisher hatten [$_C$ zwei staatliche Prüfer] [$_T$ das Zugangsalter] **festgelegt**. Für diesen Fall ... haben [$_C$ die Career Center] [$_T$ ganz klare Spielregeln] **festgelegt** ...
n$_C$ **a**$_T$ [P]	+ anordnen, bestimmen, festlegen
	Jetzt kann [$_T$ nichts] mehr [$_C$ von oben] **angeordnet** und durchgesetzt werden . [$_P$ Für weitere Krankenhäuser] wurden ... [$_T$ Sicherheitsvorkehrungen] **angeordnet**. ... wurde zunächst [$_T$ der Maßregelvollzug] **angeordnet** ... ... weshalb für diesen Freitag im ganzen Land [$_T$ Staatstrauer] **angeordnet** wurde. Auch [$_T$ die Grenzen der beiden Möchtegernstaaten] werden neu **bestimmt**: [$_T$ Präsident Bushs Politik] wird dagegen [$_C$ von dem Wunsch] **bestimmt** ... In jedem Jahr sollten [$_T$ Zuwanderungsquoten] **festgelegt** werden ... Darin sollten [$_T$ Begleitmaßnahmen wie Sprachunterricht] **festgelegt** werden.
n$_C$ **a**$_P$ **p**$_T$:Akk.auf	+ festlegen / ¬ anordnen, bestimmen
	... aber [$_C$ er] hat [$_P$ seine Partei] nicht [$_T$ auf sein Modell] **festlegen** können. ... Entscheidung, [$_P$ die Tour] [$_T$ auf den kommenden Samstag] **festzulegen** ...
n$_C$ **a**$_P$ **p**$_T$:Akk.auf [P]	+ festlegen / ¬ anordnen, bestimmen
	Da [$_P$ der Ausgabekurs der Aktien] [$_T$ auf 130,38 DM] **festgelegt** wurde ... [$_P$ Die Haltung] ... scheint weniger [$_T$ auf die ... Variante] **festgelegt** zu sein ... ... wurde [$_P$ sie] endgültig [$_T$ auf das Genre des eleganten Melodrams] **festgelegt** ...

Frame	Participating Verbs & Corpus Examples
$\mathbf{n}_C$ $\mathbf{r}_P$	+ festlegen / ¬ anordnen, bestimmen
	Für all jene, [$_C$ die] [$_P$ sich] nicht so eindeutig **festlegen** ...
	[$_C$ Sein geistiger Vater] hat [$_P$ sich] da nicht **festlegen** wollen.
$\mathbf{n}_C$ $\mathbf{r}_P$ $\mathbf{p}_T$:Akk.auf	+ festlegen / ¬ anordnen, bestimmen
	[$_T$ Auf das Wann] wolle [$_C$ er] [$_P$ sich] nicht **festlegen** ...
	Deshalb konnte [$_C$ sie] [$_P$ sich] nie [$_T$ auf ein Thema] dauerhaft **festlegen**.
$\mathbf{n}_C$ $\mathbf{i}_T$	+ anordnen, bestimmen / ¬ festlegen
	[$_C$ Außenminister Kinkel] hat **angeordnet**, [$_T$ den Fall ... zu verfolgen].
	... hatte [$_C$ Augustus] selbst **angeordnet**, [$_T$ die Mysterien ... zu feiern].
	[$_C$ Er] **bestimmt**, [$_T$ die Satzung so zu ändern, daß ... Stellvertreter zur Seite stehen].
$\mathbf{n}_C$ $\mathbf{i}_T$ [P]	+ anordnen / ¬ bestimmen, festlegen
	... oftmals wurde **angeordnet**, [$_T$ Briefe oder Eingaben nicht zu beantworten] ...
	In der Regel wurde auf Divisionsebene **angeordnet**, [$_T$ keinen Widerstand zu leisten].
$\mathbf{n}_C$ $\mathbf{r}_P$ $\mathbf{i}_T$	+ festlegen / ¬ anordnen, bestimmen
	[$_C$ Die Bundesregierung]hat [$_P$ sich] zwar **festgelegt**, [$_T$ ... ein Gebäude ... zu mieten] ...
$\mathbf{n}_C$ $\mathbf{s\text{-}dass}_T$	+ anordnen, bestimmen, festlegen
	Als [$_C$ sie] **anordneten**, [$_T$ daß Guevaras Leiche gewaschen werden sollte] ...
	[$_C$ Die Ärzte] hätten **angeordnet**, [$_T$ daß der Präsident zu Hause bleiben solle].
	In Paragraph 56 **bestimmt** [$_C$ das Gesetz], [$_T$ daß ... Bewährung stattfindet] ...
	[$_C$ Die Verfassung] **bestimme**, [$_T$ daß das alte Parlament im Amt bleibe] ...
	In einer Stellungnahme hat [$_C$ die Kommission] **festgelegt**, [$_T$ daß Menschen] ...
	1997 hatte [$_C$ die Stadt] **festgelegt**, [$_T$ daß die Geschäfte ... geöffnet bleiben dürften].
$\mathbf{n}_C$ $\mathbf{s\text{-}dass}_T$ [P]	+ anordnen, bestimmen, festlegen
	Gleichzeitig wurde jedoch **angeordnet**, [$_T$ daß ... eine Bescheinigung vorzulegen ist] ...
	Im Irak wurde ... **angeordnet**, [$_T$ daß Hausbesitzer Luftschutzkeller anlegen müssen].
	Schließlich wurde mit einer Stimme Mehrheit **bestimmt**, [$_T$ daß sie noch warten sollen].
	Im Einigungsvertrag wurde **bestimmt**, [$_T$ daß ... die Aufenthaltsbewilligung endet].
	Damals war **festgelegt** worden, [$_T$ daß diese Taten ... verjähren].
	Per Verordnung wird **festgelegt**, [$_T$ daß 500 Kalorien pro Tag ausreichend sind] ...
$\mathbf{n}_C$ $\mathbf{s\text{-}w}_T$	+ anordnen, bestimmen, festlegen
	[$_C$ Ein Computer und nicht Menschen] **bestimmen**, [$_T$ wann ... geschlossen werden].
	Doch das Gen, [$_C$ das] bei Mäuseembryonen **bestimmt**, [$_T$ wo sich ... entwickeln] ...
	... Kapazitätsverordnung, [$_C$ die] **festlege**, [$_T$ wer ... auszubilden habe].
	... [$_C$ die] exakt **festlegen**, [$_T$ wie die musikalischen Daten auf den Diskus gelangen].
$\mathbf{n}_C$ $\mathbf{s\text{-}w}_T$ [P]	+ anordnen, bestimmen, festlegen
	Es wurde [$_C$ von oben] **angeordnet**, [$_T$ wie die Tarife auszusehen haben].
	[$_C$ In dem Vertrag] ... wird dann **bestimmt**, [$_T$ wann ... der Bauer ... zu mähen hat].
	Mitte dieses Jahres wird ... **festgelegt**, [$_T$ welche Länder ... importieren dürfen].
$\mathbf{n}_C$ $\mathbf{r}_P$ $\mathbf{s\text{-}w}_T$	+ festlegen / ¬ anordnen, bestimmen
	[$_C$ Schmidt] wollte [$_P$ sich] nicht **festlegen**, [$_T$ wann das Unternehmen] ...
	... daß [$_P$ sich] [$_C$ die Regierung] nicht **festlege**, [$_T$ wie die Wirtschaftsförderung] ...
$\mathbf{n}_C$ $\mathbf{s\text{-}ob}_T$	+ bestimmen / ¬ anordnen, festlegen
	[$_C$ Vermieter] könnten dann selbst **bestimmen**, [$_T$ ob sie eine Wohnung ... wollten].
	[$_C$ Die Frauen] **bestimmen**, [$_T$ ob der nächste Kongreß ... stattfindet oder nicht].
$\mathbf{n}_C$ $\mathbf{r}_P$ $\mathbf{s\text{-}ob}_T$	+ festlegen / ¬ anordnen, bestimmen
	Ebenso will [$_C$ sie] [$_P$ sich] nicht **festlegen**, [$_T$ ob die Gründe] ...
	[$_C$ Die Contra-Chefs] konnten [$_P$ sich] nicht **festlegen**, [$_T$ ob die Attacke] ...

Class 23: *Statement* → *Promise*

Verbs: *versichern, versprechen, zusagen*

Ambiguity: *versichern* has a commercial sense of 'to insure'. *zusagen* has a sense of 'to appeal'.

Scene: [$_T$ A promise] is given by [$_P$ somebody or something who or which is the position to make a promise]. There might be a [$_R$ receiver] of the promise.

Frame Roles: P(romiser), T(heme)

Modification Roles: R(eceiver)

Levin class: 13.3 *(Verbs of Change of Possession → Verbs of Future Having)*

Schumacher class: 5 *(Verben des Handlungsspielraums, Kausative Handlungsverben)*
Schumacher's class 5 contains my classes 22 and 23. *versprechen* and *zusagen* are mentioned as subgroup of this class, but not characterised.

Frame	Participating Verbs & Corpus Examples
n$_P$	+ zusagen / ¬ versichern, versprechen
	... doch [$_P$ acht Länder] hätten **zugesagt** ...
	Ganz spontan haben [$_P$ alle] **zugesagt** ...
n$_P$ **a**$_T$	+ versichern, versprechen, zusagen
	[$_T$ Das] **versicherten** [$_P$ die beiden Politiker] vor ihrem Treffen ...
	In der "Erklärung von Hanoi" **versicherten** [$_P$ die Teilnehmer] zudem [$_T$ ihren Willen] ...
	[$_T$ Genußradeln am Bodensee] **verspricht** [$_P$ der Autor] ...
	[$_P$ Der Markt] indessen **verspricht** [$_T$ Freiheit] ...
	[$_P$ Die New World] ... hat bereits [$_T$ ihre Zustimmung zu der Transaktion] **zugesagt**.
	[$_T$ Mehr als die 10,5 Millionen Mark] ... wollte [$_P$ Braun] nicht **zusagen**.
n$_P$ **a**$_T$ [P]	+ versprechen, zusagen / ¬ versichern
	[$_R$ Für die Spandauer] ... wird seit Jahren [$_T$ der Baubeginn] lediglich **versprochen**.
	Wo [$_T$ das besondere Naturerlebnis] **versprochen** wird ...
	... daß in die neuen Bundesländer [$_T$ mehr Kredite] **zugesagt** worden sind.
	... besonders wenn [$_T$ Charterraten] **zugesagt** worden seien ...
n$_P$ **a**$_T$ **d**$_R$	+ versichern, versprechen, zusagen
	[$_T$ Das] **versicherte** [$_P$ Samaranch] [$_R$ der Präsidentin des Weltrates] ...
	[$_P$ Die Marketingspezialisten] **versicherten** [$_R$ der Geschichte] [$_T$ ihren Respekt] ...
	... **verspricht** [$_P$ der frühere Nationalspieler] [$_T$ seinem Arbeitgeber] ... [$_T$ Loyalität] ...
	[$_R$ Den Kunden] **versprechen** [$_P$ die Anbieter] nicht nur [$_T$ ein vorgewärmtes Auto] ...
	... [$_T$ die] [$_P$ man] [$_R$ Tschechow] **zugesagt** hatte ...
	[$_P$ Ein reicher Graf in Madrid] hat [$_T$ seine Tochter] ... [$_R$ Don Pinto] **zugesagt**.
n$_P$ **a**$_T$ **d**$_R$ [P]	+ versprechen, zusagen / ¬ versichern
	... [$_R$ denen] 1988 [$_T$ politisches Asyl] **versprochen** wurde.
	... Juristen, [$_R$ denen] unter anderem [$_T$ das Gespräch mit Häftlingen] **versprochen** wird.
	... [$_R$ denen] bereits [$_T$ ABM-Stellen beim Hausbau] **zugesagt** worden waren ...
	[$_R$ Ihm] sei [$_T$ Unterstützung] **zugesagt** worden ...

Frame	Participating Verbs & Corpus Examples
n$_P$ **i**$_T$	+ versichern, versprechen, zusagen
	[$_P$ Maupertuis] **versichert** aber auch, [$_T$ die Grenzen ... nicht zu überschreiten]. Damit **versichert** [$_P$ Spanien], [$_T$ ... auf einem strengen ... Kurs zu bleiben]. [$_P$ Der Donnerstag] **verspricht** [$_T$ ein vielseitiger Börsentag zu werden]. Aller Mystik beraubt ... **verspricht** [$_P$ sie], [$_T$ eine neue Eroberung ... zu werden] ... [$_P$ Die Firma Hotel Docs] **verspricht**, [$_T$ ... einen Mediziner ... zu schicken]. [$_P$ Amex und die Regionalbörsen] haben **zugesagt**, [$_T$ ... den Handel] ... [$_P$ Die Länder] hätten schon **zugesagt**, [$_T$ diesen Weg mitzugehen] ...
n$_P$ **d**$_R$ **i**$_T$	+ versichern, versprechen, zusagen
	[$_P$ Er] ... **versicherte** [$_R$ auch der ... Fronde], [$_T$ niemand brauche sich zu sorgen]. [$_P$ Trautvetter] hat [$_R$ mir] **versichert**, [$_T$ beim ... Verband darauf hinzuweisen] ... [$_P$ Sie] **versprach** [$_R$ dem Senat], [$_T$ ... auf eine innere Reform ... zu dringen] ... [$_P$ Er] **versprach** [$_R$ ihr], [$_T$ Kontakt zu ihrem verstorbenen Mann herzustellen]. ... weil [$_P$ Maier] [$_R$ ihm] angeblich **zugesagt** hatte, [$_T$ im Bundesrat ... zu stimmen] ... [$_P$ Sie] hatten [$_R$ ihr] **zugesagt**, [$_T$ den Aufbau ... mitzufinanzieren] ...
n$_P$ **s-2**$_T$	+ versichern, versprechen / ¬ zusagen
	[$_T$ Diese Investition zahle sich schnell aus], **versichert** [$_P$ Pinchev]. [$_P$ Vizepräsident Al Gore] **versicherte**, [$_T$ das Gespenst der Inflation sei gebannt] ... ... aber [$_P$ er] **versprach**, [$_T$ sein Zug würde ... das Staunen schon wieder lehren]. In einer Erklärung **versprach** [$_P$ er] ..., [$_T$ ... legale Proteste würden ... nicht verboten].
n$_P$ **s-dass**$_T$	+ versichern, versprechen, zusagen
	[$_P$ Topp] **versicherte**, [$_T$ daß die Marke Trix ... ihre Eigenständigkeit erhalte] ... [$_P$ Hoyer] **versicherte**, [$_T$ daß dies nichts ... zu tun habe]. [$_P$ Wolle und sein Freund] **versprachen**, [$_T$ daß seine Freundin ... gelangen würde] ... ... [$_P$ ich] habe **versprochen**, [$_T$ daß die Qualität ... ein zentrales Thema sein würde]. [$_P$ Chefminister] ... habe **zugesagt**, [$_T$ daß die Regierung nichts unternehmen werde] ... ... aber [$_P$ Ford] hat bereits **zugesagt**, [$_T$ daß es ... keine Unterschiede ... geben soll].
n$_P$ **s-dass**$_T$ [P]	+ versichern, versprechen, zusagen
	[$_P$ Von seiten Apples] wird ... **versichert**, [$_T$ daß man mit dem PowerPC Pläne ... habe]. Immer wieder wird deshalb auch **versichert**, [$_T$ daß China ... festhalten werde]. Es wurde feierlich **versprochen**, [$_T$ daß keines der Länder ... behindern wird]. Es wurde **versprochen**, [$_T$ daß er auf die Tagesordnung kommt]. [$_P$ Von den Verbänden] wurde **zugesagt**, [$_T$ daß ... Sicherheitsbestimmungen] ... Jetzt wurde **zugesagt**, [$_T$ daß nach der Antwort auch die Fragen veröffentlicht werden].
n$_P$ **d**$_R$ **s-dass**$_T$	+ versichern, versprechen, zusagen
	[$_P$ Schwarz-Schilling] **versichert** [$_R$ Beckstein], [$_T$ daß er mit ihm ... übereinstimme] ... [$_P$ Ich] kann [$_R$ Ivan Pedroso] **versichern**, [$_T$ daß er ... Olympiasieger wird]. ... bis [$_P$ sein Vater] [$_R$ ihm] **verspricht**, [$_T$ daß sie ... in einen Freizeitpark fahren]. [$_P$ Ich] hatte [$_R$ meinen Kindern] ... **versprochen**, [$_T$ daß ich sie zurückhole] ... [$_P$ Kohl] habe ... [$_R$ Jelzin] **zugesagt**, [$_T$ daß eine Ausdehnung ... beabsichtigt sei]. [$_P$ Der Präsident] hat [$_R$ jeder Familie] **zugesagt**, [$_T$ daß sie künftig ... schicken darf].
n$_P$ **d**$_R$ **s-dass**$_T$ [P]	+ versichern, versprechen, zusagen
	... und [$_R$ ihm] wird **versichert**, [$_T$ daß Fabios Küche traditioneller ... sein könnte] ... [$_P$ Aus dem ... Außenministerium] wurde [$_R$ den Balten] **versichert**, [$_T$ daß die Nato] ... Außerdem wurde [$_R$ ihnen] **versprochen**, [$_T$ daß sie den Wagen bald fahren dürften]. Bei den Fundamentalisten wird [$_R$ den Leuten] **versprochen**, [$_T$ daß sie erlöst werden]. [$_R$ Den Geiselnehmern] sei ... **zugesagt** worden, [$_T$ daß der Autohändler ...]. Gestern wurde [$_R$ ihnen] [$_P$ vom Sozialressort] **zugesagt**, [$_T$ daß sie ... können].

Class 24: *Observation*

Verbs: *bemerken, erkennen, erfahren*$_2$*, feststellen, realisieren, registrieren*

Ambiguity: *bemerken* has a sense of 'to remark', closely related to the verbs in class 21. The verb *erfahren* is also classified in class 17, which refers to the specific body senses for perception, whereas this class refers to realising an observation. *realisieren* has a sense of roughly 'to carry a plan into effect'. *registrieren* has a sense of 'to record'.

Scene: [$_E$ Somebody] experiences [$_T$ an observation] which has an effect on the knowledge state, changing or increasing it.

Frame Roles: E(xperiencer), T(heme)

Levin class: 30.1 *(Verbs of Perception → See Verbs)*
Levin's class 30.1 is the closest pendant to my class 24, but rather captures the verbs' sense of perception, which shows the similarity to class 17.

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_E$ **a**$_T$	+ bemerken, erkennen, erfahren, feststellen, realisieren, registrieren
	[$_T$ Das] **bemerkte** [$_E$ kaum jemand], weshalb der Beifall gering blieb.
	Wahrscheinlich **bemerken** [$_E$ sie] [$_T$ eine allzu ähnliche genetische Ausstattung] ...
	Oftmals **bemerkten** [$_E$ die Westdeutschen] [$_T$ die eigene Häme] nicht einmal ...
	[$_T$ Dies] haben [$_E$ die westlichen Führer] verspätet **erkannt** ...
	[$_E$ Sie] **erkannten** [$_T$ den Buchbinder Pfaff, den Magazineur Bienert], ...
	[$_E$ Jeder Soldat] kann [$_T$ die Realität der Kräfteverhältnisse] **erkennen**.
	[$_T$ Inneres Wachstum und Vergnügen] **erfährt** [$_E$ man] beim Zuschauen ...
	So **erfährt** [$_E$ der Leser] [$_T$ einiges über Rußlands Mafia] ...
	... hatten [$_E$ die Bewohner] schon vor einigen Tagen [$_T$ Gasgeruch] **festgestellt** ...
	... um "hundertprozentig" [$_T$ die Identität] **festzustellen**.
	Zwar habe [$_E$ das Papier] [$_T$ das Nachholpotential] schnell **realisiert** ...
	... [$_T$ die sogenannte leichte Muse] zu **realisieren** als die klassische Kultur.
	[$_E$ Ich] brauche einige Tage, um [$_T$ diesen Erfolg] zu **realisieren** ...
	[$_E$ Die Grundkreditbank] **registriert** im Immobiliengeschäft [$_T$ mehr Interessenten].
	[$_T$ Was] [$_E$ man] **registrierte**, war Galileos einzigartige Fähigkeit ...
n$_E$ **a**$_T$ [P]	+ bemerken, erkennen, feststellen, registrieren / ¬ erfahren, realisieren
	Selbst wenn [$_T$ der Einstich] **bemerkt** wird ...
	[$_T$ Die Fehlfunktion von "Spartan"] war am Freitag **bemerkt** worden ...
	[$_T$ Das Problem] ist längst **erkannt** ...
	Wenn überhaupt [$_T$ etwas] auf den Fotos zu **erkennen** ist ...
	[$_T$ Das Dasein] soll als Illusion **erkannt** werden ...
	[$_T$ Das Ausmaß der tatsächlichen Schäden] kann nicht genau **festgestellt** werden ...
	[$_T$ Die genauen Prioritäten der Regierungsarbeit] würden erst **festgestellt** ...
	[$_T$ Das] ist hierzulande eher am Rande, in Holland aber sehr genau **registriert** worden.
	[$_T$ Das] wird lediglich **registriert**, als Problem ist es längst abgehakt.

Frame	Participating Verbs & Corpus Examples
$\mathbf{n}_E$ $\mathbf{p}_T$:Dat.von	+ erfahren / ¬ bemerken, erkennen, feststellen, realisieren, registrieren
	... als [$_E$ sie] [$_T$ von seinem Tod] **erfährt**. [$_E$ Der 59 Jahre alte Bill Cosby] ... **erfuhr** in New York [$_T$ von der Tragödie].
$\mathbf{n}_E$ **s-dass**$_T$	+ bemerken, erkennen, erfahren, feststellen, realisieren, registrieren
	In der Tat hatte [$_E$ Isabel] **bemerkt**, [$_T$ daß Warburton noch immer in sie verliebt war]. Jetzt erst **bemerkt** [$_E$ er], [$_T$ daß die Pfeife im Regen längst ausgegangen ist]. Dabei **bemerkte** [$_E$ er], [$_T$ daß sein Haus die Sicht noch versperrte] ... ... als [$_E$ sie] **erkannten**, [$_T$ daß die Familie nicht so vermögend war] ... Vielmehr kann [$_E$ man] **erkennen**, [$_T$ daß die Märkte ... zusammenwachsen] ... ... Delphine, [$_E$ die] mittlerweile **erfahren** hat, [$_T$ daß sie de Sades Tochter ist] ... [$_E$ Wir] haben dabei **erfahren**, [$_T$ daß die Japaner viel großherziger ... sind] ... ... sobald [$_E$ das System] **festgestellt** hat, [$_T$ daß daheim niemand "abhebt"]. Vermutlich hat [$_E$ man] inzwischen **festgestellt**, [$_T$ daß das Problem ... ihre Qualität war] ... [$_E$ Man] **realisiert** gar nicht, [$_T$ daß es zu Ende ist]. [$_E$ Er] scheint zu **realisieren**, [$_T$ daß dort der Stärkste dieser Tour fährt]. Voller Stolz haben [$_E$ sie] **registriert**, [$_T$ daß ihr Spiel ... annonciert wird] ... [$_E$ Aufmerksame Beobachter] haben **registriert**, [$_T$ daß der eine ... vor Kraft strotzt] ...
$\mathbf{n}_E$ **s-dass**$_T$ [P]	+ bemerken, erfahren, erkennen, feststellen, realisieren, registrieren
	... wurde voller Bewunderung **bemerkt**, [$_T$ daß ... noch 15 Fans übrig waren]. Dort wurde schnell **bemerkt**, [$_T$ daß ich anders sprach]: ... werde jetzt offenbar **erkannt**, [$_T$ daß das NDR-Angebot den Interessen ... mehr diene] ... ... war zu **erfahren**, [$_T$ daß Kvaerner 150 Millionen Kronen ... zahlen soll] ... In einer Kosten-Nutzen-Rechnung wurde **festgestellt**, [$_T$ daß ... Therapie billiger ist] ... Trotzdem wurde **festgestellt**, [$_T$ daß viele der Greifvögel ... zugrunde gegangen waren]. ... wurde mit Sorge **registriert**, [$_T$ daß sich die Rahmenbedingungen ... ändern würden]. Erleichtert wurde **registriert**, [$_T$ daß es keine gibt].
$\mathbf{n}_E$ **s-ob**$_T$	+ bemerken, erkennen, erfahren, feststellen / ¬ realisieren, registrieren
	... daß [$_E$ die einsteigenden Herrschaften] nicht **bemerken** konnten, [$_T$ ob der Chauffeur] ... [$_E$ Man] wird bald **erkennen**, [$_T$ ob Netanjahu Gesprächserfolge wirklich will]. [$_E$ Ich] wollte **erfahren**, [$_T$ ob ich akzeptiert werde] ... [$_T$ Ob es Überkapazitäten gebe], müsse [$_E$ jedes Land] **feststellen**.
$\mathbf{n}_E$ **s-ob**$_T$ [P]	+ feststellen / ¬ bemerken, erfahren, erkennen, realisieren, registrieren
	Bisher wird in einer ersten Anhörung **festgestellt**, [$_T$ ob überhaupt ein Asylantrag vorliegt].
$\mathbf{n}_E$ **s-w**$_T$	+ bemerken, erkennen, erfahren, feststellen, realisieren, registrieren
	... daß [$_E$ kaum jemand] **bemerkte**, [$_T$ wie flexibel er auch im Grundsätzlichen war]. Seit [$_E$ die Russen] **bemerkten**, [$_T$ wieviele Eintrittsgelder die Juwelen ... versprechen] ... [$_E$ Er] hat **erkannt**, [$_T$ wie wichtig die Testarbeit ist] ... So möchte [$_E$ man] einerseits **erfahren**, [$_T$ wo der Schuh drückt] ... ... ohne daß [$_E$ man] **erführe**, [$_T$ warum und wozu]. Jetzt gelte es **festzustellen**, [$_T$ wer von dem Video gewußt habe] ... [$_E$ Mark Twain] mußte **feststellen**, [$_T$ wie grotesk falsch die Erziehung ... gewesen war]. Erst nach dem Gespräch in Bonn habe [$_E$ ich] **realisiert**, [$_T$ wie wichtig es war] ... ... und erst als [$_E$ Trainerin und Manager] **realisierten**, [$_T$ wie schnell das Rennen war] ... [$_E$ Sie] **registrieren**, [$_T$ wohin sich die Warenwelt bewegt].
$\mathbf{n}_E$ **s-w**$_T$ [P]	+ erkennen, feststellen, registrieren / ¬ bemerken, erfahren, realisieren
	Bei IBM wurde erst spät **erkannt**, [$_T$ wie gefährlich ihm ... werden könnte]. ... wird **festgestellt**, [$_T$ wie wenig man sich doch als "Bruder und Schwester" zu sagen hat] ... An der Kasse wird ... **registriert**, [$_T$ welcher Haushalt wie oft zu der Marktneuheit greift] ...

Class 25: *Description*

Verbs: *beschreiben, charakterisieren, darstellen*$_1$, *interpretieren*

Ambiguity: Next to a sense of ‘to present’ (class 26), *darstellen* has a sense of ‘to constitute, to be’.

Scene: [$_D$ A descriptor] describes [$_T$ something] and possibly adds [$_I$ an interpretation] of the description.

Frame Roles: D(escriptor), T(heme)

Modification Roles: I(nterpretation)

Levin class: 29.2 *(Verbs with Predicative Complements* $\rightarrow$ *Characterize Verbs)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_D$ **a**$_T$	+ beschreiben, charakterisieren, darstellen, interpretieren
	Eine Innenansicht, [$_D$ die] [$_T$ das Lebens-und Familiengefühl] ... **beschreibt** ... Und [$_D$ sie] **beschreibt** offenkundig [$_T$ kein Komplott]. Der Entwicklung der Kunst folgend, **beschreibt** [$_D$ er] [$_T$ den Weg] ... ... Kennlinien, [$_D$ die] [$_T$ das System] **charakterisieren**. [$_D$ Heinrich von Kleist] **charakterisierte** ... [$_T$ seine Reiseeindrücke] ... In seinem Buch **charakterisiert** [$_D$ Reemtsma] [$_T$ den "Engländer"] ... ... um [$_T$ den wirtschaftlichen Erfolg] [$_I$ größer] **darzustellen**, als er eigentlich ist. ... [$_T$ die Jagd] umfassend in Wort und Bild **darzustellen**. ... daß [$_D$ die Szene] [$_T$ nichts weiter] **darstellt** als ein frivoles Vorspiel ... [$_D$ Er] wollte [$_T$ diese Position] jedoch nicht **interpretieren**. ... [$_T$ die angeblichen Manipulationen] [$_I$ derart] zu **interpretieren**. [$_D$ Er] brauchte [$_T$ Homer] nicht zu **interpretieren** ...
n$_D$ **a**$_T$ [P]	+ beschreiben, charakterisieren, darstellen, interpretieren
	[$_T$ Die einzelnen Häuser] sind detailliert **beschrieben** ... [$_I$ So] wird [$_T$ der Traum eines Migränekranken] **beschrieben** ... [$_T$ Die Ästhetik seiner Aufnahmen] wurde mit einem Slogan **charakterisiert** ... Erst jüngst ist wieder [$_T$ ein neuer Typ] **charakterisiert** worden. [$_T$ Duffi] wird immer[$_I$ so locker] **dargestellt**, aber das ist er nicht mehr. [$_T$ Jüdisches Bürgertum] wird hier [$_I$ anschaulich und kenntnisreich] **dargestellt**. ... wenn [$_T$ eine Inschrift] richtig **interpretiert** worden ist ... [$_I$ So] allerdings ist [$_T$ sie] in Deutschland nie **interpretiert** worden. [$_T$ Das Günstigkeitsprinzip] muß auch in diesem Sinne **interpretiert** werden dürfen.
n$_D$ **a**$_T$ **p**$_I$:Akk.vgl	+ beschreiben, charakterisieren, darstellen, interpretieren
	[$_T$ Einen von ihnen] **beschreibt** [$_D$ er] [$_I$ als einen "weltoffenen Geistlichen"]. [$_D$ Ärzte] **beschrieben** [$_T$ ihren Gesundheitszustand] [$_I$ als vergleichsweise gut] ... [$_D$ Er] hätte [$_T$ die Tat] [$_I$ als Mord] **charakterisiert** ... [$_T$ Die Mitteilung Nikosias] ... versucht [$_D$ er] [$_I$ als Erfolg] ... **darzustellen**. ... obwohl auch [$_D$ er] [$_T$ sich] [$_I$ als Anhänger der Unabhängigkeit] **darstellt**. Und [$_D$ Porsche] **interpretiert** [$_T$ diese Motorbauform] [$_I$ als Teil seiner Identität].
n$_D$ **a**$_T$ **p**$_I$:Akk.vgl [P]	+ beschreiben, darstellen, charakterisieren, interpretieren
	Daß Neptun, [$_T$ der] schon in Vergils "Aeneis" [$_I$ als Rhetor] **beschrieben** wird ... [$_T$ Er] wird [$_D$ von italienischen Beobachtern] [$_I$ als fähiger Sanierer] **beschrieben** ... Daß [$_T$ "Tödliche Ahnungen"] ... [$_I$ als abregend] **charakterisiert** werden kann ... [$_T$ Haider] hatte nichts dagegen, [$_I$ als Populist] **charakterisiert** zu werden. ... wonach [$_T$ Wehrmachtsangehörige] [$_I$ als Mörder] **dargestellt** würden. ... [$_T$ die] es sich gefallen lassen müssen, [$_I$ als Sexsymbole] **dargestellt** zu werden? ... Gestalt mit nacktem Bauch, [$_T$ die] gerne [$_I$ als "Indianer"] **interpretiert** wird ... [$_T$ Er] ist aber zunehmend auch **interpretiert** worden [$_I$ als Lizenz zur Ignoranz] ...
n$_D$ **s-w**$_T$	+ beschreiben, darstellen / ¬ charakterisieren, interpretieren
	[$_D$ Physiker] **beschreiben**, [$_T$ wie sich Systeme mit ... vielen Teilchen verhalten] ... [$_D$ Er] **beschreibt**, [$_T$ wie Musik in den Augen der Komponistin funktioniert]. ... in der [$_D$ sie] **darstellen**, [$_T$ wie eine Finanzkrise ... vermieden werden kann]. ... [$_D$ das] **darstellte**, [$_T$ wie man ein erfolgreicher Geschäftsmann ... sein könne] ...
n$_D$ **s-w**$_T$ [P]	+ beschreiben, darstellen / ¬ charakterisieren, interpretieren
	In ihm wird **beschrieben**, [$_T$ wie der schwedische Wallenberg-Konzern ... handelte]. Es wird ... **beschrieben**, [$_T$ wie man die ZWEI und die DREI auszusprechen hat]. Auf archaischen Vasenbildern wird häufig **dargestellt**, [$_T$ wie Achill ... Waffen erhält].

Class 26: *Presentation*

Verbs: *darstellen*$_2$*, demonstrieren, präsentieren, veranschaulichen, vorführen*

Ambiguity: As well as a sense of 'to describe' (class 25), *darstellen* has a sense of 'to constitute, to be'. *demonstrieren* has a more specific sense within the area of presentation as 'take part in a (political) demonstration'.

Scene: [$_D$ A demonstrator or a means of presentation] shows [$_T$ something]. There might be [$_R$ an audience] for the presentation. The focus of the verbs is on presentation, but not interpretation.

Frame Roles: D(emonstrator), T(heme)

Modification Roles: R(eceiver)

Levin class: 37.1 / 48.1.2 *(Verbs of Communication → Verbs of Transfer of a Message / Verbs of Appearance, Disappearance, and Occurrence → Reflexive Verbs of Appearance)*
The verbs in this class show properties of the two Levin classes 37.1 and 48.1.2. Concerning class 37.1, they are similar to the teaching verbs in class 29.

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_D$ **a**$_T$	+ darstellen, demonstrieren, präsentieren, veranschaulichen, vorführen
	[$_F$ Es] kann [$_T$ nichts] **darstellen**, was ein bißchen plebejisch ... ist.
	... Reichtum, [$_T$ den] [$_D$ die Satellitenbilder von SPOT] **darstellen** ...
	Noch einmal **demonstriert** [$_D$ das Haus] [$_T$ seinen künstlerischen Anspruch] ...
	[$_T$ Diese Kunst] **demonstrierte** [$_D$ er] so augenfällig und auffällig ...
	Stolz **präsentiert** [$_D$ er] [$_T$ den Vorspann des Programms] ...
	In Leverkusen **präsentiert** [$_D$ das Landesmuseum Bonn] [$_T$ Höhepunkte aus seiner Sammlung].
	... und [$_D$ die Diva] **präsentiert** [$_T$ es] unter eigenhändigem Kastagnettenklappern.
	[$_D$ Er] **veranschaulicht** [$_T$ die soziale Umwelt] ...
	[$_T$ Diesen Schaden] will [$_D$ Michael Crichton] **veranschaulichen**.
	Als [$_D$ der Professor] gar [$_T$ einen Handkuß] **vorführte** ...
	... [$_D$ die] [$_T$ ihre ans Pornographische grenzenden Kreationen] **vorführten**.
n$_D$ **a**$_T$ [P]	+ darstellen, demonstrieren, präsentieren, veranschaulichen, vorführen
	Ich finde, [$_T$ der Nikolaus] ist nicht richtig **dargestellt** worden ...
	... [$_T$ die Form eines potentiellen Anti-Krebs-Moleküls] wurde ... **dargestellt** ...
	[$_T$ Dieser Führungswechsel] muß auch [$_R$ gegenüber den Mitarbeitern] **demonstriert** werden.
	Deutlicher kann [$_T$ der Ernst der Lage] kaum **demonstriert** werden.
	Dort wird [$_T$ ein Komponist] **präsentiert** ...
	[$_T$ Seine bösartigen ... Züge] ... werden aber eher von ihrer grotesken Seite **präsentiert**.
	[$_T$ Sie] wird [$_D$ durch 155 Abbildungen] **veranschaulicht** ...
	[$_T$ Die wachsende Bedeutung] ... wird hier [$_D$ durch einige Beispiele] **veranschaulicht** ...
	Im gleichen Stil wird [$_T$ das gesamte Romanpersonal] **vorgeführt** ...
	[$_T$ Das Wort von der "Aufarbeitung"] wird in seiner ganzen Effizienz **vorgeführt** ...
	Hier wird nicht nur [$_T$ ein Genre] **vorgeführt** ...

Frame	Participating Verbs & Corpus Examples
n$_D$ **a**$_T$ **d**$_R$	+ darstellen, demonstrieren, präsentieren, veranschaulichen, vorführen
	... so wie [$_D$ sie] [$_R$ uns] [$_T$ die Berliner Freunde] **darstellten** ... ... wenn [$_R$ ihm] [$_D$ die Experten] [$_T$ den angeschwollenen Wissensstand] **demonstrieren** ... ... um [$_R$ dem eigenen Verein] [$_T$ ihre Unzufriedenheit] zu **demonstrieren**. [$_D$ Wir] würden [$_R$ ihnen] [$_T$ unsere Zahlen] **präsentieren** ... [$_R$ Den russischen Journalisten] **präsentierte** [$_D$ er] [$_T$ sich] ... mit den Worten: ... [$_T$ das] [$_D$ er] dann [$_R$ herbeigeeilten Polizisten] **präsentierte**. [$_D$ Er] **veranschaulichte** [$_R$ uns] [$_T$ das Absenken eines Tunnelabschnitts] ... [$_D$ Ich] **veranschauliche** [$_R$ mir] [$_T$ das] mit folgendem Beispiel: [$_T$ Das] haben [$_R$ uns] [$_D$ die Ölkrisen] **vorgeführt**. Ein rechter Literatenspaß war es, [$_R$ dem Publikum] [$_T$ verklemmte Typen] **vorzuführen** ... ... um [$_T$ es] [$_R$ dem Kaiser Franz Joseph] **vorzuführen**.
n$_D$ **a**$_T$ **d**$_R$ [P]	+ darstellen, präsentieren, vorführen / ¬ demonstrieren, veranschaulichen
	So wie [$_R$ uns] [$_T$ das Aussteigerprogramm] **dargestellt** worden ist ... ... damit [$_T$ sie] [$_R$ dem Kunden] in perfekter Qualität **präsentiert** werden können. ... bevor [$_T$ sie] [$_T$ der Weltöffentlichkeit] **präsentiert** werden können. Als [$_T$ die Fallstudie in Nagano] [$_R$ einer Hundertschaft Journalisten] **vorgeführt** wird ... [$_T$ Sie] sollte im Lauf des Tages [$_R$ einem Amtsarzt] **vorgeführt** werden ...
n$_D$ **s-dass**$_T$	+ darstellen, demonstrieren, veranschaulichen, vorführen / ¬ präsentieren
	[$_D$ Sein Kollege Klein] hatte ... **dargestellt**, [$_T$ daß Naturheilmittel ... wirken]. [$_D$ Ich] versuchte **darzustellen**, [$_T$ daß Kurdistan noch nicht einmal eine Kolonie ist]. [$_D$ Viele Beiträge] sollen **demonstrieren**, [$_T$ daß der Horizont sich geweitet hat] ... In eigener Person wollte [$_D$ er] **demonstrieren**, [$_T$ daß eine Verschmelzung ...] [$_D$ Eine Reihe von Modellen] **veranschaulicht**, [$_T$ daß der Formenreichtum ...] Erst [$_D$ die "Philosophischen Untersuchungen"] werden **vorführen**, [$_T$ daß man ...]
n$_D$ **s-dass**$_T$ [P]	+ darstellen, demonstrieren, vorführen / ¬ präsentieren, veranschaulichen
	Endlich wurde auch einmal **dargestellt**, [$_T$ daß ... plattgemacht wird]. [$_D$ In dem ... Bericht] wird **dargestellt**, [$_T$ daß die ... Häftlinge ... nicht wußten] ... [$_D$ Damit] wurde ... **demonstriert**, [$_T$ daß der Abzug vom originalen Negativ stammt] ... Hierbei wurde **demonstriert**, [$_T$ daß allein die Masse ... entscheidend ist]. Nachdrücklicher ... wurde **vorgeführt**, [$_T$ daß die Innenminister ... haben]. ... wird hier **vorgeführt**, [$_T$ daß ... sowohl Recherche- als auch Darstellungsform ist].
n$_D$ **s-w**$_T$	+ darstellen, demonstrieren, präsentieren, veranschaulichen, vorführen
	[$_D$ Das Zellstoffwerk Pirna] konnte **darstellen**, [$_T$ wie es die Chloranwendung ...] [$_D$ Ich] habe versucht **darzustellen**, [$_T$ was wir auf verschiedenen Wegen getan haben] ... ... [$_D$ die Szene aus dem Badehaus] **demonstriere**, [$_T$ wie gesittet es ... zugegangen sei]. ... [$_D$ Initiativen] **präsentieren**, [$_T$ wie sie Frauen fördern] ... ... hat [$_D$ die Regierung] jetzt **präsentiert**, [$_T$ wie es im Inneren ... zugehen soll] ... [$_T$ Wo der Unterschied liegt], **veranschaulicht** kurz darauf [$_D$ ein törichtes junges Paar] ... ... [$_D$ welches] **veranschaulicht**, [$_T$ wie sich ... die Schichten ... überlagerten]. ... Gesellschaftsmodell, [$_D$ das] **vorführt**, [$_T$ wie Gruppenarbeit aussehen sollte]. Indem [$_D$ Erhart] an Beispielen **vorführte**, [$_T$ wie Heine die Rhetorik benutzt] ...
n$_D$ **s-w**$_T$ [P]	+ darstellen, demonstrieren, präsentieren, vorführen / ¬ veranschaulichen
	... vielmehr wird [$_D$ im Buch] **dargestellt**, [$_T$ wie Baumgarten versuchte] ... Aber [$_D$ nirgends] wird **dargestellt**, [$_T$ was heute ins Auge springt]: ... denn aus der Sicht des Juristen wird einmal mehr **demonstriert**, [$_T$ wie "souverän" ...] Routiniert wurde **präsentiert**, [$_T$ was neu entstanden war] ... Es wird **vorgeführt**, [$_T$ wie vieler Vehikel es bedarf, daß wir miteinander reden]. Schließlich wurde **vorgeführt**, [$_T$ was man versäumt, wenn man einfach nur Auto fährt].

Frame	Participating Verbs & Corpus Examples
n$_D$ **d**$_R$ **s-dass**$_T$	+ demonstrieren, vorführen, veranschaulichen / ¬ darstellen, präsentieren
	... um [$_R$ den Kollegen] ... zu **demonstrieren**, [$_T$ daß auch ein schwer Stotternder ...] [$_D$ Das Beschäftigungskapitel] ... sollten [$_R$ den Bürgern] **veranschaulichen**, [$_T$ daß ...] ... wenn [$_D$ er] [$_R$ uns] **vorführt**, [$_T$ daß er noch immer sein Handwerk beherrscht]. ... weil [$_D$ er] [$_R$ den Zuschauern] **vorführt**, [$_T$ daß sie selbst potentielle Opfer sind]. ... [$_D$ die] [$_R$ dem Zuschauer] täglich **vorführt**, [$_T$ daß er ja blöd ist] ...
n$_D$ **d**$_R$ **s-dass**$_T$ [P]	+ demonstrieren, vorführen / ¬ darstellen, präsentieren, veranschaulichen
	[$_D$ Damit] werde [$_R$ der Öffentlichkeit] **demonstriert**, [$_T$ daß Umweltschutz ...] ... und wieder einmal wird [$_R$ uns] **vorgeführt**, [$_T$ daß die Zerstörungen ... sind]: Hier wurde [$_R$ dem Staatsbürger] offenbar **vorgeführt**, [$_T$ daß man auch ...]
n$_D$ **d**$_R$ **s-w**$_T$	+ demonstrieren, präsentieren, vorführen, veranschaulichen / ¬ darstellen
	... daß [$_D$ der Künstler] [$_R$ uns] **demonstrieren** möchte, [$_T$ wie Kunst zur Kunst wird] ... [$_D$ Er] möge [$_R$ mir] **demonstrieren**, [$_T$ wie er ein alkoholisches Getränk ...] ... um [$_R$ ihnen] zu **präsentieren**, [$_T$ wie nackter Urlaub à la France aussieht]. ... können [$_D$ wir] [$_R$ uns] **veranschaulichen**, [$_T$ was es bedeutet ...]; Zudem kann [$_D$ die serbische Seite] [$_R$ der Welt] so **vorführen**, [$_T$ wer ...] [$_D$ Annan] hat [$_R$ der Welt] **vorgeführt**, [$_T$ woran sie kaum noch glauben mochte] ...
n$_D$ **d**$_R$ **s-w**$_T$ [P]	+ demonstrieren, präsentieren, vorführen / ¬ darstellen, veranschaulichen
	Besonders [$_R$ den Anlegern] wird **demonstriert**, [$_T$ wie man sein Vermögen ...] ... wird [$_R$ dem Leser] **demonstriert**, [$_T$ was im Laufe der Jahrhunderte ...] Erst dann wird [$_R$ ihr] **präsentiert**, [$_T$ worauf sie ein Recht hat] ... Besonders gern wird [$_R$ der Öffentlichkeit] **präsentiert**, [$_T$ was intensiv ...] Endlich wird [$_R$ der Öffentlichkeit] **vorgeführt**, [$_T$ wie sich deutsche Winzer ...] Allein [$_D$ im ORF] wurde [$_R$ uns] **vorgeführt**, [$_T$ was ... zu verstehen ist].

Class 27: ***Speculation***

Verbs: *grübeln, nachdenken, phantasieren, spekulieren*

Scene: [$_E$ Somebody] speculates about [$_T$ something].

Frame Roles: E(xperiencer), T(heme)

Levin class: -

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_E$	+ grübeln, nachdenken, phantasieren, spekulieren
	... obwohl [$_E$ er] lange **grübelt** ... Einer, [$_E$ der] nicht lamentiert und **grübelt**, sondern zupackt. ... [$_E$ die] viel **nachdenkt** und sich den Kopf zerbricht. [$_E$ Ich] habe jetzt viel Zeit **nachzudenken**. [$_E$ Die Schriftsteller] **phantasieren**. Mein Job ist es nicht, zu **phantasieren**, sondern zu antizipieren. [$_E$ Der Mann] **phantasiert**, quasselt, weint ... Es sei eine interessante Frage, aber [$_E$ ich] möchte nicht **spekulieren**. [$_E$ Wer] es dennoch tut, **spekuliert**.
n$_E$ **a**$_T$	+ phantasieren / ¬ grübeln, nachdenken, spekulieren
	Als Alternative **phantasiert** [$_E$ er] dazu [$_T$ eine Kurzgeschichte] ... [$_E$ Man] **phantasiert** [$_T$ die wildesten Liebesspiele] unterm Bettlaken ... ... indem [$_E$ sie] [$_T$ eine jüdische Identität] **phantasierte** ... Daß [$_E$ Frauen] immer wieder von sich aus [$_T$ Hexerei] **phantasierten** ...
n$_E$ **a**$_T$ [P]	+ phantasieren / ¬ grübeln, nachdenken, spekulieren
	... werden [$_T$ Frauen] als hinterhältige Wesen **phantasiert** ... [$_T$ Der Körper] wird genauso gefährlich und schmutzig **phantasiert** ... Im Kampf ... werden [$_T$ bedrohlichste Entwicklungen] **phantasiert**.
n$_E$ **p**$_T$:Akk.über	+ grübeln, nachdenken, phantasieren, spekulieren
	[$_E$ Ein Labour-Taktiker] **grübelt** [$_T$ über eine andere Zermürbungsmethode]: [$_T$ Über dieses Phänomen] habe [$_E$ ich] lange **gegrübelt** ... [$_E$ Lundgaard] will nun [$_T$ über einen neuen Modus der Indizierung] **nachdenken**. ... erst dann kann [$_E$ man] [$_T$ über Änderungen] **nachdenken**. So **phantasierte** etwa [$_E$ ein Abgeordneter] ... [$_T$ über das Reich des Bösen]. [$_E$ Er] **phantasiert** [$_T$ über das Thema: Die Konferenz der Internationale]. In Ruanda **spekulierte** [$_E$ man] daher [$_T$ über die Hintergründe] ... [$_T$ Über solche] **spekulieren** [$_E$ die Forscher] seit geraumer Zeit.
n$_E$ **p**$_T$:Akk.über [P]	+ grübeln, nachdenken, phantasieren, spekulieren
	Lange ist [$_T$ über das Geheimnis Stradivaris] **gegrübelt** worden ... Es müsse mehr [$_T$ über neue Instrumente] ... **nachgedacht** werden. [$_T$ Über weitere Projekte] wird **nachgedacht**. An Kneipentheken ... wird ... [$_T$ darüber] **phantasiert** ... Um so mehr darf [$_T$ über die wirklichen Beweggründe] **spekuliert** werden. ... [$_T$ über die] bereits zuvor in der russischen Presse **spekuliert** worden war.

Frame	Participating Verbs & Corpus Examples
n$_E$ **p**$_T$:Akk.von	+ phantasieren / ¬ grübeln, nachdenken, spekulieren
	Danach **phantasiert** [$_E$ er] [$_T$ von erotischen Begegnungen mit Männern] ...
	[$_E$ Er] **phantasiert** [$_T$ von Abhörgeräten] ...
	... [$_T$ vom Abriß der ganzen Halle] zu **phantasieren**.
n$_E$ **p**$_T$:Akk.auf	+ spekulieren / ¬ grübeln, nachdenken, phantasieren
	[$_E$ Die Börse] ist bereit, [$_T$ auf Veränderungen] ... zu **spekulieren**.
	[$_E$ Leverkusen] **spekulierte** vergebens [$_T$ auf eine Abseitsentscheidung] ...
n$_E$ **s-2**$_T$	+ phantasieren, spekulieren / ¬ grübeln, nachdenken
	... [$_T$ bald werden rote Würgearme hineinlangen], **phantasierte** [$_E$ der Knabe].
	[$_T$ Insgesamt], **phantasiert** [$_E$ er], [$_T$ gebe es auf der Welt eine Milliarde Juden].
	[$_E$ Die 'Jerusalem Post'] **spekulierte**, [$_T$ Bonn werde sich ... auf Geldmangel herausreden].
	[$_E$ Sie] **spekulierten**, Bourassa hoffe, [$_T$ daß die Nationalisten ...]
n$_E$ **s-2**$_T$ [P]	+ spekulieren / ¬ grübeln, nachdenken, phantasieren
	[$_T$ Ein Angriff auf Israel], so wird in der irakischen Führung **spekuliert**, [$_T$ könnte ...]
n$_E$ **s-dass**$_T$	+ phantasieren, spekulieren / ¬ grübeln, nachdenken
	[$_E$ Sie] **phantasiert**, [$_T$ daß das Kind sie ... ertappt habe] ...
	... [$_E$ die beiden] **phantasieren**, ... [$_T$ daß sie ... aufsteigen wollen] ...
	[$_E$ Wieder andere] **spekulierten**, [$_T$ daß der Euro ... "schwach" sein werde] ...
	[$_E$ Sie] **spekulieren**, [$_T$ daß man auch Stammzellen ... immunisieren könnte].
n$_E$ **s-dass**$_T$ [P]	+ spekulieren / ¬ grübeln, nachdenken, phantasieren
	Hier wird **spekuliert**, [$_T$ daß man aus dem Abraum ... Profit schlagen will] ...
	Hier wurde **spekuliert**, [$_T$ daß Cronenberg das "Neue Fleisch" bisexuell betrachtet] ...
	Es wird **spekuliert**, [$_T$ daß die Familie Maxwell ... diese Gruppe verkaufen muß] ...
n$_E$ **s-w**$_T$	+ grübeln, nachdenken, phantasieren, spekulieren
	Und wenn [$_E$ man] anfängt zu **grübeln**, [$_T$ warum es nicht geht] ...
	[$_E$ Stoiber] **grübelt**, [$_T$ wo sich dieser dann entladen würde].
	... anfangen **nachzudenken**, [$_T$ wer unsere Lebensgrundlage ... zerstört].
	... wenn [$_E$ sie] **phantasiert**, [$_T$ wie sie auf ihrer eigenen kleinen feinen Demo] ...
	So kann [$_E$ man] **spekulieren**, [$_T$ welches die Gründe für den Verkauf ... sind] ...
	Natürlich kann [$_E$ man] **spekulieren**, [$_T$ warum Scarpia so ist, wie er ist].
n$_E$ **s-w**$_T$ [P]	+ grübeln, spekulieren / ¬ nachdenken, phantasieren
	Derzeit wird an der Börse **gegrübelt**, [$_T$ an welchen Konzern der Vorstand ...]
	... und schon jetzt wird **gegrübelt**, [$_T$ wie "die Rechten" ... fernzuhalten seien].
	In Washington wird nun darüber **spekuliert**, [$_T$ was mögliche neue Kompromißlinien ...]
	Unterdessen wurde **spekuliert**, [$_T$ welchen Zweck Saddam Hussein ... verfolgt].
n$_E$ **s-ob**$_T$	+ grübeln, nachdenken, spekulieren / ¬ phantasieren
	... während [$_E$ Eleonora Herdt] weiter **grübelt**, [$_T$ ob sie dorthin ausreisen soll] ...
	... und manchmal muß [$_E$ ich] ... **grübeln**, [$_T$ ob ich ... was vergessen habe].
	[$_E$ Alle] müssen noch einmal **nachdenken**, [$_T$ ob es Sinn macht ...].
	... Männern, [$_E$ die] laut **nachdenken**, [$_T$ ob sie handgreiflich werden sollen].
	[$_E$ Die amerikanische Presse] hatte ... **spekuliert**, [$_T$ ob der US-Geheimdienst] ...
	Dann hätte [$_E$ ich] **spekuliert**, [$_T$ ob die drei ihre erste LP ... finanziert haben] ...
n$_E$ **s-ob**$_T$ [P]	+ spekulieren / ¬ grübeln, nachdenken, phantasieren
	... und jahrelang wurde **spekuliert**, [$_T$ ob es die Gruppe überhaupt noch gebe].
	Von verschiedenen Seiten wurde **spekuliert**, [$_T$ ob die Hilfspfeiler ... stabilisieren] ...

Class 28: *Insistence*

Verbs: *beharren, bestehen*$_1$*, insistieren, pochen*

Ambiguity: *bestehen* also refers to (i) the existence (cf. class 37) and (ii) the consistence of something; (iii) in addition, it has a sense of passing an exam. *pochen* has a sense of ‘to knock’.

Scene: [$_I$ An agent] insists on [$_T$ something].

Frame Roles: I(nsistor), T(heme)

Levin class: -

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_I$	+ insistieren / ¬ beharren, bestehen, pochen
	... [$_I$ sie] hätten Donnerstag früh nicht **insistieren** wollen.
	... und [$_I$ wer] zu **insistieren** wagt, steht vor einer Mauer.
	Erst nachdem [$_I$ die Staatsanwälte] **insistierten** ...
	Als [$_I$ Jung] bei seinen Vorgesetzten **insistierte**, wurde ihm der Fall entzogen:
n$_I$ **p**$_T$:Dat.auf	+ beharren, bestehen, insistieren / ¬ pochen
	[$_I$ Solms und Westerwelle] **beharrten** am Wochenende [$_T$ auf dem Plan ihrer Partei] ...
	[$_I$ Der Bankenverband] **beharrt** [$_T$ darauf, Eigenkapital koste seine Mitgliedsinstitute]...
	[$_I$ Peking] **beharrte** demgegenüber [$_T$ auf seinem Mitspracherecht] ...
	[$_I$ Sie] **beharren** jedoch [$_T$ auf ihrer Eigenständigkeit] ...
	Doch [$_I$ Prinz Charles] **beharrte** [$_T$ auf seiner Wahl].
	[$_I$ Die FDP] **bestehe** [$_T$ auf der Abschaffung der Gewerbekapitalsteuer] ...
	[$_I$ Arafat] **bestehe** weiter [$_T$ auf einem detaillierten Zeitplan] ...
	... [$_I$ sie] **besteht** [$_T$ darauf, daß es nichts zu bereuen gebe] ...
	... [$_I$ er] **bestand** [$_T$ darauf, bei Diners am Kopf der Tafel plaziert zu werden] ...
	Auch [$_I$ Leute wie Anwar] freilich **insistieren** [$_T$ auf der Bedeutung] ...
	[$_I$ Sie] **insistieren** [$_T$ auf dem negativen Gebrauch des Wortes Hacker].
	Mit Recht jedoch **insistiert** [$_I$ er] [$_T$ darauf, daß wir von den Urfiguren des Lebens ...]
	Denn [$_I$ beide Positionen] **insistierten** [$_T$ auf der ... absoluten Stellung des Fürsten] ...
	[$_I$ Sie] **insistieren** zu Recht [$_T$ auf Beweisen].
n$_I$ **p**$_T$:Akk.auf	+ pochen / ¬ beharren, bestehen, insistieren
	Da **pocht** [$_I$ einer] despotisch [$_T$ aufs Patriarchat] ...
	[$_I$ Sie] **pocht** [$_T$ darauf, daß der Vertrag ... strikt eingehalten werde] ...
	[$_I$ Die organisierten Lehrer] **pochen** [$_T$ auf die ererbten Privilegien ihres Standes] ...
	[$_I$ Kiew] **pocht** [$_T$ auf die Einhaltung des westlichen Versprechens].
n$_I$ **s-2**$_T$	+ insistieren / ¬ beharren, bestehen, pochen
	[$_I$ Justizminister Klaus Kinkel] **insistierte**, [$_T$ der ... habe keinen "Racheakt" zu erwarten] ...
	[$_I$ Corsten] hat auf der Veranstaltung **insistiert**, [$_T$ die taz habe keine Chefredaktion]!
	[$_T$ "Wo ist der Rest?"] **insistiert** [$_I$ Heß].
n$_I$ **s-dass**	+ insistieren / ¬ beharren, bestehen, pochen
	Hatte [$_I$ Washington] doch wiederholt **insistiert**, [$_T$ daß China keine Raketen ... liefert].
	[$_I$ Kvint] **insistiert**, [$_T$ daß ... zu einem rationalen Interessenausgleich führen könne].

Class 29: *Teaching*

Verbs: *beibringen, lehren, unterrichten, vermitteln*$_2$

Ambiguity: *beibringen* has a sense of ‘to bring’. *unterrichten* has a sense of ‘to notify’. Next to the sense of *vermitteln* as a (material) supply verb, it has a sense of ‘to mediate’.

Scene: [$_T$ A teacher] teaches [$_L$ learners] some kind of [$_M$ material].

Frame Roles: T(eacher), L(earner)

Modification Roles: M(aterial)

Levin class: 37.1 *(Verbs of Communication → Verbs of Transfer of a Message)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_T$	+$_{adv}$ lehren, unterrichten / ¬ beibringen, vermitteln
	[$_T$ Heller] **lehrte** bis 1987 an der Sorbonne in Paris.
	[$_T$ Einige, halbwegs privilegiert,] **lehren** an amerikanischen Universitäten.
	Fachlehrer, [$_T$ die] in Parallelklassen **unterrichten** ...
	Nach dem Studium **unterrichtete** [$_T$ ich] an einer Pädagogischen Hochschule ...
n$_T$ **a**$_M$	+ lehren, unterrichten, vermitteln / ¬ beibringen
	[$_T$ Der Verfasser] **lehrt** [$_M$ Sozialgeschichte der Medizin] ...
	Leider **lehren** [$_T$ die Erfahrungen in Deutschland] [$_M$ das Gegenteil] ...
	Indem [$_T$ sie] [$_M$ Sprachen] **lehre** ...
	... Lehrer, [$_T$ die] bald [$_M$ ein bis zwei Pflichtstunden] mehr **unterrichten** müssen.
	[$_T$ Herr Mau] **unterrichtet** [$_M$ Theatertheorie und Dramaturgiearbeit] ...
	Denn [$_T$ sie] vermag wenig mehr zu **vermitteln** als [$_M$ die Ambitionen eines Sammlers] ...
	Zugleich **vermitteln** [$_T$ sie] [$_M$ etwas vom Entstehungsprozeß dieser Musik] ...
	[$_T$ Dieses Bild der Südländer] **vermittelt** [$_M$ Vorstellungen von Trinkfestigkeit] ...
n$_T$ **a**$_L$	+ unterrichten / ¬ beibringen, lehren, vermitteln
	Inzwischen **unterrichtet** [$_T$ Pilch] [$_L$ Kinder von umliegenden Bauernhöfen] [$_M$ im Lautenspiel] ...
	In Boston **unterrichtete** [$_T$ er] [$_L$ taubstumme Kinder und Heranwachsende] ...
n$_T$ **a**$_M$ [P]	+ lehren, unterrichten, vermitteln / ¬ beibringen
	Vieles von dem, [$_M$ was] heute an den Universitäten **gelehrt** werde ...
	Im übrigen wird [$_M$ Musikologie] an Konservatorien **gelehrt** ...
	In den Schulen wird [$_M$ die allgemein übliche Rechtsschreibung] **unterrichtet**.
	Es soll wieder [$_M$ mehr und auch schwierige Literatur] **unterrichtet** werden ...
	... Kurssystem, in dem [$_M$ alle Fächer] in englischer Sprache **unterrichtet** werden ...
	In den Schulbüchern sollte [$_M$ mehr Vaterlandsliebe] **vermittelt** werden ...
	... daß [$_M$ die figurativen Bedeutungen] nur sekundär **vermittelt** werden.
n$_T$ **a**$_L$ [P]	+ unterrichten / ¬ beibringen, lehren, vermitteln
	... Parallelklassen, [$_L$ die] [$_T$ von demselben Lehrer] **unterrichtet** werden ...
	Nebenan ... werden [$_L$ Studentinnen und Studenten] **unterrichtet** ...
	... daß ... [$_L$ Schulklassen im afrikanischen Busch] ... **unterrichtet** werden können.

Frame	Participating Verbs & Corpus Examples
n$_T$ **a**$_L$ **a**$_M$[3]	+ lehren / ¬ beibringen, unterrichten, vermitteln
	[$_M$ Was] **lehrt** [$_T$ er] [$_L$ uns]? Die Tugenden, [$_M$ die] [$_T$ Madeira] [$_L$ seine Besucher] **lehrt**, scheinen unzeitgemäß. Nun **lehrten** [$_T$ die Finanzmärkte] [$_L$ sie] [$_M$ Mores] ...
n$_T$ **a**$_L$ **a**$_M$ [P]	+ lehren / ¬ beibringen, unterrichten, vermitteln
	... [$_L$ sie] mußte erst [$_M$ das Fürchten] **gelehrt** werden, um sie weise zu machen ...
n$_T$ **a**$_M$ **d**$_L$	+ beibringen, vermitteln / ¬ lehren, unterrichten
	... daß [$_T$ ihre früheren Lehrer] [$_L$ ihnen] [$_M$ nichts oder nicht genug] **beigebracht** hätten. [$_T$ George III.] soll versucht haben, [$_L$ den Vögeln] [$_M$ das Singen] **beizubringen**. ... [$_M$ was] [$_T$ Alkuin] [$_L$ den Franken] alles **beigebracht** hat ... ... als [$_T$ er] [$_L$ einer Jugendgruppe] [$_M$ das Klettern] **beibringen** wollte. [$_M$ Das Lesen] hatte [$_T$ er] [$_L$ sich] selbst **beigebracht** ... Zwar **vermittele** [$_T$ die Hochschule] [$_L$ ihren Studenten] [$_M$ die Theorie] ... So **vermittelte** [$_T$ er] [$_L$ auch den Studenten] [$_M$ nicht nur Sprachkenntnisse] ... [$_T$ Ich] würde sehr gern [$_L$ jungen Spielern] [$_M$ etwas] **vermitteln**.
n$_T$ **a**$_M$ **d**$_L$ [P]	+ beibringen, vermitteln / ¬ lehren, unterrichten
	Deshalb muß [$_M$ es] [$_L$ ihnen] alle vier Jahre im Schnellkurs **beigebracht** werden. [$_M$ Religiöser Fanatismus] wurde [$_L$ ihm] nicht **beigebracht**. ... [$_M$ was] [$_L$ ihnen] laut Rahmenplan **vermittelt** werden müßte ... [$_L$ Dem betroffenen Bürger] ist [$_M$ das] nicht zu **vermitteln** ... ... soll [$_L$ den Teilnehmern] darin [$_M$ das notwendige Handwerkszeug] **vermittelt** werden ... ... bei der [$_L$ den Auszubildenden] [$_M$ die ... Erkenntnisse] **vermittelt** würden.
n$_T$ **a**$_L$ **i**$_M$	+ lehren / ¬ beibringen, unterrichten, vermitteln
	Längst hat [$_T$ die Erfahrung] [$_L$ die Progressiven] **gelehrt**, [$_M$ sich ... anzueignen]. [$_T$ Die schwierige Zeit] habe [$_L$ ihn]... **gelehrt**, [$_M$ bescheiden zu sein]. [$_L$ Unternehmen] will [$_T$ er] **lehren**, [$_M$ nicht mehr nach Gewinn zu streben] ... Nun hat [$_T$ die Partei] [$_L$ ihn] **gelehrt**, [$_M$ ... Bescheidenheit zu üben].
n$_T$ **d**$_L$ **i**$_M$	+ beibringen, vermitteln / ¬ lehren, unterrichten
	... wie [$_T$ er] [$_L$ den afrikanischen Fans] **beibringt** [$_M$ zu rufen] ... Um [$_L$ Kindern] **beizubringen**, [$_M$ sich als Fußgänger ... verkehrsgerecht zu verhalten] ... [$_T$ Man] hatte [$_L$ ihnen] **beigebracht**, [$_M$ in Deckung zu gehen] ... ... möchte [$_L$ ihren Schülern] ... **vermitteln**, [$_M$ fremde Kulturen ... anzuerkennen].
n$_T$ **d**$_L$ **i**$_M$ [P]	+ beibringen, vermitteln / ¬ lehren, unterrichten, vermitteln
	... soll [$_L$ den Jugendlichen] **beigebracht** werden, [$_M$ Verantwortung zu übernehmen] ... [$_L$ Deutschen Sängern] werde **beigebracht**, [$_M$ ihre Stimme zu "decken"] ... [$_L$ Dem Volk] solle aber auch **beigebracht** werden, [$_M$ das Gesetz zu respektieren].
n$_T$ **s-2**$_M$	+ lehren / ¬ beibringen, unterrichten, vermitteln
	Bis dahin hatte [$_T$ das Weistum] **gelehrt**, [$_M$ Loyalität sei die Geheimwaffe] ...
n$_T$ **s-dass**$_M$	+ lehren, vermitteln / ¬ beibringen, unterrichten
	... daß [$_T$ die Erfahrung] hingegen **lehre**, [$_M$ daß auch ein Arzt ...] [$_T$ Der Islam] **lehrt** doch, [$_M$ daß Geld und Gut keine Wichtigkeit haben]. [$_T$ Die Erfahrung] **lehrt** jedoch, [$_M$ daß es zahlreiche Gründe dafür geben kann] ... Nicht erst [$_T$ der Anrufbeantworte]r hat **gelehrt**, [$_M$ daß unglaublich viele Leute ...] Oder [$_T$ sie] wollten **vermitteln**, [$_M$ daß divergierende politische Ansätze ... führten]. [$_T$ Wir] müssen **vermitteln**, [$_M$ daß das Geld notwendig ist] ...
n$_T$ **s-dass**$_M$ [P]	+ lehren, vermitteln / ¬ beibringen, unterrichten
	... und es wird **gelehrt**, [$_M$ daß man nur die bekannten Exemplare verzehren solle]. Denn zum einen wird schön suggestiv **vermittelt**, [$_M$ daß die kulturelle Betätigung ...]

Frame	Participating Verbs & Corpus Examples
n$_T$ **s-w**$_M$	+ lehren, vermitteln / ¬ beibringen, unterrichten [$_T$ Der Film] **lehrt**, [$_M$ was wir schon wissen]: ... [$_M$ wie] [$_T$ das Beispiel von Ländern wie Chile] **lehrt** ... [$_T$ Die Lektüre] ... **lehrt**, [$_M$ warum sie nicht geschrieben werden] ... [$_T$ Er] muß **vermitteln**, [$_M$ was die neue Regierung ... zu beschließen hat] ... ... warum nicht spielerisch **vermitteln**, [$_M$ was ... gespielt werden soll]? Um der Bundesregierung zu **vermitteln**, [$_M$ wie tödlich ernst es ihnen damit war] ...
n$_T$ **s-w**$_M$ [P]	+ lehren, vermitteln / ¬ beibringen, unterrichten Es wird unter anderem **gelehrt**, [$_M$ was Theorie eigentlich bedeutet] ... Hier wird versucht zu **vermitteln**, [$_M$ was diese Volunteers in jener Nacht lernten].
n$_T$ **s-ob**$_M$	+ lehren / ¬ beibringen, unterrichten, vermitteln Nur [$_T$ die Zeit] wird **lehren**, [$_M$ ob Clinton eben diese Eigenschaften ...]
n$_T$ **a**$_L$ **s-2**$_M$	+ lehren / ¬ beibringen, unterrichten, vermitteln [$_T$ Erhard] **lehrt** [$_L$ uns]: [$_M$ Nur die Freiheit aller kann Wohlstand für alle erzeugen].
n$_T$ **a**$_L$ **s-dass**$_M$	+ lehren / ¬ beibringen, unterrichten, vermitteln [$_T$ Die] **lehrt** [$_L$ uns], [$_M$ daß die Jagd die Mutter der Yacht ist] ... Und [$_T$ die Statistik] **lehrt** [$_L$ uns], [$_M$ daß der Anteil ... gestiegen ist] ... [$_M$ Daß das ganze System ...], **lehrt** [$_L$ uns] [$_T$ die Gegenwart].
n$_T$ **a**$_L$ **s-w**$_M$	+ lehren / ¬ beibringen, unterrichten, vermitteln [$_T$ Rose] ... **lehrt** [$_L$ ihn], [$_M$ wie man Studenten zum Zuhören bringt]. [$_T$ Die Regierung] hat [$_L$ uns] oft **gelehrt**, [$_M$ welche andere Methoden ...]
n$_T$ **d**$_L$ **s-2**$_M$	+ beibringen, vermitteln / ¬ lehren, unterrichten ... als [$_T$ der König] [$_L$ seinen Untertanen] ... **beibrachte**, [$_M$ er werde ... müssen]. ... [$_L$ dem Mädchen] zu **vermitteln**, [$_M$ das Gericht gehe davon aus ...]
n$_T$ **d**$_L$ **s-2**$_M$ [P]	+ beibringen, vermitteln / ¬ lehren, unterrichten [$_L$ Dem kleinen Louis] wird **beigebracht**, [$_M$ er sei ein adoptiertes Waisenkind] ... [$_L$ Den Menschen] ... wurde immer **beigebracht**, [$_M$ sie seien überall zu Hause]. [$_L$ Uns] wurde **vermittelt**, [$_M$ wir müßten dankbar dafür sein ...]
n$_T$ **d**$_L$ **s-dass**$_M$	+ beibringen, vermitteln / ¬ lehren, unterrichten ... [$_L$ Mama] schonend **beizubringen**, [$_M$ daß man den Milchkrug ...] Muß [$_T$ ich] [$_L$ ihr] ja schonend **beibringen**, [$_M$ daß ich wieder zu ihr ziehe]. ... [$_T$ die] [$_L$ den Leuten] **vermittelten**, [$_M$ daß sie Rechte und Kräfte hatten]. [$_T$ Ich] versuche [$_L$ meinen Spielern] zu **vermitteln**, [$_M$ daß es ... Leben danach gibt]. ... [$_T$ die] [$_L$ den Leuten] **vermittelten**, [$_M$ daß sie Rechte und Kräfte hatten].
n$_T$ **d**$_L$ **s-dass**$_M$ [P]	+ beibringen, vermitteln / ¬ lehren, unterrichten [$_L$ Dem deutschen Steuerzahler] sei es kaum **beizubringen**, [$_M$ daß er ... solle]. [$_L$ Der Kundschaft] ist schwer zu **vermitteln**, [$_M$ daß zwei Athleten ...] Es wird [$_L$ uns] **vermittelt**, [$_M$ daß wir Menschen sind].
n$_T$ **d**$_L$ **s-w**$_M$	+ beibringen, vermitteln / ¬ lehren, unterrichten [$_T$ Sie] wolle [$_L$ den Russen] **beibringen**, [$_M$ was Schönheit ist] ... ... [$_L$ den Schülern] **beizubringen**, [$_M$ wie sie den Einfluß der Reize mindern können]. [$_T$ Es] versucht [$_L$ den Leuten] zu **vermitteln**, [$_M$ wie man sparsam ... fahren kann]. ... [$_T$ der] [$_L$ ihnen] **vermittelt**, [$_M$ wer sie sind] ...
n$_T$ **d**$_L$ **s-w**$_M$ [P]	+ beibringen, vermitteln / ¬ lehren, unterrichten In der Schule werde [$_L$ den Kindern] nicht mehr **beigebracht**, [$_M$ wie die Kommata] ... ... während [$_L$ ihnen] **beigebracht** werden müßte, [$_M$ wer die Mülleimer ausleert] ... Zunächst wird [$_L$ dem Leser] **vermittelt**, [$_M$ was Ängste sind] ...

[3]This frame is not coded in the grammar.

Class 30: ***Position*** $\rightarrow$ ***Bring into Position***

Verbs: *legen, setzen, stellen*

Ambiguity: All verbs participate in a number of collocations with high frequencies, e.g. *Wert legen* 'to set value on', *in Bewegung setzen* 'to actuate', *zur Verfügung stellen* 'to provide'.

Scene: [$_M$ A person or some circumstances] bring [$_P$ something] into [$_C$ a spatial configuration].

Frame Roles: M(over), P(atient)

Modification Roles: C(onfiguration)

Levin class: 9.2 *(Verbs of Putting* $\rightarrow$ *Verbs of Putting in a Spatial Configuration)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_M$ **a**$_P$	+$_{adv}$ legen, setzen, stellen
	[$_M$ Der israelische Geheimdienst Mossad] habe [$_P$ einen Sprengkörper] [$_C$ in den Bus] **gelegt** ... [$_M$ Der Plan] **legt** daher [$_P$ den Finger] [$_C$ in die Wunden] ... [$_M$ Sie] wollten mir [$_P$ keine Steine] [$_C$ in den Weg] **legen** ... [$_M$ Ich] mußte [$_P$ 120 Mark] [$_C$ auf den Tisch] **legen** als Ostzonaler. [$_M$ Clinton] habe sie zu küssen versucht, [$_P$ seine Hand] [$_C$ auf ihre Brust] **gelegt** ... Während [$_M$ die Italiener] [$_P$ das Ohr] [$_C$ auf die Straße] **legten** ... Mögen [$_M$ die Anhänger] ... [$_P$ anderslautende Gerüchte] [$_C$ in die Welt] **setzen** ... In einem Zug **setzt** [$_M$ sie] [$_P$ zwei feinbestrumpfte ... Frauenbeine] [$_C$ aufs Papier]. Dabei **setzt** [$_M$ man] besonders [$_C$ auf Deutschland] [$_P$ große Hoffnungen]. ... [$_M$ der] [$_P$ sich] [$_C$ in die Natur] **setzte** ... [$_M$ Der Autor] **setzt** [$_P$ die wüsten Beschimpfungen] ... [$_C$ in Szene] ... [$_M$ Serbiens orthodoxe Kirche] **stellte** [$_P$ sich] [$_C$ hinter die Forderung der Bürger] ... ... [$_M$ die Iren] **stellten** [$_P$ die Niederländer] [$_C$ vor große Schwierigkeiten] ... [$_M$ Die orthodoxe Kirche Serbiens] **stellt** [$_P$ sich] [$_C$ gegen das Regime in Belgrad] ... [$_M$ Die Gentechnik] **stellt** [$_P$ strenggläubige Juden] [$_C$ vor Schwierigkeiten]:

Class 31: *Position → Be in Position*

Verbs: *liegen, sitzen, stehen*

Ambiguity: As in class 30, all verbs participate in a number of collocations with high frequencies, e.g. *liegen an* 'to be caused by', or *um etwas gut/schlecht stehen* 'to be in good/bad condition'.

Scene: As the inchoative pendant of class 30, [$_P$ something] is in [$_C$ a spatial configuration].

Frame Roles: P(atient)

Modification Roles: C(onfiguration)

Levin class: 47.6 *(Verbs of Existence → Verbs of Spatial Configuration)*

Schumacher class: 2.1 *(Verben der speziellen Existenz → Verben der Existenzsituierung)*
Only *liegen* is classified, together with some of my aspect verbs (class 1) and my verbs of existence (class 37).

Frame	Participating Verbs & Corpus Examples
n$_P$	+$_{adv}$ liegen, sitzen, stehen
	In der Gesamtwertung **liegt** [$_P$ Thoma] nun [$_C$ auf Platz fünf].
	[$_C$ Dazwischen] **liegen** [$_P$ die Fachleute der anderen großen Investmentbanken].
	Deshalb **liegt** bei Dor [$_P$ Augsburg] [$_C$ an der Donau] ...
	Dann **liegt** [$_P$ sie], [$_C$ eisgekühlt], [$_C$ von Fliegen umschwärmt], [$_C$ auf dem Tisch].
	[$_P$ Weitere Schulungsbetriebe] **liegen** [$_C$ in der Gegend von Valdivia] ...
	[$_P$ Die Schlafsäcke] **liegen** [$_C$ in der Apsis] ...
	Allerdings **liegen** [$_P$ die Preise] ... [$_C$ über denen des Supermarktes].
	... Gefängnisse anzusehen, [$_C$ in denen] [$_P$ gefangene Rebellen] **sitzen**.
	[$_P$ Die profiliertesten Zungenredner] **sitzen** [$_C$ in den Banken] ...
	[$_C$ In all dem Durcheinander] **sitzt** [$_P$ eine alte Frau].
	[$_C$ Auf einer Bank] **sitzen** [$_P$ junge Leute in Geschäftsanzügen].
	[$_P$ Ganz Israel] **saß** [$_C$ zu Hause vor dem Fernseher] ...
	[$_P$ Man] **saß** ein wenig eingemauert [$_C$ zwischen der Tür und einem Gebirgszug] ...
	[$_P$ Ein Mitfahrer] **sitzt** eher [$_C$ unbequem] ...
	Dieser Rüstungshilfe **stehe** [$_P$ nichts mehr] [$_C$ im Wege] ...
	Dabei **stehen** [$_P$ vier Themenbereiche] [$_C$ im Mittelpunkt].
	[$_P$ Der Große Wagen] **steht** nun [$_C$ nahezu senkrecht] [$_C$ im nordöstlichen Himmelsareal] ...
	... dort, [$_C$ am Eingangstor], **steht** [$_P$ die Büste des Bernhardiners Barry] ...
	[$_P$ Er] **steht** [$_C$ an der Spitze eines exzellenten Teams] ...
	Alle paar Minuten **standen** [$_P$ wir] [$_C$ vor Wasserrinnen] ...

Class 32: *Production*

Verbs: *bilden, erzeugen, herstellen, hervorbringen, produzieren*

Scene: [$_P$ A producer] creates [$_R$ a product].

Frame Roles: P(roducer), R(esult)

Levin class: 26.4 *(Verbs of Creation and Transformation → Create Verbs)*

Schumacher class: 1.3 *(Verben der allgemeinen Existenz → Kausative Verben der allgemeinen Existenz)*

Frame	Participating Verbs & Corpus Examples
n$_P$	$+_{adv}$ produzieren / ¬ bilden, erzeugen, herstellen, hervorbringen
	[$_P$ Mercedes-Benz] **produziert** in East London ...
	Im Gegensatz zu Rheiner Moden **produziert** [$_P$ Artländer] zu rund 80 Prozent in eigenen Werken.
	Beim derzeitigen Goldpreis **produzieren** [$_P$ 14 Minen] mit Verlust ...
n$_P$ **a**$_R$	+ bilden, erzeugen, herstellen, hervorbringen, produzieren
	Dann **bilden** [$_P$ wir] eben [$_R$ eine andere Plattform] und gleichen dieses Manko aus.
	... [$_P$ die Abgase der Autos und Motorräder] **bilden** [$_R$ dichte Wolken] ...
	Nicht nur darstellerisch **bildet** [$_P$ das Intermezzo] [$_R$ den Höhepunkt des Abends]:
	Zwar habe [$_P$ man] mit Weißrußland [$_R$ eine Gemeinschaft] **gebildet** ...
	Nun **bilden** [$_P$ die Flußquanten, die sich alle frei bewegen,] [$_R$ ein Bose-Einstein-Kondensat].
	[$_P$ Sie] **erzeugen** [$_R$ einen feinen Strahl] ...
	[$_P$ Gary Hume] **erzeugt** aus verständlichen Symbolen [$_R$ verschlossen wirkende Gemälde] ...
	... [$_P$ die] [$_R$ atomwaffenfähiges Plutonium] **erzeugen** können.
	[$_P$ Daß sie Mantel und Hut tragen], **erzeugt** [$_R$ alles andere als den Eindruck von Bürgerlichkeit] ...
	[$_P$ Bartolis perlende Skalen und gleißende Schleifen] **erzeugen** [$_R$ besonderes Glücksgefühl] ...
	[$_P$ Wir] werden vor allem in Pattaya wieder [$_R$ die alte Ordnung] **herstellen** ...
	er besitzt eine Privatbrauerei, in der [$_P$ die Gäste] [$_R$ ihr Bier] selbst **herstellen**.
	... [$_P$ der] [$_R$ auch Golfsportgeräte] **herstellt** ...
	[$_R$ Babywindeln] könne [$_P$ man] damit **herstellen** ...
	[$_P$ Citroën] hat 1996 [$_R$ mehr Automobile] **hergestellt** ...
	Durch die Mode hat [$_P$ sie] [$_R$ eine Stimmung] **hervorbringen** lassen ...
	[$_P$ Deutschland] habe [$_R$ wenige Gelehrte] **hervorgebracht** ...
	... daß [$_P$ sie] [$_R$ gemeinsamen Nachwuchs] **hervorbringen** können.
	[$_P$ Die wenigen Inselquadratkilometer] ... haben [$_R$ zahlreiche Kapitäne] ... **hervorgebracht**.
	... System, [$_P$ dessen relative Gleichförmigkeit] [$_R$ sie] **hervorgebracht** hat.
	[$_R$ Den einzigen Mißton] ... **produzierte** bezeichnenderweise [$_P$ ein gehobener Militär] ...
	[$_P$ Wer] im mentalen Bereich [$_R$ etwas] bearbeiten oder **produzieren** will ...
	Denn [$_P$ Bakker] **produziert** [$_R$ seine Aquanaut-Yachten] ...
	[$_P$ Die BASF] **produziert** jetzt noch [$_R$ Düngemittel von knapp 1 Million Tonnen Stickstoff].
	[$_P$ Sie] hat [$_R$ Berge von Papier] **produziert** ...
	[$_P$ Es] **produziert** [$_R$ seine Mythen] inzwischen für internationale Marken ...
n$_P$ **a**$_R$ [P]	+ bilden, erzeugen, herstellen, hervorbringen, produzieren
	[$_R$ Die Fonds] müßten aus unversteuertem Einkommen **gebildet** werden ...
	[$_R$ Das andere Enzym, Cox 2,] wird nur bei Bedarf **gebildet** ...
	... daß in Bulgarien [$_R$ eine neue Regierung] **gebildet** wird.
	[$_R$ Er] wird [$_P$ von Blasenkäfern] **gebildet** und ausgeschieden.
	... daß in dieser Gesellschaft [$_R$ soviel menschliches Strandgut] **erzeugt** wird ...
	... daß in jedem Fall [$_R$ menschliche Nähe] **erzeugt** wird ...
	... weil nur von außen [$_R$ der nötige Schub] **erzeugt** werden konnte.
	Zwar werde [$_R$ ein gewisser Handlungsdruck] **erzeugt**, der den Ermessensspielraum ...
	... Gabbehs, [$_R$ die] [$_P$ von nomadischen Familien] **hergestellt** werden ...
	... daß [$_R$ ein Einvernehmen] **hergestellt** werden könne.
	... waren in Weihnachtskarten versteckt, [$_R$ die] aus einer Art Packpapier **hergestellt** waren.
	[$_R$ Keine Wirkung der Posse] wird [$_P$ durch Ironie] **hervorgebracht** ...
	Dennoch wird so [$_R$ Volkskunst] **hervorgebracht** ...
	[$_R$ Sie] sind ideenreich gestaltet, aufwendig **produziert** ...
	Dort werden [$_R$ ABC-Waffen und Raketen] **produziert** ...
	Produziert werden [$_R$ Haus- und Heimtextilien, technische Gewebe] ...

Class 33: *Renovation*

Verbs: *dekorieren, erneuern, renovieren, reparieren*

Scene: [$_R$ A renovator] renews [$_T$ something]. The context might mention [$_D$ some kind of decoration].

Frame Roles: R(enovator), T(heme)

Modification Roles: D(ecoration)

Levin class: -
The English counterparts of the renovation verbs are not classified. The closest class to the verbs is 26.4 as for class 32 above.

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_R$ **a**$_T$	+ dekorieren, erneuern, renovieren, reparieren
	[$_R$ Lady Diana] ..., um [$_T$ ihr New Yorker Appartement] zu **dekorieren**. Allerdings **dekoriert** [$_R$ jeder] erst dann [$_T$ sein Haus] ... [$_R$ Er] illustriert literarische Vorlagen und **dekoriert** mit Vergnügen [$_T$ die Opernbühne]. ... [$_R$ er] habe ... [$_T$ die Ecken des Rahmens] [$_D$ mit Kaktusblüten] **dekoriert**. [$_R$ Er] **dekorierte** [$_T$ seinen grünen Traum] [$_D$ mit einer Renaissance-Loggia] ... [$_R$ Er] **erneuerte** [$_T$ die Zusage] ... [$_R$ Der Minister] **erneuerte** [$_T$ das Angebot an die Hornoer] ... Jeden Tag **erneuert** [$_R$ die Künstlerin] [$_T$ die Pfützen im Raum], kleine Seen: [$_R$ Die IBA] will alte Industrieanlagen umbauen, [$_T$ Landschaften] **erneuern** ... [$_R$ Deren Mitglieder] ... schworen sich, ... [$_T$ die lateinamerikanische Poesie] zu **erneuern** ... [$_R$ Der hitzköpfige Dauerquatscher] ... **erneuert** [$_T$ seine Substanz] ... [$_D$ mit Schokolade] ... [$_R$ Tschechen und Deutsche] **renovierten** gemeinsam [$_T$ die verfallene Kirche] ... Um [$_T$ das Vorschlagswesen] zu **renovieren**, werden die Aufgaben neu verteilt. [$_R$ Die neuen Eigentümer] **renovieren** [$_T$ Fassaden, Keller- und Erdgeschosse]. ... daß nach Umfragen [$_R$ viele Bundesbürger] [$_T$ ihr Bad] **renovieren** wollen. ... Kunstverein Lingen, [$_R$ der] [$_T$ ihn] zu einer Kunsthalle **renoviert** hat ... Seit 1988 **renoviert** [$_R$ sie] [$_T$ Saal für Saal] [$_D$ mit Originalmaterialien] ... [$_R$ Die Reform] versucht, [$_T$ etwas] zu **reparieren** ... In dieser Werkstatt **reparierte** [$_R$ er] mit seinen Freunden ... in der Nacht [$_T$ Autos]. [$_R$ Ich] **repariere** [$_T$ die Welt], ich sammle die Scherben auf. Vorher schon hat [$_R$ er] [$_T$ Geigen] **repariert**, indem er sie kaputtzerlegte. [$_R$ Ein Pflaumenbaum] **repariert** [$_T$ Elektrogeräte]?
n$_D$ **a**$_T$	+ dekorieren / ¬ erneuern, renovieren, reparieren
	Gleich daneben **dekorieren** [$_D$ mehr als dreihundert Ansichtskarten] [$_T$ die Wand] ... [$_D$ Magenta- oder purpurrote pralle Blüten] **dekorieren** [$_T$ die neuen Leinwände] ... [$_D$ Stahlträger und Vorhänge, ... Nebelschwaden und Gewehrsalven] **dekorieren** [$_T$ die Bühne].
n$_R$ **a**$_T$ [P]	+ dekorieren, erneuern, renovieren, reparieren
	Innenräume konnten raffiniert gewölbt, [$_T$ Fassaden] aufwendig **dekoriert** werden ... Zurück in Washington, wird [$_T$ der Kommandeur] **dekoriert**. [$_T$ Der Mexikaner] ist [$_D$ mit Preisen] reich **dekoriert** ... Sie sind mit einem Etagenbett ausgestattet, [$_T$ das] als Pirateninsel ... **dekoriert** ist. ... auf denen [$_T$ Tierschädel] [$_D$ mit einer Rose] **dekoriert** werden ... In Versailles wurde [$_T$ das Hemd] [$_D$ mit feinen Spitzen] **dekoriert** ... [$_T$ Die Waggons] seien modernisiert und **erneuert** worden, und die Züge führen häufiger. Noch in diesem Jahr soll auch [$_T$ die Weberei in Rheine] **erneuert** werden. Nach und nach werden [$_T$ alle Zellen] **erneuert**, das Ensemble aber bleibt. Drumherum wurden [$_T$ Fassaden] **erneuert** ... [$_T$ Der Boiler] war vor kurzem **erneuert** worden. [$_T$ Die Stätte] wurde nach der Wende so gründlich **renoviert**, daß sie innen wie neu wirkt ... [$_T$ Das Gebäude] muß jedoch grundlegend **renoviert** werden. [$_T$ Die überwiegend traditionellen Häuser] ... sind ... [$_D$ mit viel Holz] **renoviert**. [$_T$ Diese Strecke] wird derzeit noch **renoviert**, in Kürze soll sie vollständig befahrbar sein. [$_T$ Die Uffizien in Florenz] werden **renoviert** und erweitert. Seither wurde [$_T$ die Wandmalerei in der Eingangshalle] **repariert** ... [$_T$ Die Brücke] ist in den letzten Jahren **repariert** worden ... ... [$_T$ die Telefonleitungen] müßten **repariert** ... werden. [$_T$ Die gebrochenen Deiche] seien nicht zu **reparieren** gewesen ... [$_T$ Das undichte Schieferdach] ... durfte ... [$_D$ mit Kunststoff-Schindeln] **repariert** werden.

Class 34: *Support*

Verbs: *dienen, folgen*$_1$*, helfen, unterstützen*

Ambiguity: *folgen* has a sense of 'to result' (class 42); in addition, *folgen* has a sense of 'to follow (in space or in time)'.

Scene: [$_S$ A supporter] helps [$_R$ somebody or a situation in need]. The scene might define [$_P$ a purpose] for the support.

Frame Roles: S(upporter), R(eceiver)

Modification Roles: P(urpose)

Levin class: -

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_S$	+ helfen / ¬ dienen, folgen, unterstützen
	Auch [$_S$ das Ausland] muß **helfen**:
	[$_S$ Jeder einzelne] kann aber ein bißchen **helfen** - und sei es nur als zufälliger Beobachter.
	[$_S$ Auch Postsäcke] **helfen** [$_P$ gegen die Flut].
	Da **helfen** [$_S$ auch die Informationsbroschüren] wenig ...
n$_S$ **a**$_R$	+ unterstützen / ¬ dienen, folgen, helfen
	[$_S$ Gaddafi] **unterstützt** [$_R$ die Idee] ...
	[$_S$ Deutschland] **unterstützt** [$_R$ beide Länder] [$_P$ bei der Heranführung] ...
	... [$_R$ die Rebellen in Ostzaire] zu **unterstützen**.
	... [$_S$ der] [$_R$ ihn] auch [$_P$ beim Bau eines Eigenheims] **unterstützen** will ...
	... aus Liebe **unterstützen** [$_S$ Tierschützer und Vegetarier] [$_R$ eine ... Kultur] ...
	... [$_S$ die Hausfrau] möchte [$_R$ den Aufstand] **unterstützen**.
	[$_S$ Sie] wolle [$_R$ das kleine und mittlere Unternehmertum] **unterstützen** ...
n$_S$ **a**$_R$ [P]	+ unterstützen / ¬ dienen, folgen, helfen
	[$_R$ Gotovac] wird [$_S$ von acht Parteien] **unterstützt** ...
	Mit dem Verkauf ... soll jetzt [$_R$ der Wiederaufbau] ... **unterstützt** werden.
	... [$_R$ dieses Konzept] wird [$_S$ von der Union] **unterstützt**.
	[$_R$ Der Staranwalt] wird **unterstützt** [$_S$ von dem ehemaligen Clinton-Berater] ...

Frame	Participating Verbs & Corpus Examples
$\mathbf{n}_S$ $\mathbf{d}_R$	+ dienen, folgen, helfen / ¬ unterstützen
	[$_S$ Nichtrostende Stähle] **dienten** [$_R$ der Wirtschaftlichkeit] ... [$_S$ Der Beamte] **dient** [$_R$ dem ganzen Volk, nicht einer Partei]. So **diente** [$_S$ er] einst [$_R$ seinem Landsmann Pedro Delgado] ... [$_S$ Die knapp einwöchige Reise] **dient** [$_R$ sowohl dem Besuch ... als auch zu Gesprächen] ... [$_S$ Wir] wollen ja [$_R$ den Kunden] **dienen** ... [$_S$ Sie] will nicht mehr [$_R$ der Flucht vor der Gegenwart] **dienen**. [$_S$ Der Dichter] **folgte** [$_R$ seinem poetischen Auftrag in Srebrenica]. ... will [$_S$ das Unternehmen] nicht [$_R$ der Strategie großer Konzerne] **folgen** ... [$_R$ Dem] **folgte** [$_S$ das Gericht] nicht ... Lieber **folge** [$_S$ ich] [$_R$ Magister Albinus ... als dem Leiter der Duden-Redaktion]. [$_S$ Giannotti] **folgt** [$_R$ Machiavelli] nicht darin ... ... also [$_R$ Göttern] zu **folgen**, von denen sie ahnen, daß sie keine sind. In der Sache **folgen** [$_S$ die Kommission und der Gerichtshof] ... [$_R$ der ... Forderung] ... [$_R$ Dem Standort Deutschland] könnten [$_S$ einzig Innovationen] **helfen**. ... um [$_R$ frierenden Menschen] aus den Wagen zu **helfen**. [$_S$ Er] wollte [$_R$ den arbeitenden Menschen] **helfen**; das Theoretisieren lag ihm nicht. Vielleicht **helfen** [$_S$ diese Erläuterungen] [$_R$ dem Ahnungslosen] ...
$\mathbf{n}_S$ $\mathbf{d}_R$ [P]	+ dienen, folgen, helfen / ¬ unterstützen
	[$_R$ Dem] werde durch neue Behörden nicht **gedient**. Mit dem Beharren auf 3,0 Prozent werde nicht [$_R$ der Währungsstabilität] **gedient** ... ... daß vorrangig [$_R$ den Interessen des Großkonzerns] **gedient** werde. Daß [$_R$ der ... amerikanischen Studie] ... nicht **gefolgt** worden sei ... ... wie [$_R$ dem Verbrecher und "Führer" Hitler] **gefolgt** wurde ... ... daß [$_R$ der reformierten Rechtschreibung] **gefolgt** werde ... [$_R$ Den neuen Bundesländern] muß in der Tat noch lange **geholfen** werden. Dabei soll vor allem [$_R$ den Zulieferern des Stahlproduzenten] **geholfen** werden ... ... daß zur Zeit niemand zu sagen vermag, wie [$_R$ dem Land] **geholfen** werden kann.
$\mathbf{n}_S$ $\mathbf{p}_P$:Nom.vgl	+ dienen / ¬ helfen, folgen, unterstützen
	... solange ihnen [$_S$ der Negativismus] ... [$_P$ als Vorwand] **dient**. [$_S$ Ein Hotel] **diente** jahrzehntelang [$_P$ als Tarnung für einen Atombunker der Regierung]. [$_S$ Sie] **diente** seit der Kolonisierung der Kapkolonie 1652 [$_P$ als Exilplatz] ... [$_S$ Eine Walzerszene] **dient** [$_P$ als Hintergrund für ein Tête-à-tête-Picknick]. ... [$_P$ als Hilfe zur Umstrukturierung sowie zur betrieblichen Fortbildung] **dienen**.
$\mathbf{n}_S$ $\mathbf{p}_P$:Dat.zu	+ dienen / ¬ helfen, folgen, unterstützen
	[$_S$ Diese Algorithmen] **dienen** [$_P$ zur Simulation der Evolution]. ... ob [$_S$ Klonen] [$_P$ dazu] **dienen** könnte, [$_P$ kinderlosen Paaren zu helfen]. Offiziell **dient** [$_S$ ein Sprachproblem] [$_P$ zur Begründung]: [$_S$ Das Herbizid] **diene** [$_P$ zur Bekämpfung von Schadgräsern im Nachauflauf] ...
$\mathbf{n}_S$ $\mathbf{i}_P$	+ helfen / ¬ dienen, folgen, unterstützen
	... denn [$_S$ überdurchschnittliche Löhne] **helfen** auch, ... [$_P$ Arbeitskräfte zu gewinnen] ... [$_S$ Ein Begleitbuch] **hilft**, [$_P$ die unübersichtliche Fülle der Wortmeldungen zu ordnen] ... [$_S$ Die Tiere] sollen **helfen**, [$_P$ die ... Erbanlagen zu identifizieren]. [$_S$ Es] **hilft**, [$_P$ die Kinder von der Straße zu holen] ...
$\mathbf{n}_S$ $\mathbf{d}_R$ $\mathbf{i}_P$	+ helfen / ¬ dienen, folgen, unterstützen
	[$_S$ Ich] habe [$_R$ ihr] **geholfen**, [$_P$ aus dem Wrack zu klettern] ... ... [$_R$ ihrem begleitenden Partner] zu **helfen**, [$_P$ auch einen Job zu finden]. ... um [$_R$ den Menschen] zu **helfen**, [$_P$ die Welt verwandelt zu sehen] ...!

Class 35: *Quantum Change*

Verbs: *erhöhen, erniedrigen, senken, steigern, vergrößern, verkleinern*

Ambiguity: *erniedrigen* has a (related) sense of 'to humiliate'.

Scene: [$_C$ Somebody or something] causes an increase or decrease of [$_T$ something]. The context might mention the [$_M$ magnitude of the change], or [$_S$ the starting or resulting state].

Frame Roles: C(hanger), T(heme)

Modification Roles: M(agnitude), S(tate)

Levin class: 45.6 *(Verbs of Change of State → Verbs of Calibratable Changes of State)*

Schumacher class: 3.2/3 *(Verben der Differenz → Einfache Änderungsverben / Kausative Änderungsverben)*

Frame	Participating Verbs & Corpus Examples
n$_C$ **a**$_T$	+ erhöhen, erniedrigen, senken, steigern, vergrößern, verkleinern
	Wenn [$_C$ wir] [$_T$ die Sozialbeiträge] nicht **erhöhen** wollen, bleiben die Mittel begrenzt.
	Ob [$_C$ man] denn erst [$_T$ die Erbschaftsteuer] [$_M$ kräftig] **erhöhen** ...
	[$_C$ Die Deutsche Bahn AG] **erhöht** zum 1. April [$_T$ die Fahrpreise im Personenverkehr].
	[$_C$ Das] kann [$_T$ die Dichte] **erhöhen**, auf Kosten der Vielfalt.
	Im zweiten Drittel **erhöhten** [$_C$ die Adler] [$_T$ den Druck] ...
	... daß [$_C$ sie] [$_T$ die Spannungen an beiden Drähten] [$_M$ leicht] **erniedrigten**.
	Als [$_C$ die Wissenschaftler] [$_T$ die Temperatur] [$_S$ auf minus 270 Grad] **erniedrigten** ...
	... [$_C$ man] müsse [$_T$ den Taglohn] [$_S$ auf fünfzehn Sous] **erniedrigen**.
	[$_C$ Die Kulturdeputation] aber hat ... [$_T$ den Betrag] ... [$_S$ auf 250.000 Mark] **erniedrigt**:
	[$_C$ Air France] will [$_T$ Kosten] **senken**.
	... ob [$_C$ die Notenbank] [$_T$ die Zinsen] weiter **senken** wird oder nicht.
	Deutlich **senken** will [$_C$ der neue Vorstandsvorsitzende] [$_T$ den Materialaufwand] ...
	... daß [$_C$ Prävention] [$_T$ die Ausgaben für medizinische Behandlung] **senke**.
	... wonach [$_C$ das Aufsagen niederschmetternder Texte] [$_T$ die Schmerztoleranz] **senkt** ...
	[$_C$ Braun] **steigerte** die Beschäftigtenzahl [$_M$ um etwa 200] [$_S$ auf knapp 23000 Mitarbeiter].
	... würde [$_C$ dies] [$_T$ ihr Selbstvertrauen] noch **steigern**.
	[$_C$ Man] will [$_T$ die Produktivität] **steigern** ...
	Auch [$_C$ der Verband] hat [$_T$ seinen Anteil] **gesteigert**.
	Um [$_T$ den Wert] aus Sicht des Anlegers zu **steigern** ...
	Macht man sie darauf aufmerksam, **vergrößern** [$_C$ sie] [$_T$ die Schritte] ...
	... [$_C$ die Streitkräfte] haben [$_T$ ihren Einfluß] **vergrößert** ...
	... [$_C$ sein Unternehmen] wolle nicht um jeden Preis [$_T$ das Volumen] **vergrößern** ...
	... nämlich [$_T$ den Kreis der Zwangsversicherten] zu **vergrößern** ...
	[$_C$ Neuzugänge] **vergrößern** [$_T$ die Vielfalt an Branchen] ...
	[$_C$ Das daraus resultierende elektrische Feld] **verkleinert** [$_T$ die Schwelle] ...
	Um [$_T$ den Berg der gesamten Staatsschuld] zu **verkleinern** ...
	Nur so besteht Hoffnung, [$_T$ die Populationen] dauerhaft zu **verkleinern**.
	[$_C$ Die Deutsche Börse] **verkleinert** [$_T$ die Kontraktgrößen für Aktienoptionen] ...

Frame	Participating Verbs & Corpus Examples
n$_C$ **a**$_T$ [P]	+ erhöhen, erniedrigen, senken, steigern, vergrößern, verkleinern
	Soll also [$_T$ das Haushaltsdefizit] **erhöht** werden? ... daß [$_T$ die Rentenbesteuerung] bereits 1995 **erhöht** wurde. ... daß [$_T$ der Marktanteil des Airbus] ... **erhöht** werden müßte. [$_T$ Die Investitionszulagen] würden [$_M$ deutlich] **erhöht**. Dazu soll einerseits [$_T$ das Kindergeld] **erhöht** werden ... Wird [$_T$ die Unfallgefahr] auch wieder **erniedrigt**? Regierung und Opposition wollen, daß [$_T$ die Steuersätze] **gesenkt** werden ... ... daß [$_T$ die Fangquoten] immer weiter **gesenkt** werden. [$_T$ Die Bodenpreise in Deutschland] müßten jedoch [$_M$ erheblich] **gesenkt** werden ... ... daß [$_T$ der Preis für einfache Arbeit] so **gesenkt** wird ... Daß [$_T$ die Zahl der 1,8 Millionen Gäste] rasch **gesteigert** werden kann, gilt als sicher. Bei Produktionssteigerung wurde zuerst [$_T$ die Intensität] **gesteigert** ... [$_T$ Die Kapazitäten sollen] **gesteigert**, Gastronomie und Tourismus erschlossen werden. ... konnte [$_T$ das Konzernergebnis] [$_M$ beträchtlich] **gesteigert** werden. Im Ergebnis würden [$_T$ die Verlustvorträge] **vergrößert** ... So wurde [$_T$ Jerusalem] [$_M$ deutlich] **vergrößert**. [$_T$ Der Profit] werde künstlich **vergrößert** ... Zudem ist [$_T$ die Gruppe] ... [$_S$ auf acht Tiere unterschiedlichen Alters] ... **vergrößert** worden ... Mit dem Kristall wird ... [$_T$ der Abstand zwischen Linse und Scheibe] **verkleinert**. Nun wird [$_T$ das Filialnetz] drastisch **verkleinert**; [$_T$ Das Angebot] wurde [$_S$ von 3000] [$_S$ auf 1200 Artikel] **verkleinert**. Wie mitgeteilt wird, soll [$_T$ der Vorstand] **verkleinert** werden;
n$_T$ **r**	+ erhöhen, senken, steigern, vergrößern, verkleinern / ¬ erniedrigen
	Zudem **erhöhe** sich [$_T$ der Verhandlungsaufwand bei Kliniken und Kassen] [$_M$ immens]. [$_T$ Die Grunderwerbsteuer] **erhöht** sich [$_S$ von 2] [$_S$ auf 3,5 Prozent]. [$_T$ Ihr Durchschnittspreis] **erhöhte** sich 1996 [$_M$ um etwa acht Prozent] [$_S$ auf 1500 Franken]. ... **erhöhte** sich im vergangenen Jahr [$_T$ der Erdgas-Import aus den Niederlanden] ... ... wenn am späten Nachmittag [$_T$ die Dämmerung] sich [$_S$ über das Hochtal] **senkt**. [$_T$ Die Nacht] **senkt** sich [$_S$ über die schwedische Wohnungspolitik]. [$_T$ Stagnation und Repression] **senkten** sich wie Mehltau [$_S$ über das geistige ... Leben]; [$_T$ Die Peinlichkeitsschwellen] haben sich **gesenkt** ... Gleichzeitig **steigert** sich [$_T$ die Helligkeit des roten Planeten] [$_M$ immer mehr] ... ... [$_T$ die] sich erst im Auge gegenseitig **steigerten**. Ein Rennläufer, [$_T$ der] sich im zweiten Durchgang **steigern** kann, gilt als hoffnungsvoll. [$_T$ Die Eintracht] **steigerte** sich aber im weiteren Verlauf ... Steigt ihre Zahl, **vergrößert** sich auch [$_T$ die Aufnahmekapazität]. Die Menschen sind grundrechtsbewußter geworden, [$_T$ die Bundesrepublik] hat sich **vergrößert**. Gleichzeitig **vergrößert** sich [$_T$ der Abstand zwischen Erde und Mond] ... [$_T$ Die Internet-Gemeinde] **vergrößert** sich nach Expertenschätzungen ... [$_T$ Die Dauerausstellung der Sammlung] **verkleinert** sich in absehbarer Zeit ... ... "Titanschrumpfschild", mit dem [$_T$ er] sich [$_M$ so weit] **verkleinert** ... ... [$_T$ die Belegschaft] **verkleinerte** sich [$_M$ drastisch]. Insgesamt hätten sich [$_T$ die Zugangsmöglichkeiten] nicht **verkleinert** ...

Class 36: *Opening*

Verbs: *öffnen*, *schließen*$_1$

Ambiguity: *schließen* has a sense of inference (class 41); in addition, *schließen* has a specific sense of 'to end' (such as books, or talks), and of 'to make, to arrange' (such as friendships, contracts).

Scene: [$_A$ Somebody or certain circumstances] changes the opening status of [$_T$ something].

Frame Roles: A(ctor), T(heme)

Levin class: 45.4 *(Verbs of Change of State → Other Alternating Verbs of Change of State)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
$\mathbf{n}_A$ $\mathbf{a}_T$	+ öffnen, schließen
	Aber das bedeutet nicht, daß [$_A$ ich] [$_T$ die Wundertüte] **öffne**.
	Die Großmutter unterbricht ihre Darstellung, als [$_A$ eine Frau] [$_T$ die Tür] **öffnet**.
	Dann **öffnet** [$_A$ das Deutsche Sportmuseum] [$_T$ seine Pforten].
	[$_A$ Sein Abgang] **öffnet** [$_T$ die Tür zur Shoah]:
	[$_A$ Das] muß oft bitter gewesen sein, **öffnete** aber auch [$_T$ die Möglichkeit] ...
	Es wurde vereinbart, [$_T$ weitere Flughäfen] in beide Entitäten zu **öffnen** ...
	[$_T$ Das Fenster] **schließen** - [$_T$ die Tür] **öffnen**?
	[$_A$ Die Händler in Hebron] **schlossen** [$_T$ ihre Läden].
	[$_A$ Der Pfarrer] ließ die Tore **schließen**, als der Kirchenraum voll besetzt war.
	[$_A$ Wir] wollen [$_T$ keine Türen] endgültig **schließen**.
	Erleichtert **schließt** [$_A$ die Wirtin] hinter uns [$_T$ die Tür].
	[$_A$ Sie] **schlossen** [$_T$ die Königliche Akademie], wo das Khmer-Drama gelehrt wurde ...
	[$_A$ Der alte Mann] nimmt einen Zug, **schließt** [$_T$ die Augen] ...
$\mathbf{n}_A$ $\mathbf{a}_T$ [P]	+ öffnen, schließen
	Die Bomben sollten explodieren, wenn [$_T$ die Weihnachtskarte] **geöffnet** würde.
	Am Donnerstag wurde [$_T$ ein weiterer Polder] **geöffnet**.
	[$_T$ Das Wertpapiergeschäft] soll für neue Anbieter **geöffnet** ... werden.
	... da [$_T$ viele Kindergärten] nur vier Stunden **geöffnet** seien;
	Da [$_T$ die Wohnung des mutmaßlichen Mörders] nicht **geöffnet** wurde ...
	Manche Kästchen verbergen gut, daß [$_T$ sie] zu **öffnen** sind. [$_T$ Viele] sind perfekt zu **schließen**.
	... wenn [$_T$ das ausführende Institut] **geschlossen** wird ...
	[$_T$ Die Liste wurde] bereits 1960 **geschlossen**, und 1996 wurden nur fünf Namen gestrichen.
	[$_T$ Dieser Abstand] muß **geschlossen** werden, und [$_T$ er] muß unverzüglich **geschlossen** werden.
	In der Bodenbelagsparte werde [$_T$ kein Betrieb] **geschlossen** ...
	Als der letzte sein Kärtchen ... abgeholt hat, werden [$_T$ die Türen] **geschlossen** ...
	Die Haltung unnatürlich, [$_T$ die Augen] oft **geschlossen** ...
	[$_T$ Das Frankfurter Büro] war zuvor **geschlossen** worden.
$\mathbf{n}_T$ $\mathbf{r}$	+ öffnen, schließen
	Mühelos **öffneten** sich für Tresko [$_T$ alle Türen] ...
	Und immer [$_T$ neue Folianten] **öffnen** sich voll spätgotischer Buchstaben ...
	[$_T$ Die Tür] **öffnet** sich erneut.
	... bevor sich [$_T$ die Blütenknospe] **öffnet**.
	... als sich [$_T$ die Grenzen] **öffneten** ...
	Unlängst erst **öffneten** sich [$_T$ die Archive der ehemaligen Sowjetunion].
	... daß das Tor, [$_T$ das] sich so unerwartet **geöffnet** hatte, wieder zufallen könnte ...
	Erst allmählich ... **öffnen** und **schließen** sich [$_T$ Formationen und Bildräume] ...
	... in denen [$_T$ Räume] sich **öffnen** und **schließen** ...
	[$_T$ Diese Lücke] ist jetzt mit Hipparcos **geschlossen** worden.
	Manchmal **schließt** sich [$_T$ der Kreis] in einem Spiel ...
	[$_T$ Die letzte dieser Öffnungen] **schloß** sich vor etwa 4,6 Millionen Jahren ...
	[$_T$ Die Schere zwischen den alten und neuen Ländern] hat sich also nicht **geschlossen** ...
	Doch [$_T$ die Haustür] hat sich schon wieder **geschlossen**.

Class 37: *Existence*

Verbs: *bestehen$_2$, existieren, leben*

Ambiguity: *bestehen* also refers to (i) an insistence (cf. class 28) and (ii) the consistence of something; (iii) in addition, it has a sense of passing an exam.

Scene: [$_E$ Somebody or something] does exist.

Frame Roles: E(xistence)

Levin class: 47.1 *(Verbs of Existence → Exist Verbs)*

Schumacher class: 1.1 *(Verben der allgemeinen Existenz → Zustandsverben der allgemeinen Existenz)*

Frame	Participating Verbs & Corpus Examples
n$_E$	+ bestehen, existieren / +$_{adv}$ leben
	Allerdings **bestünden** noch [$_E$ diplomatische Beziehungen] ...
	In der Nähe des Tunnels **besteht** nach Angaben der Behörden weiterhin [$_E$ Lawinengefahr].
	Am Erfolg einer Zwangstherapie **bestünden** [$_E$ erhebliche Zweifel].
	Weder für die Post noch für die Bank **bestand** [$_E$ die Möglichkeit, aus dem gemeinsamen Boot auszusteigen].
	Für die Bundesbank **bestehe** gegenwärtig [$_E$ kein Handlungsbedarf].
	Zwischen Gesellschaft und Täter **besteht** [$_E$ ein höchst komplexes Zusammenwirken von tausend Faktoren].
	[$_E$ Ein Zusammenhang] **bestehe** daher nicht.
	[$_E$ Ein größeres Ensemble] **existiert** nicht mehr ...
	Es **existiert** [$_E$ die Meinung, daß man das Gemälde Malewitschs in alle großen Museen ... schicken solle].
	Es sollen [$_E$ nur noch Republiken] **existieren** ...
	Sogar [$_E$ eine Zeichnung] **existiert** von diesem Superwesen.
	[$_E$ Die Neue Welt] **existiert** von jeher in schroffen Gegensätzen und harten Spannungen ...
	[$_E$ Sarah] hatte die seltene Gabe zu **existieren**, ohne zu **leben** ...
	[$_E$ 400 jüdische Siedler] **leben** hier unter 100000 Muslimen.
	[$_E$ Wir] wollen alle in Frieden und Versöhnung **leben**.
	[$_E$ Die Fixer] **leben** in der Gosse ...
	[$_E$ Eule] **lebt** für seine Jungs ...
	[$_E$ Die Bewohner] **leben** von Muschel- und Fischfang, von der Lachszucht.
	... wo [$_E$ der ehemalige Präsident der bosnischen Serben] **lebt**.
	151 Jahre lang **lebte** [$_E$ Ulm] nach den Regeln des Schwörbriefs und blühte auf.

Class 38: ***Consumption***

Verbs: *essen, konsumieren, lesen, saufen, trinken*

Scene: [$_C$ A living entity] consumes [$_T$ something]. The consumption might be described by [$_M$ a certain manner], or with [$_R$ a specific result].

Frame Roles: C(onsumer), T(heme)

Modification Roles: M(anner), R(esult)

Levin class: 39.1/4 *(Verbs of Ingesting → Eat Verbs / Devour Verbs)*
The English pendant verb of *lesen* is not included in the respective Levin classes.

Schumacher class: 7.2 *(Verben der vitalen Bedürfnisse → Verben der Konsumation)*
All verbs except for *lesen* are classified.

Frame	Participating Verbs & Corpus Examples
n$_C$	+ essen, konsumieren, lesen, saufen, trinken
	... ob [$_C$ sie] lieber heizen oder **essen** wollte ... Doch **essen** muß [$_C$ man] immer. Auch im Himmel muß [$_C$ man] schlafen und **essen** ... [$_C$ Er] **ißt** und **trinkt**, wäscht und rasiert sich, wenn sie es will ... In der Vergangenheit haben [$_C$ wir] immer nur **konsumiert** ... Die Japaner zahlen hohe Mieten, [$_C$ sie] **konsumieren**, sie sichern Arbeitsplätze ... [$_C$ Wer] heute **konsumiert**, trägt geringere Steuerlasten ... Mit Staunen **liest** [$_C$ man] [$_T$ von der Existenz eines umfänglichen Handbuches] ... Doch [$_C$ der] **liest**, durchaus verständlich, lieber in dieser Zeitung ... ... daß [$_C$ er] lieber Musik hört und **liest** ... [$_C$ Sie] küssen und sie schlagen sich post festum, **saufen** und raufen ... [$_C$ Dessen Papa] **säuft**, Mama ist plemplem und geht ins Wasser ... [$_C$ Die] haben dort zusammen **gesoffen** und Spektakel gemacht ... [$_C$ Er] **trank** in Maßen, dafür **soff** [$_C$ die Frau]. [$_C$ Wer] **trinkt** jetzt noch, um seine Sorgen zu vergessen? Er rauchte nicht, [$_C$ er] **trank** nicht.
n$_C$ [P]	+ essen, konsumieren, lesen, saufen, trinken
	Während der Aufführung wurde **gegessen**, mehr aber noch **getrunken** ... Wo aber wurde **gegessen** - und was? ... in den Gaststätten weniger **gegessen** wurde ... Wenn im Inland nicht **konsumiert** wird, so sagen die Ökonomen, kann die Rezession nicht weichen. Da wird **gelesen**, disputiert und gefeiert. Im Herbst wird wieder **gelesen** ... Es wird geschubst und gedrängelt, geschmatzt und **gesoffen**. ... wo **gesoffen**, gegrölt und Schwachsinn geredet wird ... Da wird gefickt und gefressen, **gesoffen** und gerülpst. Wenn früher [$_R$ bis zur Besinnungslosigkeit] **gesoffen** wurde ... Nach der Eröffnung wurde vor der Akademie ... gefeiert, getanzt, gelacht und **getrunken**. Erst wird heftig **getrunken**, dann kommt die Prohibition ...

Frame	Participating Verbs & Corpus Examples
n$_C$ **a**$_T$	+ essen, konsumieren, lesen, saufen, trinken
	Am Samstag **aß** [$_C$ der Rabbi] [$_T$ nichts] ... Aber vorher **ißt** [$_C$ Zach] [$_T$ einen Hot dog]. ... daß heute [$_C$ ein Chinese] im Jahr [$_T$ ein Kilogramm Rindfleisch] **esse** ... [$_C$ Er] weiß, daß er auch Kohlehydrate braucht, und **ißt** [$_T$ Pfannkuchen]. [$_C$ Ein Benzinmotor] hätte [$_T$ etwa ein Drittel mehr] **konsumiert**. ... daß [$_C$ Abhängige] [$_T$ mitgebrachtes Rauschgift] **konsumierten** ... [$_C$ Die Privathaushalte] **konsumieren** per saldo [$_T$ mehr, als sie einnehmen]. [$_C$ Der Träumer] **konsumiert** [$_T$ den Traum anstelle der erträumten Dinge]. Daher **liest** [$_C$ man] in ihren Blättern [$_T$ eher negativ gefärbte Berichte] ... [$_C$ Er] **liest** [$_T$ die Gedanken der Menschen seiner Umgebung] ... [$_C$ Vier von fünf Deutschen] **lesen** [$_T$ Zeitung]. [$_C$ Pferde], die laufen, müssen auch [$_T$ was] zu **saufen** haben. ... und [$_C$ er] **säuft** [$_T$ Kellers Wein]. Kette rauchend und [$_T$ Whiskey] **saufend** ... [$_C$ Er] will [$_T$ seinen Cappuccino] wirklich **trinken** ... [$_C$ Wer] nichts riskiert, wird auch [$_T$ keinen Champagner] **trinken**. In der Not **trinkt** [$_C$ der Hamborner] [$_T$ das Bier] auch aus der Flasche. ... wo [$_C$ sie] [$_T$ Schokolade] **trinkt** und die Herren der Schöpfung beobachtet.
n$_C$ **a**$_T$ [P]	+ essen, konsumieren, lesen, saufen, trinken
	Wenn es kalt ist, wird [$_T$ mehr] **gegessen** und **getrunken**. ... wo [$_T$ die Kringel] in Schokolade getaucht und dann **gegessen** werden. ... heute werde [$_T$ zwischen zwei und vier Jahre alter Käse] am liebsten **gegessen**. Selbst [$_T$ Waschpulver] wird mitunter **gegessen** ... Denn [$_T$ die Zuschüsse aus Brüsseler Fleischtöpfen] können **konsumiert** werden ... ... sofern [$_T$ die Produkte] in ihren Provinzen **konsumiert** werden. Während im Umlageverfahren [$_T$ die Einnahmen] sofort **konsumiert** werden ... Wenn [$_T$ kein Ecstasy] mehr **konsumiert** wurde, seien die Depressionen nicht mehr aufgetreten. Häufig werden [$_T$ solche Konvolute] dann kaum **gelesen** und noch weniger befolgt. ... müßten doch [$_T$ die Akten] ausführlich **gelesen** und bewertet werden. Nun wurde [$_T$ es] **gelesen** und von prominenten Kritikern besprochen. [$_T$ Die Geschichte] kann in drei Sprachen ... gehört und **gelesen** werden. ... aber bei uns wurde eben zuviel Wasser gepredigt und [$_T$ Wein] **gesoffen**. Da in der Region meist [$_T$ Rotwein] **getrunken** werde ... Zu Rod Powells Country-music wird [$_T$ Bier] **getrunken** ... [$_T$ Liköre] werden eher [$_C$ von Älteren] **getrunken**.
n$_C$ **r**	$+_{adv}$ essen, saufen, trinken / ¬ konsumieren, lesen
	... wenn [$_C$ er] sich [$_T$ daran] [$_R$ satt] **essen** durfte. Wie sich [$_C$ die Aborigines] hingegen im Dschungel [$_R$ gesund] **aßen** ... Kurz darauf **säuft** [$_C$ der Vater] sich endgültig [$_R$ zu Tode] ... ... **säuft** [$_C$ der Bräutigam] sich am Vorabend ... [$_R$ so richtig die Hucke voll]. [$_C$ Die letzten Mucker und Philister] **saufen** sich drei Tage lang [$_R$ bewußtlos] ... ... [$_C$ mein Blick] **trank** sich [$_R$ satt] an den sich sanft hinziehenden Hügeln ...
n$_T$ **r**	$+_{adv}$ konsumieren, lesen / ¬ essen, saufen, trinken
	Von Freunden ... läßt sich [$_T$ der Film] etwa so **konsumieren** [$_M$ wie ein Fußballmatch] ... [$_M$ Genauso eingängig, wie] sich [$_T$ seine Erfahrungsberichte] **konsumieren** lassen ... [$_T$ Seine Webart] **liest** sich [$_M$ schwierig], ist aber denkbar simpel. [$_T$ Airframe] **liest** sich [$_M$ wie ein Fall für den Verbraucheranwalt]. Nunmehr **liest** [$_T$ es] sich [$_M$ anders, aber noch nicht richtig].

Class 39: *Elimination*

Verbs: *eliminieren, entfernen, exekutieren, töten, vernichten*

Ambiguity: The reflexive usage of *entfernen* means ‘to move away’.

Scene: [$_E$ Somebody or some circumstances] destroys the existence of [$_D$ something].

Frame Roles: E(liminator), D(estroyed)

Levin class: 42.1 *(Verbs of Killing → Murder Verbs)*

Schumacher class: 1.3 *(Verben der allgemeinen Existenz → Kausative Verben der allgemeinen Existenz)*
Only *vernichten* is classified.

Frame	Participating Verbs & Corpus Examples
n$_E$ **a**$_D$	+ eliminieren, entfernen, exekutieren, töten, vernichten
	... Wahrheiten, [$_D$ die] [$_E$ er] systematisch aus seiner Biographie **eliminiert** hatte.
	Denn [$_E$ das statistische Bereinigungsverfahren] **eliminiere** nur [$_D$ Witterungseinflüsse] ...
	Eliminiert [$_E$ man] [$_D$ den Wechselkurseffekt der Dollar-Aufwertung] ...
	[$_D$ Die Feinde des Volkes] zu **eliminieren** ist keine Straftat.
	[$_E$ Die Terrormaschinerie der Bolschewiken] **eliminierte** [$_D$ die Fähigsten] ...
	... indem [$_E$ er] [$_D$ die verrotteten Teile] **entfernte** ...
	[$_E$ Soldaten mit Schutzmasken] **entfernten** [$_D$ die defekten Container] ...
	[$_E$ Wer] mir [$_D$ mein Ohr] **entfernt** hat, möchtet Ihr wissen?
	[$_E$ Der Göttinger Botaniker] **entfernte** bei zahlreichen Blüten [$_D$ die Staubfäden] ...
	... Geschwür, [$_D$ das] [$_E$ man] mit chirurgischen Methoden **entfernen** könne ...
	Anstatt [$_D$ das Material] aus einem massiven Kristall zu **entfernen** ...
	Es sei nicht ihre Absicht, [$_D$ Gefangene] zu **exekutieren** ...
	[$_E$ Reiter] **exekutiere** [$_D$ den politischen Willen einiger Landesregierungen] ...
	... desto braver **exekutiert** [$_E$ sie] [$_D$ die konventionellen Arrangements].
	Von der Großzügigkeit, mit der [$_E$ Strauss] selber [$_D$ seine Partituren] **exekutierte** ...
	Doch gerade in ihrer Alltäglichkeit **exekutieren** [$_E$ sie] [$_D$ die Geschichte].
	[$_E$ Das Gedanken-Ich] **tötet** [$_D$ Josef].
	Nach dem Eintreffen der alarmierten Polizei drohte [$_E$ er] [$_D$ sich] zu **töten** ...
	[$_E$ Sie] entführen Manager von Teeplantagen und **töten** [$_D$ Elefanten und Nashörner] ...
	Man zwingt [$_E$ sie], [$_D$ ihren Vater] zu **töten**.
	... um [$_D$ seine wohlhabende Ehefrau und seine vier Kinder] zu **töten** ...
	[$_E$ Aids] **tötet** [$_D$ Kinder in Entwicklungsländern] ... schneller als in den Industriestaaten ...
	Dürfen [$_E$ wir] am Kunstwerk [$_D$ Zeit] **vernichten** ...?
	[$_E$ Die Araber] versuchten sofort, [$_D$ sie] zu **vernichten**, und scheiterten.
	Alle fünfzig Jahre versuchen [$_E$ die Russen], [$_D$ uns] zu **vernichten**.
	[$_E$ Viele] **vernichteten** [$_D$ ihre Ausweispapiere] ...
n$_E$ **a**$_D$ [P]	+ eliminieren, entfernen, exekutieren, töten, vernichten
	Erst in Landnähe ... kann [$_D$ der Fehler] **eliminiert** werden.
	[$_D$ Alle Reste der spätgotischen und barocken Bauphase] wurden **eliminiert**.
	... daß [$_D$ die Inflation] praktisch **eliminiert** wurde.
	Dabei wird auch [$_D$ deren Erbgut] **eliminiert**.
	Das signalisiert der Zelle, daß [$_D$ diese Stoffe] zu **entfernen** sind.
	... selbst wenn [$_D$ die Geschwülste] zuerst vollständig **entfernt** wurden.
	An 200 Kranken, denen [$_D$ ein Stück Darm] **entfernt** werden mußte ...
	Als bei Umbauarbeiten dann noch [$_D$ Wände] **entfernt** wurden, stürzte das Gebäude ein.
	Allerdings wirkt es sich nachteilig aus, daß auch [$_D$ das Pigmentepithel] **entfernt** wird.
	... so daß zumindest [$_D$ der gröbste Dreck] **entfernt** wird.
	[$_D$ Ein Mann] wurde in Riad nach dem Freitagsgebet **exekutiert**.
	... bevor [$_D$ er] auf offener Szene **exekutiert** wird ...
	... wonach [$_D$ Kasi] einen "Hirnschaden" habe und deshalb nicht **exekutiert** werden solle ...
	... daß [$_D$ die Nachbarn] in ihrem Haus **exekutiert** wurden ...
	Bei Schießereien wurde [$_D$ ein Mann] **getötet** ...
	Zu Recht, denn durch den Eingriff wird aus ihrer Sicht [$_D$ der Spender] **getötet**.
	Dabei wurde [$_D$ sie] von einer S-Bahn erfaßt und **getötet**.
	[$_D$ Die mit Taschenlampen geblendeten Kaimane] waren ... leicht zu **töten**.
	[$_D$ Die sichergestellten Waren] sollen **vernichtet** werden.
	... Einrichtungen, in denen [$_D$ chemische Kampfstoffe] gelagert und **vernichtet** werden ...
	... [$_D$ was] durch den Nationalpark an Arbeitsplätzen **vernichtet** wurde ...
	Er galt als überführt, [$_D$ sein Ruf] war **vernichtet**.

Class 40: *Basis*

Verbs: *basieren, beruhen, gründen, stützen*

Ambiguity: *gründen* has a sense of 'to found, to establish'. *stützen* has a related sense of 'to sustain'.

Scene: [$_T$ Something] refers to [$_B$ a basis]. There might be [$_S$ somebody who defines the basis].

Frame Roles: T(heme), B(asis)

Modification Roles: S(etter)

Levin class: -

Schumacher class: 4.5 *(Verben der Relation und des geistigen Handelns* $\rightarrow$ *Verben der Grundlegung)*

Frame	Participating Verbs & Corpus Examples
n$_T$ **p**$_B$:Dat.auf	+ basieren, beruhen, gründen / ¬ stützen
	Technisch **basiert** [$_T$ der Lupino] [$_B$ auf einer verkürzten Plattform des Polo] ... [$_T$ Der Qualifizierungsweg] ... **basiert** [$_B$ auf ... Grundveranstaltungen] ... [$_T$ Das Stück] **basiert** [$_B$ auf Tatsachen]: Der Text, [$_T$ der] [$_B$ auf einer älteren Übersetzung] **basiert**, wurde zur Grundlage ... [$_T$ Die neuen Demokratien] ... **basieren** nicht [$_B$ auf Systemen] ... [$_T$ Die Grundstückswerte] **basieren** [$_B$ auf Bodenrichtwerten] ... ... [$_T$ das] [$_B$ auf der Orthographie des BGB] **basierte** ... [$_T$ Alle Berichte] **basieren** offenbar [$_B$ auf Gesprächen mit Mitarbeitern] ... Aber [$_T$ dieses Argument] **beruht** [$_B$ auf einem Irrtum]. ... [$_T$ es] soll [$_B$ auf Tatsachen] **beruhen** ... ... da [$_T$ sein Wert] nicht [$_B$ auf Volumen oder Gewicht] **beruhe**. [$_T$ Das solide Wirtschaftswachstum] **beruht** ... vor allem [$_B$ auf der Landwirtschaft] ... ... Prinzipien, [$_B$ auf denen] [$_T$ jenes ideale Gemeinwesen] **beruht**. [$_T$ Das Verfahren] **beruht** im Prinzip [$_B$ auf der Photoionisation] ... In diesem Fall jedoch **beruht** [$_T$ die Distanz] ... [$_B$ auf einer Entscheidung] ... [$_T$ Dieses französische Modell] **beruht** [$_B$ auf der Partizipation der Staatsbürger] ... [$_T$ Die Leistungsfähigkeit der Nato] **beruhe** [$_B$ auf der Kraft Amerikas]; [$_T$ Sie] **beruht** [$_B$ darauf, daß das Vakuum nicht wirklich "leer" ist]. [$_T$ Sein OEuvre] **gründet** nämlich [$_B$ auf seinen freundschaftlichen Kontakten] ... ... Staatsbürgerdemokratie, [$_T$ die] [$_B$ auf einem einheitlichen Raum] ... **gründet** ... ... **gründet** [$_T$ der Reiz der Reise im Netz] [$_B$ auf den Verlust von Normalität] ... [$_T$ Die europäische Gemeinschaft] **gründet** letztlich [$_B$ auf der ... Tradition] ... ... [$_T$ sein neuer Ruhm] **gründet** [$_B$ auf dem ... herrlichen Garten] ... ... Mangels, [$_B$ auf dem] [$_T$ der Kommunismus] nachher **gründen** sollte.
n$_S$ **a**$_T$ **p**$_B$:Akk.auf	+ gründen, stützen / ¬ basieren, beruhen
	[$_B$ Auf drei Pfeiler] **gründet** [$_S$ jeder der Spitzenkandidaten] [$_T$ seinen Wahlkampf]: [$_S$ Sie] **stützt** [$_T$ ihre Beschuldigungen] [$_B$ auf persönliche Aufzeichnungen Ekincis] ... Denn [$_S$ die Rezensentin] **stützt** [$_T$ ihr Urteil] ... [$_B$ auf unvollständige Zitate] ... [$_S$ Sie] **stützen** [$_T$ ihre Prognosen] [$_B$ auf eine Reihe von Studien] ...
n$_T$ **r** **p**$_B$:Akk.auf	+ gründen, stützen / ¬ basieren, beruhen
	[$_T$ Unsere Weltzivilisation] **gründet** sich [$_B$ auf dieses Werkzeug]. [$_T$ Die Ansprüche] **gründen** sich [$_B$ auf Recherchen in Akten und Archiven] ... ... [$_B$ worauf] [$_T$ alle Begriffe und Reden der Christen] sich **gründeten** ... [$_B$ Worauf] sich [$_T$ diese Zuversicht] **gründet**, verrät er nicht. ... da [$_T$ die Analyse] sich [$_B$ auf Regel- und Unregelmäßigkeiten] ... **gründet**. ... und [$_B$ auf diese Eigenart] habe sich [$_T$ alles] zu **gründen** ... [$_T$ Die Angst der Hongkonger] **gründet** sich vor allem [$_B$ auf die Zeitbombe] ... [$_T$ Ihre Argumente] **stützen** sich [$_B$ auf die unterschiedlichen Betriebsgrößen] ... [$_T$ Die Vereinbarung] **stützt** sich in erster Linie [$_B$ auf den Interimsvertrag] ... ... Gerechtigkeitsgefühls, [$_B$ auf das] sich [$_T$ die Demokratie] **stützt**. [$_T$ Die Erwartung] **stütze** sich vor allem [$_B$ auf den ... Auftragseingang] ... [$_T$ Achtung und Ansehen] ... **stützen** sich wesentlich [$_B$ auf diese Darlegung] ... [$_T$ Die Bedenken des Gerichts] **stützen** sich [$_B$ darauf, daß in dem ... Fall] ... [$_T$ Die Entlassung] **stützt** sich [$_B$ auf angebliche Äußerungen] ...

Class 41: *Inference*

Verbs: *folgern, schließen*$_2$

Ambiguity: *schließen* has a sense of closing something open (class 36); in addition, *schließen* has a specific sense of 'to end' (such as books, or talks), and of 'to make, to arrange' (such as friendships, contracts).

Scene: [$_A$ Somebody or a personified affair] performs [$_I$ an inference], possibly activated by [$_C$ an underlying clue].

Frame Roles: A(rguer), I(nference)

Modification Roles: C(lue)

Levin class: -

Schumacher class: 4.6 *(Verben der Relation und des geistigen Handelns)*
Schumacher provides one verb class for the inference (this class) and result (my class 42) verbs.

Frame	Participating Verbs & Corpus Examples
n$_A$ **s-2**$_I$	+ folgern, schließen
	[$_I$ Ein Krieg ist ... notwendig], so kann [$_A$ ich] nur [$_C$ aus seinem Artikel] **folgern**.
	Also **folgern** [$_A$ wir] glasklar ... [$_I$ gleich muß Depardieu auf den Plan kommen].
	[$_I$ Ein Christ], **folgerten** [$_A$ mittelalterliche Theologen], [$_I$ betrachtet jeden Menschen] ...
	... daß [$_A$ gut unterrichtete Kreise] **folgern**, [$_I$ er lege sich eine Speckschwarte zu] ...
	Zynisch will [$_A$ ich] **schließen**: [$_I$ Der Werbeslogan ...]
	Kann [$_A$ man] [$_C$ daraus] **schließen**, [$_I$ das Universum habe einen Sinn]?
	[$_I$ Der "Amerikanische Traum"], **schließt** [$_A$ Lipset] [$_C$ daraus], [$_I$ könne ... behalten].
	[$_C$ Aus alledem] zu **schließen**, [$_I$ man sei an einen ungarischen Spaßvogel geraten] ...
n$_A$ **s-2**$_I$ [P]	+ folgern / ¬ schließen
	... weil sonst vielleicht **gefolgert** worden wäre, [$_I$ er wolle ihn stürzen].
	[$_C$ Daraus] ist ... zu Recht **gefolgert** worden, [$_I$ die Landwirtschaft ...]
	[$_C$ Daraus] kann aber nicht **gefolgert** werden, [$_I$ die ... Eingriffe seien losgelöst] ...
n$_A$ **s-dass**$_I$	+ folgern, schließen
	... aber läßt sich [$_C$ daraus] **folgern**, [$_I$ daß du wach bist, während du es liest]?
	[$_A$ Er] **folgerte**, [$_I$ daß vielleicht die "Wächter" ... ihm helfen könnten] ...
	[$_A$ Wir] **folgern** [$_C$ daraus], [$_I$ daß es ... keine ethischen und moralischen Grundsätze gibt] ...
	[$_A$ Die Presse] **folgerte** natürlich, [$_I$ daß die Stones ... auf die Beatles reagierten] ...
	[$_A$ Sie] **schließen**, [$_I$ daß das Fach "auf sich selbst" zurückfiel] ...
	... dann müssen [$_A$ wir] [$_C$ daraus] **schließen**, [$_I$ daß dies ... die Ursache ist].
	[$_C$ Daraus] **schließt** [$_A$ Diller], [$_I$ daß der Haushalt ... aus den Fugen geraten werde].
	[$_A$ Die Forscher] **schließen** [$_C$ daraus], [$_I$ daß es noch ein ... infektiöses Agens geben muß].
n$_A$ **s-dass**$_I$ [P]	+ folgern, schließen
	Ist [$_C$ daraus] zu **folgern**, [$_I$ daß Abschreckung durch Strafe insgesamt nicht funktioniert]?
	[$_C$ Daraus] wird messerscharf **gefolgert**, [$_I$ daß man nun den Angriff ... organisieren müsse].
	[$_C$ Daraus] wird **geschlossen**, [$_I$ daß die Regierung ... bereit sein könnte].
	[$_C$ Daraus] wird **geschlossen**, [$_I$ daß es kaum möglich sein würde] ...

Class 42: *Result*

Verbs: *ergeben, erwachsen, folgen*$_2$, *resultieren*

Ambiguity: In its reflexive usage, *ergeben* has a sense of ‘to capitulate’. *folgen* has a sense of ‘to follow, to obey’ (class 34); in addition, *folgen* has a sense of ‘to follow (in space or in time)’.

Scene: [$_R$ Something] results from [$_S$ something], which might be the cause, a source, an analysis, or the origin of the result.

Frame Roles: R(esult), S(tarting point)

Levin class: -

Schumacher class: 4.6 *(Verben der Relation und des geistigen Handelns)*
Schumacher provides one verb class for the inference (my class 41) and result (this class) verbs.

Frame	Participating Verbs & Corpus Examples
n$_R$	+$_{adv}$ erwachsen, folgen, resultieren / ¬ ergeben
	[$_R$ Die Unterstützung] ... **erwächst** fast ausschließlich [$_S$ aus dem Gefühl einer Zwangslage] ... Für die Risiken, [$_R$ die] [$_S$ aus dem Bruch zwischen Norden und Süden] **erwachsen** ... [$_S$ Daraus] **erwuchs** schließlich [$_R$ jene ... Tragödie] ... [$_S$ Aus ihm] **erwuchsen** [$_R$ ebenso bittere Enttäuschungen] wie [$_S$ aus der Versicherung] ... Die Gewinne, [$_R$ die] ... [$_S$ aus steuerlichen Vergünstigungen und anderen Privilegien] **folgen** ... Für den gewöhnlichen Leser **folgt** [$_S$ daraus] ... [$_R$ gesteigertes Verständnis]; [$_S$ Aus einer Behinderung] **folge** nicht etwa [$_R$ die Minderung der Menschenwürde]. [$_R$ Sie] **resultierten** vor allem [$_S$ aus der Strukturreform des Stützpunktsystems]. Gewiß **resultiert** [$_R$ dieser Mangel] auch [$_S$ aus dem Werkstatt-Charakter der Publikation] ... [$_R$ Das Wachstum] **resultiert** zu 15 Prozent [$_S$ aus der Übernahme] ... [$_S$ Aus dieser größten Freiheit] **resultierte** [$_R$ der höchste Symboldruck]: [$_R$ Seine erhöhten Testosteron-Werte] **resultierten** [$_S$ aus körpereigener Überproduktion] ...
n$_S$ **a**$_R$	+ ergeben / ¬ erwachsen, folgen, resultieren
	[$_S$ Meinungsumfragen] **ergeben** [$_R$ eine moderat optimistische Stimmung]. [$_R$ Das] **ergab** [$_S$ die Obduktion des Leichnams] im niederländischen Den Haag ... [$_S$ Moos, Spinnenweben, Federn und Tierhaare] **ergeben** [$_R$ eine weiche Polsterung] ...
n$_R$ **d**$_S$	+ erwachsen / ¬ ergeben, folgen, resultieren
	[$_R$ Die Rückkehr historischen Denkens] **erwächst** ... [$_S$ dem Niedergang einer Epoche] ... Daß jedoch [$_S$ dem Einzelhandel] [$_R$ mächtige Konkurrenz] **erwachsen** könnte ... Und [$_R$ Konkurrenz] **erwächst** [$_S$ dem ... vorhandenen Start-System] nun überall:
n$_S$ **p**$_R$:Dat.in	+ resultieren / ¬ ergeben, erwachsen, folgen
	In einer Zelle **resultiert** [$_S$ das Ganze] [$_R$ in einem höchst kohärenten Stoffwechsel] ... [$_S$ Die Spezialbehandlung] ... **resultiere** [$_R$ in geringerer Aggressivität] ...
n$_R$ **r**	+ ergeben / ¬ erwachsen, folgen, resultieren
	Da **ergeben** sich nun auf einmal [$_R$ äußerst reizvolle Einblicke]: [$_S$ Aus dem Archivmaterial] **ergibt** sich also [$_R$ ein genau umgekehrtes Bild] ... Für den Schuldner **ergibt** sich [$_R$ der Vorteil einer dauerhaften Schuldensenkung].
n$_S$ **s-dass**$_R$	+ ergeben / ¬ erwachsen, folgen, resultieren
	[$_S$ Die Stella-Marktforschung] hat **ergeben**, [$_R$ daß die Schulbildung ... entspreche]. [$_S$ Eine Umfrage] habe **ergeben**, [$_R$ daß ... einen fleischlosen Burger haben wollen]. [$_S$ Eine eigene Analyse] habe **ergeben**, [$_R$ daß der Koffeinanteil nicht höher ... sei].
n$_S$ **s-ob**$_R$	+ ergeben / ¬ erwachsen, folgen, resultieren
	Im Herbst wird [$_S$ ein Volksentscheid] **ergeben**, [$_R$ ob der Senat ... bleiben wird]. Hier müsse [$_S$ die Auslegung der Parteivereinbarung] **ergeben**, [$_R$ ob die Erfüllung ...] [$_S$ Die sichergestellten Geschäftsunterlagen] sollen **ergeben**, [$_R$ ob der Vorwurf zutrifft].
x s-dass$_R$	+ folgen, resultieren / ¬ ergeben, erwachsen
	Schon [$_S$ hieraus] **folgt**, [$_R$ daß ... eine Organentnahme von vornherein ausscheidet]. [$_S$ Daraus] **folge**, [$_R$ daß notwendige Veränderungen unterblieben]. [$_S$ Daraus] **resultiert**, [$_R$ daß es sich um "Zufälle" handelt] ... [$_S$ Aus dem Gesagten] **resultiert**, [$_R$ daß der Prozeß der Vereinigung Deutschlands] ...
x r s-dass$_R$[4]	+ ergeben / ¬ erwachsen, folgen, resultieren
	... habe sich jedoch **ergeben**, [$_R$ daß der ... Personenkreis größer gefaßt werden müsse]. Es hat sich **ergeben**, [$_R$ daß dieses Theater sehr gut ist, daß es verwandt ist mit uns]. Wenn sich **ergibt**, [$_R$ daß die Clearingstelle abgeraten hat] ...

[4] This frame is not coded in the grammar.

Class 43: ***Weather***

Verbs: *blitzen, donnern, dämmern, nieseln, regnen, schneien*

Scene: The weather is described, typically without any agentive involvement. But prototypical properties (such as look, sound) of the weather verbs are specified by or transfered to [$_T$ themes]. Depending on the property transfer, there might be [$_P$ a path] for the (weather) movement, [$_E$ an experiencer], or [$_C$ a content].

Frame Roles: T(heme)

Modification Roles: P(ath), C(ontent), E(xperiencer)

Levin class: 57 *(Weather Verbs)*

Schumacher class: -

Frame	Participating Verbs & Corpus Examples
n$_T$	+ blitzen, donnern, dämmern / +$_{adv}$ nieseln, regnen / ¬ schneien
	... [$_T$ Schwerter] **blitzen** ...
	... daß [$_T$ die Juwelen] ... gar so hell **blitzten**.
	Und es **blitzten** [$_T$ die Sterne].
	... hier und da **blitzt** [$_T$ ein Siegelring] ...
	Seit [$_T$ die russischen Schützenpanzer] nicht mehr [$_P$ durch Grosnyj] **donnern** ...
	[$_T$ Die Eismassen] **donnerten** [$_P$ durch das letzte nicht blockierte Wehr der Staustufe].
	[$_T$ Das Boot] **donnert** [$_P$ in die Reihe der im Schleusenbecken ankernden Schiffe] und versinkt.
	Schon **donnert** [$_T$ das Blech] ...
	... **donnert** [$_T$ ein Tanklastzug] [$_P$ über die Kreuzung] und verschwindet wie ein Spuk.
	... wenn [$_T$ die Kanonen] **donnern** ...
	Wobei [$_T$ die Erkenntnis] **dämmert**:
	Hier **dämmern** bereits [$_T$ die mysteriösen Gefilde für Künstler von Munch bis Beuys].
	... **dämmert** plötzlich [$_T$ die wahre Bedeutung]:
	[$_P$ Vom dämmrigen Himmel] **nieselt** [$_T$ kalter Regen] ...
	[$_T$ Blumen] **regnen** [$_P$ vom Himmel] ...
	Auf Rubens Reis **regnen** [$_T$ aufgeblasene Kondome] ...
	[$_T$ Wolken weißlich-glühender Asche] **regneten** [$_P$ auf den Ampato] ...
n$_T$ **s-2**$_C$	+ donnern / ¬ blitzen, dämmern, nieseln, regnen, schneien
	[$_T$ Kohl] **donnerte**, [$_C$ die CDA-Leute müßten wissen] ...
	[$_C$ "Ich bin Künstler, kein Buchhalter"], **donnert** [$_T$ der Theaterdirektor].
x	+ blitzen, donnern, dämmern, nieseln, regnen, schneien
	Jede Sekunde **blitzt** es weltweit rund siebzig- bis hundertmal.
	Es **blitzt** viermal.
	... und manchmal **donnert** es ...
	Es **donnerte** und blitzte ...
	... es **dämmerte** schon wieder ...
	Nieselregen fällt und es **dämmert** bereits.
	Nur über den Bergen **nieselte** es ...
	Mal **nieselte** es leicht ...
	... an deren Küsten es auch im Sommer regelmäßig **regnet**.
	Dann **regnet** es derart heftig ...
	Es **regnet** in Strömen ...
	Am meisten **schneite** es in der Nähe von Mannheim.
	In den Alpen **schneite** es sogar.
x a$_T$	+ regnen, schneien / ¬ blitzen, donnern, dämmern, nieseln
	Es **regnete** [$_T$ Blumen] ...
	Auf der Berlinale **regnet** es [$_T$ nur Wasser].
	Es **regnet** [$_T$ Splitter] [$_P$ vom Himmel].
	Bei uns **schneit**'s jede Nacht [$_T$ neuen Pulverschnee] aus 35 Schneelaternen.
	Es **schneit** [$_T$ viel Papierschnipsel].
x d$_E$	+ dämmern / ¬ blitzen, donnern, nieseln, regnen, schneien
	Allmählich **dämmert** es [$_E$ den inländischen Anlegern] ...
	Möglicherweise **dämmert** es [$_E$ auch dem Landesvorsitzenden Beck] ...
	... und so **dämmerte** es [$_E$ mir] allmählich ...

2.3 Usage of Verb Classes

As said in the beginning of this chapter, the purpose of the manual classification within this thesis is to evaluate the reliability and performance of the clustering experiments, both with respect to (i) the underlying relationship between verb meaning components and verb behaviour, and (ii) the usage of clustering methodologies for the automatic acquisition of a high-quality and large-scale lexical semantic resource. But there are various other possibilities for how one uses an existing manual verb classification within the area of Natural Language Processing. This holds for a preliminary and restricted classification as presented in this thesis, and even more with respect to a large-scale lexical semantic resource.

For theoretical linguistic research, a verb classification at the syntax-semantic interface represents a valuable resource. On basis of the verb classes, the researcher can verify hypotheses concerning the relationship between verb meaning and verb behaviour, with respect to symbolic or statistical properties. For example, this is important within the research in language acquisition, concerning the question how a child learns to generalise (and restrict) the usage of verbs.

For computational linguistic research, possible areas of NLP application include the following:

- Parsing:

  Providing lexical information about verbs by verb class labels serves two purposes in parsing: (i) On the one hand the class information restricts possible parses and decreases parse ambiguities, since the class labels implicitly define the range of possible syntactic and semantic verb environments. (ii) On the other hand the class information supplies additional information on the syntax-semantic embedding for verbs which are defined vaguely. Combining both uses, the parsing quality might be improved.

- Language Modelling:

  Replacing verbs in a language model by the respective verb classes might improve a language model's robustness and accuracy, since the class information provides more stable syntactic and semantic information than the individual verbs. For example, the probability of the preposition *nach* following any manner of motion verb is comparably high, since (among other senses) it indicates a path. Nevertheless, the model might provide less reliable information on the individual manner of motion verbs, especially in low frequent cases such as *rasen* 'to speed'. The verb class information contributes this missing information by generalising over the verbs within one class, and is therefore able to predict a *nach*-PP for *rasen*.

- Information Extraction:

  A user query which is meant to extract information from documents can be extended with respect to its syntactic and especially semantic information (e.g. complement realisation) by adding the existing class information of the query predicate to the query description, in addition to the individual verb information.

- Machine Translation:

  Assuming that a similar system of verb classes exists in various languages, the problem that the translation of a verb from one language into another activates several verbs in the target language can be solved by filtering the correct translation with respect to the source verb class. For example, the verb *bestehen* has at least four different senses, each coupled with a preferred subcategorisation behaviour: (i) *bestehen* meaning 'to insist' subcategorises `np` with auf_{Dat}, (ii) *bestehen* meaning 'to consist' subcategorises `np` with aus_{Akk}, (iii) *bestehen* meaning 'to exist, to survive' subcategorises `n` or `np` with in_{Akk}, and (iv) *bestehen* meaning 'to pass' (e.g. of an exam) subcategorises `na`. With respect to the source context (the syntactico-semantic embedding), the verb class of the source verb is determined, and based on the source class the target verb is filtered. In addition, missing information concerning the source or target verb with respect to its syntactic and semantic embedding might be added by the respective class and refine the translation.

- Smoothing:

  Smoothing is a meta-usage of the verb classes which is inherent in most of the preceding applications. The verb classes can be used to smooth the syntactic and semantic information of individual verbs by the information provided by verbs in the same class. The smoothing creates more uniform distributions over the syntactico-semantic verb features.

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

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

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

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

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

Document Classification Klavans and Kan (1998) use Levin's verb classes to discriminate article types within the news domain of the *Wall Street Journal (WSJ)* corpus. They consider the nouns in a document as the conceptual entities, and the verbs as the conceptual events and actions within the documents. The paper focuses on the role of verbs in document analysis.

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

Word Sense Disambiguation in Target Word Selection Prescher, Riezler, and Rooth (2000) present an approach for disambiguation in target word selection. Given a translation produces multiple equivalences of a source word, a disambiguation model selects the target word. The core part of the disambiguation system is represented by a probabilistic class-based lexicon, which is induced in an unsupervised manner (via the EM algorithm) from unannotated newspaper corpus data. The lexicon provides estimated frequencies for English verb-noun pairs with respect to a grammatical relationship. For example, Table 2.1 presents the 10 most frequent nouns which are learned as direct objects of the verb *to cross*.

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

Table 2.1: Class-based estimated frequencies of direct object nouns

Given that in a translation process a decision has to be made concerning which of a set of alternative target nouns is the most appropriate translation of an ambiguous source noun, the target nouns are looked up in the probabilistic lexicon with respect to the grammatical relationship to the (already translated) target verb. For example, in *eine Grenze überschreiten* possible English target nouns for the German source noun *Grenze* are *border, frontier, boundary, limit, periphery, edge*. But with respect to the direct object relationship to the verb *to cross* which is the translation of *überschreiten*, the lexicon determines *border* as the most probable translation, cf. Table 2.1.

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

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

2.4 Summary

I have begun this chapter by introducing the general idea of verb classifications, an artificial construct of natural language which generalises over verbs. I manually classified 168 German verbs into 43 verb classes, primarily based on semantic intuition. According to the meaning-behaviour relationship of verbs at the syntax-semantic interface, the verbs grouped in one class show to a certain extent agreement in their behaviour. German verbs and verb classes have been described in detail in relation to the semantic framework of *FrameNet*, providing a conceptual description and syntax-semantics embeddings. In the experiments on cluster induction, the behavioural verb properties will be utilised in order to probe for class membership: lexical properties at the syntax-semantic interface which are induced from corpus data and computational resources represent a basis for the automatic induction of semantic verb classes.

I described and compared related work on verb classes for various frameworks and languages. My approach on verb classification is closest to

- a German dictionary of verb valency and distribution by Helbig and Schenkel (1969) with respect to the syntactico-semantic subcategorisation information they provide; but they do not perform a classification of verbs;
- Levin's English semantic verb classes (Levin, 1993) which are based on the alternation behaviour of the verbs; but Levin (i) concentrates on the syntactic components of alternations and (ii) does not provide empirical evidence;
- the German semantic fields of Schumacher (1986) who formulates a conceptual description for verb classes and illustrated the verbs by semantic and syntactic properties; the thematic variety of classes and verbs is more restricted than in my classification;
- the FrameNet project which provides conceptual embeddings, syntax-semantic combinatorics, and corpus evidence for a frame-semantic verb classification.

In the remain of the chapter, I illustrated possible applications in Natural Language Processing on the basis of German verb classes. And I described already existing applications based on existing verb classifications such as Levin's classes and WordNet and thereby demonstrated the usefulness of the verb class construct.

Chapter 3

Statistical Grammar Model

This chapter describes the implementation, training and lexical exploitation of a German statistical grammar model. The model provides empirical lexical information, specialising on but not restricted to the subcategorisation behaviour of verbs. It serves as source for the German verb description at the syntax-semantic interface, which is used within the clustering experiments.

Before going into the details of the grammar description I introduce the definition of subcategorisation as used in the German grammar. The subcategorisation of the verbs distinguishes between obligatory and facultative verb complements.[1] The subcategorisation is defined by the arguments of a verbs, i.e. only obligatory complements are considered. A problem arises, because both in theory and in practice there is no clear-cut distinction between arguments and adjuncts. (a) Several theoretical tests have been proposed to distinguish arguments and adjuncts on either a syntactic or semantic basis, cf. Schütze (1995, pages 98–123) for an overview of such tests for English. But different tests have different results with respect to a dividing line between arguments and adjuncts, so the tests can merely be regarded as heuristics. I decided to base my judgement regarding the argument-adjunct distinction on the optionality of a complement: If a complement is optional in a proposition it is regarded as adjunct, and if a complement is not optional it is regarded as argument. I am aware that this distinction is subjective, but it is sufficient for my needs. (b) In practice, a statistical grammar would never learn the distinction between arguments and adjuncts in a perfect way, even if there were theoretically exact definitions. In this sense, the subcategorisation definition of the verbs in the German grammar is an approximation to the distinction between obligatory and facultative complements.

The chapter introduces the theoretical background of lexicalised probabilistic context-free grammars (Section 3.1) describes the German grammar development and implementation (Section 3.2), and the grammar training (Section 3.3). The empirical lexical information in the resulting statistical grammar model is illustrated (Section 3.4), and the core part of the verb information, the subcategorisation frames, are evaluated against manual dictionary definitions (Section 3.5).

[1] I use the term *complement* to subsume both arguments and adjuncts, and I refer to *arguments* as obligatory complements and *adjuncts* as facultative complements.

3.1 Context-Free Grammars and their Statistical Extensions

> *At one level of description, a natural language is a set of strings – finite sequences of words, morphemes, phonemes, or whatever.*
>
> Partee, ter Meulen, and Wall (1993, page 431)

Regarding natural language as a set of strings, a large part of language structures can be modelled using context-free descriptions. For that reason, context-free grammars have become a significant means in the analysis of natural language phenomena. But context-free grammars fail in providing structural and lexical preferences in natural language; therefore, a probabilistic environment and a lexicalisation of the grammar framework are desirable extensions of the basic grammar type.

This section describes the theoretical background of the statistical grammar model: Section 3.1.1 introduces context-free grammars, Section 3.1.2 introduces probabilistic context-free grammars, and Section 3.1.3 introduces an instantiation of lexicalised probabilistic context-free grammars. Readers familiar with the grammar formalisms might want to skip the respective parts of this section.

3.1.1 Context-Free Grammars

Context-free grammars can model the most natural language structure. Compared to linear language models –such as n-grams– they are able to describe recursive structures (such as complex nominal phrases).

Definition 3.1 *A context-free grammar CFG is a quadruple $\langle N, T, R, S \rangle$ with*

N	*finite set of non-terminal symbols*
T	*finite set of terminal symbols,* $T \cap N = \emptyset$
R	*finite set of rules* $C \rightarrow \gamma$,
	$C \in N$ *and* $\gamma \in (N \cup T)^*$
S	*distinguished start symbol,* $S \in N$

As an example, consider the context-free grammar in Table 3.1. The grammar unambiguously analyses the sentences *John loves Mary* and *John loves ice-cream* as represented in Figure 3.1. If there were ambiguities in the sentence, the grammar would assign multiple analyses, without defining preferences for the ambiguous readings.

N	S, NP, PN, CN, VP, V
T	*John, Mary, ice-cream, loves*
R	S $\rightarrow$ NP VP,
	NP $\rightarrow$ PN,
	NP $\rightarrow$ CN,
	VP $\rightarrow$ V NP,
	PN $\rightarrow$ *John*,
	PN $\rightarrow$ *Mary*,
	CN $\rightarrow$ *ice-cream*,
	V $\rightarrow$ *loves*
S	S

Table 3.1: Example CFG

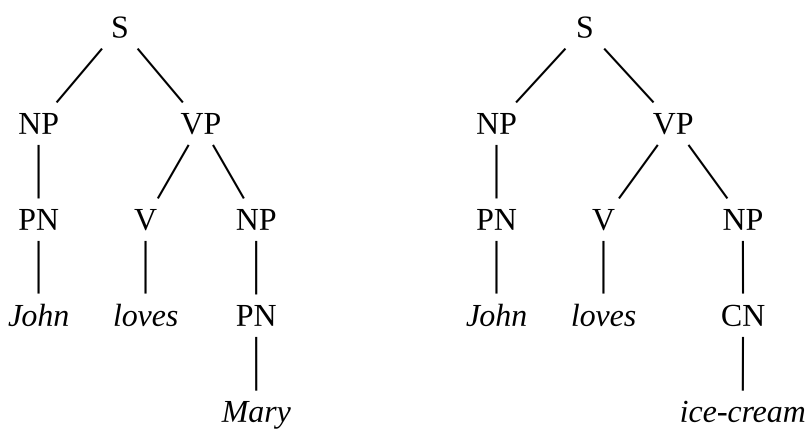


Figure 3.1: Syntactic analyses for *John loves Mary* and *John loves ice-cream*

The example is meant to give an intuition about the linguistic idea of context-free grammars. For details about the theory of context-free grammars and their formal relationship to syntactic trees, the reader is referred to Hopcroft and Ullman (1979, chapter 4) and Partee *et al.* (1993, chapter 16).

To summarise, context-free grammars can model the a large part of natural language structure. But they cannot express preferences or degrees of acceptability and therefore cannot resolve ambiguities.

3.1.2 Probabilistic Context-Free Grammars

Probabilistic context-free grammars (PCFGs) are an extension of context-free grammars which model preferential aspects of natural language by adding probabilities to the grammar rules.

Definition 3.2 *A probabilistic context-free grammar $PCFG$ is a quintuple $\langle N, T, R, p, S \rangle$ with*

N	*finite set of non-terminal symbols*
T	*finite set of terminal symbols, $T \cap N = \emptyset$*
R	*finite set of rules $C \rightarrow \gamma$,* *$C \in N$ and $\gamma \in (N \cup T)^*$*
p	*corresponding finite set of probabilities on rules,* $(\forall r \in R) \ : \ 0 \leq p(r) \leq 1$ *and* $(\forall C \in N) \ : \ \sum_\gamma p(C \rightarrow \gamma) = 1$
S	*distinguished start symbol, $S \in N$*

The probability of a syntactic tree analysis $p(t)$ for a sentence is defined as the product of probabilities for the rules r applied in the tree. The frequency of a rule r in the respective tree is given by $f_t(r)$. On the basis of parse tree probabilities for sentences or parts of sentences, PCFGs rank syntactic analyses according to their plausibility.

$$p(t) = \prod_{r \ in \ R} p(r)^{f_t(r)} \tag{3.1}$$

As an example, consider the probabilistic context-free grammar in Table 3.2. The grammar assigns ambiguous analyses to the sentence *John ate that cake*, as in Figure 3.2. (The rule probabilities are marked as subscripts on the respective parent categories.) According to the grammar rules, the demonstrative pronoun can either represent a stand-alone noun phrase or combine with a common noun to form a noun phrase. Assuming equal probabilities of 0.5 for both verb phrase types ⟨ V NP ⟩ and ⟨ V NP NP ⟩ and equal probabilities of 0.3 for both noun phrase types ⟨ N ⟩ and ⟨ DEM N ⟩ , the probabilities for the complete trees are 0.045 for the first analysis compared to 0.0045 for the second one. In this example, the probabilistic grammar resolves the structural noun phrase ambiguity in the desired way, since the probability for the preferred first (transitive) tree is larger than for the second (ditransitive) tree.

N	S, NP, PN, N, DEM, VP, V	
T	*John, cake, ate, that*	
R,p	S $\rightarrow$ NP VP,	p(S $\rightarrow$ NP VP) = 1,
	NP $\rightarrow$ PN,	p(NP $\rightarrow$ PN) = 0.3,
	NP $\rightarrow$ N,	p(NP $\rightarrow$ N) = 0.3,
	NP $\rightarrow$ DEM,	p(NP $\rightarrow$ DEM) = 0.1,
	NP $\rightarrow$ DEM N,	p(NP $\rightarrow$ DEM N) = 0.3,
	VP $\rightarrow$ V NP,	p(VP $\rightarrow$ V NP) = 0.5,
	VP $\rightarrow$ V NP NP,	p(VP $\rightarrow$ V NP NP) = 0.5,
	PN $\rightarrow$ *John*,	p(PN $\rightarrow$ *John*) = 1,
	N $\rightarrow$ *cake*,	p(N $\rightarrow$ *cake*) = 1,
	V $\rightarrow$ *ate*,	p(V $\rightarrow$ *ate*) = 1,
	DEM $\rightarrow$ *that*	p(DEM $\rightarrow$ *that*) = 1
S	S	

Table 3.2: Example PCFG (1)

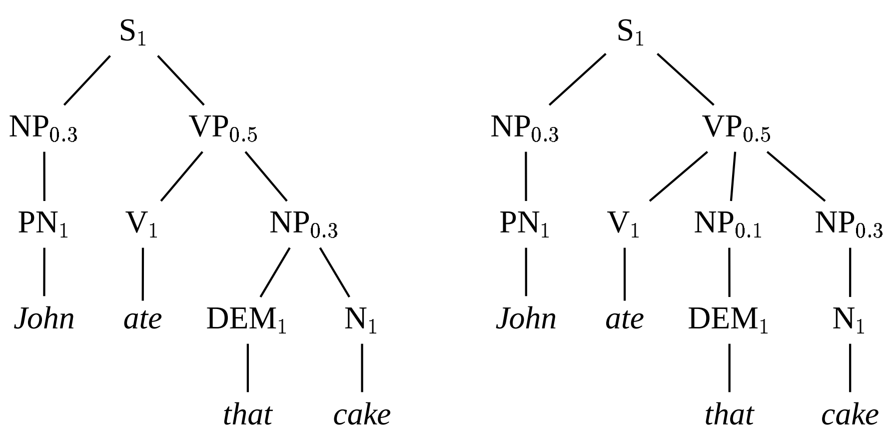


Figure 3.2: Syntactic analyses for *John ate that cake*

Now consider the probabilistic context-free grammar in Table 3.3. The grammar is ambiguous with respect to prepositional phrase attachment: prepositional phrases can either be attached to a noun phrase by NP $\rightarrow$ NP PP or to a verb phrase by VP $\rightarrow$ VP PP. The grammar assigns ambiguous analyses to the sentence *John eats the cake with a spoon*[2] as illustrated in Figure 3.3.

N	S, NP, PN, N, VP, V, PP, P, DET	
T	*John, cake, icing, spoon, eats, the, a, with*	
R,p	S $\rightarrow$ NP VP,	p(S $\rightarrow$ NP VP) = 1,
	NP $\rightarrow$ PN,	p(NP $\rightarrow$ PN) = 0.3,
	NP $\rightarrow$ N,	p(NP $\rightarrow$ N) = 0.25,
	NP $\rightarrow$ DET N,	p(NP $\rightarrow$ DET N) = 0.25,
	NP $\rightarrow$ NP PP,	p(NP $\rightarrow$ NP PP) = 0.2,
	VP $\rightarrow$ V NP,	p(VP $\rightarrow$ V NP) = 0.7,
	VP $\rightarrow$ VP PP,	p(VP $\rightarrow$ VP PP) = 0.3,
	PP $\rightarrow$ P NP,	p(PP $\rightarrow$ P NP) = 1,
	PN $\rightarrow$ *John*,	p(PN $\rightarrow$ *John*) = 1,
	N $\rightarrow$ *cake*,	p(N $\rightarrow$ *cake*) = 0.4,
	N $\rightarrow$ *icing*,	p(N $\rightarrow$ *icing*) = 0.3,
	N $\rightarrow$ *spoon*,	p(N $\rightarrow$ *spoon*) = 0.3,
	V $\rightarrow$ *eats*,	p(V $\rightarrow$ *eats*) = 1,
	P $\rightarrow$ *with*,	p(P $\rightarrow$ *with*) = 1,
	DET $\rightarrow$ *the*,	p(DET $\rightarrow$ *the*) = 0.5,
	DET $\rightarrow$ *a*	p(DET $\rightarrow$ *a*) = 0.5
S	S	

Table 3.3: Example PCFG (2)

The analyses show a preference for correctly attaching the prepositional phrase *with a spoon* as instrumental modifier to the verb phrase instead of the noun phrase: the probability of the former parse tree is $2.36 * 10^{-4}$ compared to the probability of the latter parse tree $1.58 * 10^{-4}$. This preference is based on the rule probabilities in the grammar which prefer verb phrase attachment (0.3) over noun phrase attachment (0.2).

The same grammar assigns ambiguous analyses to the sentence *John eats the cake with icing* as in Figure 3.4. In this case, the preferred attachment of the prepositional phrase *with icing* would be as modifier of the noun phrase *the cake*, but the grammar assigns a probability of $3.15 * 10^{-4}$ to the noun phrase attachment (first analysis) compared to a probability of $4.73 * 10^{-4}$ for the attachment to the verb phrase (second analysis). As in the preceding example, the structural preference for the verb phrase attachment over the noun phrase attachment is based on the attachment probabilities in the grammar.

[2]The two example sentences in Figures 3.3 and 3.4 are taken from Manning and Schütze (1999, page 278).

The examples illustrate that probabilistic context-free grammars realise PP-attachment structurally, without considering the lexical context. PCFGs assign preferences to structural units on basis of grammar rule probabilities, but they do not distinguish rule applications with reference to the lexical heads of the rules. With respect to the examples, they either have a preference for PP-attachment to the verb or to the noun, but they do not recognise that *spoon* is an instrument for *to eat* or that *icing* describes the topping of the *cake*.

In addition to defining structural preferences, PCFGs can model degrees of acceptability. For example, a German grammar might define preferences on case assignment; genitive noun phrases are nowadays partly replaced by dative noun phrases: (i) A genitive noun phrase subcategorised by the preposition *wegen* 'because of' is commonly replaced by a dative noun phrase, cf. *wegen des Regens*$_{Gen}$ and *wegen dem Regen*$_{Dat}$ 'because of the rain'. (ii) Genitive noun phrases subcategorised by the verb *gedenken* 'commemorate' are often replaced by dative noun phrases, cf. *der Menschen*$_{Gen}$ *gedenken* and *den Menschen*$_{Dat}$ *gedenken* 'commemorate the people', but the substitution is less common than in (i). (iii) Genitive noun phrases modifying common nouns cannot be replaced by dative noun phrases, cf. *der Hut des Mannes*$_{Gen}$ and **der Hut dem Mann*$_{Dat}$ 'the hat of the man'. Concluding the examples, PCFGs can define degrees of case acceptability for noun phrases depending on their structural embedding.

To summarise, PCFGs are an extension of context-free grammars in that they can model structural preferences (as for noun phrase structure), and degrees of acceptability (such as case assignment). But PCFGs fail when it comes to lexically sensitive phenomena such as PP-attachment, or selectional preferences of individual verbs, since they are based purely on structural factors.

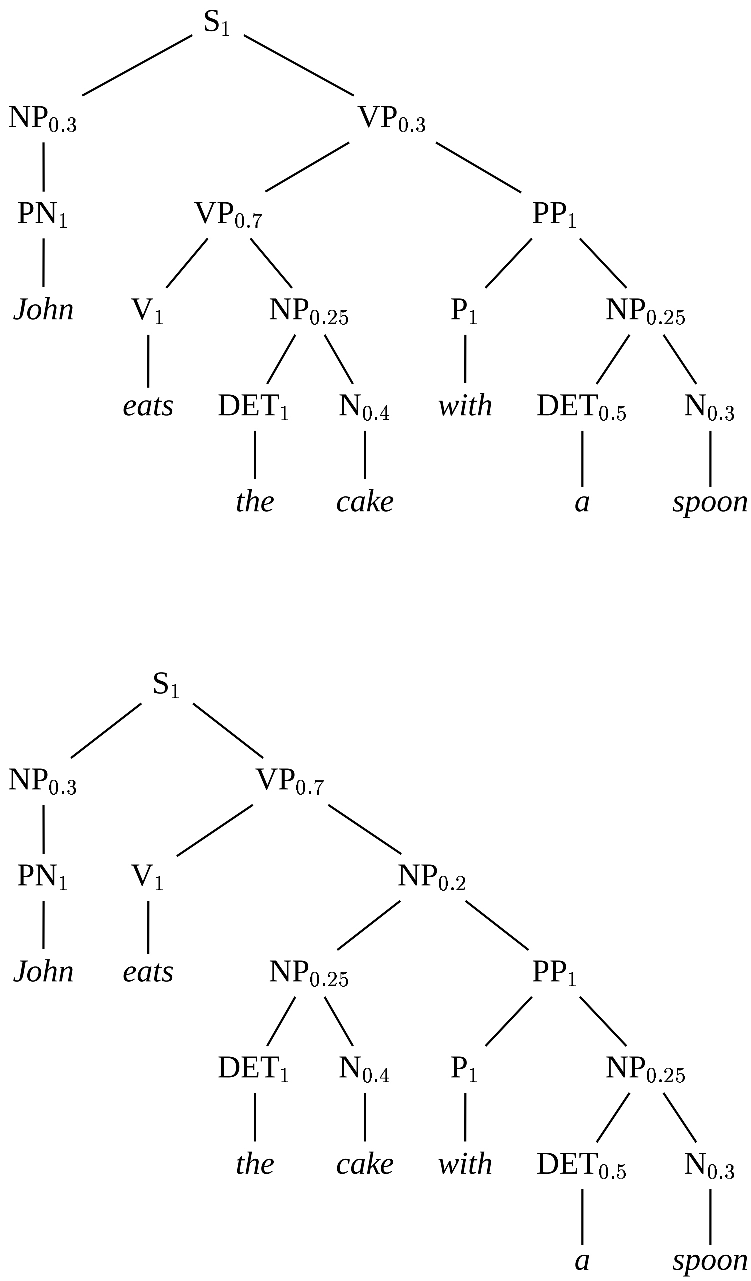


Figure 3.3: Syntactic analyses for *John eats the cake with a spoon*

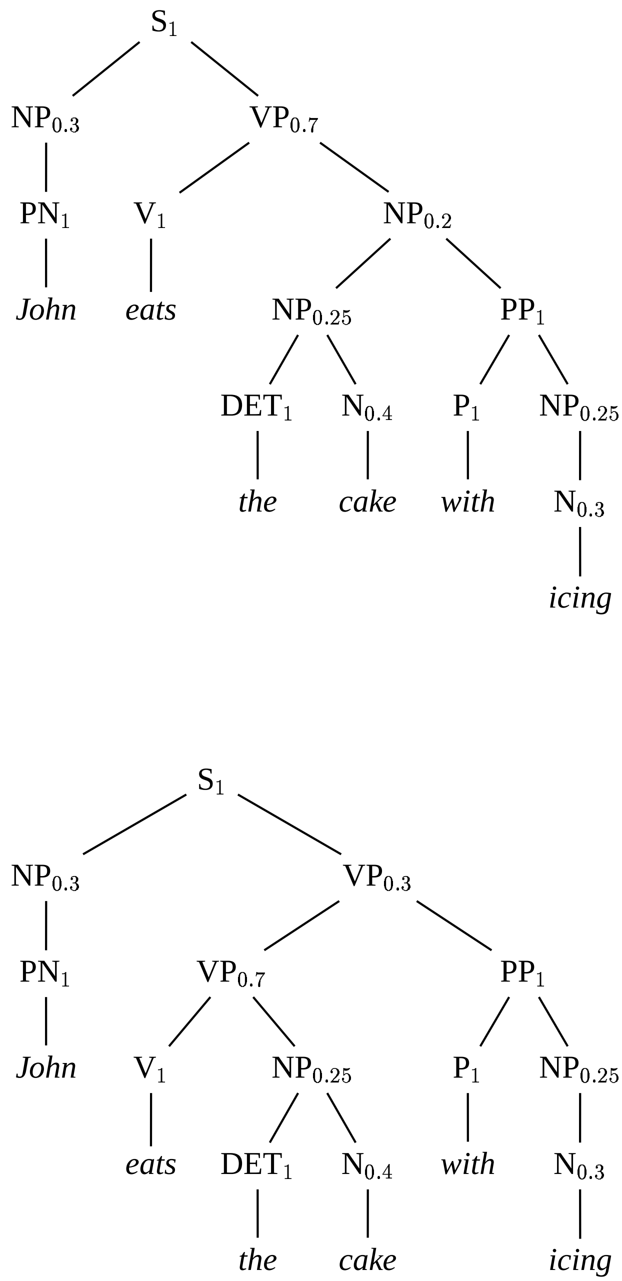


Figure 3.4: Syntactic analyses for *John eats the cake with icing*

3.1.3 Head-Lexicalised Probabilistic Context-Free Grammars

Various extensions of PCFGs are possible. Since the main drawback of PCFGs concerns their inability of modelling lexical dependencies, a common idea behind PCFG extensions is their expansion with lexical information. Examples are the decision trees in Magerman (1995), parsing models in Collins (1997), bilexical grammars in Eisner and Satta (1999), and maximum entropy modelling in Charniak (2000).

The approach as used in this thesis defines head-lexicalised probabilistic context-free grammars (H-L PCFGs) as a lexicalised extension of PCFGs. The idea of the grammar model originates from Charniak (1995) and has been implemented at the IMS Stuttgart by Carroll (1997) to learn valencies for English verbs (Carroll and Rooth, 1998). This work uses a re-implementation by Schmid (2000). Like other approaches, H-L PCFGs extend the idea of PCFGs by incorporating the lexical head of each rule into the grammar parameters. The lexical incorporation is realised by marking the head category on the right hand side of each context-free grammar rule, e.g. VP $\rightarrow$ V' NP. Each category in the rule bears a lexical head, and the lexical head from the head child category is propagated to the parent category. The lexical head of a terminal category is the respective full or lemmatised word form.

The lexical head marking in the grammar rules enables the H-L PCFG to instantiate the following grammar parameters, as defined by Schmid (2000):

- $p_{start}(s)$ is the probability that s is the category of the root node of a parse tree.
- $p_{start}(h|s)$ is the probability that a root node of category s bears the lexical head h.
- $p_{rule}(r|C,h)$ is the probability that a (parent) node of category C with lexical head h is expanded by the grammar rule r.
- $p_{choice}(h_C|C_P,h_P,C_C)$ is the probability that a (non-head) child node of category C_C bears the lexical head h_C, the parent category is C_P and the parent head is h_P.

In case a H-L PCFG does not include lemmatisation of its terminal symbols, either the lexical head h of a terminal node and the full word form $w \in T$ are identical and $p_{rule}(C \rightarrow w|C,h)$ is 1 (e.g. $p_{rule}(C \rightarrow runs|C,runs) = 1$), or the lexical head differs from the word form and $p_{rule}(C \rightarrow w|C,h)$ is 0 (e.g. $p_{rule}(C \rightarrow runs|C,ran) = 0$). In case a grammar does include lemmatisation of its terminal symbols, the probability $p_{rule}(C \rightarrow w|C,h)$ is distributed over the different word forms w with common lemmatised lexical head h (e.g. $p_{rule}(C \rightarrow runs|C,run) = 0.3$, $p_{rule}(C \rightarrow run|C,run) = 0.2$, $p_{rule}(C \rightarrow ran|C,run) = 0.5$).

The probability of a syntactic tree analysis $p(t)$ for a sentence is defined as the product of the probabilities for the start category s, the rules r, and the relevant lexical heads h which are included in the tree, cf. Equation 3.2. R refers to the set of rules established by the grammar, N to the set of non-terminal categories, and T to the set of terminal categories. Frequencies in the tree analysis are referred to by $f_t(r,C,h)$ for lexical rule parameters and $f_t(h_C,C_P,h_P,C_C)$

for lexical choice parameters. H-L PCFGs are able to rank syntactic analyses including lexical choices.

$$
\begin{aligned}
p(t) = \; & p_{start}(s) \; * \\
& p_{start}(h|s) \; * \\
& \prod_{r \in R, C \in N, h \in T} p_{rule}(r|C,h)^{f_t(r,C,h)} \; * \\
& \prod_{C_P, C_C \in N; h_P, h_C \in T} p_{choice}(h_C|C_P,h_P,C_C)^{f_t(h_C,C_P,h_P,C_C)}
\end{aligned}
\tag{3.2}
$$

As example, consider the head-lexicalised probabilistic context-free grammar in Tables 3.4 and 3.5. Table 3.4 defines the grammar rules, with the heads of the rules marked by an apostrophe. The probability distributions on the lexicalised grammar parameters are given in Table 3.5. To distinguish terminal symbols and lexical heads (here: lemmatised word forms), the terminal symbols are printed in *italic* letters, the lexical heads in `typewriter` font.

N	S, NP, PN, N, VP, V, PP, P, POSS
T	*John, Mary, anger, smile, blames, loves, for, her*
R	S → NP VP’,
	NP → PN’, NP → POSS N’,
	VP → VP’ PP, VP → V’ NP, VP → V’ NP PP,
	PP → P’ NP,
	PN → *John*’, PN → *Mary*’,
	N → *anger*’, N → *smile*’,
	V → *blames*’, V → *loves*’,
	P → *for*’, POSS → *her*’
S	S

Table 3.4: Example H-L PCFG (rules)

According to the maximum probability parse, the H-L PCFG analyses the sentence *John blames Mary for her anger* as in Figure 3.5, with the prepositional phrase *for her anger* correctly analysed as argument of the verb. The sentence *John loves Mary for her smile* is analysed as in Figure 3.6, with the prepositional phrase *for her smile* correctly analysed as adjunct to the verb phrase. In the trees, the lexical heads of the grammar categories are cited as superscripts of the categories. p_{start} is quoted on the left of the root nodes S. For each node in the tree, p_{rule} is quoted on the right of the category, and p_{choice} is quoted on the right of each child category.

Multiplying the probabilities in the trees results in a probability of $8.7 * 10^{-3}$ for *John blames Mary for her anger* in Figure 3.5 and a probability of $1.9 * 10^{-3}$ for *John loves Mary for her smile* in Figure 3.6. If the *blame* sentence had been analysed incorrectly with the prepositional phrase *for her anger* as adjunct to the verb phrase, or the *love* sentence with the prepositional phrase *for her smile* as argument of the verb, the probabilities would have been $4.3 * 10^{-4}$ and $1.1 * 10^{-3}$

p_{start}	$p_{start}(\mathrm{S}) = 1,$	
	$p_{start}(\texttt{blame} \mid \mathrm{S}) = 0.5,$	$p_{start}(\texttt{love} \mid \mathrm{S}) = 0.5$
p_{rule}	$p_{rule}(\mathrm{S} \rightarrow \mathrm{NP\ VP'} \mid \mathrm{S}, \texttt{blame}) = 1,$	$p_{rule}(\mathrm{S} \rightarrow \mathrm{NP\ VP'} \mid \mathrm{S}, \texttt{love}) = 1,$
	$p_{rule}(\mathrm{NP} \rightarrow \mathrm{PN'} \mid \mathrm{NP}, \texttt{John}) = 0.9,$	$p_{rule}(\mathrm{NP} \rightarrow \mathrm{POSS\ N'} \mid \mathrm{NP}, \texttt{John}) = 0.1,$
	$p_{rule}(\mathrm{NP} \rightarrow \mathrm{PN'} \mid \mathrm{NP}, \texttt{Mary}) = 0.9,$	$p_{rule}(\mathrm{NP} \rightarrow \mathrm{POSS\ N'} \mid \mathrm{NP}, \texttt{Mary}) = 0.1,$
	$p_{rule}(\mathrm{NP} \rightarrow \mathrm{PN'} \mid \mathrm{NP}, \texttt{anger}) = 0.1,$	$p_{rule}(\mathrm{NP} \rightarrow \mathrm{POSS\ N'} \mid \mathrm{NP}, \texttt{anger}) = 0.9,$
	$p_{rule}(\mathrm{NP} \rightarrow \mathrm{PN'} \mid \mathrm{NP}, \texttt{smile}) = 0.1,$	$p_{rule}(\mathrm{NP} \rightarrow \mathrm{POSS\ N'} \mid \mathrm{NP}, \texttt{smile}) = 0.9,$
	$p_{rule}(\mathrm{VP} \rightarrow \mathrm{VP'\ PP} \mid \mathrm{VP}, \texttt{blame}) = 0.1,$	$p_{rule}(\mathrm{VP} \rightarrow \mathrm{VP'\ PP} \mid \mathrm{VP}, \texttt{love}) = 0.3,$
	$p_{rule}(\mathrm{VP} \rightarrow \mathrm{V'\ NP} \mid \mathrm{VP}, \texttt{blame}) = 0.3,$	$p_{rule}(\mathrm{VP} \rightarrow \mathrm{V'\ NP} \mid \mathrm{VP}, \texttt{love}) = 0.6,$
	$p_{rule}(\mathrm{VP} \rightarrow \mathrm{V'\ NP\ PP} \mid \mathrm{VP}, \texttt{blame}) = 0.6,$	$p_{rule}(\mathrm{VP} \rightarrow \mathrm{V'\ NP\ PP} \mid \mathrm{VP}, \texttt{love}) = 0.1,$
	$p_{rule}(\mathrm{PN} \rightarrow \mathit{John}' \mid \mathrm{PN}, \texttt{John}) = 1$	$p_{rule}(\mathrm{PN} \rightarrow \mathit{Mary}' \mid \mathrm{PN}, \texttt{John}) = 0,$
	$p_{rule}(\mathrm{PN} \rightarrow \mathit{Mary}' \mid \mathrm{PN}, \texttt{Mary}) = 1$	$p_{rule}(\mathrm{PN} \rightarrow \mathit{John}' \mid \mathrm{PN}, \texttt{Mary}) = 0,$
	$p_{rule}(\mathrm{N} \rightarrow \mathit{anger}' \mid \mathrm{N}, \texttt{anger}) = 1,$	$p_{rule}(\mathrm{N} \rightarrow \mathit{smile}' \mid \mathrm{N}, \texttt{anger}) = 0,$
	$p_{rule}(\mathrm{N} \rightarrow \mathit{smile}' \mid \mathrm{N}, \texttt{smile}) = 1,$	$p_{rule}(\mathrm{N} \rightarrow \mathit{anger}' \mid \mathrm{N}, \texttt{smile}) = 0,$
	$p_{rule}(\mathrm{V} \rightarrow \mathit{blames}' \mid \mathrm{V}, \texttt{blame}) = 1,$	$p_{rule}(\mathrm{V} \rightarrow \mathit{loves}' \mid \mathrm{V}, \texttt{blame}) = 0,$
	$p_{rule}(\mathrm{V} \rightarrow \mathit{loves}' \mid \mathrm{V}, \texttt{love}) = 1,$	$p_{rule}(\mathrm{V} \rightarrow \mathit{blames}' \mid \mathrm{V}, \texttt{love}) = 0,$
	$p_{rule}(\mathrm{PP} \rightarrow \mathrm{P'\ NP} \mid \mathrm{PP}, \texttt{for}) = 1,$	$p_{rule}(\mathrm{P} \rightarrow \mathit{for}' \mid \mathrm{P}, \texttt{for}) = 1,$
	$p_{rule}(\mathrm{POSS} \rightarrow \mathit{her}' \mid \mathrm{POSS}, \texttt{she}) = 1$	
p_{choice}	$p_{choice}(\texttt{John} \mid \mathrm{S}, \texttt{blame}, \mathrm{NP}) = 0.4,$	$p_{choice}(\texttt{Mary} \mid \mathrm{S}, \texttt{blame}, \mathrm{NP}) = 0.4,$
	$p_{choice}(\texttt{anger} \mid \mathrm{S}, \texttt{blame}, \mathrm{NP}) = 0.1,$	$p_{choice}(\texttt{smile} \mid \mathrm{S}, \texttt{blame}, \mathrm{NP}) = 0.1,$
	$p_{choice}(\texttt{John} \mid \mathrm{S}, \texttt{love}, \mathrm{NP}) = 0.4,$	$p_{choice}(\texttt{Mary} \mid \mathrm{S}, \texttt{love}, \mathrm{NP}) = 0.4,$
	$p_{choice}(\texttt{anger} \mid \mathrm{S}, \texttt{love}, \mathrm{NP}) = 0.1,$	$p_{choice}(\texttt{smile} \mid \mathrm{S}, \texttt{love}, \mathrm{NP}) = 0.1,$
	$p_{choice}(\texttt{she} \mid \mathrm{NP}, \texttt{John}, \mathrm{POSS}) = 1,$	$p_{choice}(\texttt{she} \mid \mathrm{NP}, \texttt{anger}, \mathrm{POSS}) = 1,$
	$p_{choice}(\texttt{she} \mid \mathrm{NP}, \texttt{Mary}, \mathrm{POSS}) = 1,$	$p_{choice}(\texttt{she} \mid \mathrm{NP}, \texttt{smile}, \mathrm{POSS}) = 1,$
	$p_{choice}(\texttt{for} \mid \mathrm{VP}, \texttt{blame}, \mathrm{PP}) = 1,$	$p_{choice}(\texttt{for} \mid \mathrm{VP}, \texttt{love}, \mathrm{PP}) = 1,$
	$p_{choice}(\texttt{John} \mid \mathrm{VP}, \texttt{blame}, \mathrm{NP}) = 0.4,$	$p_{choice}(\texttt{Mary} \mid \mathrm{VP}, \texttt{blame}, \mathrm{NP}) = 0.4,$
	$p_{choice}(\texttt{anger} \mid \mathrm{VP}, \texttt{blame}, \mathrm{NP}) = 0.1,$	$p_{choice}(\texttt{smile} \mid \mathrm{VP}, \texttt{blame}, \mathrm{NP}) = 0.1,$
	$p_{choice}(\texttt{John} \mid \mathrm{VP}, \texttt{love}, \mathrm{NP}) = 0.3,$	$p_{choice}(\texttt{Mary} \mid \mathrm{VP}, \texttt{love}, \mathrm{NP}) = 0.3,$
	$p_{choice}(\texttt{anger} \mid \mathrm{VP}, \texttt{love}, \mathrm{NP}) = 0.2,$	$p_{choice}(\texttt{smile} \mid \mathrm{VP}, \texttt{love}, \mathrm{NP}) = 0.2,$
	$p_{choice}(\texttt{John} \mid \mathrm{PP}, \texttt{for}, \mathrm{NP}) = 0.25,$	$p_{choice}(\texttt{Mary} \mid \mathrm{PP}, \texttt{for}, \mathrm{NP}) = 0.25,$
	$p_{choice}(\texttt{anger} \mid \mathrm{PP}, \texttt{for}, \mathrm{NP}) = 0.25,$	$p_{choice}(\texttt{smile} \mid \mathrm{PP}, \texttt{for}, \mathrm{NP}) = 0.25$

Table 3.5: Example H-L PCFG (lexicalised parameters)

respectively, i.e. the correct analyses of the sentences in Figures 3.5 and 3.6 are more probable than their incorrect counterparts. This distinction in probabilities results from the grammar parameters which reflect the lexical preferences of the verbs, in this example concerning their subcategorisation properties. For *blame*, subcategorising the transitive ⟨ V NP PP ⟩ including the PP is more probable than subcategorising the intransitive ⟨ V NP ⟩ , and for *love* the lexical preference is vice versa.

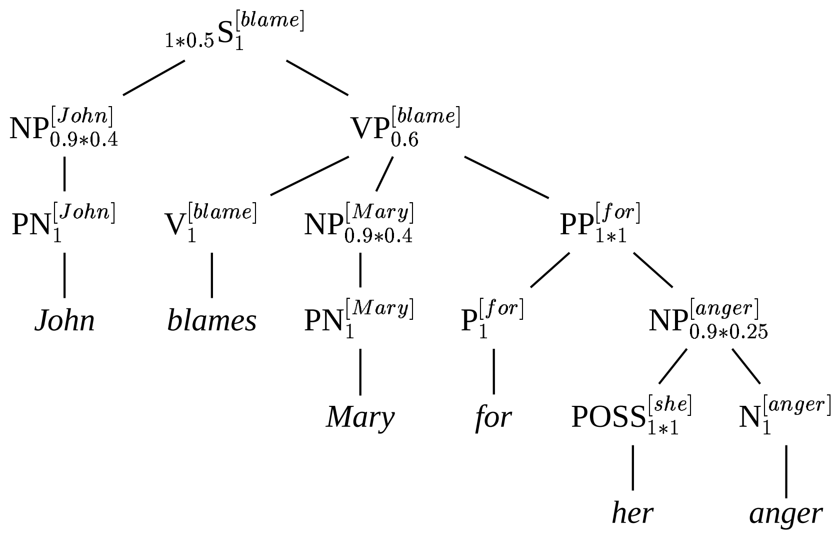


Figure 3.5: Syntactic analysis for *John blames Mary for her anger*

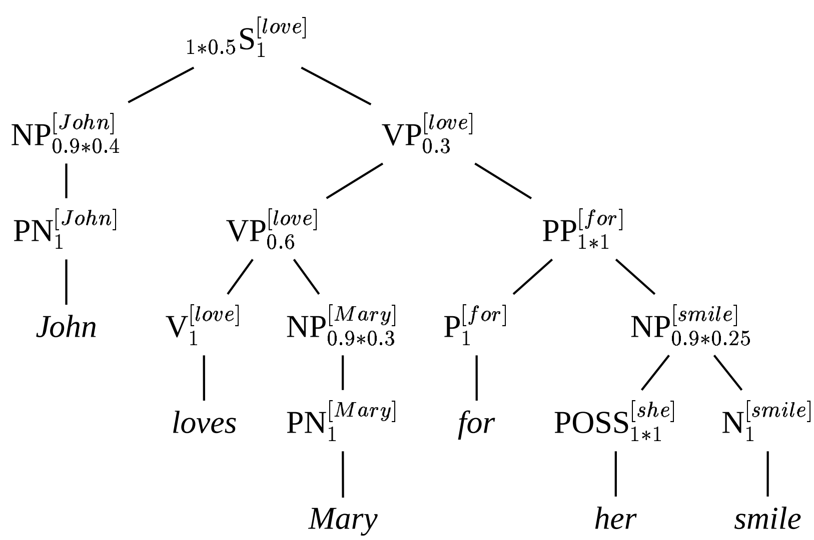


Figure 3.6: Syntactic analysis for *John loves Mary for her smile*

To summarise, H-L PCFGs are a further extension of context-free grammars in that they can model structural preferences including lexical selection, such as PP-attachment and selectional argument preferences of individual verbs. According to Manning and Schütze (1999), main problems of H-L PCFGs concern (i) the assumption of context-freeness, i.e. that a certain subtree in a sentence analysis is analysed in the same way no matter where in the sentence parse it is situated; for example, noun phrase formation actually differs according to the position, since noun phrases tend to be pronouns more often in sentence initial position than elsewhere. And (ii) for discriminating the large number of parameters in a H-L PCFG, a sufficient amount of linguistic data is required. The detailed linguistic information in the grammar model is of large value, but effective smoothing techniques are necessary to overcome the sparse data problem.

3.1.4 Summary

This section has introduced the theoretical background of context-free grammars and their statistical extensions. Context-free grammars (CFGs) can model a large part of natural language structure, but fail to express preferences. Probabilistic context-free grammars (PCFGs) are an extension of context-free grammars which can model structural preferences (as for noun phrase structure) and degrees of acceptability (such as case assignment), but they fail when it comes to lexically sensitive phenomena. Head-lexicalised probabilistic context-free grammars (H-L PCFGs) are a further extension of context-free grammars in that they can model structural preferences including lexical selection, such as PP-attachment and selectional argument preferences of individual verbs.

My statistical grammar model is based on the framework of H-L PCFGs. The development of the grammar model is organised in three steps, according to the theoretical grammar levels.

1. Manual definition of CFG rules with head-specification,
2. Assigning probabilities to CFG rules (extension of CFG to PCFG),
3. Lexicalisation of the PCFG (creation of H-L PCFG).

The following Section 3.2 describes the manual definition of the CFG rules (step 1) in detail, and Section 3.3 describes the grammar extension and training with respect to steps 2 and 3.

3.2 Grammar Development and Implementation

This section describes the development and implementation of the German context-free grammar. As explained above, the context-free backbone is the basis for the lexicalised probabilistic extension which is used for learning the statistical grammar model. Section 3.2.1 introduces the specific aspects of the grammar development which are important for the acquisition of lexicon-relevant verb information. Section 3.2.2 then describes the German context-free grammar rules.

3.2.1 Grammar Development for Lexical Verb Information

The context-free grammar framework is developed with regard to the overall goal of obtaining reliable lexical information on verbs. This goal influences the development process in the following ways:

- To provide a sufficient amount of training data for the model parameters, the grammar model should be robust, since the grammar needs to cover as much training data as possible. The robustness is important (i) to obtain lexical verb information for a large sample of German verbs, and (ii) to learn the grammar parameters to a reliable degree. To give an example, (i) in contrast to a former version of the German grammar by Beil *et al.* (1999) where only verb final clauses are regarded, the grammar covers all German sentence types in order to obtain as much information from the training corpus as possible. (ii) For fine-tuning the grammar parameters with regard to reliable verb subcategorisation, no restriction on word order is implemented, but all possible scrambling orders of German clauses are considered.

- Infrequent linguistic phenomena are disregarded if they are likely to confuse the learning of frequent phenomena. For example, coherent clauses might be structurally merged, such that it is difficult to distinguish main and subcategorised clause without crossing edges. Example (3.3) shows a merging of a non-finite and a relative clause. *sie* is the subject of the control verb *versprochen* and also embedded in the non-finite clause *den zu lieben* subcategorised by *versprochen*. Implementing the phenomenon in the grammar would enable us to parse such sentences, but at the same time include an enormous source for ambiguities and errors in the relatively free word order language German, so the implementation is ignored. The mass of training data is supposed to compromise for the parsing failure of infrequent phenomena.

(3.3)	*den*	*sie*	*zu*	*lieben*	*versprochen*	*hat*
	whom	she	to	love	promised	has

  'whom she has promised to love'

- Work effort concentrates on defining linguistic structures which are relevant to lexical verb information, especially subcategorisation. On the one hand, this results in fine-grained structural levels for subcategorisation. For example, for each clause type I define an extraordinary rule level

  `C-<type> → S-<type>.<frame>`

  where the clause level `C` produces the clause category `S` which is accompanied by the subcategorisation frame for the clause. A lexicalisation of the grammar rules with their verb heads automatically leads to a distribution over frame types. In addition, the parsing strategy is organised in an exceptional way: Since the lexical verb head as the bearer of the clausal subcategorisation needs to be propagated through the parse tree, the grammar structures are based on a so-called 'collecting strategy' around the verb head, no matter in

which topological position the verb head is or whether the verb head is realised as a finite or non-finite verb.
On the other hand, structural levels for constituents outside verb subcategorisation are ignored. For example, adjectival and adverbial phrases are realised by simple lists, which recognise the phrases reliably, but disregard fine-tuning of their internal structure.

- The grammar framework needs to control the number of parameters, especially when it comes to the lexicalised probabilistic extension of the context-free grammar. This is realised by keeping the category features in the grammar to a minimum. For example, the majority of noun phrases is recognised reliably with the case feature only, disregarding number and gender. The latter features are therefore disregarded in the context-free grammar.

The above examples concerning the grammar development strategy illustrate that the context-free grammar defines linguistic structures in an unusual way. This is so because the main goal of the grammar is the reliable definition of lexical verb information, and we need as much information on this aspect as possible to overcome the problem of data sparseness.

3.2.2 The German Context-Free Grammar

The German context-free grammar rules are manually written. The manual definition is supported by the grammar development environment of `YAP` (Schmid, 1999), a feature based parsing framework, which helps the grammar developer with managing rules and features. In addition, the statistical parser `LoPar` (Schmid, 2000) provides a graphical interface to control the grammar development. Following, I describe the grammar implementation, starting with the grammar terminals and then focusing on the grammar rules.

Grammar Terminals

The German grammar uses morpho-syntactic terminal categories as based on the dictionary database `IMSLex` and the morphological analyser `AMOR` (Lezius *et al.*, 1999, 2000): Each word form is assigned one or multiple part-of-speech tags and the corresponding lemmas. I have adopted the morphological tagging system with task-specific changes, for example ignoring the features *gender* and *number* on verbs, nouns and adjectives. Table 3.6 gives an overview of the terminal categories to which the `AMOR` tags are mapped as basis for the grammar rules, Table 3.7 lists the relevant feature values, and Table 3.8 gives examples for tag-feature combinations.

Terminal Category		Features	Tag Example
attributive adjective	ADJ	case	ADJ.Akk
indeclinable adjective	ADJ-invar		ADJ-invar
predicative adjective	ADJ-pred		ADJ-pred
adverb	ADV		ADV
article	ART	case	ART.Dat
cardinal number	CARD		CARD
year number	CARD-time		CARD-time
demonstrative pronoun	DEM	distribution, case	DEM.subst.Nom
expletive pronoun	ES		ES
indefinite pronoun	INDEF	distribution, case	INDEF.attr.Dat
interjection	INTJ		INTJ
conjunction	KONJ	conjunction type	KONJ.Sub
proper name	NE	case	NE.Nom
common noun	NN	case	NN.Gen
ordinal number	ORD		ORD
possessive pronoun	POSS	distribution, case	POSS.attr.Akk
postposition	POSTP	case, postposition	POSTP.Dat.entlang
reflexive pronoun	PPRF	case	PPRF.Dat
personal pronoun	PPRO	case	PPRO.Nom
reciprocal pronoun	PPRZ	case	PPRZ.Akk
preposition	PREP	case, preposition	PREP.Akk.ohne
preposition + article	PREPart	case, preposition	PREPart.Dat.zu
pronominal adverb	PROADV	pronominal adverb	PROADV.dazu
particle	PTKL	particle type	PTKL.Neg
relative pronoun	REL	distribution, case	REL.subst.Nom
sentence symbol	S-SYMBOL	symbol type	S-SYMBOL.Komma
truncated word form	TRUNC		TRUNC
finite verb	V*X*FIN *X* = { B(leiben), H(aben), M(odal), S(ein), V(oll), W(erden) }		VMFIN
finite verb (part of separable verb)	VVFINsep		VVFINsep
infinitival verb	V*X*INF *X* = { B(leiben), H(aben), M(odal), S(ein), V(oll), W(erden) }		VWINF
infinitival verb (incorporating *zu*)	VVIZU		VVIZU
past participle	V*X*past *X* = { B(leiben), M(odal), S(ein), V(oll), W(erden) }		VVpast
verb prefix	VPRE		VPRE
interrogative adverb	WADV	interrogative adverb	WADV.wann
interrogative pronoun	WPRO	distribution, case	WPRO.attr.Gen

Table 3.6: Terminal grammar categories

Feature	Feature Values
case	Nom, Akk, Dat, Gen
distribution	attr, subst
symbol type	Komma, Norm
conjunction type	Inf, Kon, Sub, Vgl, dass, ob
particle type	Adj, Ant, Neg, zu
preposition	[Akk] ab, an, auf, außer, bis, durch, entlang, für, gegen, gen, hinter, in, je, kontra, neben, ohne, per, pro, um, unter, versus, via, vor, wider, zwischen, über
	[Dat] ab, an, anstatt, auf, aus, außer, außerhalb, bei, binnen, dank, einschließlich, entgegen, entlang, entsprechend, exklusive, fern, gegenüber, gemäß, gleich, hinter, in, inklusive, innerhalb, laut, längs, mangels, mit, mitsamt, mittels, nach, nah, nahe, neben, nebst, nächst, samt, seit, statt, trotz, unter, von, vor, wegen, während, zu, zunächst, zwischen, ähnlich, über
	[Gen] abseits, abzüglich, anfangs, angesichts, anhand, anläßlich, anstatt, anstelle, aufgrund, ausschließlich, außer, außerhalb, beiderseits, beidseits, bezüglich, binnen, dank, diesseits, eingangs, eingedenk, einschließlich, entlang, exklusive, fern, hinsichtlich, infolge, inklusive, inmitten, innerhalb, jenseits, kraft, laut, links, längs, längsseits, mangels, minus, mittels, nahe, namens, nordwestlich, nordöstlich, nördlich,ob, oberhalb, rechts, seiten, seitens, seitlich, statt, südlich, südwestlich, südöstlich,trotz, um, unbeschadet, unerachtet, ungeachtet, unterhalb, unweit, vermittels,vermöge, orbehaltlich, wegen, westlich, während, zeit, zufolge, zugunsten, zuungunsten,zuzüglich, zwecks, östlich
postposition	[Akk] entlang, exklusive, hindurch, inklusive
	[Dat] entgegen, entlang, entsprechend, gegenüber, gemäß, nach, zufolge, zugunsten, zuliebe, zunächst, zuungunsten, zuwider
	[Gen] halber, ungeachtet, wegen, willen
pronominal adverb	dabei, dadurch, dafür, dagegen, daher, dahin, dahinter, damit, danach, daneben, daran, darauf, daraufhin, daraus, darin, darum, darunter, darüber, davon, davor, dazu, dazwischen, dementsprechend, demgegenüber, demgemäß, demnach, demzufolge, deshalb, dessenungeachtet, deswegen, dran, drauf, draus, drin, drum, drunter, drüber, hieran, hierauf, hieraufhin, hieraus, hierbei, hierdurch, hierfür, hiergegen, hierher, hierhin, hierin, hiermit, hiernach, hierum, hierunter, hiervon, hiervor, hierzu, hierüber, seitdem, trotzdem, währenddessen
interrogative adverb	wann, warum, weshalb, weswegen, wie, wieso, wieviel, wieweit, wo, wobei, wodurch, wofür, wogegen, woher, wohin, wohinein, wohinter, womit, wonach, woran, worauf, woraufhin, woraus, worein, worin, worum, worunter, worüber, wovon, wovor, wozu

Table 3.7: Terminal features

Terminal Category	Examples
ADJ.Akk	*kleine, riesiges, schönen*
ADJ-invar	*lila, mini, relaxed*
ADJ-pred	*abbruchreif, dauerhaft, schlau*
ADV	*abends, fast, immer, ratenweise*
ART.Gen	*des, einer, eines*
CARD	*0,080 5,8,14/91 dreizehn 28*
CARD-time	*1543 1920 2021*
DEM.attr.Dat / DEM.subst.Nom	*denselben, dieser, jenem* / *dasjenige, dieselben, selbige*
ES	*es*
INDEF.attr.Gen / INDEF.subst.Akk	*irgendwelcher, mehrerer* / *ebensoviele, irgendeinen, manches*
INTJ	*aha, hurra, oh, prost*
KONJ.Inf / KONJ.Kon KONJ.Sub / KONJ.Vgl KONJ.dass / KONJ.ob	*anstatt, um, ohne* / *doch, oder, und* *dass, sooft, weil* / *als, wie* *dass* / *ob*
NE.Nom	*Afrika, DDR, Julia*
NN.Dat	*ARD, C-Jugend, Häusern*
ORD	*3. 2704361.*
POSS.attr.Nom / POSS.subst.Dat	*ihr, meine, unsere* / *eurem, unseren*
POSTP.Dat.entsprechend	*entsprechend*
PPRF.Akk	*sich, uns*
PPRO.Nom	*du, ich, ihr*
PPRZ.Akk	*einander*
PREP.Akk.für	*für*
PREPart.Dat.zu	*zum*
PROADV.dadurch	*dadurch*
PTKL.Adj / PTKL.Ant PTKL.Neg / PTKL.zu	*allzu, am* / *bitte, nein* *nicht* / *zu*
REL.attr.Gen / REL.subst.Nom	*deren, dessen* / *das, der, die*
S-SYMBOL.Komma S-SYMBOL.Norm	*,* *! . : ; ?*
TRUNC	*ARD- Doktoranden- Jugend-*
VBFIN VHFIN VMFIN VSFIN VVFIN VWFIN	*bleibe, blieben* *hast, hatte* *dürftest, könnte, möchten* *sind, war, wären* *backte, ranntet, schläft* *werden, wird, würde*
VVFINsep	*gibt, rennen, trennte*
VVINF	*abblocken, eilen, schwimmen*
VVIZU	*dabeizusein, glattzubügeln*
VBpast	*geblieben*
VPRE	*ab, her, hinein, zu*
WADV.warum	*warum*
WPRO.attr.Akk / WPRO.subst.Dat	*welche, welches* / *welchen, wem*

Table 3.8: Examples of grammar terminals

Grammar Rules

The following paragraphs provide an overview of the German context-free grammar rules. Preferably the grammar code is omitted, and the rules are illustrated by syntactic trees and example sentences. Features which are irrelevant for the illustration of specific grammar rules may be left out. Explanations should help to grasp the intuition behind the rule coding strategies, cf. Section 3.2.1. The total number of context-free grammar rules is 35,821.

Sentence Structure The grammar distinguishes six finite clause types:

- `C-1-2` for verb first and verb second clauses,
- `C-rel` for relative clauses,
- `C-sub` for non-subcategorised subordinated clauses,
- `C-dass` for subcategorised subordinated *dass*-clauses ('that'-clauses),
- `C-ob` for subcategorised subordinated *ob*-clauses ('whether'-clauses),
- `C-w` for subcategorised indirect *wh*-questions.

The clause types differ with respect to their word order and their function. `C-1-2` clauses have the main verb in the first or second position of the clause, and all other claus types have the main verb in clause final position. The final clause types are distinguished, because `C-dass, C-ob` and `C-w` can represent arguments which are subcategorised by the verb, but `C-rel` and `C-sub` cannot. In addition, `C-rel` and `C-sub` have different distributions (i.e. `C-rel` typically modifies a nominal category, `C-sub` a clause), and the possible clausal arguments `C-dass, C-ob, C-w` and also `C-1-2` may be subcategorised by different verbs and verb classes.

The clause level `C` produces another the clause category `S` which is accompanied by the relevant subcategorisation frame type dominating the clause. As said before, this extraordinary rule level is provided, since the lexicalisation of the grammar rules with their verb heads will automatically lead to a distribution over frame types. The effect of this set of grammar rules will be illustrated in detail in Section 3.4 which describes the empirical lexical acquisition as based on the grammar.

```
C-<type>  ->  S-<type>.<frame>
```

In order to capture a wide range of corpus data, all possibly non-subcategorised clause types (verb first and verb second clauses, relative clauses, and non-subcategorised subordinated clauses) generate `S-top` and can be combined freely by commas and coordinating conjunctions.

```
S-top -> S-top KONJ.Kon       S-top
S-top -> S-top S-SYMBOL.Komma S-top
```

`S-top` are terminated by a full stop, question mark, exclamation mark, colon, or semicolon. `TOP` is the overall top grammar category.

```
TOP -> S-top S-SYMBOL.Norm
```

Figure 3.7 illustrates the top-level clause structure by combining a matrix clause and a non-subcategorised causal clause. The example sentence is *Peter kommt zu spät, weil er verschlafen hat* ‘Peter is late, because he overslept’.

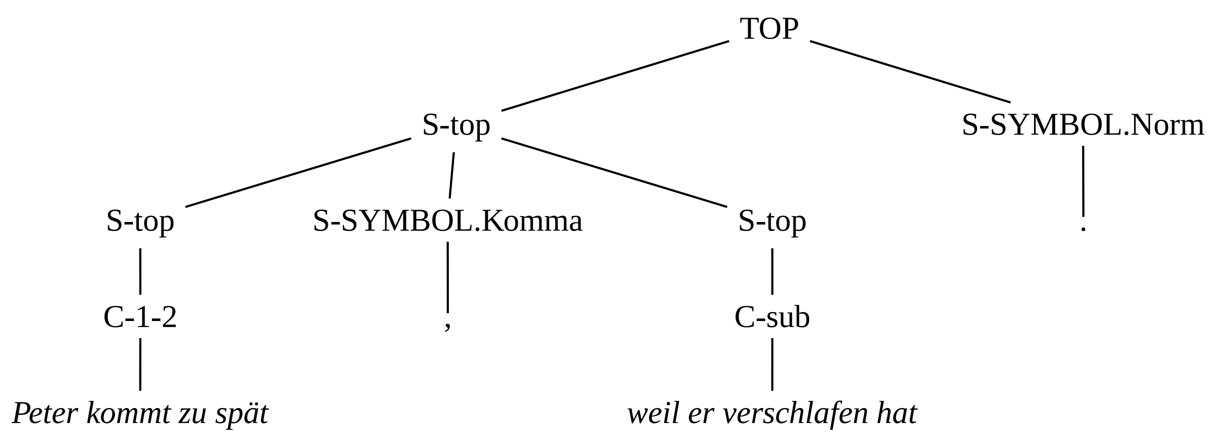


Figure 3.7: Top-level clause construction

Verb Phrases The clausal categories S-<type>.<frame> below C are generated by verb phrases which determine the clause type and the frame type. The verb phrases are the core part of the German grammar and therefore designed with special care and attention to detail. A verb phrase is defined as the verb complex which collects preceding and following arguments and adjuncts until the sentence is parsed. The resulting S-frame distinguishes verb arguments and verb adjuncts; it indicates the number and types of the verb arguments, verb adjuncts are not marked.

Four types of verb phrases are distinguished: active (VPA), passive (VPP), non-finite (VPI) verb phrases, and copula constructions (VPK). Each verb phrase type is accompanied by the frame type which may have maximally three arguments. Any verb can principally occur with any frame type. Possible arguments in the frames are nominative (n), dative (d) and accusative (a) noun phrases, reflexive pronouns (r), prepositional phrases (p), non-finite verb phrases (i), expletive *es* (x), and subordinated finite clauses (s-2 for verb second clauses, s-dass for *dass*-clauses, s-ob for *ob*-clauses, s-w for indirect *wh*-questions). Prepositional phrases in VPP, which are headed by the prepositions *von* or *durch* and indicate the deep structure subject in passive constructions, are marked by ‘P’ instead of ’p’. The frame types indicate the number and kind of subcategorised arguments, but they generalise over the argument order. For example, the verb phrase VPA.nad describes an active ditransitive verb phrase with a nominative, an accusative and a dative noun phrase (with any scrambling order); VPA.ndp describes an active verb phrase with a nominative and a dative noun phrase plus a prepositional phrase (with any scrambling order); VPP.nP describes a passive verb phrase with a nominative noun phrase and a prepositional phrase headed by *von* or *durch* (with any scrambling order).

The combinations of verb phrases and frame types are listed in Tables 3.9 to 3.12; the active frame types in Table 3.9 generalise over the subcategorisation behaviour of the verbs[3] and have already been introduced in Appendix A. The frame types are developed with reference to the standard German grammar by Helbig and Buscha (1998). The total of 38 frame types covers the vast majority of the verb structures, only few infrequent frame types such as `naa` or `nag` have been ignored.

Active and passive verb phrases are abstracted from their voice by introducing a generalising level. For example, the clause category `S.na`, a transitive type subcategorising a direct object, produces `VPA.na` in active voice and `VPP.n` and `VPP.nP` in passive voice. This treatment is justified by argument agreement of the frame types on the deep structure level, e.g. the surface structure subject in `VPP.n` and `VPP.nP` agrees with the surface structure object in `VPA.na`, and the prepositional phrase in `VPP.nP` agrees with the surface structure subject in `VPA.na`. With 'agreement' I refer to the selectional preferences of the verbs with respect to a frame type and the frame arguments. In addition to generalising over voice, the different kinds of copula constructions in Table 3.12 are generalised to the frame type 'k'. The generalisation is performed for all `S`-types. Table 3.13 provides a list of all generalised frame descriptions. `VPI` do not represent finite clauses and therefore do not generate `S`, but are instead arguments within the `S` frame types.

[3]This idea will be explained in detail below.

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

Table 3.9: Subcategorisation frame types: VPA

Frame Type	Example
n	*Peter$_n$ wird betrogen.*
nP	*Peter$_n$ wird von seiner Freundin$_P$ betrogen.*
d	*Dem Vater$_d$ wird gehorcht.*
dP	*Dem Vater$_d$ wird von allen Kindern$_P$ gehorcht.*
p	*An die Vergangenheit$_p$ wird appelliert.*
pP	*Von den alten Leuten$_P$ wird immer an die Vergangenheit$_p$ appelliert.*
nd	*Ihm$_d$ wurde die Verantwortung$_n$ übertragen.*
ndP	*Ihm$_d$ wurde von seinem Chef$_P$ die Verantwortung$_n$ übertragen.*
np	*Anna$_n$ wurde nach ihrer Großmutter$_p$ benannt.*
npP	*Anna$_n$ wurde von ihren Eltern$_P$ nach ihrer Großmutter$_p$ benannt.*
dp	*Der Organisatorin$_d$ wird für das Essen$_p$ gedankt.*
dpP	*Der Organisatorin$_d$ wird von ihren Kollegen$_P$ für das Essen$_p$ gedankt.*
i	*Pünktlich zu gehen$_i$ wurde versprochen.*
iP	*Von den Schülern$_P$ wurde versprochen, pünktlich zu gehen$_i$.*
ni	*Der Sohn$_n$ wurde verpflichtet, seiner Mutter zu helfen$_i$.*
niP	*Der Sohn$_n$ wurde von seiner Mutter$_P$ verpflichtet, ihr zu helfen$_i$.*
di	*Dem Vater$_d$ wurde versprochen, früh ins Bett zu gehen$_i$.*
diP	*Dem Vater$_d$ wurde von seiner Freundin$_P$ versprochen, früh ins Bett zu gehen$_i$.*
s-2	*Der Chef halte einen Vortrag$_{s-2}$, wurde angekündigt.*
sP-2	*Vom Vorstand$_P$ wurde angekündigt, der Chef halte einen Vortrag$_{s-2}$.*
ns-2	*Peter$_n$ wird angeschnauzt, er sei ein Idiot$_{s-2}$.*
nsP-2	*Peter$_n$ wird von seiner Freundin$_P$ angeschnauzt, er sei ein Idiot$_{s-2}$.*
ds-2	*Dem Mädchen$_d$ wird bestätigt, sie werde reich$_{s-2}$.*
dsP-2	*Dem Mädchen$_d$ wird vom Anwalt$_P$ bestätigt, sie werde reich$_{s-2}$.*
s-dass	*Dass er den Vortrag hält$_{s-dass}$, wurde rechtzeitig angekündigt.*
sP-dass	*Dass er den Vortrag hält$_{s-dass}$, wurde rechtzeitig vom Chef$_P$ angekündigt.*
ns-dass	*Die Mutter$_n$ wurde aufgefordert, dass sie verreist$_{s-dass}$.*
nsP-dass	*Die Mutter$_n$ wurde von ihrem Freund$_P$ aufgefordert, dass sie verreist$_{s-dass}$.*
ds-dass	*Dem Mädchen$_d$ wird bestätigt, dass sie reich sein wird$_{s-dass}$.*
dsP-dass	*Dem Mädchen$_d$ wird vom Anwalt$_P$ bestätigt, dass sie reich sein wird$_{s-dass}$.*
s-ob	*Ob er den Vortrag hält$_{s-ob}$, wurde gefragt.*
sP-ob	*Ob er den Vortrag hält$_{s-ob}$, wurde vom Vorstand$_P$ gefragt.*
ns-ob	*Anna$_n$ wurde gefragt, ob sie ihren Freund liebt$_{s-ob}$.*
nsP-ob	*Anna$_n$ wurde von ihrem Freund$_P$ gefragt, ob sie ihn liebt$_{s-ob}$.*
ds-ob	*Dem Mädchen$_d$ wird bestätigt, ob sie reich sein wird$_{s-ob}$.*
dsP-ob	*Dem Mädchen$_d$ wird vom Anwalt$_P$ bestätigt, ob sie reich sein wird$_{s-ob}$.*
s-w	*Wann er den Vortrag hält$_{s-w}$, wurde gefragt.*
sP-w	*Wann er den Vortrag hält$_{s-w}$, wurde vom Vorstand$_P$ gefragt.*
ns-w	*Die Mutter$_n$ wurde gefragt, wann sie verreist$_{s-w}$.*
nsP-w	*Die Mutter$_n$ wurde von ihrem Freund$_P$ gefragt, wann sie verreist$_{s-w}$.*
ds-w	*Dem Kind$_d$ wird gesagt, wer zu Besuch kommt$_{s-w}$.*
dsP-w	*Dem Kind$_d$ wird von den Eltern$_P$ gesagt, wer zu Besuch kommt$_{s-w}$.*

Table 3.10: Subcategorisation frame types: VPP

Frame Type	Example
`-`	*zu schlafen*
`a`	*ihn*$_a$ *zu verteidigen*
`d`	*ihr*$_d$ *zu helfen*
`p`	*an die Vergangenheit*$_p$ *zu appellieren*
`ad`	*seiner Mutter*$_d$ *das Geschenk*$_a$ *zu geben*
`ap`	*ihren Freund*$_a$ *am Gehen*$_p$ *zu hindern*
`dp`	*ihr*$_d$ *für die Aufmerksamkeit*$_p$ *zu danken*
`r`	*sich*$_r$ *zu erinnern*
`ar`	*sich*$_r$ *Aufwind*$_a$ *zu erhoffen*
`dr`	*sich*$_r$ *der Kirche*$_d$ *anzuschließen*
`pr`	*sich*$_r$ *für den Frieden*$_p$ *einzusetzen*
`s-2`	*anzukündigen, er halte einen Vortrag*$_{s-2}$
`as-2`	*ihn*$_a$ *anzuschnauzen, er sei ein Idiot*$_{s-2}$
`ds-2`	*ihr*$_d$ *zu sagen, sie sei unmöglich*$_{s-2}$
`s-dass`	*anzukündigen, dass er einen Vortrag hält*$_{s-dass}$
`as-dass`	*sie*$_a$ *aufzufordern, dass sie verreist*$_{s-dass}$
`ds-dass`	*ihr*$_d$ *zu sagen, dass sie unmöglich sei*$_{s-dass}$
`s-ob`	*zu fragen, ob sie ihn liebe*$_{s-ob}$
`as-ob`	*sie*$_a$ *zu fragen, ob sie ihn liebe*$_{s-ob}$
`ds-ob`	*ihr*$_d$ *zuzurufen, ob sie verreist*$_{s-ob}$
`s-w`	*zu fragen, wer zu Besuch kommt*$_{s-w}$
`as-w`	*sie*$_a$ *zu fragen, wer zu Besuch kommt*$_{s-w}$
`ds-w`	*ihr*$_d$ *zu sagen, wann der Besuch kommt*$_{s-w}$

Table 3.11: Subcategorisation frame types: VPI

Frame Type	Example
`n`	*Mein Vater*$_n$ *bleibt Lehrer.*
`i`	*Ihn zu verteidigen*$_i$ *ist Dummheit.*
`s-dass`	*Dass ich ihn treffe*$_{s-dass}$*, ist mir peinlich.*
`s-ob`	*Ob sie kommt*$_{s-ob}$*, ist unklar.*
`s-w`	*Wann sie kommt*$_{s-w}$*, wird bald klarer.*

Table 3.12: Subcategorisation frame types: VPK

Generalised Verb Phrase	Verb Phrase Type with Frame Type
S.n	VPA.n
S.na	VPA.na, VPP.n, VPP.nP
S.nd	VPA.nd, VPP.d, VPP.dP
S.np	VPA.np, VPP.p, VPP.pP
S.nad	VPA.nad, VPP.nd, VPP.ndP
S.nap	VPA.nap, VPP.np, VPP.npP
S.ndp	VPA.ndp, VPP.dp, VPP.dpP
S.ni	VPA.ni, VPP.i, VPP.iP
S.nai	VPA.nai, VPP.ni, VPP.niP
S.ndi	VPA.ndi, VPP.di, VPP.diP
S.nr	VPA.nr
S.nar	VPA.nar
S.ndr	VPA.ndr
S.npr	VPA.npr
S.nir	VPA.nir
S.x	VPA.x
S.xa	VPA.xa
S.xd	VPA.xd
S.xp	VPA.xp
S.xr	VPA.xr
S.xs-dass	VPA.xs-dass
S.ns-2	VPA.ns-2, VPP.s-2, VPP.sP-2
S.nas-2	VPA.nas-2, VPP.ns-2, VPP.nsP-2
S.nds-2	VPA.nds-2, VPP.ds-2, VPP.dsP-2
S.nrs-2	VPA.nrs-2
S.ns-dass	VPA.ns-dass, VPP.s-dass, VPP.sP-dass
S.nas-dass	VPA.nas-dass, VPP.ns-dass, VPP.nsP-dass
S.nds-dass	VPA.nds-dass, VPP.ds-dass, VPP.dsP-dass
S.nrs-dass	VPA.nrs-dass
S.ns-ob	VPA.ns-ob, VPP.s-ob, VPP.sP-ob
S.nas-ob	VPA.nas-ob, VPP.ns-ob, VPP.nsP-ob
S.nds-ob	VPA.nds-ob, VPP.ds-ob, VPP.dsP-ob
S.nrs-ob	VPA.nrs-ob
S.ns-w	VPA.ns-w, VPP.s-w, VPP.sP-w
S.nas-w	VPA.nas-w, VPP.ns-w, VPP.nsP-w
S.nds-w	VPA.nds-w, VPP.ds-w, VPP.dsP-w
S.nrs-w	VPA.nrs-w
S.k	VPK.n, VPK.i, VPK.s-dass, VPK.s-ob, VPK.s-w

Table 3.13: Generalised frame description

Clause Type	Example
verb first clause	*Liebt* Peter seine Freundin? Hat Peter seine Freundin *geliebt*?
verb second clause	Peter *liebt* seine Freundin. Peter hat seine Freundin *geliebt*.
verb final clause	weil Peter seine Freundin *liebt* weil Peter seine Freundin *geliebt* hat
relative clause	der seine Freundin *liebt* der seine Freundin *geliebt* hat
indirect *wh*-question	wer seine Freundin *liebt* wer seine Freundin *geliebt* hat
non-finite clause	seine Freundin zu *lieben* seine Freundin *geliebt* zu haben

Table 3.14: Clause type examples

As mentioned before, the lexical verb head as the bearer of the clausal subcategorisation needs to be propagated through the parse tree, since the head information is crucial for the argument selection. The grammar structures are therefore based on a so-called ‘collecting strategy’ around the verb head: The collection of verb adjacents starts at the verb head and is performed differently according to the clause type, since the verb complex is realised by different formations and is situated in different positions in the topological sentence structure. Table 3.14 illustrates the proposition *Peter liebt seine Freundin* ‘Peter loves his girl-friend’ in the different clause types with and without auxiliary verb. For example, in a verb first clause with the verb head as the finite verb, the verb head is in sentence initial position and all arguments are to its right. But in a verb first clause with the auxiliary verb as the finite verb, the verb head is in sentence final position and all arguments are between the auxiliary and the verb head.

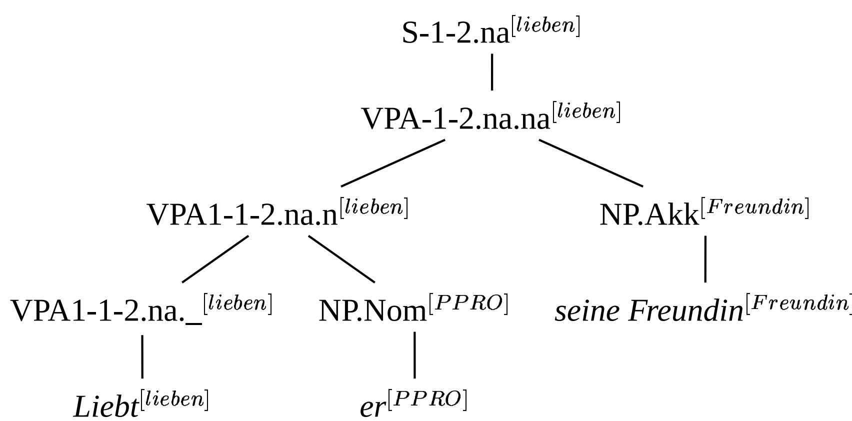
Below, A to E describe the collecting strategies in detail. Depending on the clause type, they start collecting arguments at the lexical verb head and propagate the lexical head up to the clause type level, as the head superscripts illustrate. The clause type `S` indicates the frame type of the respective sentence. Adverbial and prepositional phrase adjuncts might be attached at all levels, without having impact on the strategy or the frame type. The embedding of `S` under `TOP` is omitted in the examples.

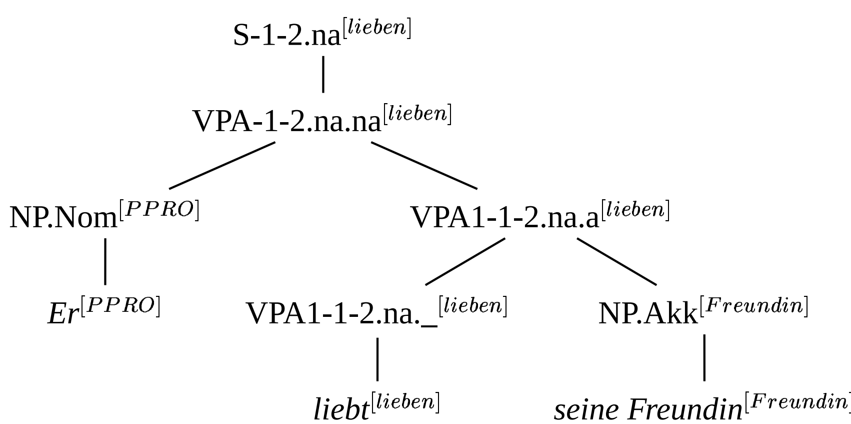
A <u>Verb First and Verb Second Clauses</u>

Verb first and verb second clauses are parsed by a common collecting schema, since they are similar in sentence formation and lexical head positions. The schema is sub-divided into three strategies:

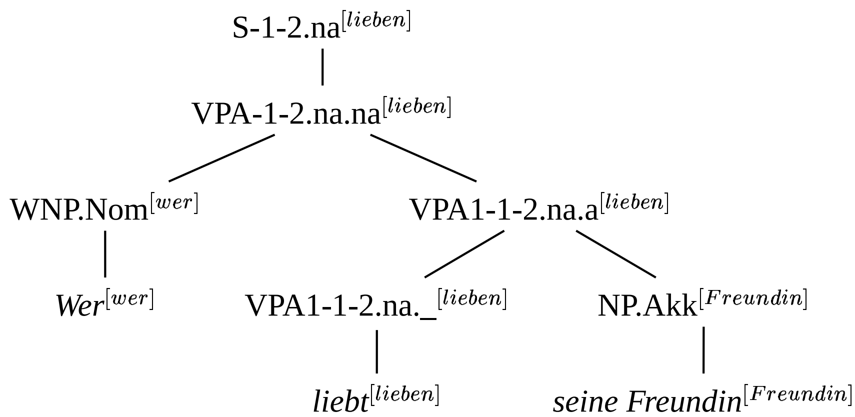
(i) In clauses where the lexical verb head is expressed by a finite verb, the verb complex is identified as this finite verb and collects first all arguments to the right (corresponding to *Mittelfeld* and *Nachfeld* constituents) and then at most one argument to the left (corresponding to the *Vorfeld* position which is relevant for arguments in verb second clauses).

Below you find examples for both verb first and verb second clause types. The verb phrase annotation indicates the verb phrase type (`VPA` in the following examples), the clause type `1-2`, the frame type (here: `na`) and the arguments which have been collected so far (`_` for none). The `1` directly attached to the verb phrase type indicates the not yet completed frame. As verb first clause example, I analyse the sentence *Liebt er seine Freundin?* 'Does he love his girl-friend?', as verb second clause example, I analyse the sentence *Er liebt seine Freundin* 'He loves his girl-friend'. The lexical head of pronouns is `PPRO`.

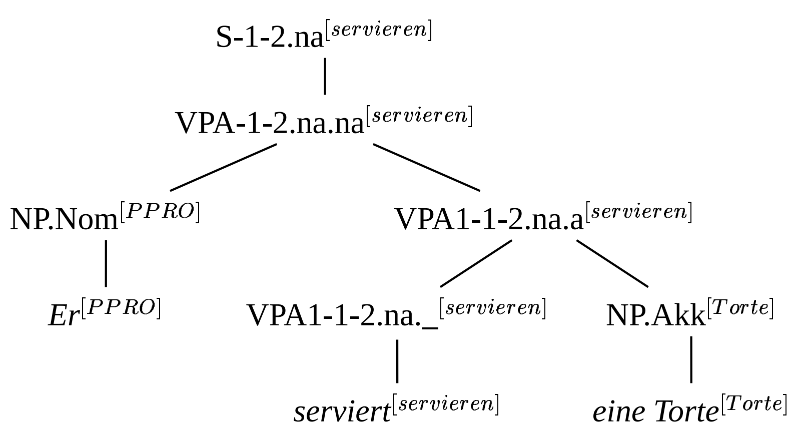




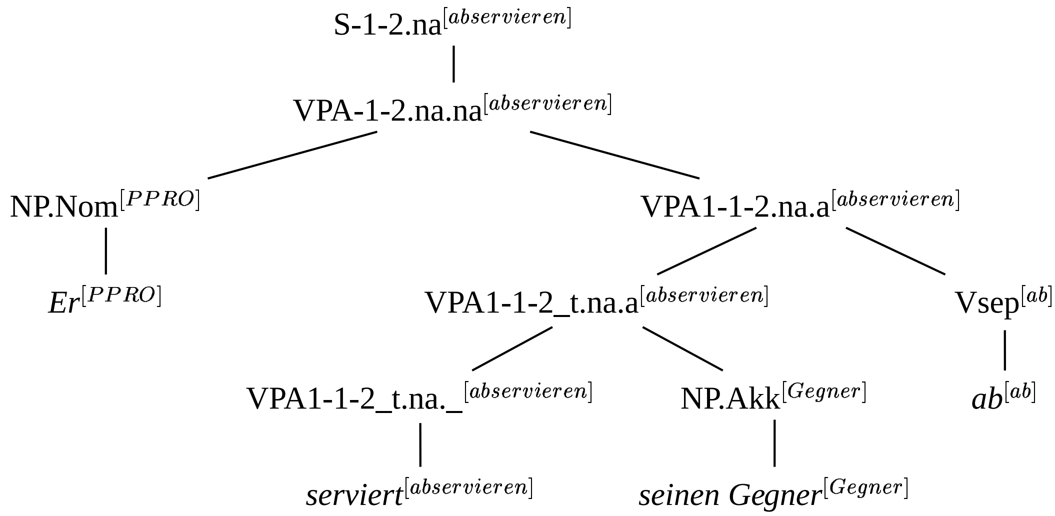
wh-questions are parsed in the same way as verb second clauses. They only differ in that the *Vorfeld* element is realised by a *wh*-phrase. The following parse tree analyses the question *Wer liebt seine Freundin?* 'Who loves his girl-friend?'. (Notice that *wh*-words in German are actually *w*-words.)



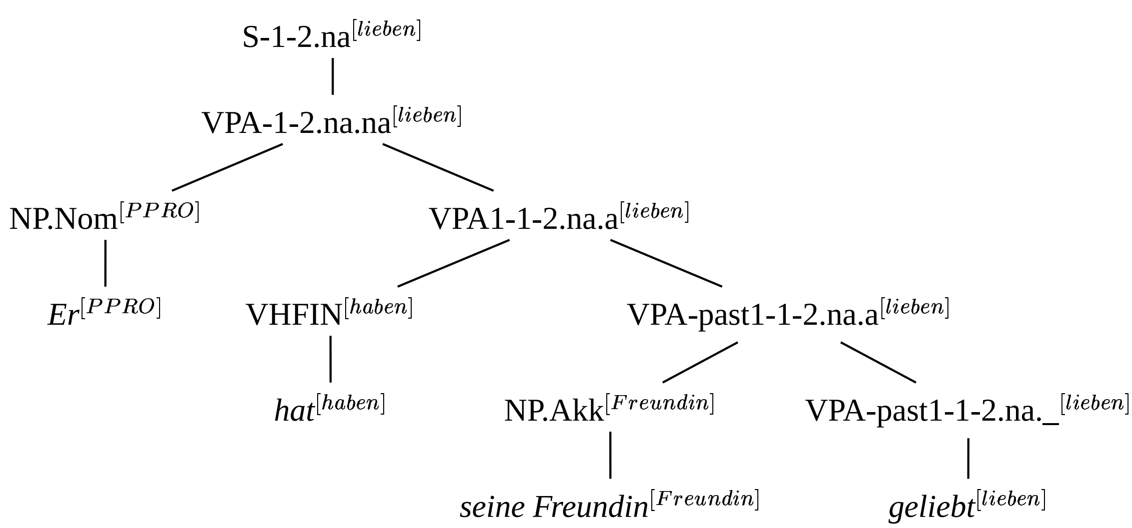
(ii) Finite verbs with separable prefixes collect their arguments in the same way. The notation differs in an additional indicator `_t` (for *trennbar* 'separable') which disappears as soon as the prefix is collected and the lexical head identified. It is necessary to distinguish verbs with separable prefixes, since the lexical verb head is only complete with the additional prefix. In this way we can, for example, differentiate the lexical verb heads *servieren* 'serve' and *abservieren* 'throw out' in *er serviert eine Torte* 'he serves a cake' vs. *er serviert seinen Gegner ab* 'he throws out his opponent'.[4] Following you find an example for the distinction. The head of the first tree is *servieren*, the head of the second tree *abservieren*:

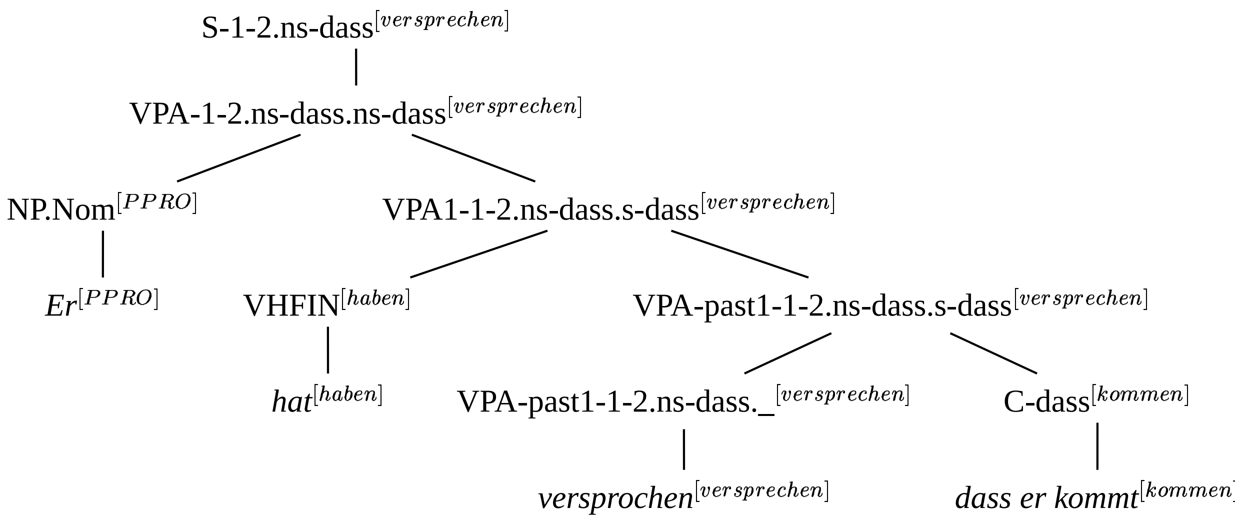


[4] `LoPar` provides a functionality to deal with particle verb lemmas.



(iii) In constructions with auxiliary verbs, the argument collection starts at the non-finite (participle, infinitival) lexical verb head, collecting arguments only to the left, since all arguments are defined in the *Vorfeld* and *Mittelfeld*. An exception to this rule are finite and non-finite clause arguments which can also appear in the *Nachfeld* to the right of the lexical verb head. The non-finite status of the verb category is marked by the low-level verb phrase types: `part` for participles and `inf` for infinitives. As soon as the finite auxiliary is found, at most one argument (to the left) is missing, and the non-finite marking on the clause category is deleted, to proceed as in (i). Below you find examples for verb second clauses: *Er hat seine Freundin geliebt* 'He has loved his girl-friend' and *Er hat versprochen, dass er kommt* 'He has promised to come'. The comma in the latter analysis is omitted for space reasons.

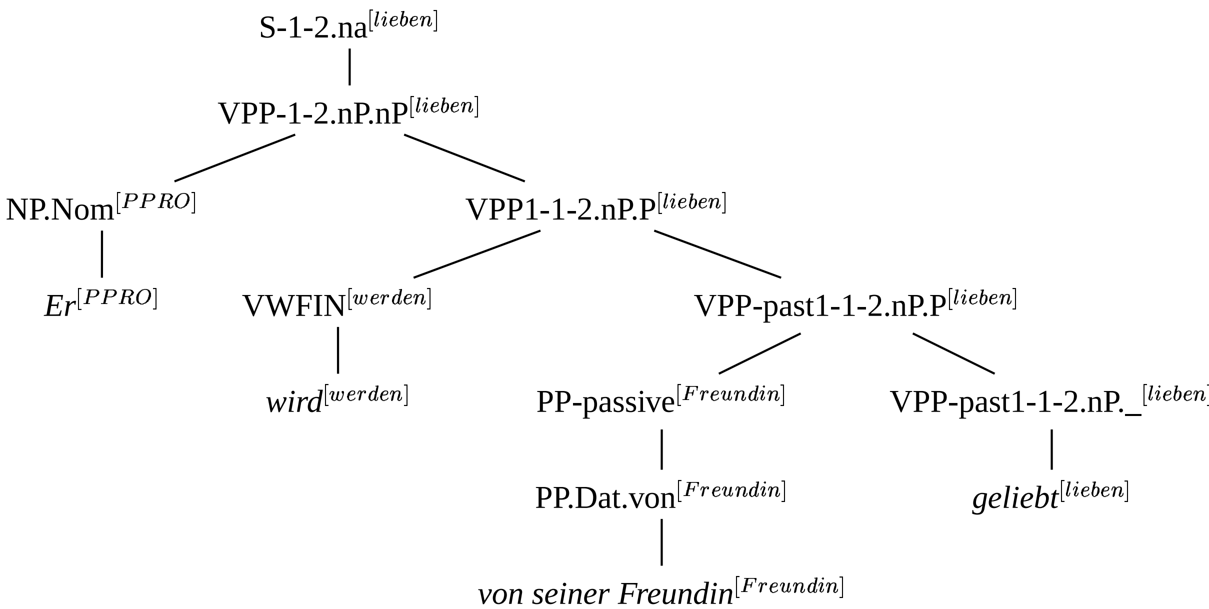




Strategies (i) and (ii) can only be applied to sentences without auxiliaries, which is a subset of VPA. Strategy (iii) can be applied to active and passive verb phrases as well as copula constructions. Table 3.15 defines the possible combinations of finite auxiliary verb and non-finite verb for the use of present perfect tense, passive voice, etc. An example analysis is performed for the sentence *Er wird von seiner Freundin geliebt* 'He is loved by his girlfriend'.

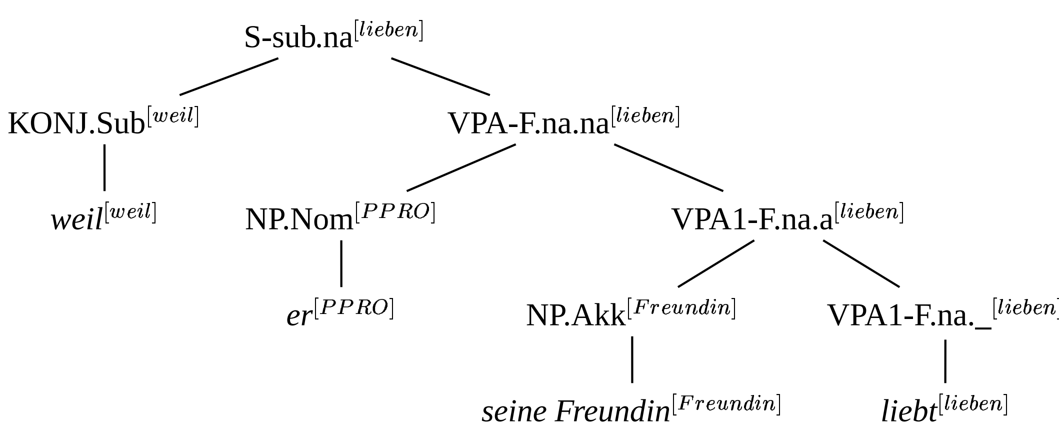
VP Type	Combination			Example
	Type	Auxiliary	Non-Finite Verb	
VPA	present perfect	VHFIN	past participle	*hat ... geliebt*
	present perfect	VSFIN	past participle	*ist ... geschwommen*
	'to have to, must'	VHFIN	infinitive	*hat ... zu bestehen*
	future tense	VWFIN	infinitive	*wird ... erkennen*
	modal construction	VMFIN	infinitive	*darf ... teilnehmen*
VPP	dynamic passive	VWFIN	past participle	*wird ... gedroht*
	statal passive	VSFIN	past participle	*ist ... gebacken*
	modal construction	VMFIN	past participle	*möchte ... geliebt werden*
VPK	'to be'	VSFIN	predicative	*ist ... im 7. Himmel*
	'to become'	VWFIN	predicative	*wird ... Lehrer*
	'to remain'	VBFIN	predicative	*bleibt ... doof*

Table 3.15: Auxiliary combination with non-finite verb forms

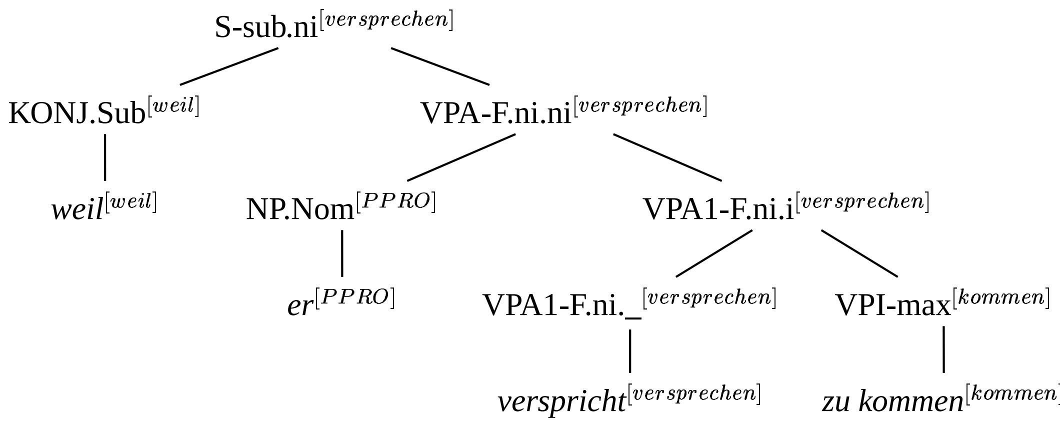


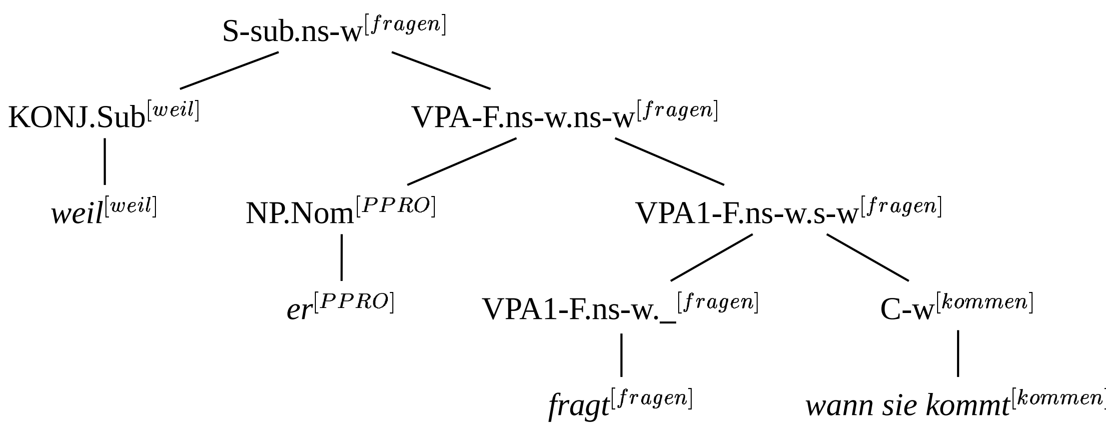
B <u>Verb Final Clauses</u>

In verb final clauses, the lexical verb complex is in the final position. Therefore, verb arguments are collected to the left only, starting from the finite verb complex. The verb final clause type is indicated by F. An example analysis for the sub-ordinated sentence *weil er seine Freundin liebt* ‘because he loves his girl-friend’ is given.



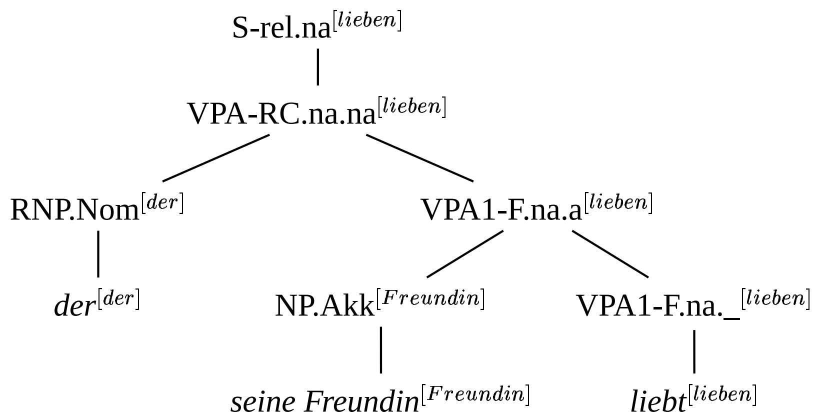
As an exception to the collecting strategy, clausal arguments might appear in the *Nachfeld* to the right of the verb complex. Below, two examples are given: In *weil er verspricht zu kommen* 'because he promises to come', *verspricht* in final clause position subcategorises a non-finite clause (`VPI-max` is a generalisation over all non-finite clauses), and in *weil er fragt, wann sie kommt* 'because he asks when she is going to come', *fragt* in clause final position subcategorises a finite *wh*-clause. The comma in the latter analysis is omitted for space reasons.





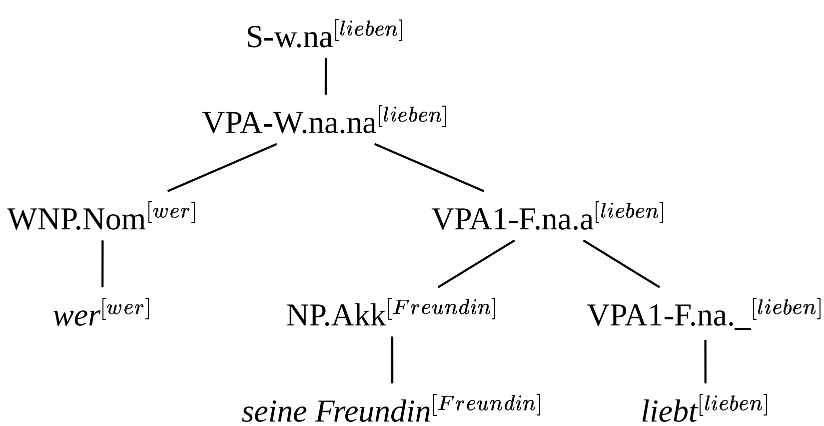
C <u>Relative Clauses</u>

Relative clauses are verb final clauses where the leftmost argument to collect is a noun phrase, prepositional phrase or non-finite clause containing a relative pronoun: `RNP`, `RPP`, `VPI-RC-max`. The clause type is indicated by `F` (as for verb final clauses) until the relative pronoun phrase is collected; then, the clause type is indicated by `RC`. An example analysis is given for *der seine Freundin liebt* 'who loves his girl-friend'. As for verb final clauses, finite and non-finite clauses might be subcategorised to the right of the finite verb.



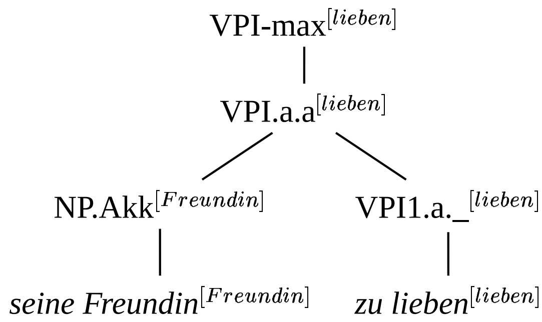
D <u>Indirect *wh*-Questions</u>

Indirect *wh*-questions are verb final clauses where the leftmost argument to collect is a noun phrase, a prepositional phrase, an adverb, or a non-finite clause containing a *wh*-phrase: `WNP`, `WPP`, `WADV`, `VPI-W-max`. The clause type is indicated by `F` (as for verb final clauses) until the *wh*-phrase is collected; then, the clause type is indicated by `W`. An example analysis is given for *wer seine Freundin liebt* 'who loves his girl-friend'. As for verb final clauses, finite and non-finite clauses might be subcategorised to the right of the finite verb.

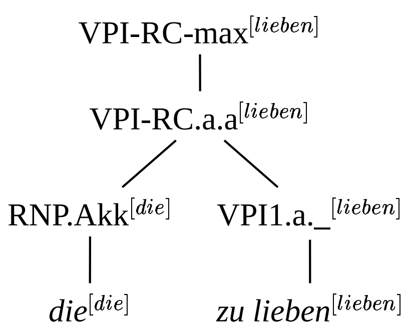


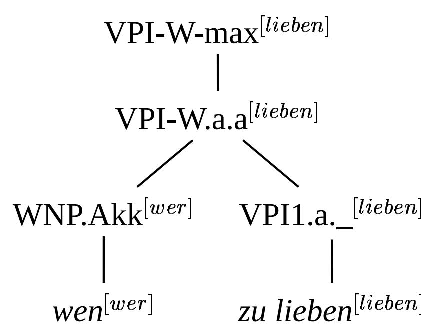
E <u>Non-Finite Clauses</u>

Non-finite clauses start collecting arguments from the non-finite verb complex and collect to the left only. As an exception, again, clausal arguments are collected to the right. An example analysis is given for *seine Freundin zu lieben* 'to love his girl-friend'. As mentioned before, `VPI-max` is a generalisation over all non-finite clauses. It is the relevant category for the subcategorisation of a non-finite clause.



Non-finite clauses might be the introductory part of a relative clause or a *wh*-question. In that case, the leftmost argument contains a relative pronoun or a *wh*-phrase, and the `VPI` category is marked by `RC` or `W`, respectively. The following examples analyse *die zu lieben* 'whom to love' and *wen zu lieben* 'whom to love'.





Noun Phrases The noun phrase structure is determined by practical needs: Noun phrases are to be recognised reliably, and nominal head information has to be passed through the nominal structure, but the noun phrase structure is kept simple without a theoretical claim. There are four nominal levels: the terminal noun `NN` is possibly modified by a cardinal number `CARD`, a genitive noun phrase `NP.Gen`, a prepositional phrase adjunct `PP-adjunct`, a proper name phrase `NEP`, or a clause `S-NN`, and is dominated by `N1`. `N1` itself may be modified by an (attributive) adjectival phrase `ADJaP` to reach `N2` which can be preceded by a determiner (`ART, DEM, INDEF, POSS`) to reach the `NP` level. All noun phrase levels are accompanied by the case feature. Figure 3.8 describes the noun phrase structure, assuming case agreement in the constituents. The clause label `S-NN` is a generalisation over all types of clauses allowed as noun modifier: `C-rel, C-dass, C-ob, C-w`. Example analyses are provided for the noun phrases *jener Mann mit dem Hut* 'that man with the hat'$_{Nom}$, *den alten Bauern Fehren* 'the old farmer Fehren'$_{Akk}$, and *der Tatsache, dass er schläft* 'the fact that he sleeps'$_{Gen}$.

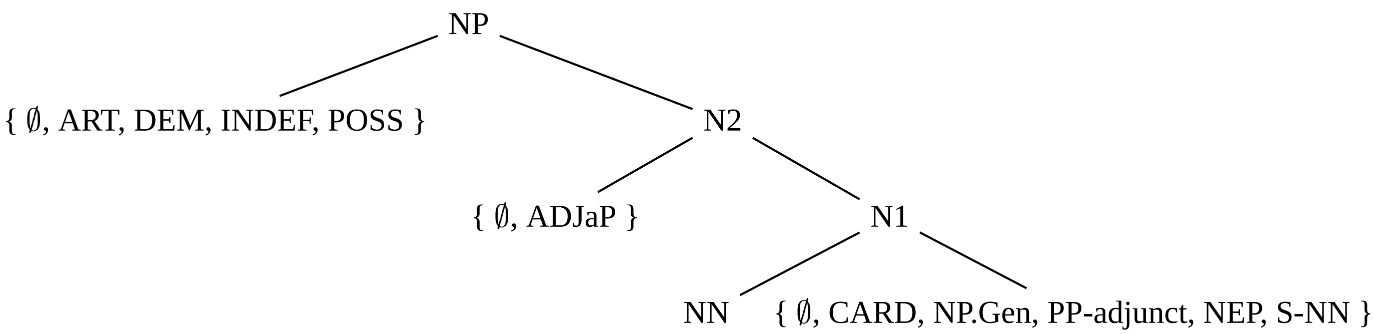
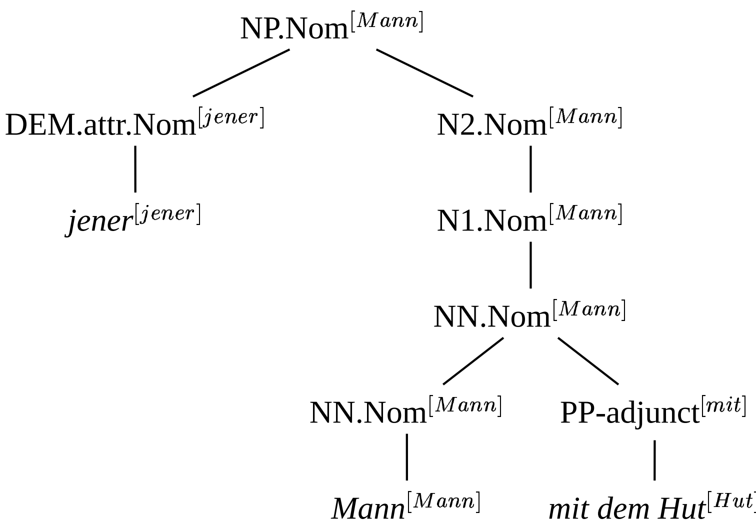
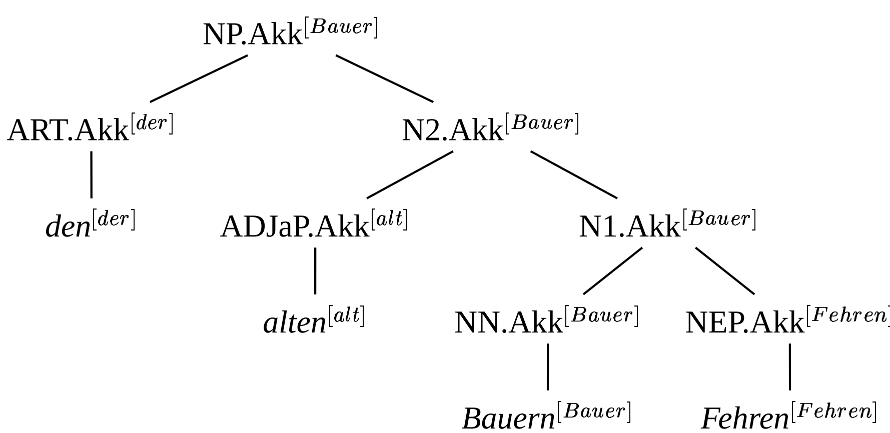


Figure 3.8: Nominal syntactic grammar categories





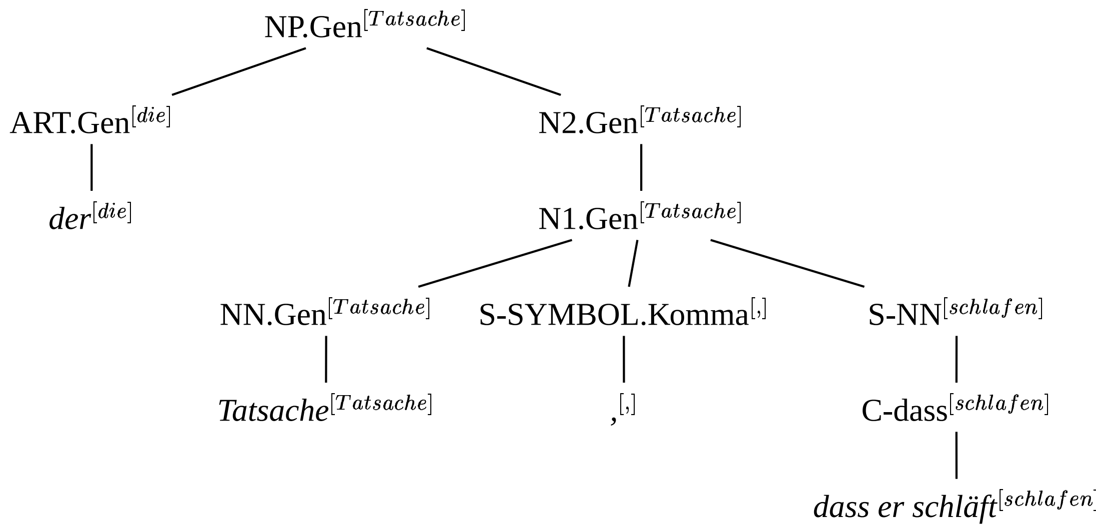


Figure 3.9 describes that proper name phrases `NEP` are simply defined as a list of proper names. As for common nouns, all levels are equipped with the case feature. Example analyses are provided for *New York*$_{Akk}$ and *der alte Peter* ‘the old Peter’$_{Nom}$.

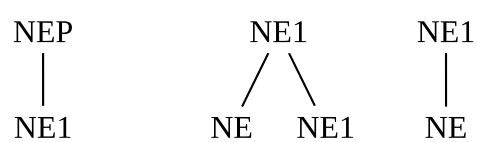
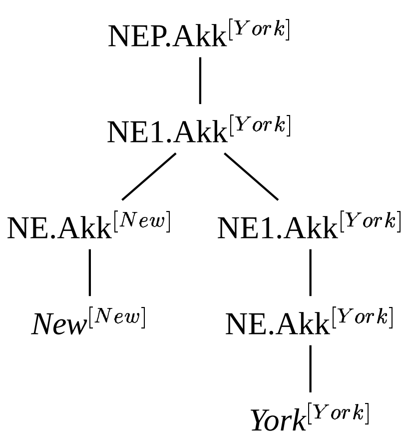


Figure 3.9: Proper names



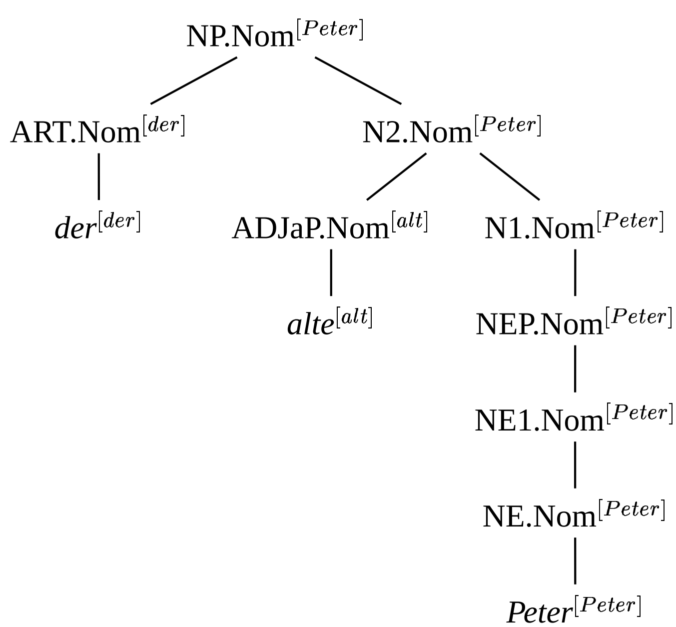


Figure 3.10 shows that noun phrases can generate pronouns and cardinal numbers, which do not allow modifiers. A number of examples is provided, illustrating the simple analyses for *ich* 'I'$_{Nom}$, *dich* 'you'$_{Akk}$, *einander* 'each other'$_{Akk}$, and *allen* 'everybody'$_{Dat}$.

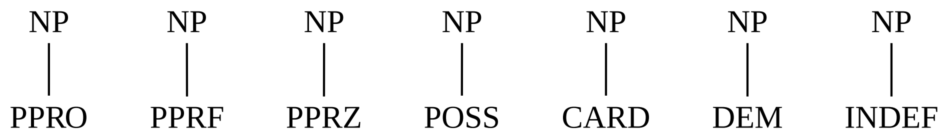
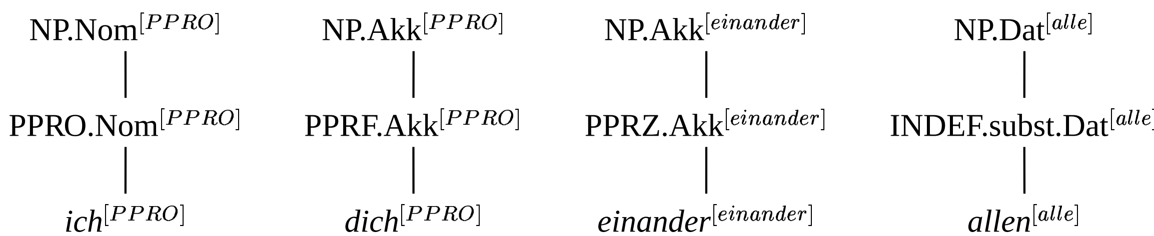


Figure 3.10: Noun phrases generating pronouns and cardinals



For relative and interrogative clauses, the specific kinds of `NP`s introducing the clause need to be defined, either as stand-alone pronoun, or attributively combined with a nominal on `N2` level. `RNP` and `WNP` are also equipped with the case feature. See the definition in Figure 3.11 and a number of example analyses for *der* 'who'$_{Nom}$, *dessen Bruder* 'whose brother'$_{Akk}$, and *wem* 'whom'$_{Dat}$.

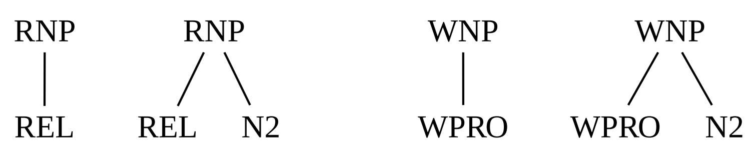
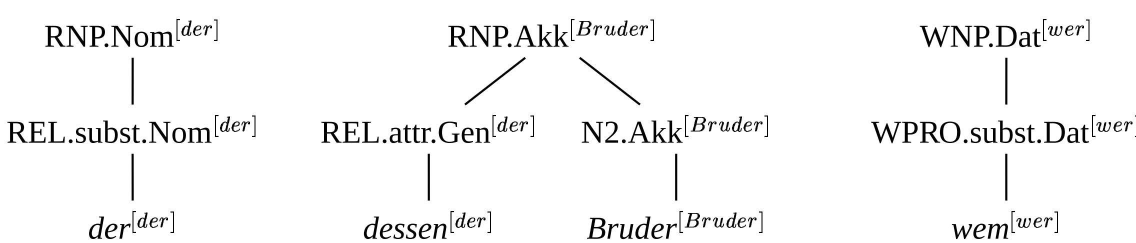


Figure 3.11: Noun phrases introducing relative and interrogative clauses



Prepositional Phrases Prepositional phrases are distinguished in their formation with respect to their syntactic function: (A) arguments vs. (B) adjuncts. By introducing both PP-arguments and PP-adjuncts I implicitly assume that the statistical grammar model is able to learn the distinction between the grammatical functions. But this distinction raises two questions:

1. Which is the distinction between PP-arguments and PP-adjuncts?

   As mentioned before, to distinguish between arguments and adjuncts I refer to the optionality of the complements. But with prepositional phrases, there is more to take into consideration. Standard German grammar such as Helbig and Buscha (1998, pages 402–404) categorise adpositions with respect to their usage in argument and adjunct PPs. With respect to PP-arguments, we distinguish verbs which are restricted to a single adposition as head of the PP (such as *achten auf* ‘to pay attention, to look after’) and verbs which require a PP of a certain semantic type, but the adpositions might vary (e.g. *sitzen* ‘to sit’ requires a local PP which might be realised by prepositions such as *auf*, *in*, etc.). Adpositions in the former kind of PP-arguments lose their lexical meaning in composition with a verb, so the verb-adposition combination acquires a non-compositional, idiosyncratic meaning. Typically, the complements of adpositions in PP-arguments are more restricted than in PP-adjuncts.

2. Is it possible to learn the distinction between PP-arguments and PP-adjuncts?

   To learn the distinction between PP-arguments and PP-adjunct is a specifically hard problem, because structurally each PP in the grammar can be parsed as argument and as adjunct, as the PP-implementation below will illustrate. The clues for the learning therefore lie in the distinction of the lexical relationships between verbs and adpositions and verbs and PP-subcategorised (nominal) head. The lexical distinction is built into the grammar rules as described below and even though not perfect actually helps the learning (cf. the grammar evaluation in Section 3.5).

A PP-Arguments

Prepositional phrase arguments combine the generated adposition with case information, i.e. `PP.<case>.<adposition>`. Basically, their syntactic structure requires an adposition or a comparing conjunction, and a noun or adverbial phrase, as Figure 3.12 shows. The head of the PP-argument is defined as the head of the nominal or adverbial phrase subcategorised by the adposition. By that, the definition of PP-arguments provides both the head information of the adposition in the category name (to learn the lexical relationship between verb and adposition) and the head information of the subcategorised phrase (to learn the lexical relationship between verb and PP-subcategorised nominal or adverbial head). Examples for the prepositional phrases in Figure 3.12 are *wie ein Idiot* ‘as an idiot’, *von drüben* ‘from over there’, *am Hafen* ‘at the port’, *meiner Mutter wegen* ‘because of my mother’.

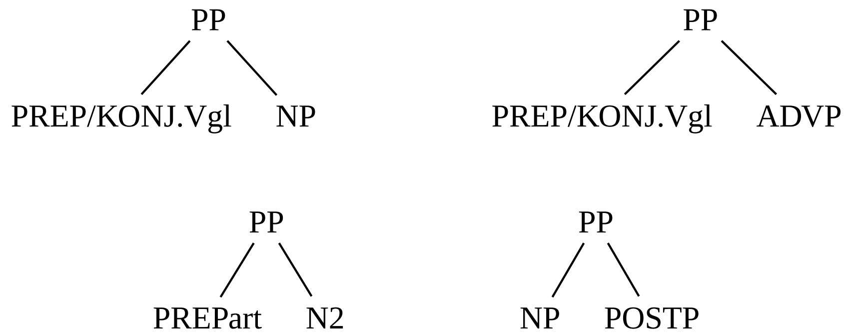
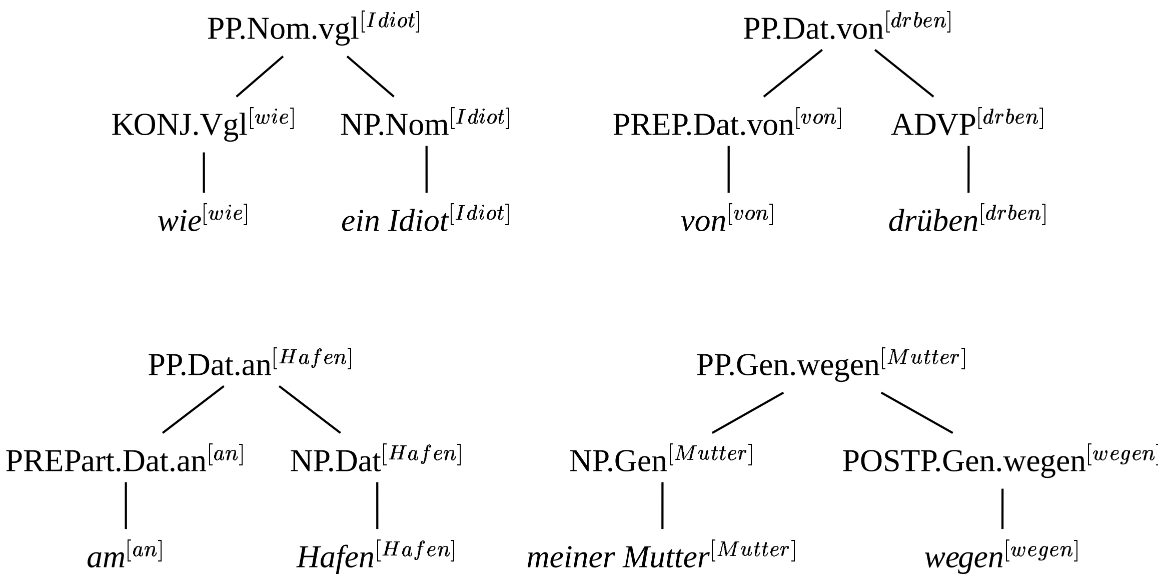


Figure 3.12: Prepositional phrase arguments



In addition, the prepositional phrases generate pronominal and interrogative adverbs if the preposition is the morphological head of the adverb, for example:

```
PP.Akk.für -> PROADV.dafür'
```

Like for noun phrases, the specific kinds of PP-arguments which introduce relative and interrogative clauses need to be defined. See the definition in Figure 3.13. Examples are given for *mit dem* ‘with whom’, *durch wessen Vater* ‘by whose father’, and *wofür* ‘for what’.

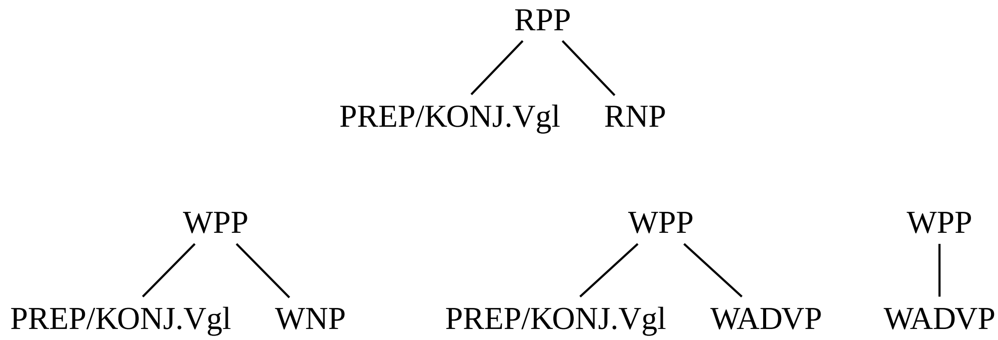
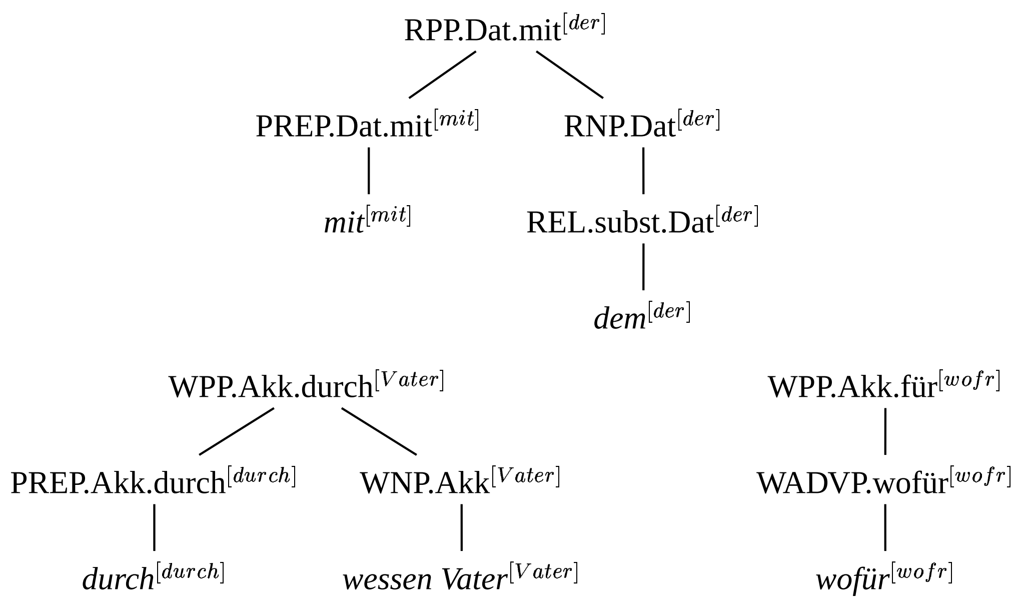
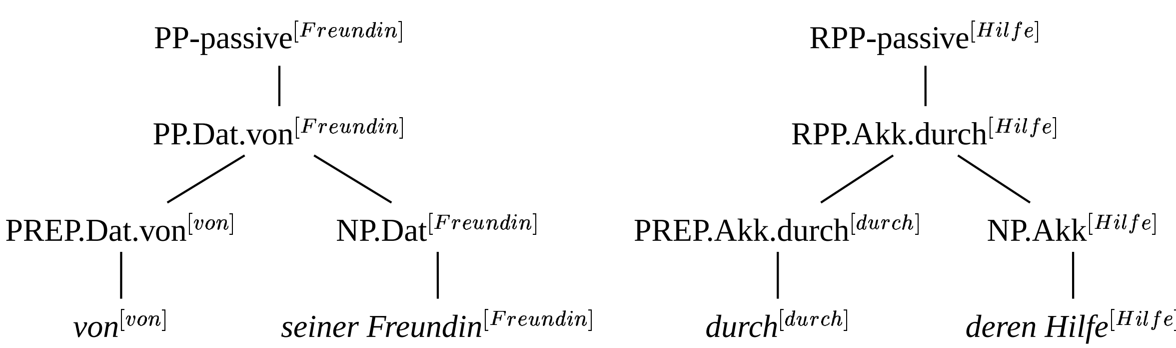


Figure 3.13: Prepositional phrase arguments in relative and interrogative clauses



Finally, a syntactically based category `(R/W)PP-passive` generates the two prepositional phrases `(R/W)PP.Akk.durch` and `(R/W)PP.Dat.von` as realisations of the deep structure subject in passive usage. See the examples for *von seiner Freundin* ‘by his girl-friend’, and *durch deren Hilfe* ‘by the help of who’.



B <u>PP-Adjuncts</u>

Prepositional phrase adjuncts are identified by the syntactic category `(R/W)PP-adjunct`. As `PP`-arguments, they require an adposition and a noun or adverbial phrase (cf. Figure 3.14), but the head of the PP-adjunct is the adposition, because the information subcategorised by the adposition is not considered relevant for the verb subcategorisation. Example analyses are provided for *bei dem Tor* 'at the gate', *nach draußen* 'to the outside', and *zu dem* 'towards who'.

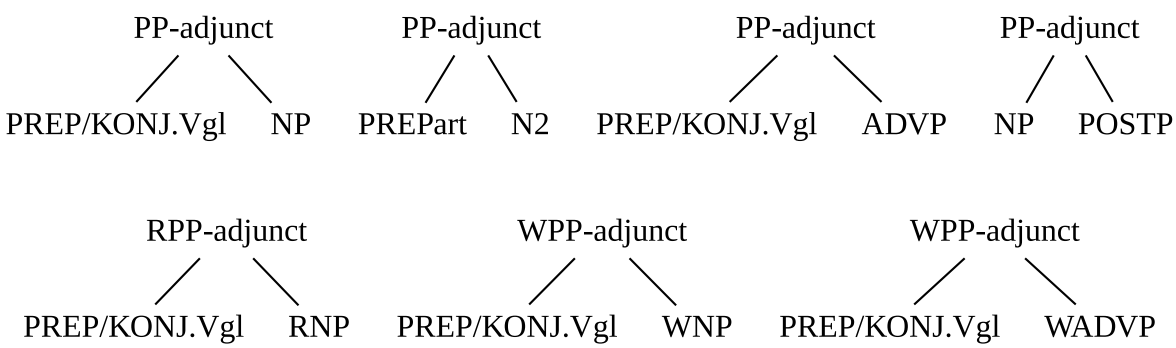
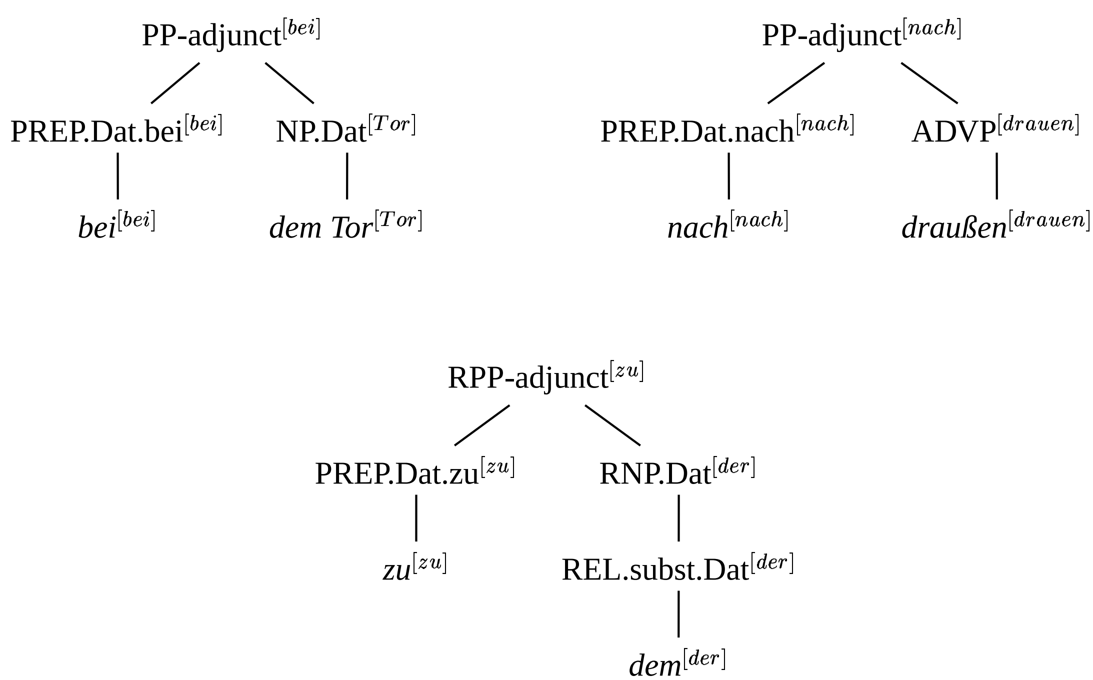


Figure 3.14: Prepositional phrase adjuncts



Adjectival Phrases Adjectival phrases distinguish between (A) an attributive and (B) a predicative usage of the adjectives.

A Attributive Adjectives

Attributive adjectival phrases are realised by a list of attributive adjectives. The adjectives are required to agree in case. Terminal categories other than declinable attributive adjectives are indeclinable adjectives, and cardinal and ordinal numbers. The attributive adjective formation is illustrated in Figure 3.15. Attributive adjectives on the bar level might be combined with adverbial adjuncts. Example analyses are provided for *tollen alten* 'great old'$_{Akk}$, and *ganz lila* 'completely pink'$_{Nom}$.

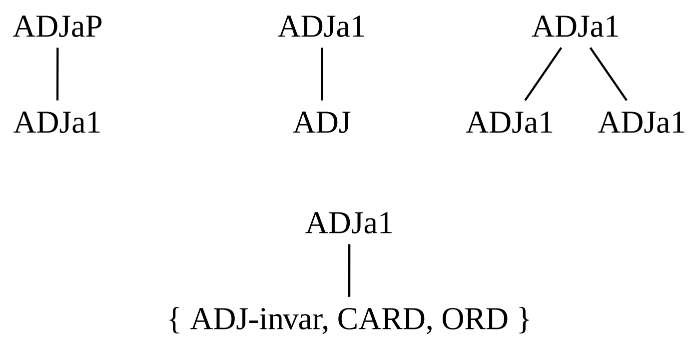
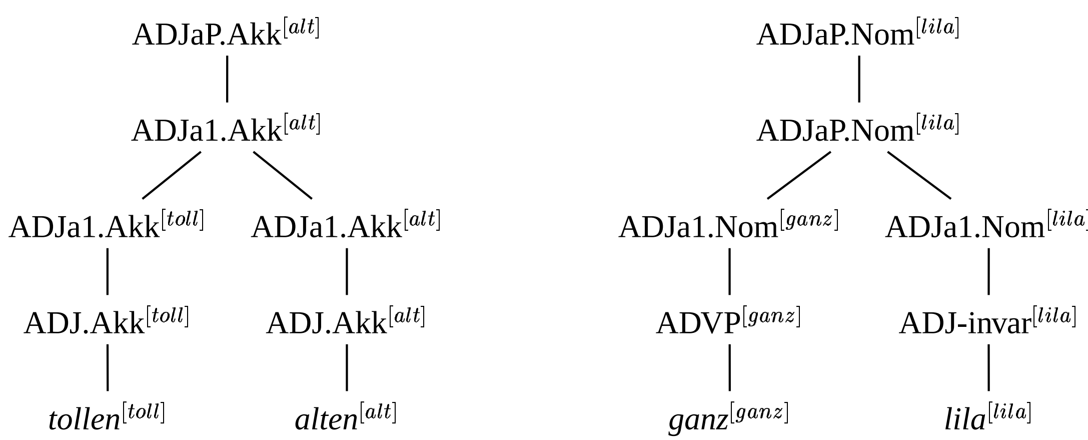


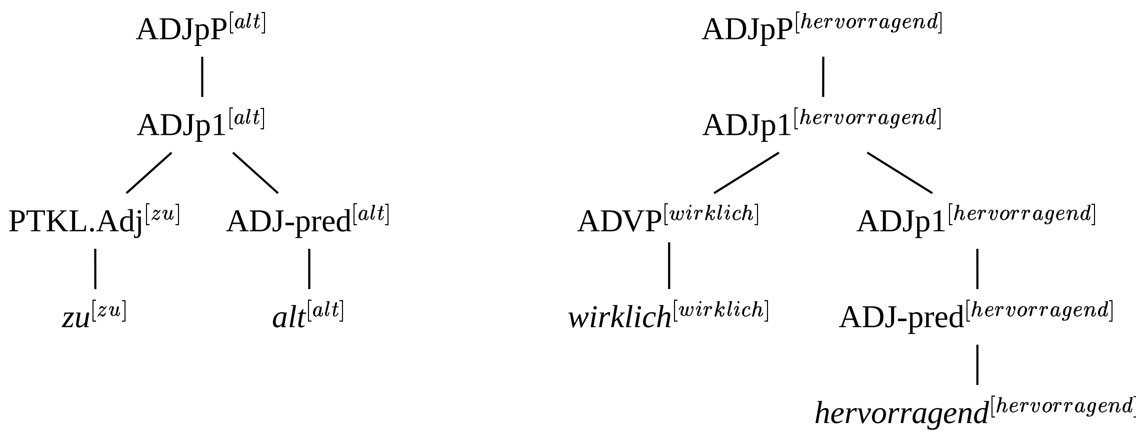
Figure 3.15: Attributive adjectival phrases



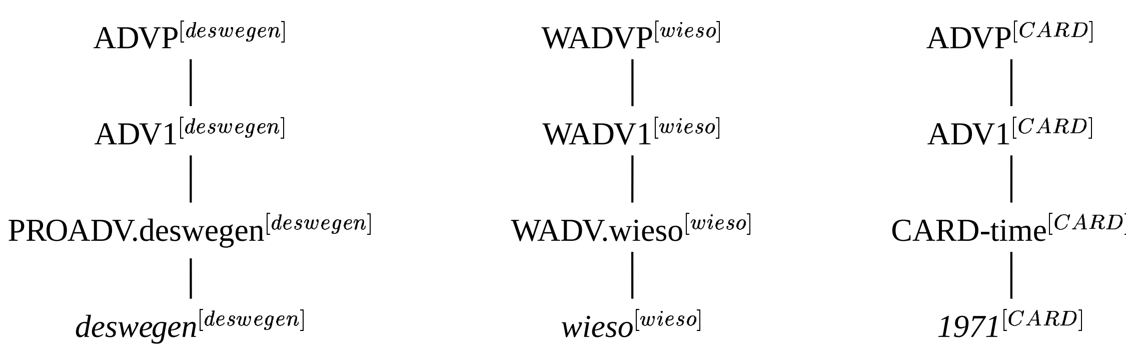
B Predicative Adjectives

Predicative adjectival phrases are realised by a predicative adjective (possibly modified by a particle), or by an indeclinable adjective, as displayed by Figure 3.16. As attributive adjectival phrases, the predicative adjectives on the bar level might be combined with adverbial adjuncts. Example analyses are given for *zu alt* ‘too old’ and *wirklich hervorragend* ‘really excellent’.

Figure 3.16: Predicative adjectival phrases



Adverbial Phrases Adverbial phrases `(W)ADVP` are realised by adverbs, pronominal or interrogative adverbs. Terminal categories other than adverbs are predicative adjectives, particles, interjections, and year numbers. The adverbial formation is illustrated in Figure 3.17, and examples are provided for *deswegen* ‘because of that’, *wieso* ‘why’, and *1971*. The lexical head of cardinal numbers is `CARD`.



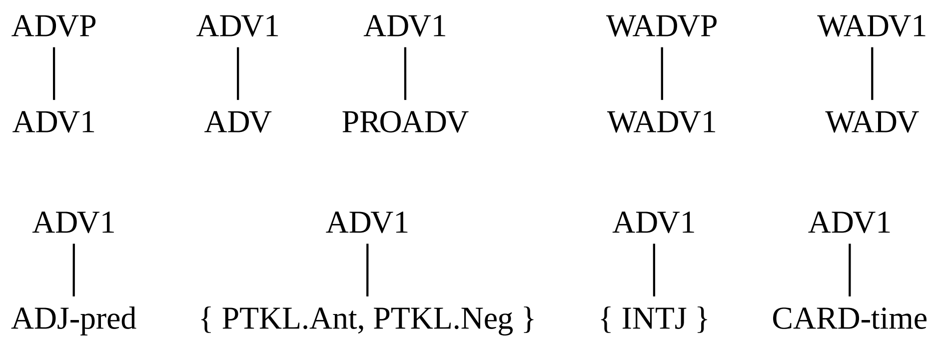
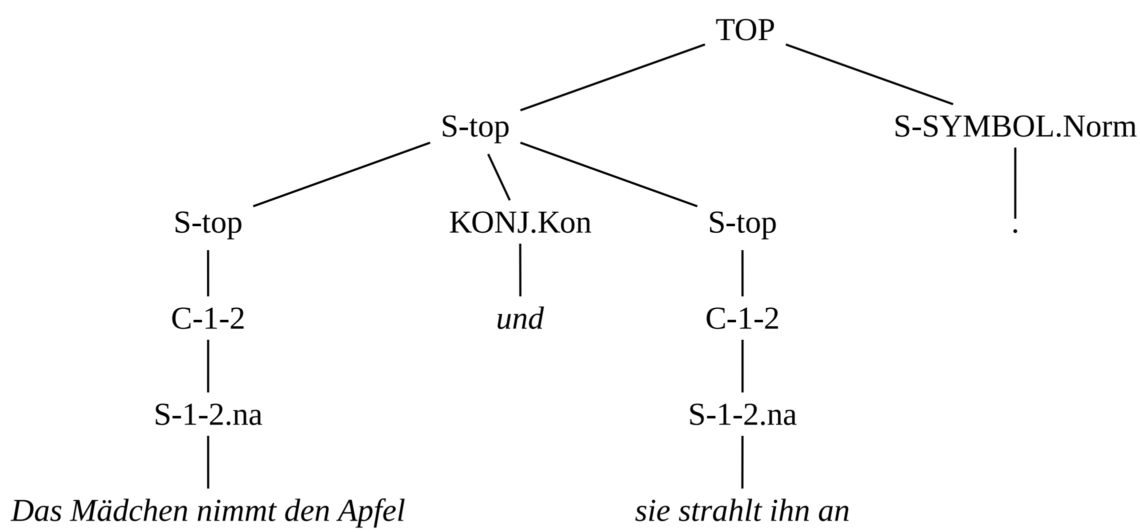


Figure 3.17: Adverbial phrases

Coordination Since coordination rules extensively inflate the grammar, coordination is only applied to specific grammar levels. Noun phrases, prepositional phrases, adjectival and adverbial phrases are combined on the phrase level only. For example, the structurally ambiguous NP *die alten Männer und Frauen* 'the old men and women' is analysed as [[die alten Männer]$_{NP}$ & [Frauen]$_{NP}$], but not as [die [alten Männer]$_{N2}$ & [Frauen]$_{N2}$] or [die alten [Männer$_{N1}$ & Frauen$_{N1}$]], since coordination only applies to NP, but not to N2 or N1. Coordination of verb units is performed on fully saturated verb phrases (via the `S-top` level) and on verb complexes. For example, the grammar fails in parsing *Das Mädchen nimmt den Apfel und strahlt ihn an* 'the girl takes the apple and smiles at him', because it would need to combine a fully saturated `VPA.na` with a `VPA.na` missing the subject. In contrast, the grammar is able to parse *Das Mädchen nimmt den Apfel und sie strahlt ihn an* 'the girl takes the apple and she smiles at him', because it combines two fully saturated `VPA.na` at the `S-top` level:



The restriction on coordination is a compromise between the necessity of including coordination into the grammar and the large number of parameters resulting from integrating coordination for all possible categories and levels, especially with respect to the fine-grained subcategorisation information in the grammar.

3.3 Grammar Training

The previous section has described the development and implementation of the German context-free grammar. This section uses the context-free backbone as basis for the lexicalised probabilistic extension, to learn the statistical grammar model. The grammar training is performed by the statistical parser `LoPar` (Schmid, 2000). Section 3.3.1 introduces the key features of the parser, and Section 3.3.2 describes the training strategy to learn the statistical grammar model.

3.3.1 The Statistical Parser

`LoPar` is an implementation of the left-corner parsing algorithm. Its functionality comprises symbolic parsing with context-free grammars, and probabilistic training and parsing with probabilistic context-free grammars and head-lexicalised probabilistic context-free grammars. In addition, the parser can be applied for Viterbi parsing, tagging and chunking.

`LoPar` executes the parameter training of the probabilistic context-free grammars by the *Inside-Outside Algorithm* (Lari and Young, 1990), an instance of the *Expectation-Maximisation (EM) Algorithm* (Baum, 1972). The EM-algorithm is an unsupervised iterative technique for maximum likelihood approximation of training data. Each iteration in the training process consists of an estimation (E) and a maximisation (M) step. The E-step evaluates a probability distribution for the data given the model parameters from the previous iteration. The M-step then finds the new parameter set that maximises the probability distribution. So the model parameters are improved by alternately assessing frequencies and estimating probabilities. The EM-algorithm is guaranteed to find a local optimum in the search space. EM is sensitive to the initialisation of the model parameters. For the *Inside-Outside Algorithm*, the EM-parameters refer to grammar-specific training data, i.e. how to determine the probabilities of sentences with respect to a grammar. The training is based on the notion of grammar categories and estimates the parameters producing a category ('outside' the category with respect to a tree structure) and the parameters produced by a category ('inside' the category with respect to a tree structure), hence the name. The parameter training with `LoPar` is performed by first optimising the PCFG parameters, then using the PCFG parameters for a bootstrapping of the lexicalised H-L PCFG model, and finally optimising the H-L PCFG parameters.

According to Manning and Schütze (1999), a main problem of H-L PCFGs is that for discriminating the large number of parameters a sufficient amount of linguistic data is required. The sparse data problem is pervasive, so effective smoothing techniques are necessary. `LoPar` implements four ways of incorporating sparse data into the probabilistic model:

(i) The number of parameters is reduced by allowing lemmatised word forms instead of fully inflected word forms.

(ii) All unknown words are tagged with the single token `<unknown>` which also propagates as lexical head. A set of categories for unknown words may be determined manually before the parsing process, e.g. noun tags are assigned by default to capitalised unknown words, and verb tags or adjective tags to non-capitalised unknown words. This handling prevents the parser from failing on sentences with unknown words.

(iii) Parameter smoothing is performed by absolute discounting. The smoothing technique as defined by Ney *et al.* (1994) subtracts a fixed discount from each non-zero parameter value and redistributes the mass of the discounts over unseen events.

(iv) The parameters of the head-lexicalised probabilistic context-free grammar can be manually generalised for reduction (see 'parameter reduction' below on details).

3.3.2 Training Strategy

The training strategy is the result of experimental work on H-L PCFGs for German, since there is no 'rule of thumb' for the parameter training which is valid for all possible setups. Former versions of the training setup and process are reported by Beil *et al.* (1999), Schulte im Walde (2000b) and Schulte im Walde *et al.* (2001). The latter reference contains an evaluation of diverse training strategies.

Training Corpus As training corpus for the German grammar model, I use parts of a large German newspaper corpus from the 1990s, which is referred to as the *Huge German Corpus (HGC)*. The HGC contains approximately 200 million words of newspaper text from *Frankfurter Rundschau*, *Stuttgarter Zeitung*, *VDI-Nachrichten*, *die tageszeitung*, *German Law Corpus*, *Donaukurier*, and *Computerzeitung*.

The corpus training data should be as numerous as possible, so the training should be performed on all 200 million words accessible. On the other hand, time constraints make it necessary to restrict the amount of data. The following training parameters have been developed out of experience and as a compromise between data and time demands.

- All 6,591,340 sentences (82,149,739 word tokens) from the HGC with a length between 5 and 20 words are used for unlexicalised training. The grammar has a coverage[5] of parsing 68.03% of the sentences, so effectively the training is performed on 4,484,089 sentences.

- All 2,426,925 sentences (18,667,888 word tokens) from the HGC with a length between 5 and 10 words are used for the lexicalisation, the bootstrapping of the lexicalised grammar model. The grammar has a coverage of parsing 71.75% of the sentences, so effectively the bootstrapping is performed on 1,741,319 sentences.

- All 3,793,768 sentences (35,061,874 word tokens) from the HGC with a length between 5 and 13 words are used for lexicalised training. The grammar has a coverage of parsing 71.74% of the sentences, so effectively the training is performed on 2,721,649 sentences.

[5]The *coverage* of the grammar refers to the percentage of sentences from the corpus which are assigned at least one parse analysis. The sentences without an analysis are not taken into consideration for in training process.

Initialisation and Training Iterations The initialisation of the PCFG grammar parameters is performed by assigning the same frequency to all grammar rules. Comparable initialisations with random frequencies had no effect on the model development (Schulte im Walde, 2000b). The parameter estimation is performed within one iteration for unlexicalised training of the PCFG, and three iterations for lexicalised training of the H-L PCFG. The overall training process takes 15 days on a Sun Enterprise 450 with 296 MHz CPU.

Parameter Reduction As mentioned before, `LoPar` allows a manual generalisation to reduce the number of parameters. The key idea is that lexical heads which are supposed to overlap for different grammar categories are tied together. For example, the direct objects of *kaufen* 'to buy' are the same irrespective of the degree of saturation of a verb phrase and also irrespective of the clause type. Therefore, I can generalise over the transitive verb phrase types `VPA1-1-2.na._, VPA1-1-2.na.n, VPA1.1-2.na.a, VPA-1-2.na.na` and include the generalisation over the different clause types `1-2, rel, sub, dass, ob, w`. In addition, we can generalise over certain arguments in active and passive and in finite and non-finite verb phrases, for example the accusative object in an active finite clause `VPA` for frame type `na` and the accusative object in an active non-finite clause `VPI` for frame type `a`. The generalisation is relevant for the lexicalised grammar model and is performed for all verb phrase types. The parameter reduction in the grammar is especially important because of the large number of subcategorisation rules.

Summary We can summarise the process of grammar development and training strategy in the following steps.

1. Manual definition of CFG rules with head-specification,
2. Assigning uniform frequencies to CFG rules (extension of CFG to PCFG),
3. Unlexicalised training of the PCFG: one iteration on approx. 82 million words,
4. Manual definition of grammar categories for parameter reduction,
5. Lexicalisation of the PCFG (bootstrapping of H-L PCFG) on approx. 19 million words,
6. Lexicalised training of the H-L PCFG: three iterations on approx. 35 million words.

3.4 Grammar-Based Empirical Lexical Acquisition

The previous sections in this chapter have introduced the German grammar implementation and training. The resulting statistical grammar model provides empirical lexical information, specialising on but not restricted to the subcategorisation behaviour of verbs. In the following, I present examples of such lexical information. The examples are selected with regard to the lexical verb descriptions at the syntax-semantic interface which I will use in the clustering experiments.

Section 3.4.1 describes the induction of subcategorisation frames for the verbs in the German grammar model, and Section 3.4.2 illustrates the acquisition of selectional preferences. In Section 3.4.3 I present related work on the automatic acquisition of lexical information within the framework of H-L PCFGs.

3.4.1 Subcategorisation Frames

The acquisition of subcategorisation frames is directly related to the grammar implementation. Recall the definition of clause types: The clause level `C` produces the clause category `S` which is accompanied by the relevant subcategorisation frame dominating the clause. Each time a clause is analysed by the statistical parser, a clause level rule with the relevant frame type is included in the analysis.

$$\texttt{C-<type>} \rightarrow \texttt{S-<type>.<frame>}$$

The PCFG extension of the German grammar assigns frequencies to the grammar rules according to corpus appearance and is able to distinguish the relevance of different frame types. The usage of subcategorisation frames in the corpus is empirically trained.

$$\begin{array}{ll}
\texttt{freq}_1 & \texttt{C-<type>} \rightarrow \texttt{S-<type>.<frame}_1\texttt{>} \\
\texttt{freq}_2 & \texttt{C-<type>} \rightarrow \texttt{S-<type>.<frame}_2\texttt{>} \\
\texttt{freq}_{...} & \texttt{C-<type>} \rightarrow \texttt{S-<type>.<frame}_{...}\texttt{>} \\
\texttt{freq}_n & \texttt{C-<type>} \rightarrow \texttt{S-<type>.<frame}_n\texttt{>}
\end{array}$$

But we are interested in the idiosyncratic, lexical usage of the verbs. The H-L PCFG lexicalisation of the grammar rules with their verb heads leads to a lexicalised distribution over frame types.

$$\begin{array}{ll}
\texttt{freq}_1 & \texttt{C-<type>}^{[verb]} \rightarrow \texttt{S-<type>.<frame}_1\texttt{>} \\
\texttt{freq}_2 & \texttt{C-<type>}^{[verb]} \rightarrow \texttt{S-<type>.<frame}_2\texttt{>} \\
\texttt{freq}_{...} & \texttt{C-<type>}^{[verb]} \rightarrow \texttt{S-<type>.<frame}_{...}\texttt{>} \\
\texttt{freq}_n & \texttt{C-<type>}^{[verb]} \rightarrow \texttt{S-<type>.<frame}_n\texttt{>}
\end{array}$$

Generalising over the clause type, the combination of grammar rules and lexical head information provides distributions for each verb over its subcategorisation frame properties.

$$
\begin{array}{lll}
\texttt{freq}_1 & \texttt{C}^{[verb]} & \rightarrow \texttt{S.<frame}_1\texttt{>} \\
\texttt{freq}_2 & \texttt{C}^{[verb]} & \rightarrow \texttt{S.<frame}_2\texttt{>} \\
\texttt{freq}_{...} & \texttt{C}^{[verb]} & \rightarrow \texttt{S.<frame}_{...}\texttt{>} \\
\texttt{freq}_n & \texttt{C}^{[verb]} & \rightarrow \texttt{S.<frame}_n\texttt{>}
\end{array}
$$

An example of such a purely syntactic subcategorisation distribution is given in Table 3.16. The table lists the 38 subcategorisation frame types in the grammar sorted by the joint frequency with the verb *glauben* 'to think, to believe'. In this example as well as in all following examples on frequency extraction from the grammar, the reader might wonder why the frequencies are real values and not necessarily integers. This has to do with the training algorithm which splits a frequency of 1 for each sentence in the corpus over all ambiguous parses. Therefore, rule and lexical parameters might be assigned a fraction of 1.

In addition to a purely syntactic definition of subcategorisation frames, the grammar provides detailed information about the types of argument PPs within the frames. For each of the prepositional phrase frame types in the grammar (`np, nap, ndp, npr, xp`), the joint frequency of a verb and the PP frame is distributed over the prepositional phrases, according to their frequencies in the corpus. For example, Table 3.17 illustrates the subcategorisation for *reden* 'to talk' and the frame type `np` whose total joint frequency is 1,121.35.

3.4.2 Selectional Preferences

The grammar provides selectional preference information on a fine-grained level: it specifies the possible argument realisations in form of lexical heads, with reference to a specific verb-frame-slot combination. I.e. the grammar provides frequencies for heads for each verb and each frame type and each argument slot of the frame type. The verb-argument frequencies are regarded as a particular strength of the statistical model, since the relationship between verb and selected subcategorised head refers to fine-grained frame roles. For illustration purposes, Table 3.18 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 3.19 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.

3.4.3 Related Work on H-L PCFGs

There is a large amount of work on the automatic induction of lexical information. In this section, I therefore concentrate on the description of related work within the framework of H-L PCFGs.

With reference to my own work, Schulte im Walde (2002b) presents a large-scale computational subcategorisation lexicon for 14,229 German verbs with a frequency between 1 and 255,676.

The lexicon is based on the subcategorisation frame acquisition as illustrated in Section 3.4.1. Since the subcategorisation frames represent the core part of the verb description in this thesis, the lexicon is described in more detail and evaluated against manual dictionary definitions in Section 3.5. The section also describes related work on subcategorisation acquisition in more detail.

Schulte im Walde (2003a) presents a database of collocations for German verbs and nouns. The collocations are induced from the statistical grammar model. 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 noun phrase 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 presented. All collocation types are combined by a perl script which can be queried by the lexicographic user in order to extract relevant co-occurrence information on a specific lexical item. The database is ready to be used for lexicographic research and exploitation.

Zinsmeister and Heid (2002, 2003b) utilise the same statistical grammar framework for lexical induction: Zinsmeister and Heid (2002) perform an extraction of noun-verb collocations, whose results represent the basis for comparing the collocational preferences of compound nouns with those of the respective base nouns. The insights obtained in this way are used to improve the lexicon of the statistical parser. Zinsmeister and Heid (2003b) 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.

Frame Type	Freq
ns-dass	1,928.52
ns-2	1,887.97
np	686.76
n	608.05
na	555.23
ni	346.10
nd	234.09
nad	160.45
nds-2	69.76
nai	61.67
ns-w	59.31
nas-w	46.99
nap	40.99
nr	31.37
nar	30.10
nrs-2	26.99
ndp	24.56
nas-dass	23.58
nas-2	19.41
npr	18.00
nds-dass	17.45
ndi	11.08
nrs-w	2.00
nrs-dass	2.00
ndr	2.00
nir	1.84
nds-w	1.68
xd	1.14
ns-ob	1.00
nas-ob	1.00
x	0.00
xa	0.00
xp	0.00
xr	0.00
xs-dass	0.00
nds-ob	0.00
nrs-ob	0.00
k	0.00

Table 3.16: Subcategorisation frame distribution for *glauben*

Refined Frame Type	Freq
np:Akk.über	479.97
np:Dat.von	463.42
np:Dat.mit	279.76
np:Dat.in	81.35
np:Nom.vgl	13.59
np:Dat.bei	13.10
np:Dat.über	13.05
np:Dat.an	12.06
np:Akk.für	9.63
np:Dat.nach	8.49
np:Dat.zu	7.20
np:Dat.vor	6.75
np:Akk.in	5.86
np:Dat.aus	4.78
np:Gen.statt	4.70
np:Dat.auf	4.34
np:Dat.unter	3.77
np:Akk.vgl	3.55
np:Akk.ohne	3.05
np:Dat.hinter	3.00
np:Dat.seit	2.21
np:Dat.neben	2.20
np:Dat.wegen	2.13
np:Akk.gegen	2.13
np:Akk.an	1.98
np:Gen.wegen	1.77
np:Akk.um	1.66
np:Akk.bis	1.15
np:Akk.ab	1.13
np:Dat.laut	1.00
np:Gen.hinsichtlich	1.00
np:Gen.während	0.95
np:Dat.zwischen	0.92
np:Akk.durch	0.75

Table 3.17: Refined np distribution for *reden*

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 3.18: 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 3.19: Nominal arguments for *reden über*$_{Akk}$ ‘to talk about’

3.5 Grammar Evaluation

This final part of the grammar chapter describes an evaluation performed on the core of the grammar, its subcategorisation frames. I evaluated the verb subcategorisation frames which are learned in the statistical grammar framework against manual definitions in the German dictionary *Duden – Das Stilwörterbuch*. The work was performed in collaboration with *Bibliographisches Institut & F. A. Brockhaus AG* who provided a machine readable version of the dictionary. The evaluation is published by Schulte im Walde (2002a).

Section 3.5.1 describes the definition of verb subcategorisation frames (i) in the large-scale computational subcategorisation lexicon based on the statistical grammar model and (ii) in the manual dictionary *Duden*. In Section 3.5.2 the evaluation experiment is performed, Section 3.5.3 contains an interpretation of the experiment results, and Section 3.5.4 compares them with related work on English and German subcategorisation induction.

3.5.1 Subcategorisation Lexica for Verbs

Learning a Verb Subcategorisation Lexicon

Schulte im Walde (2002b) presents a large-scale computational subcategorisation lexicon. The lexicon is based on the empirical subcategorisation frame acquisition as illustrated in Section 3.4.1. The induction of the subcategorisation lexicon uses the trained frequency distributions over frame types for each verb. The frequency values are manipulated by squaring them, in order to achieve a more clear-cut threshold for lexical subcategorisation. The manipulated values are normalised and a cut-off of 1% defines those frames which are part of the lexical verb entry.

The manipulation is no high mathematical transformation, but it has the following impact on the frequency distributions. Assume verb v_1 has a frequency of 50 for the frame f_a and a frequency of 10 for frame f_b; verb v_2 has a frequency of 500 for the frame f_a and a frequency of 10 for frame f_b. If we set the cut-off to a frequency of 10, for example, then for both verbs both frames f_a and f_b are listed in the subcategorisation lexicon (but note that f_b is empirically less confirmed for v_2 than for v_1). If we set the cut-off to a frequency of 50, for example, then v_1 would have no frame listed at all. It is difficult to find a reliable cut-off. If we based the decision on the respective probability values p_a and p_b ($\langle v_1, p_a \rangle = 0.83$, $\langle v_1, p_b \rangle = 0.17$, $\langle v_2, p_a \rangle = 0.98$, $\langle v_2, p_b \rangle = 0.02$) it is easier to find a reliable cut-off, but still difficult for a large number of examples. But if we first square the frequencies ($\langle v_1, f'_a \rangle = 250$, $\langle v_1, f'_b \rangle = 100$, $\langle v_2, f'_a \rangle = 250,000$, $\langle v_2, f'_b \rangle = 100$), the respective probability values ($\langle v_1, p'_a \rangle = 0.71$, $\langle v_1, p'_b \rangle = 0.29$, $\langle v_2, p'_a \rangle = 0.9996$, $\langle v_2, p'_b \rangle = 0.0004$) are stretched, and it is not as difficult as before to find a suitable cut-off.

Tables 3.20 and 3.21 cite the (original and manipulated) frequencies and probabilities for the verbs *befreien* 'to free' and *zehren* 'to live on, to wear down' and mark the demarcation of lexicon-relevant frames by an extra line in the rows on manipulated numbers. The set of marked frames corresponds to the lexical subcategorisation for the respective verb.

Frame	Freq (orig)	Prob (orig)	Freq (mani)	Prob (mani)
na	310.50	0.43313	96,410.25	0.74293
nr	137.14	0.19130	18,807.38	0.14493
nap	95.10	0.13266	9,044.01	0.06969
n	59.04	0.08236	3,485.72	0.02686
nad	29.62	0.04132	877.34	0.00676
npr	23.27	0.03246	541.49	0.00417
np	15.04	0.02098	226.20	0.00174
nd	11.88	0.01657	141.13	0.00109
ndr	11.87	0.01656	140.90	0.00109
ns-2	7.46	0.01041	55.65	0.00043
nar	3.00	0.00418	9.00	0.00007
nrs-2	3.00	0.00418	9.00	0.00007
nds-2	2.94	0.00418	8.64	0.00007
nai	2.01	0.00280	4.04	0.00003
nir	2.00	0.00279	4.00	0.00003
ni	2.00	0.00279	4.00	0.00003
nas-2	1.00	0.00139	1.00	0.00001

Lexical subcategorisation: { n, na, nr, nap }

Table 3.20: Lexical subcategorisation for *befreien*

Frame	Freq (orig)	Prob (orig)	Freq (mani)	Prob (mani)
n	43.20	0.47110	1866.24	0.54826
np	38.71	0.42214	1498.46	0.44022
na	4.79	0.05224	22.94	0.00674
nap	3.87	0.04220	14.98	0.00440
nd	1.13	0.01232	1.28	0.00038

Lexical subcategorisation: { n, np }

Table 3.21: Lexical subcategorisation for *zehren*

A refined version of subcategorisation frames includes the specific kinds of prepositional phrases for PP-arguments. The frame frequency values and the PP frequency values are also manipulated by squaring them, and the manipulated values are normalised. The product of frame probability and PP probability is calculated, and a cut-off of 20% defines those PP frame types which are part of the lexical verb entry. The resulting lexical subcategorisation for *befreien* would be { n, na, nr, nap:Dat.von, nap:Dat.aus }, for *zehren* { n, np:Dat.von, np:Dat.an }.

I collected frames for all lexical items that were identified as verbs in the training corpus at least once, according to the definitions in the German morphological analyser AMOR underlying the grammar terminals. The resulting verb lexicon on subcategorisation frames contains 14,229 German verbs with a frequency between 1 and 255,676. Examples for lexical entries in the subcategorisation are given by Table 3.22 on the purely syntactic frame types, and by Table 3.23 on the PP-refined frame types.

Lexicon Entry			
Verb		Freq	Subcategorisation
aufregen	'to get excited'	135	na, nr
beauftragen	'to order', 'to charge'	230	na, nap, nai
bezweifeln	'to doubt'	301	na, ns-dass, ns-ob
bleiben	'to stay', 'to remain'	20,082	n, k
brechen	'to break'	786	n, na, nad, nar
entziehen	'to take away'	410	nad, ndr
irren	'to be mistaken'	276	n, nr
mangeln	'to lack'	438	x, xd, xp
scheinen	'to shine', 'to seem'	4,917	n, ni
sträuben	'to resist'	86	nr, npr

Table 3.22: Examples for purely syntactic lexical subcategorisation entries

Lexicon Entry			
Verb		Freq	Subcategorisation
beauftragen	'to order', 'to charge'	230	na, nap:Dat.mit, nai
denken	'to think'	3,293	n, na, np:Akk.an, ns-2
enden	'to end'	1,900	n, np:Dat.mit
ernennen	'to appoint'	277	na, nap:Dat.zu
fahnden	'to search'	163	np:Dat.nach
klammern	'to cling to'	49	npr:Akk.an
schätzen	'to estimate'	1,357	na, nap:Akk.auf
stapeln	'to pile up'	137	nr, npr:Dat.auf, npr:Dat.in
sträuben	'to resist'	86	nr, npr:Akk.gegen
tarnen	'to camouflage'	32	na, nr, npr:Nom.vgl

Table 3.23: Examples for PP-refined lexical subcategorisation entries

Manual Definition of Subcategorisation Frames in Dictionary *Duden*

The German dictionary *Duden – Das Stilwörterbuch* (Dudenredaktion, 2001) describes the stylistic usage of words in sentences, such as their syntactic embedding, example sentences, and idiomatic expressions. Part of the lexical verb entries are frame-like syntactic descriptions, such as `<jmdn. befreien>` ‘to free somebody’ with the direct object indicated by the accusative case, or `<von etw. zehren>` ‘to live on something$_{Dat}$’.

Duden does not contain explicit subcategorisation frames, since it is not meant to be a subcategorisation lexicon. But it does contain ‘grammatical information’ for the description of the stylistic usage of verbs; therefore, the *Duden* entries implicitly contain subcategorisation, which enables us to infer frame definitions.

Alternations in verb meaning are marked by a semantic numbering `SEMX-ID` and accompanied by the respective subcategorisation requirements (`GR` provides the subcategorisation, `DEF` provides a semantic description of the respective verb usage, and `TEXT` under `BSP` provides examples for selectional preferences). For example, the lexical verb entry for *zehren* in Figure 3.18 lists the following lexical semantic verb entries:

1. `<von etw. zehren>` ‘to live on something’
2. ‘to drain somebody of his energy’
   a) no frame which implicitly refers to an intransitive usage
   b) `<an jmdm., etw. zehren>`

Idiosyncrasies in the manual frame definitions lead to a total of 1,221 different subcategorisation frames in *Duden*:

- Subcategorised elements might be referred to either by a specific category or by a general item, for example `irgendwie` ‘somehow’ comprises the subcategorisation of any prepositional phrase:
  <irgendwie>
  But prepositional phrases might also be made explicit:
  <für etw.>
  A similar behaviour is exhibited for the *Duden* expressions `irgendwo` ‘somewhere’, `irgendwohin` ‘to some place’, `irgendwoher` ‘from some place’, `irgendwann` ‘some time’, `mit Umstandsangabe` ‘under some circumstances’.

- Identical frame definitions differ in their degree of explicitness, for example
  <[gegen jmdn., etw. (Akk.)]>
  <[gegen jmdn., etw.]>
  both refer to the potential (indicated by ‘[]’) subcategorisation of a prepositional phrase with accusative case and head *gegen* ‘against’. The former frame explicitly refers to the accusative case, the latter implicitly needs the case because the preposition demands accusative case.

- In some cases, *Duden* distinguishes between animate and non-animate selectional restrictions, for example
  <etw. auf etw. (Akk.)>
  <etw. auf jmdn.>
  <etw. auf jmdn., etw.>
  <etw. auf ein Tier>
  <jmdn. auf etw. (Akk.)>
  <jmdn. auf jmdn.>
  <jmdn. auf jmdn., etw.>
  <jmdn. auf ein Tier>
  <ein Tier auf etw. (Akk.)>
  <ein Tier auf jmdn.>
  <ein Tier auf jmdn., etw.>
  all refer to a transitive frame with obligatory prepositional phrase `Akk.auf`.

- Syntactic comments in *Duden* might refer to a change in the subcategorisation with reference to another frame, but the modified subcategorisation frame is not explicitly provided. For example, <auch mit Akk.> refers to a modification of a frame which allows the verb to add an accusative noun phrase.

Correcting and reducing the idiosyncratic frames to their common information concerning our needs results in 65 subcategorisation frames without explicit prepositional phrase definitions and 222 subcategorisation frames including them.

The lexicon is implemented in `SGML`. I defined a *Document Type Definition (DTD)* which formally describes the structure of the verb entries and extracted manually defined subcategorisation frames for 3,658 verbs from the *Duden*.

```
<D2>

  <SEM1 SEM1-ID="1">
    <DEFPHR>
      <GR><von etw. zehren> </GR>
      <DEF>etw. aufbrauchen: </DEF>
      <BSP>
        <TEXT>von den Vorräten, von seinen Ersparnissen zehren; </TEXT>
      </BSP>
     </DEFPHR>
  </SEM1>

  <SEM1 SEM1-ID="2">

  <SEM2 SEM2-ID="a">
    <DEFPHR>
      <DEF>schwächen: </DEF>
      <BSP>
        <TEXT>das Fieber, die Seeluft, die See zehrt; </TEXT>
        <TEXT>eine zehrende Krankheit; </TEXT>
      </BSP>
    </DEFPHR>
  </SEM2>

  <SEM2 SEM2-ID="b">
    <DEFPHR>
      <GR><an jmdm., etw. zehren> </GR>
      <DEF>jmdm., etw. sehr zusetzen: </DEF>
      <BSP>
        <TEXT>das Fieber, die Krankheit zehrte an seinen Kräften; </TEXT>
        <TEXT>der Stress zehrt an ihrer Gesundheit; </TEXT>
        <TEXT>die Sorge, der Kummer, die Ungewissheit hat sehr an ihr,
              an ihren Nerven gezehrt. </TEXT>
      </BSP>
    </DEFPHR>
  </SEM2>

  </SEM1>

</D2>
```

Figure 3.18: *Duden* lexical entry for *zehren*

3.5.2 Evaluation of Subcategorisation Frames

Frame Mapping Preceding the actual experiment I defined a deterministic mapping from the *Duden* frame definitions onto my subcategorisation frame style, e.g. the ditransitive frame definition `<jmdm. etw.>` would be mapped to `nad`, and `<bei jmdm. etw.>` would be mapped to `nap` without and `nap:Dat.bei` with explicit prepositional phrase definition. 38 *Duden* frames do not match anything in my frame repertoire (mostly rare frames such as `nag` *Er beschuldigt ihn des Mordes* 'He accuses him of the murder', or frame types with more than three arguments); 5 of my frame types do not appear in the *Duden* (copula constructions and some frames including finite clause arguments such as `nds-2`).

Evaluation Measures For the evaluation of the learned subcategorisation frames, the manual *Duden* frame definitions are considered as the gold standard. I calculated precision and recall values on the following basis:

$$recall = \frac{tp}{tp + fn} \tag{3.4}$$

$$precision = \frac{tp}{tp + fp} \tag{3.5}$$

tp (true positives) refer to those subcategorisation frames where learned and manual definitions agree, fn (false negatives) to the *Duden* frames not extracted automatically, and fp (false positives) to those automatically extracted frames not defined by *Duden*.

Major importance is given to the f-score which considers recall and precision as equally relevant and therefore balances the previous measures:

$$f - score = \frac{2 * recall * precision}{recall + precision} \tag{3.6}$$

Experiments The evaluation experiment has three conditions.

I All frame types are taken into consideration. In case of a prepositional phrase argument in the frame, the PP is included, but the refined definition is ignored, e.g. the frame including one obligatory prepositional phrase is referred to by `np` (nominative noun phrase plus prepositional phrase).

II All frame types are taken into consideration. In case of a prepositional phrase argument in the frame, the refined definition is included, e.g. the frame including one obligatory prepositional phrase (cf. I) is referred to by `np:Akk.für` for a prepositional phrase with head *für* and the accusative case, `np:Dat.bei` for a prepositional phrase with head *bei* and the dative case, etc.

III Prepositional phrases are excluded from subcategorisation, i.e. frames including a p are mapped to the same frame type without that argument. In this way, a decision between prepositional phrase arguments and adjuncts is avoided.

Assuming that predictions concerning the rarest events (verbs with a low frequency) and those concerning the most frequent verbs (with increasing tendency towards polysemy) are rather unreliable, I performed the experiments on those 3,090 verbs in the *Duden* lexicon with a frequency between 10 and 2,000 in the corpus. See Table 3.24 for a distribution over frequency ranges for all 3,658 verbs with frequencies between 1 and 101,003. The horizontal lines mark the restricted verb set.

Freq			Verbs
1	-	5	162
5	-	10	289
10	-	20	478
20	-	50	690
50	-	100	581
100	-	200	496
200	-	500	459
500	-	1000	251
1000	-	2000	135
2000	-	5000	80
5000	-	10000	24
	>	10000	13

Table 3.24: Frequencies of *Duden* verbs in training corpus

Baseline As baseline for the experiments, I assigned the most frequent frame types n (intransitive frame) and na (transitive frame) as default to each verb.

Results The experimental results are displayed in Table 3.25.

Experiment	Recall		Precision		F-Score	
	Baseline	Result	Baseline	Result	Baseline	Result
I	49.57%	63.91%	54.01%	60.76%	51.70%	62.30%
II	45.58%	50.83%	54.01%	65.52%	49.44%	57.24%
III	63.92%	69.74%	59.06%	74.53%	61.40%	72.05%

Table 3.25: Evaluation of subcategorisation frames

Concerning the f-score, I reach a gain of 10% compared to the baseline for experiment I: evaluating all frame definitions in the induced lexicon including prepositional phrases results in 62.30% f-score performance. Complicating the task by including prepositional phrase definitions into the frame types (experiment II), I reach 57.24% f-score performance, 8% above the baseline. Completely disregarding the prepositional phrases in the subcategorisation frames (experiment III) results in 72.05% f-score performance, 10% above the baseline.

The differences both in the absolute f-score values and the difference to the respective baseline values correspond to the difficulty and potential of the tasks. Disregarding the prepositional phrases completely (experiment III) is the easiest task and therefore reaches the highest f-score. But the baseline frames `n` and `na` represent 50% of all frames used in the *Duden* lexicon, so the potential for improving the baseline is small. Compared to experiment III, experiment I is a more difficult task, because the prepositional phrases are taken into account as well. But I reach a gain in f-score of more than 10%, so the learned frames can improve the baseline decisions. Experiment II shows that defining prepositional phrases in verb subcategorisation is a more complicated task. Still, I improve the baseline results by 8%.

3.5.3 Lexicon Investigation

Section 3.5.2 presented the figures of merit for verb subcategorisation frames which are learned in the statistical grammar framework against the manual verb descriptions in the German dictionary *Duden*. The current section discusses advantages and shortcomings of the verb subcategorisation lexica concerning the selection of verbs and detail of frame types.

The verb entries in the automatic and manual subcategorisation lexica are examined: the respective frames are compared, against each other as well as against verb entries in Helbig and Schenkel (1969) (henceforth: H/S) and corpus evidence in the German newspaper corpus *die tageszeitung (TAZ)*. In addition, I compare the set of frames in the two lexica, their intersection and differences. The result of the investigation is a description of strengths and deficiencies in the lexica.

Intransitive Verbs In the *Duden* dictionary, intransitive verb usage is difficult to extract, since it is defined only implicitly in the verb entry, such as for the verbs *glücken* ‘to succeed’, *langen* ‘to suffice’, *verzweifeln* ‘to despair’. In addition, *Duden* defines the intransitive frame for verbs which can be used intransitively in exclamations, such as *Der kann aber wetzen!* ‘Wow, he can dash!’. But the exclamatory usage is not sufficient evidence for intransitive usage. The induced lexicon, on the other hand, tends to overgenerate the intransitive usage of verbs, mainly because of parsing mistakes. Still, the intersection of intransitive frames in both lexica reaches a recall of 77.19% and a precision of 66.11%,

Transitive Verbs The usage of transitive verbs in the lexica is the most frequent occurrence and at the same time the most successfully learned frame type. *Duden* defines transitive frames for 2,513 verbs, the automatic process extracts 2,597 frames. An agreement in 2,215 cases corresponds to 88.14% recall and 85.29% precision.

Dative Constructions *Duden* verb entries are inconsistent concerning the free dative construction ('freier Dativ'). For example, the free dative is existing in the ditransitive usage for the verb *ablösen* 'to remove' (*Der Arzt löste ihm das Pflaster ab* 'The doctor removed him the plaster), but not for the verb *backen* 'to bake' (H/S: *Die Mutter backt ihm einen Kuchen* 'The mother baked him a cake'). The induced lexicon is rather unreliable on frames including dative noun phrases. Parsing mistakes tend to extract accusative constructions as dative and therefore wrongly emphasise the dative usage.

Prepositional Phrases In general, *Duden* properly distinguishes between prepositional phrase arguments (mentioned in subcategorisation) and adjuncts, but in some cases, *Duden* overemphasises certain PP-arguments in the verb frame definition, such as `Dat.mit` for the verbs *aufschließen* 'to unlock', *garnieren* 'to garnish', *nachkommen* 'to keep up', `Dat.von` for the verbs *abbröckeln* 'to crumble', *ausleihen* 'to borrow', *erbitten* 'to ask for', *säubern* 'to clean up', or `Akk.auf` for the verbs *abklopfen* 'to check the reliability', *ausüben* 'to practise', *festnageln* 'to tie down', *passen* 'to fit'.

In the induced lexicon, prepositional phrase arguments are overemphasised, i.e. PPs used as adjuncts are frequently inserted into the lexicon, e.g. for the verbs *arbeiten* 'to work', *demonstrieren* 'to demonstrate', *sterben* 'to die'. This mistake is mainly based on highly frequent prepositional phrase adjuncts, such as `Dat.in, Dat.an, Akk.in`. On the other hand, the induced lexicon does not recognise verb-specific prepositional phrase arguments in some cases, such as `Dat.mit` for the verbs *gleichstellen* 'to equate', *handeln* 'to act', *spielen* 'to play', or `Dat.von` for the verbs *abbringen* 'to dissuade', *fegen* 'to sweep', *genesen* 'to convalesce', *schwärmen* 'to romanticise'.

Comparing the frame definitions containing PPs in both lexica, the induced lexicon tends to define PP-adjuncts such as `Dat.in, Dat. an` as arguments and neglect PP-arguments; *Duden* distinguishes arguments and adjuncts more correctly, but tends to overemphasise PPs such as `Dat.mit` and `Dat.bei` as arguments. Still, there is agreement on the `np` frame with 59.69% recall and 49.88% precision, but the evaluation of `nap` with 45.95% recall, 25.89% precision and of `ndp` with 9.52% recall and 15.87% precision pinpoints main deficiencies in the frame agreement.

Reflexive Verbs *Duden* generously categorises verbs as reflexives; they appear whenever it is possible to use the respective verb with a reflexive pronoun. The procedure is valid for verbs such as *erwärmen* 'to heat', *lohnen* 'to be worth', *schämen* 'to feel ashamed', but not for verbs such as *durchbringen* 'to pull through', *kühlen* 'to cool', *zwingen* 'to force'. The automatic frame definitions, on the other hand, tend to neglect the reflexive usage of verbs and rather choose direct objects into the frames, such as for the verbs *ablösen* 'to remove', *erschießen* 'to shoot', *überschätzen* 'to overestimate'. The lexicon tendencies are reflected by the `nr, nar, npr` frame frequencies: rather low recall values between 28.74% and 45.17%, and rather high precision values between 51.94% and 69.34% underline the differences.

Adjectival Phrases The definition of adjectival phrase arguments in the *Duden* is somewhat idiosyncratic, especially as demarcation to non-subcategorised adverbial phrases. For example, an adjectival phrase for the verb *scheinen* 'to shine' as in *Die Sonne schien hell* 'The sun is bright' is subcategorised, as well as for the verb *berühren* 'to touch' as in *Seine Worte haben uns tief berührt* 'His words touched us deeply'. Concerning the induced lexicon, the grammar does not contain adjectival phrase arguments, so they could not be recognised, such as for the verbs *anmuten* 'to seem', *erscheinen* 'to seem', *verkaufen* 'to sell'.

Subcategorisation of Clauses *Duden* shows shortcomings on the subcategorisation of non-finite and finite clauses; they rarely appear in the lexicon. Only 26 verbs (such as *anweisen* 'to instruct', *beschwören* 'to swear', *versprechen* 'to promise') subcategorise non-finite clauses, only five verbs (such as *sehen* 'to see', *wundern* 'to wonder') subcategorise finite clauses. Missing verbs for the subcategorisation of finite clauses are –among others– *ausschließen* 'to rule out', *sagen* 'to say', *vermuten* 'to assume', for the subcategorisation of non-finite clauses *hindern* 'to prevent', *verpflichten* 'to commit'.

The automatic lexicon defines the subcategorisation of clauses more reliably. For example, the verbs *behaupten* 'to state', *nörgeln* 'to grumble' subcategorise verb second finite clauses, the verbs *aufpassen* 'to pay attention', *glauben* 'to think', *hoffen* 'to hope' subcategorise finite *dass*-clauses, the verb *bezweifeln* 'to doubt' subcategorises a finite *ob*-clause, the verbs *ahnen* 'to guess', *klarmachen* 'to make clear', *raffen* 'to understand' subcategorise indirect *wh*-questions, and the verbs *anleiten* 'to instruct', *beschuldigen* 'to accuse', *lehren* 'to teach' subcategorise non-finite clauses. Mistakes occur for indirect *wh*-questions which are confused with relative clauses, such as for the verbs *ausbaden* 'to pay for', *futtern* 'to eat'.

General Frame Description *Duden* defines verb usage on various levels of detail, especially concerning prepositional phrases (cf. Section 2.2). For example, `irgendwie` 'somehow' in grammatical definitions means the usage of a prepositional phrase such as for the verb *lagern* 'to store' (*Medikamente müssen im Schrank lagern* 'Drugs need to be stored in a cupboard'); `irgendwo` 'somewhere' means the usage of a locative prepositional phrase such as for the verb *lauern* 'to lurk' (*Der Libero lauert am Strafraum* 'The sweeper lies in wait in the penalty area.'). In more restricted cases, the explicit prepositional phrase is given as in `<über etw. (Akk.)>` for the verb *verzweifeln* 'to despair' (*Man könnte verzweifeln über so viel Ignoranz* 'One could despair about that ignorance').

The grammatical definitions on various levels of detail are considered as a strength of *Duden* and generally favourable for users of a stylistic dictionary, but produce difficulties for automatic usage. For example, when including PP-definitions into the evaluation (experiment II), 10% of the *Duden* frames (PP-frames without explicit PP-definition, such as `np`) could never be guessed correctly, since the automatic lexicon includes the PPs explicitly.

There are frame types in *Duden* which do not exist in the automatic verb lexicon. This mainly concerns rare frames such as `nag, naa, xad` and frame types with more than three arguments such as `napr, ndpp`. This lexicon deficiency concerns about 4% of the total number of frames in the *Duden* lexicon.

Lexicon Coverage Compared to the automatic acquisition of verbs, *Duden* misses verbs in the dictionary: frequent verbs such as *einreisen* 'to enter', *finanzieren* 'to finance', *veranschaulichen* 'to illustrate', verbs adopted from English such as *dancen, outen, tunen*, vulgar verbs such as *anpöbeln* 'to abuse', *ankotzen* 'to make sick', *pissen* 'to piss', recent neologisms such as *digitalisieren* 'to digitalise', *klonen* 'to clone', and regional expressions such as *kicken* 'to kick', *latschen* 'to walk', *puhlen* 'to pick'.

The automatic acquisition of verbs covers a larger amount of verbs, containing 14,229 verb entries, including the missing examples above. Partly, mistaken verbs are included in the lexicon: verbs wrongly created by the morphology such as **angebieten, *dortdrohen, *einkommen*, verbs which obey the old, but not the reformed German spelling rules such as *autofahren* 'to drive a car', *danksagen* 'to thank', *spazierengehen* 'to stroll', and rare verbs, such as $^?$*bürgermeistern*, $^?$*evangelisieren*, $^?$*fiktionalisieren*, $^?$*feuerwerken*, $^?$*käsen*.

Table 3.26 summarises the lexicon investigation. I blindly classified 184 frame assignments from fn and fp into correct and wrong. The result emphasises (i) unreliabilities for `n` and `nd` in both lexica, (ii) insecurities for reflexive and expletive usage in both lexica, (iii) strength of clause subcategorisation in the induced lexicon (the few assignments in *Duden* were all correct), (iv) strength of PP-assignment in the *Duden*, and (v) variability of PP-assignment in the induced lexicon.

Summary The lexicon investigation showed that

- in both lexica, the degree of reliability of verb subcategorisation information depends on the different frame types. If I tried different probability thresholds for different frame types, the accuracy of the subcategorisation information should improve once more.

- we need to distinguish between the different goals of the subcategorisation lexica: the induced lexicon explicitly refers to verb arguments which are (obligatorily) subcategorised by the verbs in the lexicon, whereas *Duden* is not intended to represent a subcategorisation lexicon but rather to describe the stylistic usage of the verbs and therefore to refer to possibly subcategorised verb arguments; in the latter case, there is no distinction between obligatory and possible verb complementation.

Frame Type	*Duden*: fn		Learned: fp	
	correct	wrong	correct	wrong
`n`	4	6	3	7
`nd`	2	8	0	10
`nr, nar, ndr`	5	5	3	7
`x, xa, xd, xr`	6	4	3	7
`ni, nai, ndi`			5	5
`ns/nas/nds-dass`			9	0
`ns/nas/nds-2`			9	1
`np/nap/ndp/npr:Dat.mit`	7	3	6	4
`np/nap/ndp/npr:Dat.von`	7	3	5	0
`np/nap/ndp/npr:Dat.in`	6	4	3	7
`np/nap/ndp/npr:Dat.an`	9	1	6	4

Table 3.26: Investigation of subcategorisation frames

- a manual lexicon suffers from the human potential of permanently establishing new words in the vocabulary; it is difficult to be up-to-date, and the learned lexical entries therefore hold a potential for adding to and improving manual verb definitions.

3.5.4 Related Work

Automatic induction of subcategorisation lexica has mainly been performed for English. Brent (1993) uses unlabelled corpus data and defines morpho-syntactic cues followed by a statistical filtering, to obtain a verb lexicon with six different frame types, without prepositional phrase refinement. Brent evaluates the learned subcategorisation frames against hand judgements and achieves an f-score of 73.85%. Manning (1993) also works on unlabelled corpus data and does not restrict the frame definitions. He applies a stochastic part-of-speech tagger, a finite state parser, and a statistical filtering process (following Brent). Evaluating 40 randomly selected verbs (out of 3,104) against *The Oxford Advanced Learner's Dictionary* (Hornby, 1985) results in an f-score of 58.20%. Briscoe and Carroll (1997) pre-define 160 frame types (including prepositional phrase definitions). They apply a tagger, lemmatiser and parser to unlabelled corpus data; from the parsed corpus they extract subcategorisation patterns, classify and evaluate them, in order to build the lexicon. The lexical definitions are evaluated against the Alvey NL Tools dictionary (Boguraev *et al.*, 1987) and the COMLEX Syntax dictionary (Grishman *et al.*, 1994) and achieve an f-score of 46.09%. The work in Carroll and Rooth (1998) is closest to ours, since they utilise the same statistical grammar framework for the induction of subcategorisation frames, but not including prepositional phrase definitions. Their evaluation for 200 randomly chosen verbs with a frequency greater than 500 against *The Oxford Advanced Learner's Dictionary* obtains an f-score of 76.95%.

For German, Eckle (1999) performs a semi-automatic acquisition of subcategorisation information for 6,305 verbs. She works on annotated corpus data and defines linguistic heuristics in the form of regular expression queries over the usage of 244 frame types including PP definitions. The extracted subcategorisation patterns are judged manually. Eckle performs an evaluation on 15 hand-chosen verbs; she does not cite explicit recall and precision values, except for a subset of subcategorisation frames. Wauschkuhn (1999) constructs a valency dictionary for 1,044 verbs with corpus frequency larger than 40. He extracts a maximum of 2,000 example sentences for each verb from annotated corpus data, and constructs a context-free grammar for partial parsing. The syntactic analyses provide valency patterns, which are grouped in order to extract the most frequent pattern combinations. The common part of the combinations define a distribution over 42 subcategorisation frame types for each verb. The evaluation of the lexicon is performed by hand judgement on seven verbs chosen from the corpus. Wauschkuhn achieves an f-score of 61.86%.

Comparing our subcategorisation induction with existing approaches for English, Brent (1993), Manning (1993) and Carroll and Rooth (1998) are more flexible than ours, since they do not require a pre-definition of frame types. But none of them includes the definition of prepositional phrases, which makes our approach the more fine-grained version. Brent (1993) outperforms our approach by an f-score of 73.85%, but the number of six frames is incomparable; Manning (1993) and Briscoe and Carroll (1997) both have f-scores below ours, even though the evaluations are performed on more restricted data. Carroll and Rooth (1998) reach the best f-score of 76.95% compared to 72.05% in our approach, but their evaluation is facilitated by restricting the frequency of the evaluated verbs to more than 500.

Concerning subcategorisation lexica for German, I have constructed the most independent approach I know of, since I do need either extensive annotation of corpora, nor restrict the frequencies of verbs in the lexicon. In addition, the approach is fully automatic after grammar definition and does not involve heuristics or manual corrections. Finally, the evaluation is not performed by hand judgement, but rather extensively on independent manual dictionary entries.

3.6 Summary

This chapter has described the implementation, training and lexical exploitation of the German statistical grammar model which serves as source for the German verb description at the syntax-semantic interface. I have introduced the theoretical background of the statistical grammar model and illustrated the manual implementation of the underlying German grammar. A training strategy has been developed which learns the large parameter space of the lexicalised grammar model. On the basis of various examples and related work, I illustrated the potential of the grammar model for an empirical lexical acquisition, not only for the purpose of verb clustering, but also for theoretical linguistic investigations and NLP applications such as lexicography and parsing improvement.

It is desirable but difficult to evaluate all of the acquired lexical information at the syntax-semantic interface. For a syntactic evaluation, manual resources such as the *Duden* dictionary are available, but few resources offer a manual definition of semantic information. So I concentrated on an evaluation of the subcategorisation frames as core part of the grammar model. The subcategorisation lexicon as based on the statistical framework has been evaluated against dictionary definitions and proven reliable: 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.

Chapter 4

Clustering Algorithms and Evaluations

There is a huge number of clustering algorithms and also numerous possibilities for evaluating a clustering against a gold standard. The choice of a suitable clustering algorithm and of a suitable measure for the evaluation depends on the clustering objects and the clustering task. The clustering objects within this thesis are verbs, and the clustering task is a semantic classification of the verbs. Further cluster parameters are to be explored within the cluster analysis of the verbs.

This chapter provides an overview of clustering algorithms and evaluation methods which are relevant for the natural language clustering task of clustering verbs into semantic classes. Section 4.1 introduces clustering theory and relates the theoretical assumptions to the induction of verb classes. Section 4.2 describes a range of possible evaluation methods and determines relevant measures for a verb classification. The theoretical assumptions in this chapter are the basis for the clustering experiments in the following Chapter 5.

4.1 Clustering Theory

The section starts with an introduction into clustering theory in Section 4.1.1. Section 4.1.2 relates the theoretical definitions of data objects, clustering purpose and object features to verbs as the clustering target within this thesis, and Section 4.1.3 concentrates on the notion of similarity within the clustering of verbs. Finally, Section 4.1.4 defines the clustering algorithms as used in the clustering experiments and refers to related clustering approaches. For more details on clustering theory and other clustering applications than the verb classification, the interested reader is referred to the relevant clustering literature, such as Anderberg (1973); Duda and Hart (1973); Steinhausen and Langer (1977); Jain and Dubes (1988); Kaufman and Rousseeuw (1990); Jain *et al.* (1999); Duda *et al.* (2000).

4.1.1 Introduction

Clustering is a standard procedure in multivariate data analysis. It is designed to explore an inherent natural structure of the data objects, where objects in the same cluster are as similar as possible and objects in different clusters are as dissimilar as possible. The equivalence classes induced by the clusters provide a means for generalising over the data objects and their features. Clustering methods are applied in many domains, such as medical research, psychology, economics and pattern recognition.

Human beings often perform the task of clustering unconsciously; for example when looking at a two-dimensional map one automatically recognises different areas according to how close to each other the places are located, whether places are separated by rivers, lakes or a sea, etc. However, if the description of objects by their features reaches higher dimensions, intuitive judgements are less easy to obtain and justify.

The term *clustering* is often confused with a *classification* or a *discriminant analysis*. But the three kinds of data analyses refer to different ideas and are distinguished as follows: Clustering is (a) different from a classification, because classification assigns objects to already defined classes, whereas for clustering no a priori knowledge about the object classes and their members is provided. And a cluster analysis is (b) different from a discriminant analysis, since discriminant analysis aims to improve an already provided classification by strengthening the class demarcations, whereas the cluster analysis needs to establish the class structure first.

Clustering is an exploratory data analysis. Therefore, the explorer might have no or little information about the parameters of the resulting cluster analysis. In typical uses of clustering the goal is to determine all of the following:

- The number of clusters,
- The absolute and relative positions of the clusters,
- The size of the clusters,
- The shape of the clusters,
- The density of the clusters.

The cluster properties are explored in the process of the cluster analysis, which can be split into the following steps.

1. Definition of objects: Which are the objects for the cluster analysis?
2. Definition of clustering purpose: What is the interest in clustering the objects?
3. Definition of features: Which are the features that describe the objects?
4. Definition of similarity measure: How can the objects be compared?
5. Definition of clustering algorithm: Which algorithm is suitable for clustering the data?
6. Definition of cluster quality: How good is the clustering result? What is the interpretation?

Depending on the research task, some of the steps might be naturally given by the task, others are not known in advance. Typically, the understanding of the analysis develops iteratively with the experiments. The following sections define a cluster analysis with respect to the task of clustering verbs into semantic classes.

4.1.2 Data Objects, Clustering Purpose and Object Features

This work is concerned with inducing a classification of German verbs, i.e. the data objects in the clustering experiments are **German verbs**, and the clustering purpose is to investigate the automatic acquisition of a linguistically appropriate **semantic classification** of the verbs. The degree of appropriateness is defined with respect to the ideas of a verb classification at the syntax-semantic interface in Chapter 2.

Once the clustering target has been selected, the objects need an attribute description as basis for comparison. The properties are grasped by the data features, which describe the objects in as many dimensions as necessary for the object clustering. The choice of features is of extreme importance, since different features might lead to different clustering results. Kaufman and Rousseeuw (1990, page 14) emphasise the importance by stating that 'a variable not containing any relevant information is worse than useless, because it will make the clustering less apparent by hiding the useful information provided by the other variables'.

Possible features to describe German verbs might include any kind of information which helps classify the verbs in a semantically appropriate way. These features include the alternation behaviour of the verbs, their morphological properties, their auxiliary selection, adverbial combinations, etc. Within this thesis, I concentrate on defining the verb features with respect to the alternation behaviour, because I consider the **alternation behaviour** a key component for verb classes as defined in Chapter 2. So I rely on the meaning-behaviour relationship for verbs and use empirical verb properties at the **syntax-semantic interface** to describe the German verbs.

The verbs are described on three levels at the syntax-semantic interface, each of them refining the previous level by additional information. 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. The exact choice of the features is presented and discussed in detail in the experiment setup in Chapter 5.

The representation of the verbs is realised by vectors which describe the verbs by distributions over their features. As explained in Chapter 1, the distributional representation of features for natural language objects is widely used and has been justified by Harris (1968). The feature values for the distributions are provided by the German grammar, as described in Chapter 3. The distributions refer to (i) real values f representing frequencies of the features with $0 \leq f$, (ii) real values p representing probabilities of the features with $0 \leq p \leq 1$, and (iii) binary values b with $b \in \{0, 1\}$. Generally speaking, a standardisation of measurement units which converts the original measurements (such as frequencies) to unitless variables (such as probabilities) on the one hand may be helpful by avoiding the preference of a specific unit, but on the other hand might dampen the clustering structure by eliminating the absolute value of the feature.

4.1.3 Data Similarity Measures

With the data objects and their features specified, a means for comparing the objects is needed. The German verbs are described by features at the syntax-semantic interface, and the features are represented by a distributional feature vector. A range of measures calculates either the distance d or the similarity sim between two objects x and y. The notions of 'distance' and 'similarity' are related, since the smaller the distance between two objects, the more similar they are to each other. All measures refer to the feature values in some way, but they consider different properties of the feature vector. There is no optimal similarity measure, since the usage depends on the task. Following, I present a range of measures which are commonly used for calculating the similarity of distributional objects. I will use all of the measures in the clustering experiments.

Minkowski Metric The *Minkowski metric* or L_q *norm* calculates the distance d between the two objects x and y by comparing the values of their n features, cf. Equation (4.1). The Minkowski metric can be applied to frequency, probability and binary values.

$$d(x, y) = L_q(x, y) = \sqrt[q]{\sum_{i=1}^{n} (x_i - y_i)^q} \tag{4.1}$$

Two important special cases of the Minkowski metric are $q = 1$ and $q = 2$, cf. Equations (4.2) and (4.3).

- *Manhattan distance* or *City block distance* or L_1 *norm*:

$$d(x, y) = L_1 = \sum_{i=1}^{n} |x_i - y_i| \tag{4.2}$$

- *Euclidean distance* or L_2 *norm*:

$$d(x, y) = L_2 = \sqrt{\sum_{i=2}^{n} (x_i - y_i)^2} \tag{4.3}$$

Kullback-Leibler Divergence The *Kullback-Leibler divergence (KL)* or *relative entropy* is defined in Equation (4.4). KL is a measure from information theory which determines the inefficiency of assuming a model distribution given the true distribution (Cover and Thomas, 1991). It is generally used for x and y representing probability mass functions, but I will also apply the measure to probability distributions with $\sum_i x_i > 1$ and $\sum_i y_i > 1$.

$$d(x,y) = D(x||y) = \sum_{i=1}^{n} x_i \ * \ log \ \frac{x_i}{y_i} \tag{4.4}$$

The Kullback-Leibler divergence is not defined in case $y_i = 0$, so the probability distributions need to be smoothed. Two variants of KL, *information radius* in Equation (4.5) and *skew divergence* in Equation (4.6), perform a default smoothing. Both variants can tolerate zero values in the distribution, because they work with a weighted average of the two distributions compared. Lee (2001) has recently shown that the skew divergence is an effective measure for distributional similarity in NLP. Related to Lee, I set the weight w for the skew divergence to 0.9.

$$d(x,y) = IRad(x,y) = D(x||\frac{x+y}{2}) \ + \ D(y||\frac{x+y}{2}) \tag{4.5}$$

$$d(x,y) = Skew(x,y) = D(x||w*y \ + \ (1-w)*x) \tag{4.6}$$

τ **coefficient** Kendall's τ *coefficient* (Kendall, 1993) compares all feature pairs of the two objects x and y in order to calculate their distance. If $\langle x_i, y_i \rangle$ and $\langle x_j, y_j \rangle$ are two pairs of the features i and j for the objects x and y, the pairs are concordant if $x_i > x_j$ and $y_i > y_j$ or if $x_i < x_j$ and $y_i < y_j$, and the pairs are discordant if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If the distributions of the two objects are similar, a large number of concordances f_c is expected, otherwise a large number of discordances f_d is expected. τ is defined in Equation (4.7), with p_c the probability of concordances and p_d the probability of discordances; τ ranges from -1 to 1. The τ coefficient can be applied to frequency and probability values. Hatzivassiloglou and McKeown (1993) use τ to measure the similarity between adjectives.

$$sim(x,y) \ = \ \tau(x,y) \ = \ \frac{f_c}{f_c \ + \ f_d} \ - \ \frac{f_d}{f_c \ + \ f_d} \ = \ p_c \ - \ p_d \tag{4.7}$$

Cosine $cos(x,y)$ measures the similarity of the two objects x and y by calculating the *cosine of the angle* between their feature vectors. The degrees of similarity range from -1 (highest degree of dissimilarity with vector angle = 180°) over 0 (angle = 90°) to 1 (highest degree of similarity with vector angle = 0°). For positive feature values, the cosine lies between 0 and 1. The cosine measure can be applied to frequency, probability and binary values.

$$sim(x,y) = cos(x,y) = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} * \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{4.8}$$

Binary Distance Measures In addition, there are specific measures for binary distributions. The following list is taken from Manning and Schütze (1999). The measures are defined on basis of the feature sets X and Y for the objects x and y, respectively. Referring to the notion of set intersection and set union, the agreement and disagreement of the feature values is measured.

- The *matching coefficient* counts the dimensions on which both vectors are non-zero.

$$sim(x,y) = match(x,y) = |X \cap Y| = \sum_{i=1}^{n} |x_i = y_i = 1| \tag{4.9}$$

- The *Dice coefficient* normalises the matching coefficient for length by dividing by the total number of non-zero entries.

$$sim(x,y) = dice(x,y) = \frac{2 * |X \cap Y|}{|X| + |Y|} = \frac{\sum_{i=1}^{n} |x_i = y_i = 1|}{\sum_{i=1}^{n} |x_i = 1| \; + \; \sum_{i=1}^{n} |y_i = 1|} \tag{4.10}$$

- The *Jaccard coefficient* or *Tanimoto coefficient* penalises a small number of shared entries (as a proportion of all non-zero entries) more than the Dice coefficient does.

$$sim(x,y) = jaccard(x,y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{\sum_{i=1}^{n} |x_i = y_i = 1|}{\sum_{i=1}^{n} |(x_i = 1) \vee (y_i = 1)|} \tag{4.11}$$

- The *overlap coefficient (ol)* has a value of 1 if every feature with a non-zero value for the first object is also non-zero for the second object or vice versa, i.e. $X \subseteq Y$ or $Y \subseteq X$.

$$sim(x,y) = ol(x,y) = \frac{|X \cap Y|}{min(|X|, |Y|)} = \frac{\sum_{i=1}^{n} |x_i = y_i = 1|}{min(\sum_{i=1}^{n} |x_i = 1|, \sum_{i=1}^{n} |y_i = 1|)} \tag{4.12}$$

4.1.4 Clustering Algorithms

Clustering is a task for which many algorithms have been proposed. No clustering technique is universally applicable, and different techniques are in favour for different clustering purposes. So an understanding of both the clustering problem and the clustering technique is required to apply a suitable method to a given problem. In the following, I describe general parameters of a clustering technique which are relevant to the task of inducing a verb classification.

- Parametric design:

  Assumptions may (but need not) be made about the form of the distribution used to model the data by the cluster analysis. The parametric design should be chosen with respect to the nature of the data. It is often convenient to assume, for example, that the data can be modelled by a multivariate Gaussian.

- Position, size, shape and density of the clusters:

  The experimenter might have an idea about the desired clustering results with respect to the position, size, shape and density of the clusters. Different clustering algorithms have different impact on these parameters, as the description of the algorithms will show. Therefore, varying the clustering algorithm influences the design parameters.

- Number of clusters:

  The number of clusters can be fixed if the desired number is known beforehand (e.g. because of a reference to a gold standard), or can be varied to find the optimal cluster analysis. As Duda *et al.* (2000) state, 'In theory, the clustering problem can be solved by exhaustive enumeration, since the sample set is finite, so there are only a finite number of possible partitions; in practice, such an approach is unthinkable for all but the simplest problems, since there are at the order of $\frac{k^n}{k!}$ ways of partitioning a set of n elements into k subsets'.

- Ambiguity:

  Verbs can have multiple senses, requiring them being assigned to multiple classes. This is only possible by using a soft clustering algorithm, which defines cluster membership probabilities for the clustering objects. A hard clustering algorithm performs a *yes/no* decision on object membership and cannot model verb ambiguity, but it is easier to use and interpret.

The choice of a clustering algorithm determines the setting of the parameters. In the following paragraphs, I describe a range of clustering algorithms and their parameters. The algorithms are divided into (A) hierarchical clustering algorithms and (B) partitioning clustering algorithms. For each type, I concentrate on the algorithms used in this thesis and refer to further possibilities.

A) Hierarchical Clustering

Hierarchical clustering methods impose a hierarchical structure on the data objects and their step-wise clusters, i.e. one extreme of the clustering structure is only one cluster containing all objects, the other extreme is a number of clusters which equals the number of objects. To obtain a certain number k of clusters, the hierarchy is cut at the relevant depth. Hierarchical clustering is a rigid procedure, since it is not possible to re-organise clusters established in a previous step. The original concept of a hierarchy of clusters creates hard clusters, but as e.g. Lee (1997) shows that the concept may be transferred to soft clusters.

Depending on whether the clustering is performed top-down, i.e. from a single cluster to the maximum number of clusters, or bottom-up, i.e. from the maximum number of clusters to a single cluster, we distinguish divisive and agglomerative clustering. Divisive clustering is computationally more problematic than agglomerative clustering, because it needs to consider all possible divisions into subsets. Therefore, only agglomerative clustering methods are applied in this thesis. The algorithm is described in Figure 4.1.

1 Given: a set of objects $O = \{o_1, ..., o_n\} \subseteq \mathbb{R}^m$;
    a function for distance measure $d : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$
2 **for** all objects $o_i \in O$ **do**
3   establish cluster $C_i = \{o_i\}$
4 **let** $C = \{C_1, ..., C_n\}$
5 **while** $|C| \neq 1$ **do**
6   **for** all pairs of clusters $\langle C_i, C_{j \neq i} \rangle \in C \times C$ **do**
7   calculate $d(C_i, C_j)$
8   **let** $best(C_i, C_j) = \forall \langle C_{k \neq i}, C_{l \neq k,j} \rangle \in C \times C : [d(C_i, C_j) \leq d(C_k, C_l)]$
9   **for** $best(C_i, C_j)$ **do**
10   **let** $C_{ij} = C_i \cup C_j$
11   **let** $C^{new} = C \setminus \{C_i, C_j\}$
12   **let** $C = C^{new} \cup C_{ij}$
13 **end**

Figure 4.1: Algorithm for agglomerative hierarchical clustering

The algorithm includes a notion of measuring and comparing distances between clusters (step 7). So far, I have introduced measures for object distance and similarity in Section 4.1.3, but I have not introduced measures for cluster distance. The concept of cluster distance is based on the concept of object distance, but refers to different ideas of cluster amalgamation. Below, five well-known measures for cluster amalgamation are introduced. All of them are used in the clustering experiments.

Nearest Neighbour Cluster Distance The distance d between two clusters C_i and C_j is defined as the minimum distance between the cluster objects, cf. Equation (4.13). The cluster distance measure is also referred to as *single-linkage*. Typically, it causes a chaining effect concerning the shape of the clusters, i.e. whenever two clusters come close too each other, they stick together even though some members might be far from each other.

$$d(C_i, C_j) = d_{min}(C_i, C_j) = min_{x \in C_i, y \in C_j} d(x, y) \qquad (4.13)$$

Furthest Neighbour Cluster Distance The distance d between two clusters C_i and C_j is defined as the maximum distance between the cluster objects, cf. Equation (4.14). The cluster distance measure is also referred to as *complete-linkage*. Typically, it produces compact clusters with small diameters, since every object within a cluster is supposed to be close to every other object within the cluster, and outlying objects are not incorporated.

$$d(C_i, C_j) = d_{max}(C_i, C_j) = max_{x \in C_i, y \in C_j} d(x, y) \qquad (4.14)$$

Distance between Cluster Centroids The distance d between two clusters C_i and C_j is defined as the distance between the cluster centroids cen_i and cen_j, cf. Equation (4.15). The centroid of a cluster is determined as the average of objects in the cluster, i.e. each feature of the centroid vector is calculated as the average feature value of the vectors of all objects in the cluster. The cluster distance measure is a natural compromise between the nearest and the furthest neighbour cluster distance approaches. Different to the above approaches, it does not impose a structure on the clustering effect.

$$d(C_i, C_j) = d_{mean}(C_i, C_j) = d(cen_i, cen_j) \tag{4.15}$$

Average Distance between Clusters The distance d between two clusters c_i and c_j is defined as the average distance between the cluster objects, cf. Equation (4.16). Like d_{mean}, the cluster distance measure is a natural compromise between the nearest and the furthest neighbour cluster distance approaches. It does not impose a structure on the clustering effect either.

$$d(C_i, C_j) = d_{avg}(C_i, C_j) = \frac{1}{|C_i| * |C_j|} \; * \sum_{x \in C_i} \sum_{y \in C_j} d(x, y) \tag{4.16}$$

Ward's Method The distance d between two clusters C_i and C_j is defined as the loss of information (or: the increase in error) in merging two clusters (Ward, 1963), cf. Equation (4.17). The error of a cluster C is measured as the sum of distances between the objects in the cluster and the cluster centroid cen_C. When merging two clusters, the error of the merged cluster is larger than the sum or errors of the two individual clusters, and therefore represents a loss of information. But the merging is performed on those clusters which are most homogeneous, to unify clusters such that the variation inside the merged clusters increases as little as possible. Ward's method tends to create compact clusters of small size. It is a least squares method, so implicitly assumes a Gaussian model.

$$d(C_i, C_j) = d_{ward}(C_i, C_j) = \sum_{x \in (C_i \cup C_j)} d(x, cen_{ij}) \; - \; [\sum_{x \in C_i} d(x, cen_i) \; + \sum_{x \in cen_j} d(x, cen_j)] \tag{4.17}$$

B) Partitioning Clustering

Partitioning clustering methods partition the data object set into clusters where every pair of object clusters is either distinct (hard clustering) or has some members in common (soft clustering). Partitioning clustering begins with a starting cluster partition which is iteratively improved until a locally optimal partition is reached. The starting clusters can be either random or the cluster output from some clustering pre-process (e.g. hierarchical clustering). In the resulting clusters, the objects in the groups together add up to the full object set.

k-Means Clustering The k-Means clustering algorithm is an unsupervised hard clustering method which assigns the n data objects $o_1, ..., o_n$ to a pre-defined number of exactly k clusters $C_1, ..., C_k$. Initial verb clusters are iteratively re-organised by assigning each verb to its closest cluster (centroid) and re-calculating cluster centroids until no further changes take place. The optimising criterion in the clustering process is the sum-of-squared-error E between the objects in the clusters and their respective cluster centroids $cen_1, ..., cen_k$, cf. Equation (4.18).

$$E = \sum_{i=1}^{k} \sum_{o \in C_i} d(o, cen_i)^2 \tag{4.18}$$

The k-Means algorithm is sensitive to the selection of the initial partition, so the initialisation should be varied. k-Means imposes a Gaussian parametric design on the clustering result and generally works well on data sets with isotropic cluster shape, since it tends to create compact clusters. The time complexity of k-Means is $O(n)$ with n the number of objects. Several variants of the k-Means algorithm exist. Within this thesis, clustering is performed by the k-Means algorithm as proposed by Forgy (1965). Figure 4.2 defines the algorithm.

Given: a set of objects $O = \{o_1, ..., o_n\} \subseteq \mathbb{R}^m$;
    a function for distance measure $d : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$;
    a (random/pre-processed) clustering partition $C = \{C_1, ..., C_k\}$
do
  for all clusters $C_i \in C$ **do**
    calculate cluster centroid $cen_i \subseteq \mathbb{R}^m$
  for all objects $o \in O$ **do**
    for all clusters $C_i \in C$ **do**
      calculate $d(o, C_i) = d(o, cen_i)$
    let $best(o, C_o) = \forall C_j \in C : [d(o, cen_{C_o}) \leq d(o, cen_{C_j})]$
  undefine C^{new}
  for all objects $o \in O$ **do**
    for $best(o, C_o)$ **do**
      let $o \in C_o^{new}$
  if $C \neq C^{new}$ **then** $change = true$
  else $change = false$
until $change = false$

Figure 4.2: Algorithm for k-Means clustering

Other Clustering Methods k-Means is a hard clustering algorithm. But some clustering problems require the clustering objects being assigned to multiple classes. For example, to model verb ambiguity one would need a soft clustering algorithm. Examples for soft clustering algorithms which are based on the same data model as k-Means are such as fuzzy clustering (Zadeh, 1965), cf. also Höppner *et al.* (1997) and Duda *et al.* (2000), and the *Expectation-Maximisation (EM) Algorithm* (Baum, 1972) which can also be implemented as a soft version of k-Means with an underlying Gaussian model.

The above methods represent a standard choice for clustering in pattern recognition, cf. Duda *et al.* (2000). Clustering techniques with different background are e.g. the *Nearest Neighbour Algorithm* (Jarvis and Patrick, 1973), *Graph-Based Clustering* (Zahn, 1971), and *Artificial Neural Networks* (Hertz *et al.*, 1991). Recently, elaborated techniques from especially image processing have been transfered to linguistic clustering, such as *Spectral Clustering* (Brew and Schulte im Walde, 2002).

C) Decision on Clustering Algorithm

Within the scope of this thesis, 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. In addition, see the following arguments.

- 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. In addition, the isotropic shape of clusters reflects the intuition of a compact verb classification.

- A variation of the clustering initialisation performs a variation of the clustering parameters such as position, size, shape and density of the clusters. Even though I assume that an appropriate parametric design for the verb classification is given by isotropic cluster formation, a variation of initial clusters investigates the relationship between clustering data and cluster formation. I will therefore apply random initialisations and hierarchical clusters as input to k-Means.

- Selim and Ismail (1984) prove for distance metrices (a subset of the similarity measures in Section 4.1.3) that k-Means finds locally optimal solutions by minimising the sum-of-squared-error between the objects in the clusters and their respective cluster centroids.

- 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. Ambiguities are a difficult problem in linguistics, and are subject to future work. I will investigate the impact of the hard clustering on polysemous verbs, but not try to model the polysemy within this work.

- As to the more general question whether to use a supervised classification or an unsupervised clustering method, this work concentrates on minimising the manual intervention in the automatic class acquisition. A classification would require costly manual labelling (especially with respect to a large-scale classification) and not agree with the exploratory goal of finding as many independent linguistic insights as possible at the syntax-semantic interface of verb classifications.

4.2 Clustering Evaluation

A clustering evaluation demands an independent and reliable measure for the assessment and comparison of clustering experiments and results. In theory, the clustering researcher has acquired an intuition for the clustering evaluation, but in practise the mass of data on the one hand and the subtle details of data representation and clustering algorithms on the other hand make an intuitive judgement impossible. An intuitive, introspective evaluation can therefore only be plausible for small sets of objects, but large-scale experiments require an objective method.

There is no absolute scheme with which to measure clusterings, but a variety of evaluation measures from diverse areas such as theoretical statistics, machine vision and web-page clustering are applicable. In this section, I provide the definition of various clustering evaluation measures and evaluate them with respect to their linguistic application. Section 4.2.1 describes the demands I expect to fulfill with an evaluation measure on verb clusterings. In Section 4.2.2 I present a range of possible evaluation methods, and Section 4.2.3 compares the measures against each other and according to the evaluation demands.

4.2.1 Demands on Clustering Evaluation

An objective method for evaluating clusterings should be independent of the evaluator and reliable concerning its judgement about the quality of the clusterings. How can we transfer these abstract descriptions to more concrete demands? Following, I define demands on the task of clustering verbs into semantic classes, with an increasing proportion of linguistic task specificity. I.e. I first define general demands on an evaluation, then general demands on a clustering evaluation, and finally demands on the verb-specific clustering evaluation.

The demands on the clustering evaluation are easier described with reference to the formal notation of clustering result and gold standard classification, so the notation is provided in advance:

Definition 4.1 *Given an object set $O = \{o_1, ..., o_n\}$ with n objects, the clustering result and the manual classification as the gold standard represent two partitions of O with $C = \{C_1, ..., C_k\}$ and $M = \{M_1, M_2, ..., M_l\}$, respectively. $C_i \in C$ denotes the set of objects in the ith cluster of partition C, and $M_j \in M$ denotes the set of objects in the jth cluster of partition M.*

General Evaluation Demands Firstly, I define a demand on evaluation in general: The evaluation of an experiment should be proceeded against a gold standard, as independent and reliable as possible. My gold standard is the manual classification of verbs, as described in Chapter 2. The classification has been created by the author. To compensate for the sub-optimal setup by a single person, the classification was developed in close relation to the existing classifications for German by Schumacher (1986) and English by Levin (1993). In addition, the complete classification was finished before any experiments on the verbs were performed.

General Clustering Demands The second range of demands refers to general properties of a cluster analysis, independent of the clustering area.

- Since the purpose of the evaluation is to assess and compare different clustering experiments and results, the measure should be applicable to all similarity measures used in clustering, but possibly independent of the respective similarity measure.
- The evaluation result should define a (numerical) measure indicating the value of the clustering. The resulting value should either be easy to interpret or otherwise be illustrated with respect to its range and effects, in order to facilitate the evaluation interpretation.
- The evaluation method should be defined without a bias towards a specific number and size of clusters.
- The evaluation measure should distinguish the quality of (i) the whole clustering partition C, and (ii) the specific clusters $C_i \in C$.

Linguistic Clustering Demands The fact that this thesis is concerned with the clustering of linguistic data sharpens the requirements on an appropriate clustering evaluation, because the demands on verb classes are specific to the linguistic background and linguistic intuition and not necessarily desired for different clustering areas. The following list therefore refers to a third range of demands, defined as linguistic desiderata for the clustering of verbs.

(a) The clustering result should not be a single cluster representing the clustering partition, i.e. $|C| = 1$. A single cluster does not represent an appropriate model for verb classes.

(b) The clustering result should not be a clustering partition with only singletons, i.e. $\forall C_i \in C : |C_i| = 1$. A set of singletons does not represent an appropriate model for verb classes either.

(c) Let C_i be a correct (according to the gold standard) cluster with $|C_i| = x$. Compare this cluster with the correct cluster C_j with $|C_j| = y > x$. The evaluated quality of C_j should be better compared to C_i, since the latter cluster was able to create a larger correct cluster, which is a more difficult task.

Example:[1] C_i = <u>*ahnen vermuten wissen*</u>
C_j = <u>*ahnen denken glauben vermuten wissen*</u>

(d) Let C_i be a correct cluster and C_j be a cluster which is identical to C_i, but contains additional objects which do not belong to the same class. The evaluated quality of C_i should be better compared to C_j, since the former cluster contains fewer errors.

Example: C_i = <u>*ahnen vermuten wissen*</u>
C_j = <u>*ahnen vermuten wissen*</u> *laufen lachen*

[1] In all examples, verbs belonging to the same gold standard class are underlined in the cluster.

(e) Let C_i be a correct cluster with $|C_i| = x$. Compare this cluster with a non-correct cluster C_j with $|C_j| = x$. The evaluated quality of C_i should be better compared to C_j, since being of the same size as C_j the proportion of homogeneous verbs is larger.

Example: C_i = *ahnen vermuten wissen*
C_j = *ahnen vermuten laufen*

(f) Let C_i be a correct cluster with $|C_i| = x$. Compare this cluster with the two correct clusters (obviously in a different partition) C_{i_1} and C_{i_2} with $C_i = C_{i_1} \cup C_{i_2}$. The evaluated quality of C_i should be better compared to the sum of qualities of C_{i_1} and C_{i_2}, since the former manages to cluster the same range of homogeneous verbs in the same cluster.

Example: C_i = *ahnen denken glauben vermuten wissen*
C_{i_1} = *ahnen denken glauben*
C_{i_2} = *vermuten wissen*

(g) Let C_{i_1} and C_{i_2} be two correct clusters. Compare these clusters with a single non-correct cluster (obviously in a different partition) C_i with $C_i = C_{i_1} \cup C_{i_2}$. The evaluated quality of C_i should be worse compared to the sum of qualities of C_{i_1} and C_{i_2}, since the smaller clusters are completely correct, whereas C_i merges the clusters into an incoherent set.

Example: C_i = *ahnen denken glauben laufen rennen*
C_{i_1} = *ahnen denken glauben*
C_{i_2} = *laufen rennen*

Some of the linguistically defined demands are also subject to general clustering demands, but nevertheless included in the more specific cases.

The linguistically most distinctive demand on the clustering evaluation deserves specific attention. It refers to the representation of verb ambiguities, both in the manual and induced classifications. Two scenarios of verb ambiguity are possible:

1. The manual classification contains verb ambiguity, i.e. there are polysemous verbs which belong to more than one verb class. The cluster analysis, on the other hand, is based on a hard clustering algorithm, i.e. each verb is only assigned to one cluster.

2. The manual classification contains verb ambiguity, and the cluster analysis is based on a soft clustering algorithm, i.e. both verb sets contain verbs which are possibly assigned to multiple classes.

The third possible scenario, that the manual classification is without verb ambiguity, but the cluster analysis is a soft clustering, is not taken into consideration, since it is linguistically uninteresting. The second scenario is relevant for a soft clustering technique, but since this thesis is restricted to a hard clustering technique, we can concentrate on scenario 1: the manual classification as defined in Chapter 2 contains polysemous verbs, but k-Means only produces hard clusters.

4.2.2 Description of Evaluation Measures

In the following, I describe a range of possible evaluation measures, with different theoretical backgrounds and demands. The overview does, of course, not represent an exhaustive list of clustering evaluations, but tries to give an impression of the variety of possible methods which are concerned with clustering and clustering evaluation. Not all of the described measures are applicable to our clustering task, so a comparison and choice of the candidate methods will be provided in Section 4.2.3.

Contingency Tables Contingency tables are a typical means for describing and defining the association between two partitions. As they will be of use in a number of evaluation examples below, their notation is given beforehand.

Definition 4.2 *A $C \times M$ contingency table is a $C \times M$ matrix with rows $C_i, 1 \leq i \leq k$ and columns $M_j, 1 \leq j \leq l$. t_{ij} denotes the number of objects that are common to the set C_i in partition C (the clustering result) and the set M_j in partition M (the manual classification). Summing over the row or column values gives the marginal values $t_{i.}$ and $t_{.j}$, referring to the number of objects in classes C_i and M_j, respectively. Summing over the marginal values results in the total number of n objects in the clustering task.*

The number of pairs with reference to a specific matrix value x is calculated by $\binom{x}{2}$; the pairs are of special interest for a convenient calculation of evaluation results. For illustration purposes, a $C \times M$ contingency table is described by an example:

$$M = \{M_1 = \{a, b, c\}, M_2 = \{d, e, f\}\}$$
$$C = \{C_1 = \{a, b\}, C_2 = \{c, d, e\}, C_3 = \{f\}\}$$

$C \times M$ contingency table:

	M_1	M_2	
C_1	$t_{11} = 2$	$t_{12} = 0$	$t_{1.} = 2$
C_2	$t_{21} = 1$	$t_{22} = 2$	$t_{2.} = 3$
C_3	$t_{31} = 0$	$t_{32} = 1$	$t_{3.} = 1$
	$t_{.1} = 3$	$t_{.2} = 3$	$n = 6$

The number of pairs within the cells of the contingency tables is as follows.

	M_1	M_2	
C_1	$\binom{t_{11}}{2} = 1$	$\binom{t_{12}}{2} = 0$	$\binom{t_{1.}}{2} = 1$
C_2	$\binom{t_{21}}{2} = 0$	$\binom{t_{22}}{2} = 1$	$\binom{t_{2.}}{2} = 3$
C_3	$\binom{t_{31}}{2} = 0$	$\binom{t_{32}}{2} = 0$	$\binom{t_{3.}}{2} = 0$
	$\binom{t_{.1}}{2} = 3$	$\binom{t_{.2}}{2} = 3$	$\binom{n}{2} = 15$

Sum-of-Squared-Error Criterion

Summing over the squared distances between the clustering objects and their cluster representatives (i.e. the respective cluster centroids) is a standard cost function. The evaluation defines a measure for the homogeneity of the clustering results with respect to the object description data, but without reference to a gold standard.

The sum-of-squared-error E originally refers to Euclidean distance, but is applicable to further distance measures. The definition was given in Equation (4.18) and is repeated in Equation (4.19), with the cluster centroid of cluster C_i abbreviated as cen_i.

$$E(C) = \sum_{i=1}^{k} \sum_{o \in C_i} d(o, cen_i)^2 \tag{4.19}$$

Silhouette Value

Kaufman and Rousseeuw (1990, pages 83ff) present the silhouette plot as a means for clustering evaluation. With this method, each cluster is represented by a silhouette displaying which objects lie well within the cluster and which objects are marginal to the cluster. The evaluation method also refers to the object data, but not to a gold standard.

To obtain the silhouette value sil for an object o_i within a cluster C_A, we compare the average distance a between o_i and all other objects in C_A with the average distance b between o_i and all objects in the neighbour cluster C_B, cf. Equations 4.20 to 4.22. For each object o_i applies $-1 \leq sil(o_i) \leq 1$. If $sil(o_i)$ is large, the average object distance within the cluster is smaller than the average distance to the objects in the neighbour cluster, so o_i is well classified. If $sil(o_i)$ is small, the average object distance within the cluster is larger than the average distance to the objects in the neighbour cluster, so o_i has been misclassified.

$$a(o_i) = \frac{1}{|C_A| - 1} \sum_{o_j \in C_A, o_j \neq o_i} d(o_i, o_j) \tag{4.20}$$

$$b(o_i) = min_{C_B \neq C_A} \frac{1}{|C_B|} \sum_{o_j \in C_B} d(o_i, o_j) \tag{4.21}$$

$$sil(o_i) = \frac{b(o_i) - a(o_i)}{max\{a(o_i), b(o_i)\}} \tag{4.22}$$

In addition to providing information about the quality of classification of a single object, the silhouette value can be extended to evaluate the individual clusters and the entire clustering. The *average silhouette width* $sil(C_i)$ of a cluster C_i is defined as the average silhouette value for all objects within the cluster, cf. Equation 4.23, and the *average silhouette width for the entire data set* with k clusters $\overline{sil(k)}$ is defined as the average silhouette value for the individual clusters, cf. Equation 4.24.

$$sil(C_i) = \frac{1}{|C_i|} \sum_{o_j \in C_i} sil(o_j) \tag{4.23}$$

$$sil(C) = \overline{sil(k)} = \frac{1}{k} \sum_{i=1}^{k} sil(C_i) \tag{4.24}$$

Class-based Precision and Recall

What I call a class-based P/R evaluation has originally been defined by Vilain *et al.* (1995) as scoring scheme for the coreference task in MUC6. The evaluation method considers both the clustering and the manual classification as equivalence classes which are defined by the particular object links which are necessary to encode the equivalence relations. The precision and recall scores are obtained by calculating the least number of object links required to align the equivalence classes.

Let $c(M_i)$ be the minimal number of correct object links which are necessary to generate the equivalence class M_i in the manual classification: $c(M_i) = |M_i| - 1$. With $|p(M_i)|$ the number of classes in the clustering partition containing any of the objects in M_i, the number of missing object links in the clustering which are necessary to fully reunite the objects of class M_i is $m(M_i) = |p(M_i)| - 1$. Recall for a single cluster is defined as the proportion of existing object links of the relevant cluster compared to the minimal number of correct object links.

$$recall(M_i) = \frac{c(M_i) \; - \; m(M_i)}{c(M_i)} = \frac{|M_i| \; - \; |p(M_i)|}{|M_i| \; - \; 1} \tag{4.25}$$

Extending the measure from a single equivalence class to the entire classification of the object set S is realised by summing over the equivalence classes:

$$recall_S(C, M) = \frac{\sum_i \; |M_i| \; - \; |p(M_i)|}{\sum_i \; |M_i| \; - \; 1} \tag{4.26}$$

In the case of precision, we consider the equivalence classes C_i in the clustering and calculate the existing and missing object links in the manual classification with respect to the clustering.

$$precision(C_i) = \frac{c(C_i) \; - \; m(C_i)}{c(C_i)} = \frac{|C_i| \; - \; |p(C_i)|}{|C_i| \; - \; 1} \tag{4.27}$$

$$precision_S(C, M) = \frac{\sum_i \; |C_i| \; - \; |p(C_i)|}{\sum_i \; |C_i| \; - \; 1} \tag{4.28}$$

Classification		Clustering		Evaluation
Class	Link	Class	Link	
$M_1 = \{a, b, c\}$	a-b, b-c	$C_1 = \{a, b\}$	a-b	$recall_S(C, M) = \frac{(3-2)\ +\ (3-2)}{(3-1)\ +\ (3-1)} = \frac{2}{4} = \frac{1}{2}$
$M_2 = \{d, e, f\}$	d-e, e-f	$C_2 = \{c, d, e\}$	c-d, d-e	$precision_S(C, M) = \frac{(2-1)\ +\ (3-2)\ +\ (1-1)}{(2-1)\ +\ (3-1)\ +\ (1-1)} = \frac{2}{3}$
		$C_3 = \{f\}$		$f - score_S(C, M) = \frac{2 * \frac{1}{2} * \frac{2}{3}}{\frac{1}{2} + \frac{2}{3}} = \frac{4}{7}$

Table 4.1: Example evaluation for class-based P/R

The $f - score_S$ as given in Equation 4.29 is the harmonic mean between $precision_S$ and $recall_S$.

$$f - score_S(C, M) = \frac{2\ *\ recall_S\ *\ precision_S}{recall_S\ +\ precision_S} \tag{4.29}$$

Pair-wise Precision and Recall

Being closest to my clustering area, Hatzivassiloglou and McKeown (1993) present an evaluation method for clustering in NLP: they define and evaluate a cluster analysis of adjectives. The evaluation is based on common cluster membership of object pairs in the clustering and the manual classification. On the basis of common cluster membership, recall and precision numbers are calculated in the standard way, cf. Equations (4.30) and (4.31). True positives tp are the number of common pairs in M and C, false positives fp the number of pairs in C, but not M, and false negatives fn the number of pairs in M, but not C. I add the f-score as harmonic mean between recall and precision, as above. Table 4.2 presents an example of pair-wise precision and recall calculation.

$$recall = \frac{tp}{fn\ +\ tp} \tag{4.30}$$

$$precision = \frac{tp}{fp\ +\ tp} \tag{4.31}$$

Adjusted Pair-wise Precision

Pair-wise precision and recall calculation (see above) shows some undesired properties concerning my linguistic needs, especially concerning the recall value. I therefore use the precision value and adjust the measure by a scaling factor based on the size of the respective cluster. The definition of the adjusted pair-wise precision is given in Equation (4.32). A correct pair refers to a verb

Classification	Clustering	Evaluation
$M_1 = \{a, b, c\}$ $M_2 = \{d, e, f\}$	$C_1 = \{a, b\}$ $C_2 = \{c, d, e\}$ $C_3 = \{f\}$	number of common pairs in M and C (tp): 2 number of pairs in classification M ($fn + tp$): 6 number of pairs in clustering C ($fp + tp$): 4
		$recall = \frac{2}{6} = \frac{1}{3}$ $precision = \frac{2}{4} = \frac{1}{2}$ $f-score = \frac{2*\frac{1}{3}*\frac{1}{2}}{\frac{1}{3}+\frac{1}{2}} = \frac{2}{5}$

Table 4.2: Example evaluation for pair-wise P/R

Classification	Clustering	Evaluation
$M_1 = \{a, b, c\}$ $M_2 = \{d, e, f\}$	$C_1 = \{a, b\}$ $C_2 = \{c, d, e\}$ $C_3 = \{f\}$	$APP(C_1) = \frac{1}{3}$ $APP(C_2) = \frac{1}{4}$ $APP(C_3) = 0$
		$APP(C) = \frac{1}{3} * (\frac{\frac{1}{3}}{2} + \frac{\frac{1}{4}}{3}) = \frac{1}{3} * (\frac{1}{6} + \frac{1}{12}) = \frac{1}{12}$

Table 4.3: Example evaluation for adjusted pair-wise precision

pair which is correct according to the gold standard. The evaluation measure of the whole clustering is calculated by taking the weighted average over the qualities of the individual clusters, as defined in Equation (4.33). By inserting $|C_i|^{-1}$ as weight for each cluster $APP(C_i)$ I calculate the average contribution of each verb to $APP(C_i)$. And since the overall sum of $APP(C, M)$ for the clustering is first summed over all clusters (and therefore over the average contributions of the verbs) and then divided by the number of clusters, I calculate the average contribution of a verb to the clustering APP. The measure is developed with specific care concerning the linguistic demands, e.g. without the addend $+1$ in the denominator of $APP(C_i)$ the linguistic demands would not be fulfilled. Table 4.3 presents an example of adjusted pair-wise precision.

$$APP(C_i) = \frac{number\ of\ correct\ pairs\ in\ C_i}{number\ of\ verbs\ in\ C_i\ +\ 1} \tag{4.32}$$

$$APP(C, M) = \frac{1}{|C|} \sum_i \frac{APP(C_i)}{|C_i|} \tag{4.33}$$

Mutual Information

The way I define mutual information between the clustering and its gold standard is borrowed from Strehl *et al.* (2000) who assess the similarity of object partitions for the clustering of web documents. Mutual information is a symmetric measure for the degree of dependency between

Classification	Clustering	Evaluation
$M_1 = \{a,b,c\}$ $M_2 = \{d,e,f\}$	$C_1 = \{a,b\}$ $C_2 = \{c,d,e\}$ $C_3 = \{f\}$	$purity(C_1) = 1$ $purity(C_2) = \frac{2}{3}$ $purity(C_3) = 1$
		$MI(C,M) =$ $\frac{1}{6} * (2 * \frac{log\frac{2*6}{2*3}}{log(2*3)} + ... + 1 * \frac{log\frac{1*6}{1*3}}{log(2*3)}) = 0.27371$

Table 4.4: Example evaluation for mutual information

the clustering and the manual classification. It is based on the notion of cluster *purity*, which measures the quality of a single cluster C_i referring to p_i^j, the largest number of objects in cluster C_i which C_i has in common with a gold standard verb class M_j, having compared C_i to all gold standard verb classes in M.

$$purity(C_i) = \frac{1}{|C_i|} \, max_j(p_i^j) \tag{4.34}$$

The mutual information score between the clustering C and the manual classification M is based on the shared object membership, with a scaling factor corresponding to the number of objects in the respective clusters, cf. Equation 4.35. The second line in Equation 4.35 relates the definitions by Strehl *et al.* to the notation in the contingency table. Table 4.4 presents an example of mutual information evaluation.

$$\begin{aligned} MI(C,M) &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} p_i^j \; \frac{log(\frac{p_i^j \, * \, n}{\sum_{a=1}^{k} p_a^j \;\; \sum_{b=1}^{l} p_l^b})}{log(k \, * \, l)} \\ &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} t_{ij} \; \frac{log(\frac{t_{ij} \, * \, n}{t_{i.} \, * \, t_{.j}})}{log(k \, * \, l)} \end{aligned} \tag{4.35}$$

Rand Index

Rand (1971) defines an evaluation measure for a general clustering problem on basis of agreement vs. disagreement between object pairs in clusterings. He states that clusters are defined as much by those points which they do not contain as by those points which they do contain. Therefore, if the elements of an object-pair are assigned to the same classes in both the clustering and the manual classification, and also if they are assigned to different classes in both partitions, this represents a similarity between the equivalence classes. The similarity evaluation is based on the overlap in class agreement A, compared to the class disagreement D, as defined by Equation (4.36), with $A + D = n$. Table 4.5 presents an example of the Rand index.

$$Rand(C,M) = \frac{\sum_{i<j}^{n} \; \gamma(o_i, o_j)}{\binom{n}{2}} \tag{4.36}$$

Classification	Clustering	Evaluation
$M_1 = \{a, b, c\}$	$C_1 = \{a, b\}$	agree : number object pairs together in both M and C: 2
$M_2 = \{d, e, f\}$	$C_2 = \{c, d, e\}$	agree : number object pairs separate in both M and C: 7
	$C_3 = \{f\}$	disagree : number object pairs mixed in M and C: 6
		$Rand(C, M) = \frac{2+7}{2+7+6} = \frac{9}{15} = 0.6$

Table 4.5: Example evaluation for Rand index

where

$$\gamma(o_i, o_j) = \begin{cases} 1 & \text{if there exist } C_A \in C \text{ and } M_B \in M \text{ such that objects } o_i \text{ and } o_j \text{ are in } C_A \text{ and } M_B, \\ 1 & \text{if there exist } C_A \in C \text{ and } M_B \in M \text{ such that } o_i \text{ is in both } C_A \text{ and } M_B \\ & \quad \text{while } o_j \text{ is in neither } C_A \text{ or } M_B, \\ 0 & \text{otherwise.} \end{cases} \tag{4.37}$$

Rand Index adjusted by Chance

Hubert and Arabie (1985) argue for a correction of the Rand index for chance, in the sense that the index would take on some constant value (e.g. zero) under an appropriate null model of how the partitions have been chosen. According to Hubert and Arabie, the most obvious model for randomness assumes that the $C \times M$ contingency table is constructed from the generalised hyper-geometric distribution, i.e. the C and M partitions are picked at random, given the original number of classes and objects.

The general form of an index corrected for chance is given in Equation (4.38).[2] The *index* refers to the observed number of object pairs on which the partitions agree. The expected number of object pairs with class agreement attributable to a particular cell in the contingency table is defined by the number of pairs in the row times the number of pairs in the column divided by the total number of pairs, cf. Equation (4.39). The maximum number of object pairs is given by the average number of possible pairs in the clustering and the manual classification. Other possibilities for the maximum index would be e.g. the minimum of the possible pairs in clustering and manual classification $min(\sum_i \binom{t_{i.}}{2}, \sum_j \binom{t_{.j}}{2})$ or simply the possible pairs in the manual classification $\sum_j \binom{t_{.j}}{2}$ when considering the manual classification as the optimum. The corrected Rand index is given in Equation (4.40). The range of R_{adj} is $0 \leq R_{adj} \leq 1$, with only extreme cases below zero. Table 4.6 presents an example.

$$Index_{adj} = \frac{Index \ - \ Expected\ Index}{Maximum\ Index \ - \ Expected\ Index} \tag{4.38}$$

[2]In psychological literature, the index is referred to as *kappa statistic* (Cohen, 1960).

Classification	Clustering	Evaluation
$M_1 = \{a, b, c\}$ $M_2 = \{d, e, f\}$	$C_1 = \{a, b\}$ $C_2 = \{c, d, e\}$ $C_3 = \{f\}$	$Rand_{adj} = \frac{2 - \frac{4*6}{15}}{\frac{1}{2}(4+6) - \frac{4*6}{15}} = \frac{2 - \frac{8}{5}}{5 - \frac{8}{5}} = 0.11765$

Table 4.6: Example evaluation for adjusted Rand index

$$Exp\binom{t_{ij}}{2} = \frac{\binom{t_{i.}}{2}\binom{t_{.j}}{2}}{\binom{n}{2}} \tag{4.39}$$

$$Rand_{adj}(C, M) = \frac{\sum_{i,j}\binom{t_{ij}}{2} - \frac{\sum_i \binom{t_{i.}}{2} \sum_j \binom{t_{.j}}{2}}{\binom{n}{2}}}{\frac{1}{2}\left(\sum_i \binom{t_{i.}}{2} + \sum_j \binom{t_{.j}}{2}\right) - \frac{\sum_i \binom{t_{i.}}{2} \sum_j \binom{t_{.j}}{2}}{\binom{n}{2}}} \tag{4.40}$$

Matching Index

Fowlkes and Mallows (1983) define another evaluation method based on contingency tables. Their motivation is to define a measure of similarity between two hierarchical clusterings, as a sequence of measures which constitute the basis for a plotting procedure, to compare different cut-combinations in the hierarchies. The measure B_k is derived from the $C \times M$ contingency table with C referring to a hierarchical clustering cut at level i, and M referring to a hierarchical clustering cut at level j. B_k compares the match of assigning pairs of objects to common clusters with the total number of possible pairs, the clustering marginals; B_k is defined as in Equation (4.41). Table 4.7 presents an example of the matching index, based on the contingency table.

$$B_k(C, M) = \frac{T_k}{\sqrt{P_k\, Q_k}} \tag{4.41}$$

where

$$T_k = \sum_{i=1}^{k} \sum_{j=1}^{l} t_{ij}^2 - n \tag{4.42}$$

$$P_k = \sum_{i=1}^{k} t_{i.}^2 - n \tag{4.43}$$

$$Q_k = \sum_{j=1}^{l} t_{.j}^2 - n \tag{4.44}$$

Classification	Clustering	Evaluation
$M_1 = \{a, b, c\}$ $M_2 = \{d, e, f\}$	$C_1 = \{a, b\}$ $C_2 = \{c, d, e\}$ $C_3 = \{f\}$	$T_k = 4 + 1 + 4 + 1 - 6 = 4$ $P_k = 4 + 9 + 1 - 6 = 8$ $Q_k = 9 + 9 - 6 = 12$
		$B_k = \frac{4}{\sqrt{8*12}} = \frac{4}{\sqrt{96}} = 0.40825$

Table 4.7: Example evaluation for matching index

4.2.3 Comparison of Evaluation Measures

Section 4.2.2 has described a variety of possible measures to evaluate the result of a cluster analysis. Following, the different measures are compared against each other and according to the demands of a clustering evaluation, as defined in Section 4.2.1. The comparison is performed in Table 4.8, which lists the evaluation methods against the demands. The demands are briefly repeated:

- Reference to gold standard (given:+ or not given:-)
- Applicable to all similarity measures (yes:+ or no:-)
- Independent of similarity measure (yes:+ or no:-)
- Value for specific cluster and whole clustering (yes:+ or no:-)
- Bias in cluster number (none:-)
- Sensibility to linguistic desiderata (list of failures; none:-), with a brief repetition of the desiderata from Section 4.2.1:

  (a) Clustering result should not be $|C| = 1$.
  (A failure of this desideratum corresponds to a bias towards few large clusters.)

  (b) Clustering result should not be singletons.
  (A failure of this desideratum corresponds to a bias towards many small clusters.)

  (c) Larger correct cluster should be better than smaller correct cluster.

  (d) Correct cluster should be better than same cluster with noise.

  (e) Correct cluster with x objects should be better than noisy cluster with x objects.

  (f) Correct union of correct clusters should be better than separate clusters.

  (g) Correct, separated clusters should be better than incorrect union.

  The success and failure of the desiderata have been evaluated on artificial clustering examples which model the diverse clustering outputs.

- Sensibility to error introduction (monotonic behaviour:+ or not:-)

  This issue refers to an experiment for illustrating the sensibility of the evaluation measures to a step-wise introduction of errors. First the manual classification is evaluated against itself, then I introduce an artificial error and evaluate the result again, etc. The error introduction is repeated 25 times, and an evaluation method sensible to the error introduction should react monotonically in its quality score. Figures 4.3 to 4.5 illustrate the error sensibility of APP, the pair-wise f-score $PairF$ and $Rand_{adj}$.

- Interpretation (minimum and maximum value, if existing, else:-)

The core distinction between the methods is their reference to the gold standard: The sum-of-squared-error and silhouette plot do not refer to the gold standard at all, they measure the quality of the cluster analysis with reference to the data definition and similarity measure. Class-based P/R's underlying idea is very different to any other evaluation method; it compares the distribution of verbs belonging to a common semantic class over the different sets in a partition. Both pair-wise P/R and the adjusted precision measure consider the verb pairs correctly formed by the cluster analysis, with APP incorporating the linguistic desiderata. All other evaluation methods concentrate on the number of verbs agreeing in the gold standard and guessed partitions, as provided by contingency tables; mutual information weights the score by the sizes of the respective sets, the Rand index by the number of possible pairs, and the adjusted Rand index and the matching index take the expected number of agreeing verbs into account.

Table 4.8 illustrates that the different methods have individual strengths and weaknesses. (a) Evaluation measures without general minimum and maximum of the quality scores are more difficult, but possible to interpret. (b) In general, it is better to have quality values for both the specific clusters and the whole clustering, but we can do without the former. (c) Not acceptable for my linguistic needs are evaluation methods which (i) do not refer to the gold standard, because I want to measure how close we come to that, (ii) are dependent on a specific similarity measure, because I want to be able to compare the clustering results based on a range of similarity measures, (iii) have a strong bias towards many small or few large clusters, (iv) fail on a variety of linguistic demands, or (v) do not behave monotonically on error introduction.

To conclude, applicable evaluation methods to my clustering task are the f-score of pair-wise P/R $PairF$, the adjusted pair-wise precision APP, the adjusted Rand index $Rand_{adj}$, and the matching index B_k. Empirically, there is no large differences in the judgement of these methods, so I decided to concentrate on three measures with different aspects on the cluster evaluation: APP as the most linguistic evaluation, $PairF$ which provides an easy to understand percentage (usually the reader is familiar with judging about percentages), and the $Rand_{adj}$ which provides the most appropriate reference to a null model.

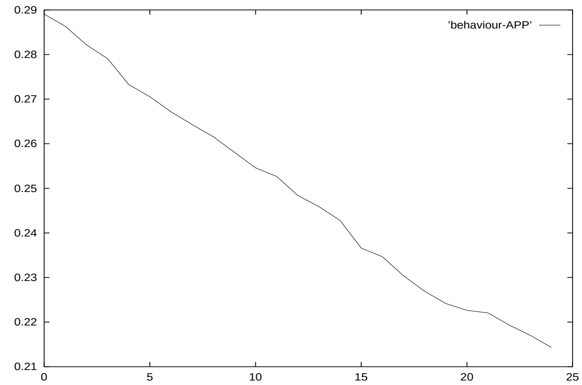


Figure 4.3: APP evaluation on introducing errors

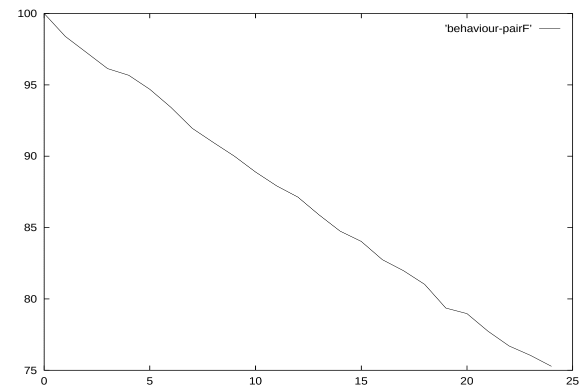


Figure 4.4: $PairF$ evaluation on introducing errors

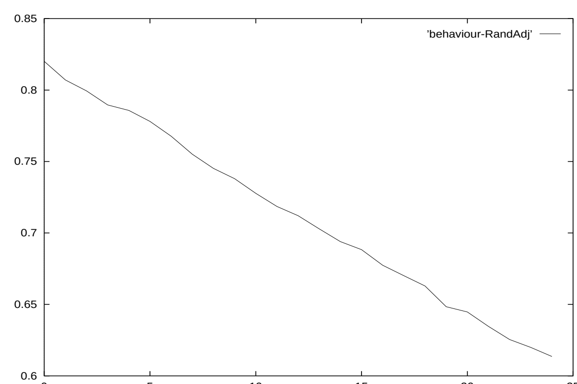


Figure 4.5: $Rand_{adj}$ evaluation on introducing errors

	gold standard	similarity measure		value	
		applicable	independent	specific	whole
Error	-	+	-	+	+
Silhouette	-	+	-	+	+
ClassR	+	+	+	-	+
ClassP	+	+	+	+	+
ClassF	+	+	+	-	+
PairR	+	+	+	-	+
PairP	+	+	+	-	+
PairF	+	+	+	-	+
APP	+	+	+	+	+
MI	+	+	+	+	+
Rand	+	+	+	-	+
Rand_{adj}	+	+	+	-	+
B-k	+	+	+	-	+

	bias	linguistics (failure)	error	interpretation	
				min	max
Error	many small	b, c, f	-	-	-
Silhouette	many small	b, f	-	-1	1
ClassR	few large	a, d, g	+	0	100
ClassP	many small	b, f	+	0	100
ClassF	few large	a	+	0	100
PairR	few large	a, d, g	+	0	100
PairP	-	c, f	+	0	100
PairF	-	-	+	0	100
APP	-	-	+	0	-
MI	many small	b	+	0	-
Rand	many small	b, d	+	0	1
Rand_{adj}	-	-	+	0	1
B-k	-	-	+	0	1

Table 4.8: Comparison of evaluation measures

4.3 Summary

This chapter has provided an overview of clustering algorithms and evaluation methods which are relevant for the natural language clustering task of clustering verbs into semantic classes. I have introduced the reader into the background of clustering theory and step-wise related the theoretical parameters for a cluster analysis to the linguistic cluster demands:

- The data objects in the clustering experiments are German verbs.

- The clustering purpose is to find a linguistically appropriate semantic classification of the verbs.

- I consider the alternation behaviour a key component for verb classes as defined in Chapter 2. The verbs are described on three levels at the syntax-semantic interface, and the representation of the verbs is realised by vectors which describe the verbs by distributions over their features.

- As a means for comparing the distributional verb vectors, I have presented a range of similarity measures which are commonly used for calculating the similarity of distributional objects.

- I have described a range of clustering techniques and argued for applying the hard clustering technique k-Means to the German verb data. k-Means will be used in the clustering experiments, initialised by random and hierarchically pre-processed cluster input.

- Based on a series of general evaluation demands, general clustering demands and specific linguistic clustering demands, I have presented a variety of evaluation measures from diverse areas. The different measures were compared against each other and according to the demands, and the adjusted pair-wise precision APP, the f-score of pair-wise P/R $PairF$, and the adjusted Rand index $Rand_{adj}$ were determined for evaluating the clustering experiments in the following chapter.

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.

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

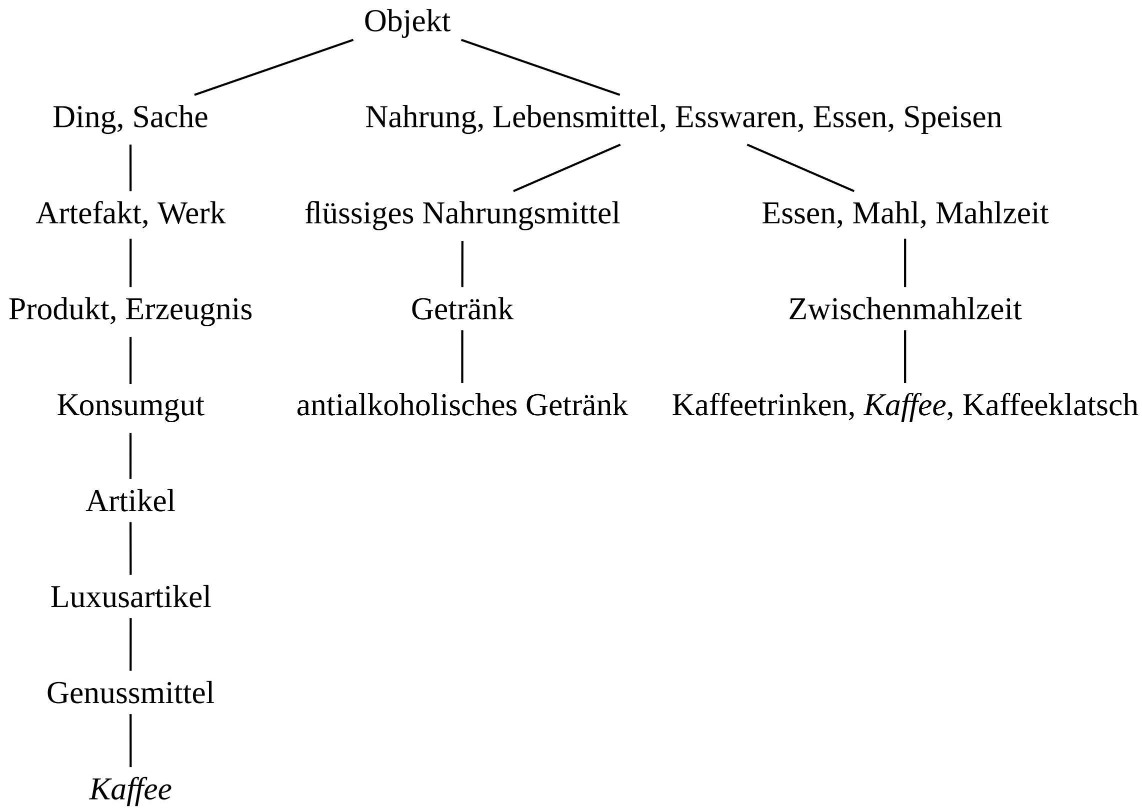


Figure 5.1: GermaNet hierarchy for noun *Kaffee* ‘coffee’

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.

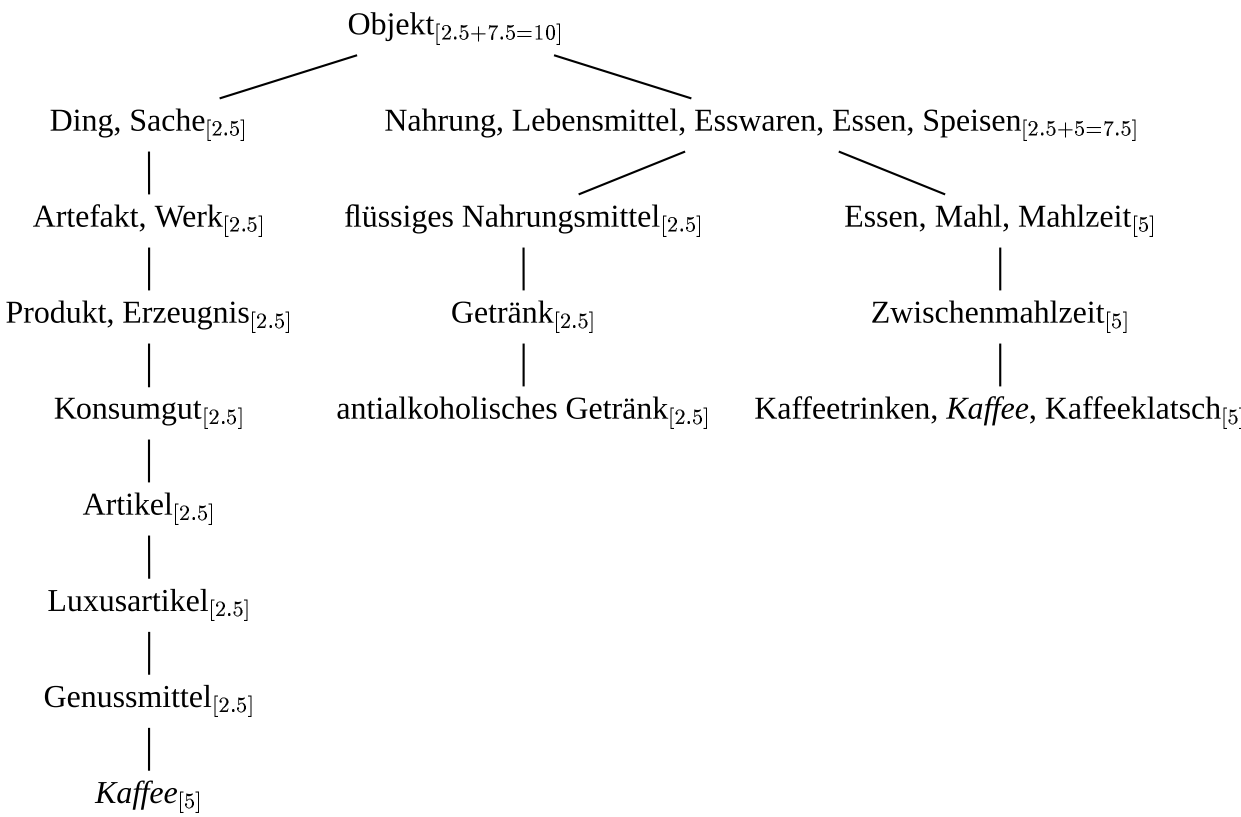


Figure 5.2: Propagating frequencies through GermaNet hierarchy

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 n<u>a</u> 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 $\mathtt{in}_{Dat}$ 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 $\underline{\mathtt{n}}$ and $\underline{\mathtt{n}}\mathtt{a}$ and *Nahrung* for $\mathtt{n}\underline{\mathtt{a}}$. *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 $\underline{\mathtt{n}}$ and in $\underline{\mathtt{n}}\mathtt{a}$, and the inchoative/causative case by *Sache* in $\underline{\mathtt{n}}$ and in $\mathtt{n}\underline{\mathtt{a}}$. 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 (<u>n</u>, n<u>a</u>, n<u>d</u>, n<u>a</u>d, <u>n</u>s-dass for prefS, and <u>n</u>, <u>na</u>, <u>nd</u>, n<u>a</u>d, <u>n</u>s-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	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	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	Random 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	Random 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	Random: 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	Random: 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 [38]	ppS [178]	ppA [183]	ppS+prefS [253]	ppA+prefA [288]	ppA+prefA_NP [906]	ppA+prefA_NP-PP [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 [38]	ppS [178]	ppA [183]	ppS+prefS [253]	ppA+prefA [288]	ppA+prefA_NP [906]	ppA+prefA_NP-PP [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 ($\rightarrow$ 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$_1$ $\underline{\text{helfen}_{10}}$

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

(7) blitzen$_{14}$ $\underline{\text{dienen}_{10}}$

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	frame+ppA	Selectional Preferences for frame+ppA+prefA					
			$\underline{\text{n}}$	$\underline{\text{n}}$a	n$\underline{\text{a}}$	$\underline{\text{n}}$ad	n$\underline{\text{a}}$d	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	frame+ppA	Selectional Preferences for frame+ppA+prefA						
			$\underline{\text{n}}$d	n$\underline{\text{d}}$	$\underline{\text{n}}$p	$\underline{\text{n}}$i	$\underline{\text{n}}$r	$\underline{\text{n}}$s-2	$\underline{\text{n}}$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	frame+ppA	Selectional Preferences ppA+prefA 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	frame+ppA	Selectional Preferences ppA+prefA 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 <u>and</u> 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	Verb Set: 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.

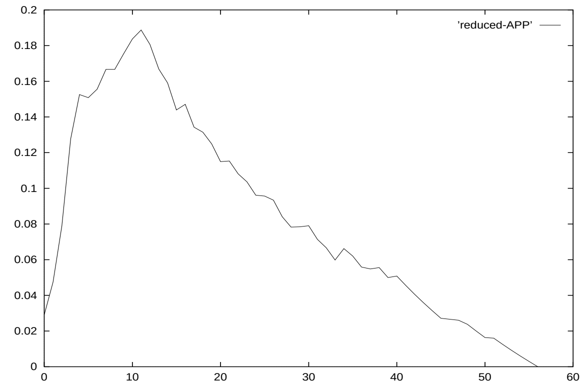


Figure 5.3: Varying the number of clusters on reduced verb set (evaluation: APP)

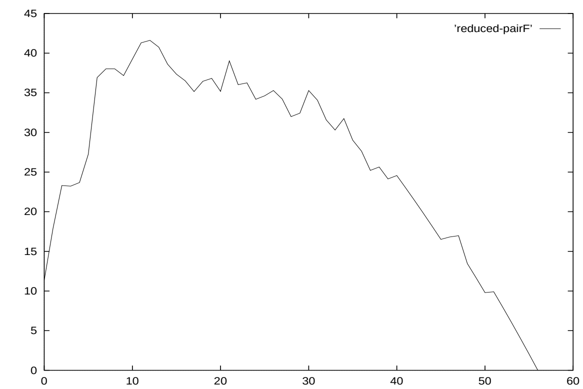


Figure 5.4: Varying the number of clusters on reduced verb set (evaluation: $PairF$)

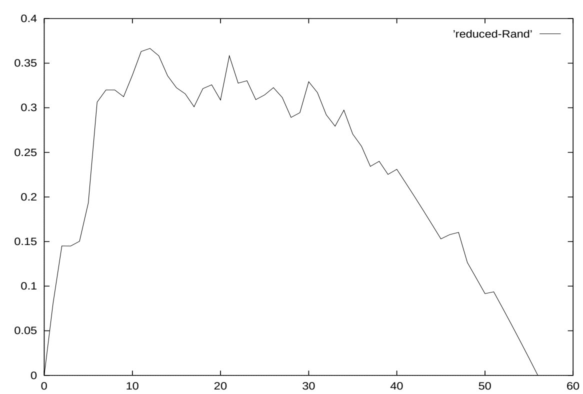


Figure 5.5: Varying the number of clusters on reduced verb set (evaluation: $Rand_{adj}$)

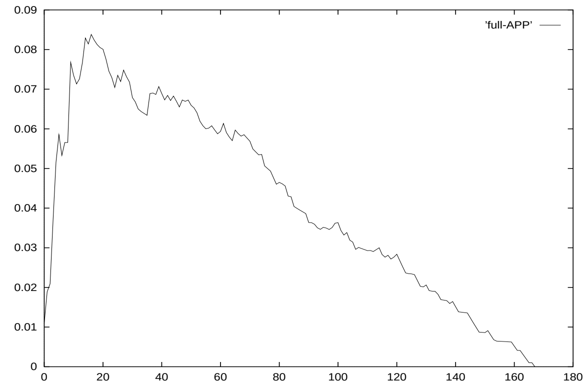


Figure 5.6: Varying the number of clusters on full verb set (evaluation: APP)

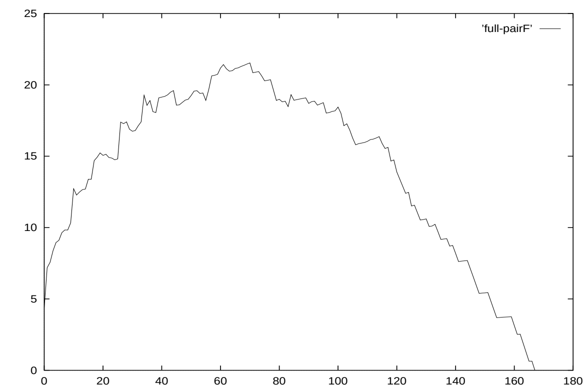


Figure 5.7: Varying the number of clusters on full verb set (evaluation: $PairF$)

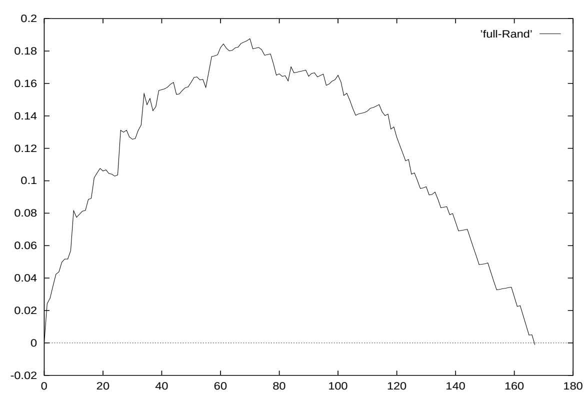


Figure 5.8: Varying the number of clusters on full verb set (evaluation: $Rand_{adj}$)

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 n/na/nd/nad/ns-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 $\uparrow$ indicating the head and $\downarrow$ 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) <u>Unergative Verbs</u>:
The horse raced past the barn.
The jockey raced the horse past the barn.

(5.4) <u>Unaccusative Verbs</u>:
The butter melted in the pan.
The cook melted the butter in the pan.

(5.5) <u>Object-Drop Verbs</u>:
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%.

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.

Appendix A

Subcategorisation Frame Types

The syntactic aspect of the German verb behaviour is captured by 38 subcategorisation frame types in the context-free German grammar, according to standard German grammar definitions, cf. Helbig and Buscha (1998). The exact choice of frame types is justified in Chapter 3 which describes the implementation of the context-free German grammar in detail.

The subcategorisation frame types comprise maximally three arguments. Possible arguments in the frames are nominative (n), dative (d) and accusative (a) noun phrases, reflexive pronouns (r), prepositional phrases (p), expletive *es* (x), subordinated non-finite clauses (i), subordinated finite clauses (s-2 for verb second clauses, s-dass for *dass*-clauses, s-ob for *ob*-clauses, s-w for indirect *wh*-questions), and copula constructions (k). The resulting frame types are listed in Table A.1, accompanied by annotated verb second example clauses.

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

Table A.1: Subcategorisation frame types

Appendix B

Corpus-Based Analysis of Subcategorisation Frames

The goal of this thesis is to find an appropriate clustering mechanism for German verbs which is able to induce semantic verb classes from verb descriptions at the syntax-semantic interface. The features which are used to describe the verbs should (a) be based on linguistic grounds and at the same time (b) have the potential to differentiate verbs with different meaning.

I performed a corpus analysis which empirically investigates the potential of the frame types in the grammar with respect to distinguishing verbs with different meanings. Tables B.1 to B.3 list all frame types, accompanied by examples of verbs which use the respective type. The verbs are roughly assigned to meaning classes m which comprise verbs with similar meanings. The aim is to illustrate which frame types are utilised by only single verbs and which are utilised by whole verb classes, and in which of the verb classes the verbs actually agree on the selectional preferences. The investigation should help to find arguments in frame types which are interesting to be refined by selectional preference information.

The corpus analysis illustrates that we can distinguish the frame types with respect to the variety of verbs using them.

- Some infrequent frame types are only utilised by verbs which are very similar in their meaning, e.g. `ndi, ns-ob`. A verb which appears with such a frame type can be assigned to the respective verb class related to the frame, e.g. class *Order* for `ndi` verbs.
- Some frame types are utilised by verbs which are semantically inhomogeneous and could also not be distinguished by preference refinements of the frame, e.g. `nr, xa`. Such frames might cause difficulties as verb features.
- Some frame types are utilised by verbs which form several semantically homogeneous classes, but there is no obvious way to distinguish the classes by selectional preference information, e.g. `nir`.

- Some frame types are utilised by verbs which form several semantically homogeneous classes, and by means of preference refinement we can distinguish the classes, e.g. `nd`, `ns-dass`. These frame types are especially interesting for the refinement of selectional preferences.

Obviously, the difference between the kinds of frame types is floating, and the distinction of verb classes by prepositional or selectional preferences refers to different levels of preferences, i.e. in some cases general labels can distinguish the classes, in other cases fine-grained information is needed. But the analysis illustrates that the prepositional phrases seem an important means to distinguish verbs with different meanings, and that certain arguments in the frames are especially important for a selectional preference refinement: n in n, n and particularly a in na, d and maybe n in nd, a in nad, and n in `ns-dass`.

Frame	Example Verbs	Comment
n	m_1: rennen, joggen, laufen m_2: muffeln, riechen, stinken m_3: gurren, kläffen, meckern, piepsen	frame distinguishes meaning classes m_i; n distinguishes selectional preferences
na	m_1: bauen, bilden, einrichten, erstellen m_2: essen, fressen, trinken, saufen m_3: ahnen, begreifen, verstehen m_4: bremsen, fahren, fliegen, lenken m_5: anfangen, beenden, beginnen, starten	frame distinguishes meaning classes m_i; a (mainly) and n distinguish preferences
nd	m_1: behagen, gefallen, guttun, mißfallen, schaden m_2: beipflichten, zustimmen, widersprechen m_3: angehören, beitreten, entstammen m_4: assistieren, dienen, helfen	frame distinguishes meaning classes m_i; d distinguishes preferences; n has less potential in distinction
np	m_1: bohren, fahnden, suchen m_2: basieren, beruhen, fußen m_3: feilschen, flehen, kämpfen m_4: liegen, lümmeln, sitzen, stehen m_5: anfangen, beginnen, enden	frame distinguishes meaning classes m_i; distinction by preposition
nad	m_1: anvertrauen, aufbürden, überlassen, zuweisen m_2: bescheinigen, nachsagen, unterstellen m_3: ankreiden, anlasten, vorwerfen	frame distinguishes meaning classes m_i; *n* and d mostly person; distinguishing argument: a

Table B.1: Corpus-based analysis of subcategorisation frames (1)

Frame	Example Verbs	Comment
nap	m_1: beziffern, dotieren, schätzen, veranschlagen m_2: lehnen, pinnen, projezieren	frame distinguishes few meaning classes; distinction by preposition
ndp	m_1: gereichen, verhelfen m_2: fehlen, mangeln	frame with few verbs, similar in meaning; subtle distinction by preposition
ni	m_1: beabsichtigen, bestreben, versuchen m_2: dürfen, können, mögen, müssen, sollen, wollen	frame filters modal verbs; n mostly person, no distinguishing argument
nai	m_1: befugen, bevollmächtigen, ermächtigen, verpflichten m_2: beschuldigen, bezichtigen, verdächtigen	frame comprises meaning classes m_i; no distinguishing argument
ndi	befehlen, freistellen, nahelegen, raten, untersagen	frame with few verbs, similar in meaning
nr	m_1: wohlfühlen, zurechtfinden m_2: abmühen, schwertun m_3: verkalkulieren, verspekulieren verirren auskennen rächen	frame comprises meaning classes m_i; verbs semantically inhomogeneous; no distinguishing argument
nar	m_1: aneignen, erschleichen, leihen, verschaffen, zuziehen m_2: erhoffen, wünschen m_3: einbilden, vormachen anschauen reiben	frame comprises meaning classes m_i; verbs semantically inhomogeneous; no distinguishing argument
ndr	nähern, hingeben, widmen, zuwenden	frame with few verbs, similar in meaning
npr	m_1: belaufen, läppern, summieren m_2: scharen, tummeln m_3: ärgern, empören, freuen, grämen m_4: befinden, ereignen	frame distinguishes meaning classes m_i; distinction by preposition and selectional preferences
nir	m_1: anschicken, bequemen m_2: bereiterklären, erbieten, weigern m_3: getrauen, scheuen	frame distinguishes meaning classes m_i; verbs similar in meaning; no distinguishing argument
ns-2	m_1: betonen, beteuern, versichern m_2: feixen, frohlocken, scherzen m_3: folgern, schließen	frame distinguishes meaning classes m_i; verbs similar in meaning; no distinguishing argument
nas-2	anschnauzen, ausschelten	frame with few verbs, similar in meaning
nds-2	m_1: nachrufen, zujubeln m_2: einschärfen, vorgaukeln, weismachen	frame with few verbs, similar in meaning; no distinguishing argument
nrs-2	entrüsten, ereifern, plustern	frame with few verbs, similar in meaning

Table B.2: Corpus-based analysis of subcategorisation frames (2)

Frame	Example Verbs	Comment
ns-dass	m_1: ahnen, argwöhnen, glauben, vermuten m_2: besagen, beweisen, hindeuten, implizieren m_3: herausfinden, merken, realisieren, registrieren m_4: andeuten, betonen, klarmachen, klarstellen	frame distinguishes meaning classes m_i; verbs similar on general meaning level; n partly distinguishes preferences
nas-dass	anflehen, beknien, überzeugen	frame with few verbs, similar in meaning
nds-dass	m_1: einreden, klarmachen, vorgaukeln, weismachen m_2: nachsehen, verübeln, zugutehalten	frame with few verbs, similar in meaning; no distinguishing argument
nrs-dass	m_1: freuen, genieren, schämen m_2: einbilden, vorstellen versteifen brüsten	frame comprises meaning classes m_i; verbs semantically inhomogeneous; no distinguishing argument
ns-ob	checken, prüfen, rausfinden, sondieren	frame with few verbs, similar in meaning
nas-ob		no corpus evidence
nds-ob		no corpus evidence
nrs-ob	fragen, vergewissern, überlegen	frame with few verbs, similar in meaning
ns-w	dahinterkommen, kapieren, rausfinden, spitzkriegen	frame with few verbs, similar in meaning
nas-w		no corpus evidence
nds-w	beibringen, nahebringen	frame with few verbs, similar in meaning
nrs-w	erkundigen, fragen, vergewissern, umhören	frame with few verbs, similar in meaning
x	m_1: frieren, gießen, nieseln, regnen, schneien m_2: grummeln, knistern, knirschen, rumoren	frame distinguishes few meaning classes m_i
xa	geben, hageln erwischen	frame with few inhomogeneous verbs
xd	dämmern schlechtgehen	frame with few inhomogeneous verbs
xp	hapern, mangeln kribbeln	frame with few inhomogeneous verbs
xr	geziemen	frame with few inhomogeneous verbs
xs-dass	stimmen, vorkommen	frame with few inhomogeneous verbs
k	bleiben, sein, werden	frame filters copula verbs

Table B.3: Corpus-based analysis of subcategorisation frames (3)

Appendix C

Large-Scale Set of German Verbs

Tables C.1 to C.8 present the 883 German verbs which are used within the large-scale clustering experiments in Section 5.5. The tables list the verbs and their frequencies in the training corpus, as estimated by the statistical grammar model.

Verb	Freq	Verb	Freq	Verb	Freq
abbauen	1,187	ankündigen	3,737	aufrufen	1,513
abbrechen	825	anlegen	1,038	aufstellen	1,353
abfahren	582	anmelden	1,317	auftauchen	1,355
abgeben	2,287	annehmen	2,447	auftreten	1,761
abhalten	649	anordnen	682	aufweisen	675
abholen	586	anpassen	577	aufzeigen	512
abhängen	887	anrufen	827	ausbauen	1,058
ablaufen	664	ansagen 8	82	ausbrechen	791
ablegen	572	anschauen	512	ausdrücken	735
ablehnen	5,608	anschließen	1,440	ausfallen	1,808
ablösen	636	ansehen	1,326	ausgeben	1,453
abnehmen	1,107	ansetzen	973	ausgehen	2,938
abreißen	701	ansprechen	992	ausgleichen	596
absagen	666	anstehen	962	auskommen	621
abschaffen	842	ansteigen	663	auslaufen	646

Table C.1: Large-scale set of German verbs (1)

Verb	Freq	Verb	Freq	Verb	Freq
abschließen	2,164	anstreben	829	auslösen	1,790
abschneiden	530	antreten	1,837	ausmachen	1,598
absehen	602	antworten	1,558	ausreichen	1,268
absetzen	549	anweisen	739	ausschließen	2,985
absolvieren	515	anwenden	667	aussehen	3,866
abstimmen	799	anzeigen	524	aussetzen	1,173
abwarten	965	anziehen	554	aussprechen	2,189
abziehen	563	appellieren	595	ausstehen	714
achten	579	arbeiten	8,761	ausstellen	802
agieren	507	argumentieren	815	ausweisen	916
ahnen	553	aufbauen	1,833	ausweiten	531
akzeptieren	1,809	aufbrechen	602	auswirken	653
alarmieren	504	aufbringen	763	auszahlen	512
anbieten	3,525	auffallen	689	auszeichnen	779
andeuten	586	auffordern	1,543	ausüben	687
anerkennen	1,225	aufführen	530	basieren	355
anfangen	2,554	aufgeben	2,092	bauen	3,878
angeben	1,425	aufgehen	972	beantragen	1,703
angebieten	1,060	aufhalten	939	beantworten	1,180
angehen	1,150	aufheben	1,577	beauftragen	520
angehören	1,206	aufhören	993	bedauern	945
angreifen	1,120	aufklären	677	bedeuten	4,858
anhalten	907	aufkommen	1,041	bedienen	803
anhören	613	auflösen	1,174	bedrohen	1,138
ankommen	1,831	aufnehmen	4,065	bedürfen	1,063
beeindrucken	525	beobachten	1,654	bewahren	637
beeinflussen	603	beraten	1,637	bewegen	2,137
beenden	1,978	bereiten	1,372	beweisen	1,686
befassen	905	bereitstellen	681	bewerben	586
befinden	4,572	bergen	951	bewerten	852
befragen	534	berichten	6,780	bewirken	600
befreien	817	berlinern	1,058	bewältigen	588
befürchten	2,203	berufen	971	bezahlen	2,948
befürworten	602	beruhen	543	bezeichnen	4,331
begegnen	928	beruhigen	724	beziehen	1,771
begehen	809	berücksichtigen	1,205	beziffern	1,054
begeistern	573	beschlagnahmen	586	bezweifeln	682
beginnen	14,612	beschleunigen	535	bieten	5,603
begleiten	1,260	beschließen	3,517	bilden	3,159
begreifen	1,024	beschreiben	1,914	billigen	690

Table C.2: Large-scale set of German verbs (2)

Verb	Freq	Verb	Freq	Verb	Freq
begrenzen	599	beschränken	939	binden	870
begründen	1,725	beschädigen	773	bitten	3,883
begrüßen	2,003	beschäftigen	3,464	blicken	1,193
behalten	1,331	beseitigen	756	blitzen	122
behandeln	1,862	besetzen	1,792	blockieren	925
beharren	348	besitzen	1,722	brauchen	10,075
behaupten	2,184	besorgen	761	brechen	1,420
beherrschen	892	bessern	532	bremsen	606
behindern	681	bestehen	7,101	brennen	851
beibringen	242	bestellen	869	bringen	12,249
beitragen	1,694	bestimmen	2,400	charakterisieren	222
bekanntgeben	1,643	bestrafen	530	dachen	867
bekennen	4,659	bestreiten	1,864	danken	1,190
beklagen	1,435	bestätigen	5,608	darstellen	2,359
bekommen	10,022	besuchen	2,185	dauern	3,466
bekräftigen	700	beteiligen	3,170	decken	802
bekämpfen	682	betonen	3,922	definieren	636
belasten	973	betrachten	1,471	dekorieren	56
belaufen	1,164	betragen	5,797	dementieren	813
belegen	1,861	betreffen	2,751	demonstrieren	1,541
bemerken	926	betreiben	2,056	denken	6,011
bemühen	1,608	betreten	644	dienen	3,121
benennen	555	betreuen	1,039	diskutieren	2,543
benutzen	1,568	beurteilen	912	dokumentieren	802
benötigen	1,268	bevorstehen	911	dominieren	717
donnern	105	ekeln	31	erinnern	3,416
drehen	1,573	eliminieren	52	erkennen	2,936
dringen	624	empfangen	893	erlangen	237
drohen	4,574	empfehlen	2,173	erlassen	785
drängen	1,309	empfinden	1,077	erlauben	1,544
drücken	976	enden	3,079	erleben	2,591
durchführen	1,441	engagieren	871	erledigen	839
durchsetzen	2,584	entdecken	2,765	erleichtern	745
dämmern	71	entfallen	1,423	erleiden	1,817
dürfen	3,543	entfernen	1,119	erläutern	1,340
ehren	813	entgegennehmen	922	ermitteln	2,123
eignen	677	entgehen	522	ermorden	820
eilen	226	enthalten	2,388	ermöglichen	1,166
einbeziehen	524	entkommen	588	ernennen	500
einbringen	1,164	entlassen	1,480	erneuern	457

Table C.3: Large-scale set of German verbs (3)

Verb	Freq	Verb	Freq	Verb	Freq
einfallen	831	entlasten	513	erniedrigen	35
einführen	1,429	entscheiden	6,746	ernten	541
eingehen	2,021	entschließen	573	erreichen	7,444
eingreifen	788	entschuldigen	629	errichten	1,082
einhalten	729	entsprechen	2,814	erscheinen	4,755
einholen	527	entstehen	7,273	erschießen	1,284
einigen	1,897	enttäuschen	617	ersetzen	1,283
einladen	2,648	entwickeln	3,425	erstatten	758
einlassen	565	entziehen	857	erstellen	639
einlegen	1,244	erarbeiten	724	erteilen	1,398
einleiten	866	ereignen	638	erwachsen	180
einnehmen	1,054	erfahren	2,632	erwarten	8,139
einrichten	2,289	erfassen	805	erweisen	1,774
einräumen	1,668	erfinden	545	erweitern	982
einschalten	762	erfolgen	2,750	erwerben	861
einschlagen	645	erfordern	851	erwischen	647
einschränken	704	erfüllen	2,285	erwägen	583
einschätzen	544	ergeben	3,118	erwähnen	751
einsetzen	3,990	ergehen	719	erzeugen	665
einsparen	614	ergreifen	709	erzielen	2,102
einsteigen	533	ergänzen	957	erzählen	3,992
einstellen	2,823	erhalten	10,716	eröffnen	2,787
eintreffen	710	erheben	1,930	essen	1,352
eintreten	1,407	erhoffen	680	exekutieren	64
einziehen	882	erhöhen	2,881	existieren	1,603
explodieren	557	gegenüberstehen	1,271	hinnehmen	1,105
fahren	5,451	gehaben	982	hinterlassen	973
fallen	5,894	gehen	33,110	hinweisen	1,671
fassen	1,588	gehören	9,686	hinzufügen	800
fehlen	8,543	gelangen	1,894	hinzukommen	798
feiern	3,799	gelingen	2,632	hoffen	4,185
fernsehen	778	genehmigen	783	holen	1,914
festhalten	1,468	genießen	1,144	hängen	2,455
festlegen	1,556	genügen	1,103	hören	5,040
festnehmen	2,692	geraten	3,132	identifizieren	558
feststehen	969	geschehen	3,680	ignorieren	601
feststellen	2,876	gestalten	1,366	informieren	2,149
filmen	693	gestehen	923	insistieren	36
finanzieren	1,871	gewinnen	5,633	inszenieren	698
fliegen	1,634	gewähren	817	interessieren	2,549

Table C.4: Large-scale set of German verbs (4)

Verb	Freq	Verb	Freq	Verb	Freq
fliehen	985	gewährleisten	520	interpretieren	512
fließen	1,512	gewöhnen	623	investieren	2,328
flüchten	1,378	glauben	7,275	jammern	128
flüstern	123	gleichen	1,542	kandidieren	658
folgen	6,570	gleiten	83	kassieren	728
folgern	102	gratulieren	615	kaufen	2,846
formulieren	1,332	greifen	1,800	kennen	5,489
fortsetzen	2,959	grinsen	168	kennenlernen	858
fragen	6,617	grübeln	66	klagen	2,093
freien	766	gründen	2,465	klappen	940
freuen	2,478	gucken	583	klettern	685
funktionieren	2,058	gähnen	63	klingen	1,788
fällen	1,534	handeln	5,121	klären	1,344
fördern	1,980	hasten	85	kochen	697
fühlen	4,506	hausen	1,122	kommentieren	1,204
füllen	1,016	heben	572	kommunizieren	91
fürchten	2,003	heiraten	657	konsumieren	126
garantieren	932	helfen	5,281	kontrollieren	1,272
garen	879	herauskommen	721	konzentrieren	993
geben	71,604	herrschen	4,369	korrespondieren	66
gebrauchen	618	herstellen	1,546	korrigieren	565
gebären	1,195	hervorbringen	166	kosen	1,213
gedenken	699	hervorgehen	1,057	kosten	6,524
gefallen	1,849	heuen	1,372	kreisen	3,491
gefährden	1,245	heulen	160	kriechen	125
kriegen	2,044	mischen	509	prüfen	1,828
kritisieren	3,354	mitbringen	1,264	raten	1,008
kämpfen	2,820	mitkündigen	1,115	reagieren	4,091
kümmern	1,566	mitmachen	1,040	realisieren	581
kündigen	723	mitnehmen	1,077	rechen	879
kürzen	611	mitspielen	566	rechnen	6,086
lachen	1,428	mitteilen	4,225	rechtfertigen	817
laden	979	mögen	3,175	reden	3,987
lagern	518	nachdenken	1,142	reduzieren	1,395
lamentieren	52	nachgeben	588	regeln	1,068
landen	1,454	nachgehen	575	regieren	864
langen	1,040	nachkommen	501	registrieren	1,591
laufen	6,904	nachweisen	654	regnen	295
lauten	3,613	nennen	8,946	reichen	3,465
leben	9,566	nieseln	8	reisen	1,137

Table C.5: Large-scale set of German verbs (5)

Verb	Freq	Verb	Freq	Verb	Freq
legen	3,013	notieren	724	reißen	672
lehren	538	nutzen	3,414	rennen	366
leiden	2,230	nähern	553	renovieren	281
leihen	139	nützen	780	reparieren	253
leisten	3,780	offenbaren	634	resultieren	215
leiten	1,208	opfern	524	retten	1,931
lernen	3,407	organisieren	2,119	richten	1,905
lesen	3,592	orientieren	725	riechen	523
lieben	2,187	packen	549	rollen	977
liefern	2,395	passen	1,671	rotieren	62
liegen	20,078	passieren	3,989	rudern	49
loben	1,011	pflegen	906	rufen	2,397
locken	751	phantasieren	26	ruhen	539
lohnen	1,194	planen	4,831	räumen	990
losgehen	837	platzen	1,049	rücken	824
lächeln	397	plädieren	1,534	sammeln	1,791
löschen	612	pochen	198	saufen	80
lösen	2,448	preisgeben	512	schaden	1,302
mahnen	512	proben	617	schaffen	7,129
malen	1,748	produzieren	1,762	schauen	1,019
mangeln	665	profitieren	1,317	scheinen	8,129
mehren	4,127	protestieren	1,389	scheitern	2,606
melden	4,407	provozieren	517	schenken	650
merken	1,105	prägen	1,044	scheuen	374
messen	941	präsentieren	2,460	schicken	2,072
schieben	663	starren	179	umfassen	1,039
schießen	1,727	starten	1,954	umgehen	1,624
schildern	723	stattfinden	7,058	umsetzen	1,400
schimpfen	597	stecken	2,939	unterbrechen	805
schlafen	885	stehen	31,710	unterbringen	1,214
schlagen	3,038	stehlen	669	unterhalten	851
schleichen	138	steigen	5,678	unterliegen	865
schließen	3,986	steigern	808	unternehmen	1,167
schneien	92	stellen	11,233	unterrichten	626
schonen	1,465	sterben	4,943	untersagen	514
schreiben	6,649	stiften	317	unterscheiden	1,324
schreien	709	stimmen	3,494	unterschreiben	751
schrumpfen	597	stoppen	1,458	unterstreichen	717
schulen	662	stoßen	2,375	unterstützen	3,722
schweigen	1,019	streichen	1,511	untersuchen	12,64

Table C.6: Large-scale set of German verbs (6)

Verb	Freq	Verb	Freq	Verb	Freq
schätzen	2,684	streiken	523	unterzeichnen	834
schützen	1,781	streiten	943	urteilen	637
segeln	98	studieren	984	verabschieden	1,710
sehen	24,862	stunden	1,011	veranschaulichen	76
senden	505	stärken	874	veranschlagen	639
senken	812	stören	1,531	veranstalten	958
setzen	7,545	stürzen	998	verantworten	501
sichern	2,021	stützen	673	verbergen	785
sicherstellen	908	suchen	6,546	verbessern	1,694
siegen	934	tagen	1,749	verbieten	2,490
signalisieren	887	tanzen	1,155	verbinden	1,767
singen	1,875	teilen	2,131	verbreiten	935
sinken	2,174	teilnehmen	2,156	verbrennen	588
sitzen	6,124	tragen	5,879	verbringen	889
sollen	9,597	trauen	857	verbuchen	615
sondern	5,555	treffen	8,040	verdienen	2,614
sorgen	5,133	treiben	1,572	verdoppeln	572
sparen	1,786	trennen	1,204	verdrängen	512
spekulieren	461	treten	2,734	vereinbaren	998
spenden	702	triefen	781	vereinen	1,022
sperren	1,209	trinken	986	verfolgen	1,732
sprechen	12,984	träumen	779	verfügen	1,875
springen	736	tunen	558	vergeben	1,381
spüren	1,706	töten	2,433	vergehen	673
stammen	2,702	umbringen	683	vergessen	2,187
vergleichen	716	verstecken	676	wechseln	1,520
vergrößern	382	verstehen	4,900	wecken	637
verhaften	1,074	verstoßen	642	wehren	1,308
verhalten	1,207	verständigen	534	weichen	560
verhandeln	1,834	verstärken	1,322	weigern	870
verhindern	3,632	versuchen	7,144	weilen	567
verhängen	539	verteidigen	1,506	weinen	452
verkaufen	3,820	verteilen	1,966	weisen	1,105
verkleinern	151	vertreiben	738	weitergehen	2,794
verkünden	1,358	vertreten	2,647	weißen	6,263
verlangen	4,460	verursachen	925	wenden	1,780
verlassen	4,063	verurteilen	2,232	werben	1,186
verlaufen	1,945	verwandeln	765	werfen	2,221
verlegen	997	verweigern	1,682	werten	1,058
verleihen	853	verweisen	1,715	widersprechen	1,534

Table C.7: Large-scale set of German verbs (7)

Verb	Freq	Verb	Freq	Verb	Freq
verletzen	4,855	verwenden	1,484	widmen	1,071
verlieren	5,959	verwirklichen	613	wiederholen	1,302
verlorengehen	720	verzeichnen	1,211	wiegen	644
verlängern	789	verzichten	2,730	wirken	3,173
vermachen	31	verändern	2,616	wissen	12,312
vermeiden	1,650	veröffentlichen	1,648	wohnen	2,140
vermissen	854	vollziehen	632	wollen	21,464
vermitteln	1,415	vorbereiten	1,881	wundern	707
vermuten	1,758	vorführen	700	wählen	4,322
vermögen	620	vorgehen	1,274	wünschen	2,534
vernichten	669	vorkommen	1,098	würdigen	596
verpassen	538	vorlegen	2,072	zahlen	4,194
verpflichten	1,089	vorliegen	3,391	zeichnen	632
verraten	1,184	vornehmen	1,340	zeihen	654
verringern	923	vorschlagen	1,791	zerstören	1,988
versagen	651	vorsehen	3,624	ziehen	6,166
versammeln	581	vorstellen	3,851	zielen	626
verschaffen	626	vorwerfen	2,214	zitieren	1,136
verschieben	979	wachsen	3,340	zugeben	875
verschwinden	2,414	wagen	1,116	zugehen	992
verschärfen	901	wahrnehmen	824	zugutekommen	505
versetzen	579	wandeln	542	zukommen	1,103
versichern	1,756	wandern	581	zulassen	1,497
versorgen	838	warnen	3,594	zulegen	778
versprechen	3,104	warten	4,210	zunehmen	1,773
zurückführen	687	zustimmen	2,469	überlegen	852
zurückgeben	541	zutreffen	563	übernehmen	5,760
zurückgehen	1,851	zwingen	1,595	überprüfen	1,225
zurückhalten	530	zählen	3,783	überraschen	972
zurückkehren	1,759	ändern	6,007	überreichen	633
zurückkommen	713	ängstigen	33	überschreiben	92
zurücknehmen	563	ärgern	627	überschreiten	863
zurücktreten	913	äußern	3,849	übersehen	629
zurückweisen	2,282	öffnen	2,643	überstehen	522
zurückziehen	1,268	üben	1,511	übertragen	902
zusagen	869	überfallen	626	überwachen	520
zusammenarbeiten	964	übergeben	903	überwinden	718
zusammenbrechen	683	übergehen	613	überzeugen	1,562
zusammenkommen	837	überlassen	984		
zustellen	127	überleben	1,334		

Table C.8: Large-scale set of German verbs (8)

Zusammenfassung

1. Motivation

Das Verb hat eine zentrale Position im Satz, da es die Struktur und die Bedeutung eines Satzes maßgeblich beeinflusst. Lexikalische Informationen zu Verben sind daher von grundlegender Bedeutung im Bereich der natürlichen Sprachverarbeitung. Es ist allerdings sehr aufwendig, die Feinheiten der natürlichen Sprache manuell zu definieren, und besonders semantische Ressourcen stellen in diesem Zusammenhang einen Engpass dar. Automatische Methoden zur Induktion von lexikalischem Wissen haben daher erheblich an Bedeutung gewonnen. Diese Arbeit stellt einen Beitrag zur automatischen Erstellung von lexikalisch-semantischem Wissen dar, der die Möglichkeiten und Grenzen einer Methode für eine automatische Induktion von semantischen Verbklassen fürs Deutsche vorstellt.

Semantische Verbklassen

Semantische Verbklassen generalisieren über Verben in Bezug auf ihre semantischen Eigenschaften. Die Klassen sind ein nützliches Mittel, umfangreiche Informationen zu Verben zu erfassen, ohne die idiosynkratischen Details für jedes Verb zu definieren. Die Namen der Verbklassen beschreiben die Verben auf einer konzeptuellen Ebene, und die idiosynkratischen lexikalischen Eigenschaften der Verben bleiben entweder unterspezifiziert oder werden hinzugefügt. Beispiele für Verbklassen auf der konzeptuellen semantischen Ebene sind *Positionsverben* (*liegen, sitzen, stehen*) oder *Bewegungsverben mit einem Fahrzeug* (*fahren, fliegen, rudern*). Für einige Sprachen sind bereits semantische Verbklassen definiert worden, z.B. Englisch (Levin, 1993; Baker *et al.*, 1998) und Spanisch (Vázquez *et al.*, 2000). Nach meinem Wissen existiert im Deutschen keine semantische Verbklassifikation, die maschinell für die natürliche Sprachverarbeitung zur Verfügung steht.

Was ist der Nutzen von Verbklassen in der natürlichen Sprachverarbeitung? Einerseits reduzieren Verbklassen Redundanz in Verbbeschreibungen, da sie sich auf gemeinsame Eigenschaften von Verben beschränken. Andererseits können Verbklassen Eigenschaften von Verben, die empirisch nur ungenügend beschrieben sind, hinzufügen oder vorhersagen, weil Eigenschaften von Verben derselben Klasse übertragen werden können. Es gibt bereits Arbeiten, die, basierend auf der englischen Verbklassifikation von Levin (1993), den Nutzen von Verbklassen in der natürlichen Sprachverarbeitung demonstriert haben: Dorr and Jones (1996) stellen einen Ansatz zur Desambiguierung von Verbbedeutungen vor, Dorr (1997) beschreibt den Nutzen der Klassifikation im

Bereich der maschinellen Übersetzung, und Klavans and Kan (1998) verwenden die Verbklassen für Dokumentenklassifikation.

Automatische Induktion von Semantischen Verbklassen im Deutschen

Wie können wir eine semantische Klassifikation von Verben erstellen, ohne sie von Hand zu definieren? Es ist schwierig, semantische Eigenschaften auf der Basis von vorhandenen Ressourcen zu lernen, sowohl in Bezug auf lexikalische Eigenschaften als auch in Bezug auf konzeptuelle Strukturen. Üblicherweise wird daher auf eine etablierte linguistische Hypothese zurückgegriffen, die einen engen Bezug zwischen den lexikalischen Bedeutungseigenschaften und dem Verhalten eines Verbs vorhersagt: Zu einem gewissen Grad bestimmt die Bedeutung eines Verbs sein Verhalten im Satz, besonders in Bezug auf die Wahl der Verbargumente, cf. Levin (1993, page 1). Diese Hypothese ist sehr nützlich bei der Induktion von semantischen Verbklassen, denn wir können eine Verbklassifikation aufgrund der Verhaltensweisen eines Verbs erstellen (die leichter zu lernen sind als semantische Eigenschaften), und diese Verhaltensklassifikation sollte zu einem gewissen Grad mit einer semantischen Klassifikation übereinstimmen.

Das Verbverhalten kann durch Verbalternationen dargestellt werden, alternative Verbkonstruktionen der Syntax-Semantik-Schnittstelle, die entweder dieselbe oder eine sehr ähnliche Bedeutung haben. Beispiel (1) stellt die häufigsten Alternationen für das Verb *fahren* dar. Die konzeptuellen Argumente, die typischerweise im Zusammenhang mit *fahren* verwendet werden, sind ein Fahrer, ein Fahrzeug, eine gefahrene Person und ein Weg. In (a) ist ein Fahrzeug das Subjekt der transitiven Verbalphrase, und eine Präpositionalphrase bezeichnet den Weg der Bewegung. In (b) ist ein Fahrer das Subjekt der transitiven Verbalphrase, und eine Präpositionalphrase bezeichnet wiederum den Weg der Bewegung. In (c) ist ein Fahrer das Subjekt der transitiven Verbalphrase, und das Akkusativobjekt bezeichnet ein Fahrzeug. In (d) ist ein Fahrer das Subjekt der ditransitiven Verbalphrase, das Akkusativobjekt bezeichnet eine gefahrene Person, und eine Präpositionalphrase bezeichnet den Weg.

(1) (a) *Der Wagen fährt in die Innenstadt.*
(b) *Die Frau fährt nach Hause.*
(c) *Der Filius fährt einen blauen Ferrari.*
(d) *Der Junge fährt seinen Vater zum Zug.*

Unter der Annahme, dass das Verbverhalten durch die Alternationen des Verbs charakterisiert werden kann, welche syntaktischen und semantischen Eigenschaften benötigt man für eine Beschreibung des Verbverhaltens? Beispiel (1) hat verdeutlicht, dass (i) die syntaktischen Strukturen relevant sind, weil sie die Funktionen der Verbargumente realisieren, (ii) Präpositionen relevant sind, um zum Beispiel Richtungsangaben von Ortsangaben zu unterscheiden, und (iii) Selektionspräferenzen relevant sind, weil sie die konzeptuellen Rollen der Arguments definieren. Diese drei Ebenen werden im Folgenden für eine Verbbeschreibung verwendet.

Wenn nun eine solche Beschreibung des Verbverhaltens vorliegt, wie kann man diese als Grundlage für die automatische Induktion von semantischen Verbklassen verwenden? Diese Arbeit wendet einen Algorithmus zum Clustering an, der die Verbbeschreibung auf der Syntax-Semantik-Schnittstelle als empirische Verbeigenschaften benutzt und auf der Basis dieser Verbeigenschaften eine semantische Klassifikation lernt. Bei der Anwendung eines Clustering-Algorithmus auf multivariante Daten kommt es allerdings oft zu Überraschungen, weil man nicht notwendigerweise die Effekte der Datendefinition und des Algorithmus unterscheiden kann und das Ergebnis der Clusteranalyse daher nicht leicht zu interpretieren ist. Meiner Meinung nach ist es aber sehr wichtig, die Daten und den Algorithmus unter Berücksichtigung des Aufgabengebietes zu definieren. Wir forschen im Bereich der Linguistik, also sollten auch die Daten linguistisch definiert sein und der Algorithmus mit linguistischen Daten umgehen können. Im Rahmen dieser Arbeit habe ich mich daher auf zwei Teilaspekte der automatischen Induktion von Verbklassen konzentriert. Ich habe empirisch untersucht, wie Verben auf der Syntax-Semantik-Schnittstelle definiert werden können, mit anderen Worten: (i) Welches sind die semantischen Eigenschaften, die semantische Verbklassen charakterisieren? (ii) Welches sind die Eigenschaften, die Verbverhalten charakterisieren? (iii) In welchem Maße kann man den Zusammenhang zwischen Verbbedeutung und Verbverhalten benutzen, um semantische Verbklassen automatisch zu lernen? Der zweite Teilaspekt beschäftigt sich mit der Entwicklung einer Clustering-Methode, die in der natürlichen Sprachverarbeitung eingesetzt werden kann. Die Parameter des Algorithmus sollen so definiert werden, dass die Clusteranalyse weitgehend mit einer natürlichsprachlichen Klassifikation übereinstimmt. Meiner Meinung nach ist es sehr wichtig, das Potential und die Einschränkungen der linguistischen und technischen Teilaspekte zu verstehen; denn nur so kann eine Methode entwickelt werden, die uneingeschränkt auf Verbdaten angewendet werden kann.

2. Clustering-Methode

Die Clustering-Methode bringt die folgenden Aspekte in einen Zusammenhang: (a) das Konzept einer semantischen Verbklassifikation im Deutschen, (b) die empirischen Daten für die Verbbeschreibung auf der Syntax-Semantik-Schnittstelle, (c) den Clustering-Algorithmus und (d) Maße für die Evaluierung von Clusteranalysen.

<u>Semantische Verbklassen im Deutschen</u>

Ich habe manuell eine Verbklassifikation mit 43 semantischen Klassen und 168 teilweise ambigen deutschen Verben erstellt. Die Konstruktion der Verbklassen hat sich primär an den semantischen Eigenschaften der Verben orientiert: Die Verben werden den Klassen aufgrund ihrer lexikalischen und konzeptuellen Bedeutung zugewiesen, und die Klassen erhalten einen konzeptuellen Namen. Die Klassen beziehen sich auf zwei konzeptuelle Ebenen; allgemeine konzeptuelle Klassen wie z.B. *Bewegungsverben* wurden in spezifischere konzeptuelle Klassen unterteilt wie z.B. *Art der Fortbewegung, Rotation, Eile, Fahrzeug, Fließen*. Da es einen Bezug zwischen Verbbedeutung und Verbverhalten gibt, zeigen die Verben, die aufgrund der konzeptuellen Bedeutung einer Klasse zugewiesen wurden, auch eine gewisse Übereinstimmung in ihrem Verhalten. Die Klassengröße in der Verbklassifikation variiert von 2 bis 7 Verben pro Klasse, was einem Durchschnitt von 3.9 Verben pro Klasse entspricht. Acht Verben sind in Bezug auf die Klassenzuweisung ambig und wurden jeweils zwei Klassen zugeordnet. Die Klassen enthalten sowohl hochfrequente als auch niedrigfrequente Verben: Die Verbfrequenz variiert von 8 bis 71.604. Die Verbklassifikation entspricht bei der Evaluierung der Clustering-Experimente dem goldenen Standard.

Jede Klasse wurde detailliert im Rahmen von Fillmore's 'Scenes-and-Frames' Semantik (Fillmore, 1977, 1982) beschrieben, die in ähnlicher Weise bereits im Rahmen von *FrameNet* verwendet wurde (Baker *et al.*, 1998; Johnson *et al.*, 2002): Eine Szenenbeschreibung charakterisiert die Klasse und definiert die Rollen für obligatorische und modifizierende Verbargumente. Die Verbrahmen, die im Rahmen der Klassenbedeutung Verwendung finden, werden aufgelistet. Die Rahmen und die Rollen wurden auf der Basis von Zeitungskorpora aus den 1990er Jahren entwickelt. Für jedes Verb und jeden Rahmen in einer Klasse wurden Beispielsätze aus den Korpora extrahiert. Da in jedem Satz Rahmen und Rollen annotiert wurden, illustrieren die Beispielsätze die Verbalternationen, denn die Rollen können über die Rahmengrenzen hinweg in Verbindung gebracht werden.

Empirische Distributionen für deutsche Verben

Ich habe ein statistisches Grammatikmodell fürs Deutsche entwickelt, implementiert und trainiert. Die Grammatik beruht auf der Idee von lexikalisierten probabilistischen kontextfreien Grammatiken, die ursprünglich von Charniak (1995) entwickelt wurde. Im Rahmen dieser Arbeit habe ich eine Implementierung von Schmid (2000) verwendet. Das statistische Grammatikmodell enthält empirische lexikalische Information, die auf die Subkategorisierung von Verben spezialisiert, aber nicht beschränkt ist.

Die empirische Information in dem Grammatikmodell wird für die Verbbeschreibung verwendet. Ich stelle die deutschen Verben durch distributionelle Vektoren dar, basierend auf der Hypothese von Harris (1968), dass 'jede Sprache durch ihre distributionelle Struktur beschrieben werden kann, d.h. durch das Vorkommen mit anderen Teilen der Sprache'. Die Verben werden distributionell auf drei Ebenen der Syntax-Semantik-Schnittstelle beschrieben, wobei jede Ebene die vorherige durch zusätzliche Information anreichert. Auf der Ebene $D1$ werden rein syntaktische Eigenschaften von Verben dargestellt, auf der Ebene $D2$ wird präpositionale Information hinzugefügt, und auf der Ebene $D3$ werden Selektionspräferenzen definiert. Ich fange so auf einer rein syntaktischen Ebene für Verbbeschreibung an und füge schrittweise semantische Information hinzu. Die umfangreichste Verbbeschreibung entspricht in etwa der Beschreibung von Verbalternation. Ich habe mich für diese Dreiteilung von Verbeigenschaften entschieden, weil die Clusteranalysen auf der Basis der verschiedenen Verbeigenschaften Einblick in den Zusammenhang zwischen Verbbedeutung und Verbverhalten geben sollen.

Die folgende Tabelle illustriert die Verbbeschreibungen auf den drei Ebenen, und zwar für drei Verben aus drei unterschiedlichen Verbklassen. Für jede Ebene der Verbbeschreibung werden die zehn wahrscheinlichsten Verbeigenschaften dargestellt. Die Rahmen der Verben setzen sich aus Abkürzungen für Verbargumente zusammen: Nominalphrasen im Nominativ (n), Akkusativ (a) und Dativ (d), Reflexivpronomen (r), Präpositionalphrasen (p), Expletive (x), nicht-finite subkategorisierte Sätze (i), finite subkategorisierte Sätze (s-2 bei Verb-Zweit-Sätzen, s-dass bei *dass*-Sätzen, s-ob bei *ob*-Sätzen und s-w bei indirekten Fragen) und Kopula-Konstruktionen (k). Der konkrete Bezug auf Präpositionalphrasen erfolgt durch Kasus und Präposition, z.B. 'mit$_{Dat}$', und 'für$_{Akk}$'. Die Selektionspräferenzen werden durch die 15 Top-Knoten in *GermaNet* (Hamp and Feldweg, 1997; Kunze, 2000) dargestellt: *Lebewesen, Sache, Besitz, Substanz, Nahrung, Mittel, Situation, Zustand, Struktur, Physis, Zeit, Ort, Attribut, Kognitives Objekt, Kognitiver Prozess*. Ich habe die Selektionspräferenzen dadurch definiert, dass ich Frequenzen von nominalen Fillern in bestimmten Argumentpositionen aufwärts durch die GermaNet-Hierarchie propagiert und dann eine Verallgemeinerung auf der höchsten Ebene abgelesen habe. Das Argument in einem Rahmen, das durch Selektionspräferenzen ergänzt wird, ist jeweils unterstrichen. Die Kerninformation bei den Verbbeschreibungen, die Rahmeninformation auf den Ebenen $D1$ und $D2$, ist evaluiert worden. In Schulte im Walde (2002b) habe ich ein Subkategorisierungslexikon für 14.229 Verben aus dem statistischen Grammatikmodell abgeleitet, mit Verbfrequenzen von 1 bis 255.676. Eine Evaluierung des Lexikons in Schulte im Walde (2002a) hat gezeigt, dass die Subkategorisierungsinformation manuelle Definitionen ergänzen und verbessern kann und daher wertvoll für einen Einsatz in der natürlichen Sprachverarbeitung ist.

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

Auf $D1$ sind die wahrscheinlichsten Rahmen für *beginnen* ‘np’ und ‘n’. $D2$ zeigt, dass auch nach dem Verteilen der Rahmenwahrscheinlichkeit für ‘np’ über die verschiedenen Arten von PPs noch eine Reihe von prominenten PPs zu finden ist, temporal um_{Akk}, an_{Dat} und $nach_{Dat}$, die Kennzeichnung eines beginnenden Ereignisses durch mit_{Dat}, lokative PPs mit in_{Dat}. $D2$ macht deutlich, dass nicht nur PPs in Argumentfunktion, sondern auch PPs in Adjunktfunktion eine wichtige Rolle im Verbverhalten darstellen. $D3$ zeigt, dass typische beginnende Ereignisse durch *Situation, Zustand, Zeit, Sache* charakterisiert werden. Außerdem kann man die implizite Definition von Alternationsverhalten erkennen, denn ‘n$\underline{\text{a}}$(Situation)’ in einem transitiven Verbrahmen bezieht sich auf die gleiche Rolle wie ‘$\underline{\text{n}}$(Situation)’ in einem intransitiven Verbrahmen. Das Verb *essen* zeigt starke Präferenzen für sowohl einen transitiven als auch einen intransitiven Rahmen. Wie gewünscht bestimmen *Lebewesen* in ‘$\underline{\text{n}}$’ und ‘$\underline{\text{n}}$a’ sowie *Nahrung* in ‘n$\underline{\text{a}}$’ die Selektionspräferenzen und das Alternationsverhalten. *fahren* taucht mit den typischen Rahmen für Bewegungsverben auf: ‘n’, ‘np’, ‘na’. Die PPs beziehen sich entweder auf eine Richtungsangabe (in_{Akk}, zu_{Dat}, $nach_{Dat}$) oder auf ein Fahrzeug (mit_{Dat}, in_{Dat}, auf_{Dat}). Auch hier weisen die Selektionspräferenzen auf das typische Alternationsverhalten hin: *Lebewesen* in ‘$\underline{\text{n}}$’ und ‘$\underline{\text{n}}$a’, *Sache* in der kausativen Alternation ‘$\underline{\text{n}}$’ und ‘n$\underline{\text{a}}$’.

Algorithmen für eine Clusteranalyse und ihre Evaluierung

Die Clusteranalyse für die deutschen Verben wurde durch den k-Means Algorithmus durchgeführt (Forgy, 1965). k-Means ist ein Standard-Algorithmus im Bereich des Clustering, der iterativ Cluster re-organisiert und Verben ihrem jeweils nächsten Cluster-Mittelpunkt zuweist, bis es keine Änderungen mehr gibt. Eine Anwendung des k-Means Algorithmus kommt den Annahmen gleich, (i) dass die Verben durch distributionelle Vektoren beschrieben werden und (ii) dass Verben, die aufgrund mathematischer Berechnung nicht weit voneinander entfernt sind, auch im linguistischen Sinn einander ähnlich sind.

k-Means erfordert die Definition von einer Reihe von Parametern: Die Anzahl der Cluster in der Clusteranalyse ist nicht gegeben und muss daher experimentell ermittelt werden. In engem Bezug zu diesem Parameter steht die Ebene der konzeptuellen Struktur: Je mehr Cluster in der Clusteranalyse vorhanden sind, desto spezifischer ist die konzeptuelle Ebene. Außerdem kann die Eingabe für den Algorithmus variiert werden: Es können Zufallscluster vorgegeben werden, oder die Eingabe kann durch eine vorhergehende Clusteranalyse vorverarbeitet werden. Die Clustereingabe ist von entscheidender Bedeutung bei k-Means, und daher habe ich sowohl Experimente mit Zufallsclustern als auch mit vorverarbeiteten hierarchischen Clustern durchgeführt. Bei hierarchischen Clustern habe ich außerdem mit verschiedenen Arten der Cluster-Zusammenführung experimentiert. Ein weiterer Parameter bei k-Means ist das Distanzmaß zwischen Verben und Clustern. Welches Maß ist optimal, um Distanzen zwischen Verben zu berechnen? Die Clustering-Methode wurde auf der Basis der Verben in der manuellen Klassifikation erarbeitet und anschließend im Hinblick auf eine große Datenmenge von Verben getestet und diskutiert.

Um die Clusteranalysen zu evaluieren, braucht man ein unabhängiges und zuverlässiges Maß für die Bewertung und den Vergleich von Experimenten und Cluster-Ergebnissen. Theoretisch hat die experimentierende Person ein Gefühl für die Güte der Ergebnisse, aber praktisch gesehen sind die Datenmengen zu groß und die Details in der Datenbeschreibung zu fein für eine objektive Bewertung. Es gibt keine Standard-Maße für die Evaluierung von Clusteranalysen, aber es gibt ähnliche Problemstellungen in anderen Forschungsbereichen wie z.B. der theoretischen Statistik, Bildverarbeitung und Clustering von Webseiten, deren Evaluierungsmethoden sich auf Verbklassen übertragen lassen. Ich habe eine Anzahl von generellen Anforderungen an eine Evaluierung, generellen Anforderungen an eine Clusteranalyse und spezifischen Anforderungen an eine linguistische Clusteranalyse formuliert und verschiedene Evaluierungsmaße im Hinblick auf diese Anforderungen bewertet. Als Ergebnis habe ich drei Evaluierungsmaße für die Experimente ausgewählt: Eine Berechnung von Precision- und Recall-Werten, die auf einer paarweisen Evaluierung von Verben beruht (Hatzivassiloglou and McKeown, 1993) und intuitiv einfach interpretiert werden kann, eine Berechnung von Precision-Werten, die ebenfalls auf einer paarweisen Evaluierung von Verben beruht, aber im Hinblick auf die linguistische Aufgabenstellung durch einen Skalierungsfaktor der Clustergröße optimiert wurde (Schulte im Walde and Brew, 2002) und die Berechnung des angepassten Rand-Index, der die Übereinstimmung und die Unterschiede von Cluster-Zuweisungen bestimmt und einen direkten Bezug zum Nullmodell beinhaltet (Hubert and Arabie, 1985).

<u>Beispiele zu Clusteranalysen</u>

Zu Illustrationszwecken stelle ich Beispiele von Clusteranalysen vor. Die erste Analyse klassifiziert die 168 deutschen Verben der manuell erstellten Klassifikation und beruht auf folgenden Parametern: Die Eingabe ist eine hierarchische Clusteranalyse, bei der die Cluster auf der Basis von *Ward's* Methode zusammengeführt wurden. Die Anzahl der Cluster ist die Anzahl der manuell erstellten Klassen (43), und Distanzmaße wurden durch die Skew-Divergenz, eine Variante der Kullback-Leibler-Divergenz, durchgeführt. Im Folgenden beschreibe ich repräsentative Teile der Clusteranalyse, die auf Verbbeschreibungen der Ebene $D3$ beruht, mit Selektionspräferenzen für die Rollen in '<u>n</u>', '<u>na</u>', '<u>nd</u>', 'n<u>a</u>d', '<u>n</u>s-dass'. Ich vergleiche die ausgewählten Cluster mit den entsprechenden Clustern auf den Ebenen $D1$ und $D2$. In jedem Cluster werden alle Verben, die zu einer gemeinsamen konzeptuellen Verbklasse gehören, in einer Zeile aufgelistet zusammen mit den Klassennamen.

(a) beginnen enden – *Aspekt*
bestehen existieren – *Existenz*
liegen sitzen stehen – *Position*
laufen – *Bewegung: Art*

(b) kriechen rennen – *Bewegung: Art*
eilen – *Bewegung: Eile*
gleiten – *Bewegung: Fließen*
starren – *Gesichtsausdruck*

(c) klettern wandern – *Bewegung: Art*
fahren fliegen segeln – *Bewegung: Fahrzeug*
fließen – *Bewegung: Fließen*

(d) festlegen – *Festlegung*
bilden – *Produktion*
erhöhen senken steigern vergrößern verkleinern – *Maßänderung*

(e) töten – *Eliminierung*
unterrichten – *Lehre*

(f) nieseln regnen schneien – *Wetter*

(g) dämmern – *Wetter*

Die Wetterverben in Cluster (f) zeigen eine starke Übereinstimmung in ihren syntaktischen Rahmen auf $D1$. Sie werden daher schon auf der Basis von $D1$ einem gemeinsamen Cluster zugewiesen und brauchen keine Verfeinerungen auf $D2$ und $D3$. *dämmern* in Cluster (g) ist ein ambiges Verb. Es ist ein Wetterverb und sollte daher mit den Verben in Cluster (f) vorkommen, aber es hat auch eine Bedeutung von Verstehen (*mir dämmert* ...), die sich in einem Rahmen wiederfindet, der *dämmern* von den anderen Wetterverben unterscheidet. Es ist daher weder auf $D1 - D3$ mit den anderen Wetterverben in einem Cluster. Die Verben aus den Klassen *Bewegung, Existenz, Position* und *Aspekt* sind sich auf der Ebene $D1$ sehr ähnlich und können dort nicht unterschieden werden. Auf $D2$ unterscheiden die klassenspezifischen Präpositionen die Verben und erstellen eigene Cluster für die jeweiligen Klassen: Bewegungsverben verwenden typischerweise direktionale PPs, Aspektverben geben das beginnende Ereignis durch mit_{Dat} an oder definieren temporale oder lokative PPs, und Existenz- und Positionsverben verwenden lokative PPs, wobei die Positionsverben eine größere Variation zeigen. Die PP-Information ist relevant für die Unterscheidung der Verben, und auf $D2$ werden die Verben erfolgreich unterschieden. Aber diese Koheränz wird teilweise durch $D3$ zerstört: Die Mehrheit der Bewegungsverben (von den verschiedenen Unterklassen) ist in den Clustern (b) und (c), aufgrund einer starken Ähnlichkeit im Alternationsverhalten. Aber die Existenz-, Positions- und Aspektverben haben idiosynkratische Anforderungen an Selektionspräferenzen, so dass ihre Ähnlichkeit auf $D2$ durch die Verfeinerung auf $D3$ zerstört wird. Sie befinden sich alle in Cluster (a), dem man tatsächlich immer noch eine gewisse Ähnlichkeit nicht verwehren kann, denn alle Verben beschreiben eine Art von Existenz. In Cluster (d) sind fast alle Verben der *Maßänderung* zusammen mit einem Produktionsverb und einem Verb der *Festlegung*. Das Cluster ist also konzeptionell sehr gut. Die Verben in dem Cluster subkategorisieren typischerweise ein direktes Objekt oder 'nr' und 'npr', im letzteren Fall mit den PPs auf_{Akk} und um_{Akk}. Die Selektionspräferenzen sind entscheidend für die Erstellung dieses Clusters: Die Verben stimmen überein in einer Sache oder Situation als Subjekt und verschiedenen Objekten wie z.B. Attributen, kognitiven Objekten, Zuständen, Strukturen oder Dingen. $D1$ und $D2$ können diese Verben nicht unterscheiden. Schließlich gibt es Cluster wie in (e), deren Verben nur auf einer ganz allgemeinen konzeptuellen Ebene übereinstimmen. Bei *töten* und *unterrichten* erfolgt beispielsweise eine Aktion, die von einer Person oder Gruppe auf eine andere Person oder Gruppe gerichtet ist.

Eine zweite Clusteranalyse hat sich derselben Parameter bedient, aber eine größere Anzahl von Verben klassifiziert: Ich habe alle deutschen Verben mit einer empirischen Frequenz zwischen 500 und 10.000 im Trainingskorpus extrahiert. Die Gesamtheit von 809 Verben schließt 94

Verben der manuellen Klassifikation mit ein. Ich habe die fehlenden Verben der manuellen Klassifikation aus Evaluierungsgründen hinzugefügt, so dass sich eine Gesamtzahl von 883 Verben ergeben hat. Die Verben wurden in 100 Cluster eingeteilt, was einer durchschnittlichen Anzahl von 8.83 Verben pro Cluster entspricht. In der Clusteranalyse sind einige semantisch sehr gute Cluster, einige Cluster enthalten semantisch ähnliche und entferntere Verben, und bei einigen Clustern ist es sehr schwierig, überhaupt eine konzeptuelle Ähnlichkeit festzustellen. Für jede Art von Clustern folgen hier einige Beispiele. Verben, die ich innerhalb eines Clusters für sehr ähnlich halte, sind durch Fettdruck markiert.

(a) *anhören*, *auswirken*, *einigen*, *lohnen*, *verhalten*, *wandeln*

(b) *beschleunigen*, ***bilden***, *darstellen*, *decken*, *erfüllen*, ***erhöhen***, *erledigen*, *finanzieren*, *füllen*, *lösen*, *rechtfertigen*, ***reduzieren***, ***senken***, ***steigern***, ***verbessern***, ***vergrößern***, ***verkleinern***, ***verringern***, ***verschieben***, ***verschärfen***, ***verstärken***, ***verändern***

(c) ***ahnen***, ***bedauern***, ***befürchten***, ***bezweifeln***, ***merken***, ***vermuten***, *weißen*, ***wissen***

(d) ***anbieten***, *angebieten*, ***bieten***, ***erlauben***, ***erleichtern***, ***ermöglichen***, ***eröffnen***, ***untersagen***, *veranstalten*, ***verbieten***

(e) ***basieren***, ***beruhen***, ***resultieren***, ***stammen***

(f) ***befragen***, ***entlassen***, ***ermorden***, ***erschießen***, ***festnehmen***, ***töten***, ***verhaften***

Cluster (a) ist ein Beispiel, in dem die Verben keinerlei semantische Ähnlichkeit aufweisen. In der Clusteranalyse sind solche semantisch inkohärenten Cluster typischerweise sehr groß, mit 15-20 Verben pro Cluster. Cluster (b) ist ein Beispiel, in dem ein Teil der Verben eine semantische Ähnlichkeit aufweist, aber auch entferntere Verben enthalten sind. Die Mehrzahl der Verben gehören zur Klasse *Maßänderung*. Cluster (c) bis (f) sind Beispiele für semantisch sehr kohärente Klassen. Cluster (c) enthält Verben, die eine propositionale Einstellung formulieren; die unterstrichenen Verben beschreiben zudem noch eine Emotion. Auch das unmarkierte Verb *weißen* hat eine Berechtigung in dem Cluster, denn es beruht auf einem Fehler in der Morphologie, die Wortformen mit den Lemmata *wissen* und *weißen* nicht immer unterscheiden kann. Cluster (d) beschreibt eine Situation, in der jemand jemandem etwas ermöglicht sowohl im negativen als auch im positiven Sinn. Neben dem auf einem Lemmatisierungsfehler beruhenden *angebieten* (kein Infinitiv, sondern das inkorrekte Partizip Perfekt von *anbieten*) ist der einzige Fehler in dem Cluster das Verb *veranstalten*. In Cluster (e) beziehen sich alle Verben auf eine Basis, und in Cluster (f) beschreibt die Gesamtheit der Verben den Prozess von der Verhaftung eines Verdächtigen bis zu seiner Bestrafung. Die Clusteranalyse könnte nach einer manuellen Korrektur als lexikalische semantische Ressource verwendet werden.

3. Folgerungen

Ich habe eine Methodik für eine Clusteranalyse von deutschen Verben vorgeschlagen, deren Ergebnis mit einer manuellen Klassifikation in großen Teilen übereinstimmt und daher als automatische Basis für eine semantische Klassifikation von Verben verwendet werden kann. Die

Kernaspekte der Methodik beziehen sich einerseits auf die linguistischen und andererseits auf die technischen Aspekte der Aufgabenstellung.

Linguistische Aspekte

Ich habe eine Strategie verwendet, die Verbverhalten durch syntaktische Subkategorisierungseigenschaften, Präpositionen und Selektionspräferenzen beschreibt. Diese Strategie hat sich als sehr erfolgreich erwiesen, da sie den engen Zusammenhang zwischen Verbverhalten und Verbbedeutung darstellt, indem sie auf der Basis der Verbbeschreibungen semantische Klassen erzeugt. Die Clusteranalysen wurden durch jede Ebene der Verbbeschreibung verbessert. Die Auswertungen der Experimente haben die praktischen Grenzen der linguistischen Eigenschaften aufgezeigt: Es gibt eine linguistische Grenze, die durch die Unterscheidung von gemeinsamen und idiosynkratischen Verbeigenschaften definiert wird. Aus theoretischer Sicht ist einleuchtend, dass die Verben in einer gemeinsamen Klasse nur aufgrund ihrer gemeinsamen Eigenschaften in ein Cluster definiert werden können. Wenn aber die Verbbeschreibung schrittweise verfeinert wird und dadurch mehr und mehr idiosynkratische Eigenschaften eingeschlossen werden, schwindet die Gemeinsamkeit der Verben. Aus praktischer Sicht ist es sehr schwierig, die Grenze zwischen gemeinsamen und idiosynkratischen Eigenschaften festzulegen, weil sie von der konzeptuellen Ebene einer semantischen Klasse abhängt und daher von Klasse zu Klasse auf unterschiedlicher Ebene liegen kann. Die herausgearbeitete Definition von Subkategorisierung, Präpositionen und einer Auswahl von Selektionspräferenzen in dieser Arbeit hat sich dabei als ein linguistisch brauchbarer Kompromiss herausgestellt.

Technische Aspekte

Ich habe den Zusammenhang zwischen der grundlegenden Idee eines Clustering-Algorithmus, dessen Parametern und den resultierenden Clusteranalysen im Hinblick auf eine natürlichsprachliche Klassifikation untersucht. Die Eingabe für den Algorithmus spielt eine große Rolle, denn k-Means benötigt für eine linguistisch aussagekräftige Analyse kompakte Cluster mit ähnlicher Größe. Die passenden Analysen durch k-Means werden auf der Basis von denjenigen vorverarbeiteten hierarchischen Clustern erstellt, die wie k-Means eine Minimierung von der Distanz zwischen Verben und Cluster-Mittelpunkten durchführen. *Ward's* Methode hat sich dabei als die beste erwiesen und ist tatsächlich ähnlich erfolgreich auch ohne Nachverarbeitung durch k-Means. Die Ähnlichkeitsmaße von Verben auf der Basis ihrer Darstellung als Vektoren haben nur sekundäre Bedeutung. Erst bei großen Mengen an Objekten zum Clustern oder Eigenschaften für die Objektbeschreibung sind Varianten der Kullback-Leibler-Distanz bevorzugt. Für die distributionelle Beschreibung der Verben eignen sich sowohl Frequenzen als auch Wahrscheinlichkeiten; beide Arten der Darstellung können durch Smoothing noch verbessert werden. Die Anzahl der Cluster in der Clusteranalyse ist nur relevant in Bezug auf die Größenordnung. Eine große Anzahl von Clustern zu verlangen ist sehr ergeizig. Bevorzugt sollten die Verben weniger Clustern mit allgemeinerem Inhalt zugeordnet werden, da das die Fehlerrate senkt und die Cluster eine gute Ebene für manuelle Korrektur darstellen.

Summary

1. Motivation

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 (NLP). But it is tedious and rather impossible to manually define the details of human language. Therefore, 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.

<u>Semantic Verb Classes</u>

Semantic verb classes are an artificial construct of natural language which generalises over verbs according to their semantic properties. They represent a practical means to capture large amounts of verb knowledge without defining the idiosyncratic details for each verb. The class labels refer to the common semantic properties of the verbs in a class at a general conceptual level, and the idiosyncratic lexical semantic properties of the verbs are either added to the class description or left underspecified. Examples for conceptual structures are *Position* verbs such as *liegen* 'to lie', *sitzen* 'to sit', *stehen* 'to stand', and *Manner of Motion with a Vehicle* verbs such as *fahren* 'to drive', *fliegen* 'to fly', *rudern* 'to row'. Semantic verb classes have been defined for several languages, the most dominant examples concerning English (Levin, 1993; Baker *et al.*, 1998) and Spanish (Vázquez *et al.*, 2000). 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.

What is the usage of verb classes in Natural Language Processing applications? On the one hand, verb classes reduce redundancy in verb descriptions, since they encode the common properties of verbs. On the other hand, verb classes can predict and refine properties of a verb that received insufficient empirical evidence, with reference to verbs in the same class: under this aspect, a verb classification is especially useful for the pervasive problem of data sparseness in NLP, where little or no knowledge is provided for rare events. Previous work on verb classes has proven their usefulness: particularly the English verb classification by Levin (1993) has been used for NLP applications such as word sense disambiguation (Dorr and Jones, 1996), machine translation (Dorr, 1997), and document classification (Klavans and Kan, 1998).

Automatic Induction of German Semantic Verb Classes

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.

A common approach to define verb behaviour is captured by the diathesis alternation of verbs. Alternations are alternative constructions at the syntax-semantic interface which express the same or a similar conceptual idea of a verb. In Example (1), the most common alternations for the *Manner of Motion with a Vehicle* verb *fahren* 'to drive' are illustrated. The participants in the conceptual structure are a vehicle, a driver, a driven person, and a direction. In (a), the vehicle is expressed as subject in a transitive verb construction, with a prepositional phrase indicating the direction. In (b), the driver is expressed as subject in a transitive verb construction, with a prepositional phrase indicating the direction. In (c), the driver is expressed as subject in a transitive verb construction, with an accusative noun phrase indicating the vehicle. In (d), the driver is expressed as subject in a ditransitive verb construction, with an accusative noun phrase indicating a driven person, and a prepositional phrase indicating the direction. Even if a certain participant is not realised within an alternation, its contribution might be implicitly defined by the verb. For example, in (a) the driver is not expressed overtly, but we know that there is a driver, and in (b) and (d) the vehicle is not expressed overtly, but we know that there is a vehicle.

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

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

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

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

Assuming that the verb behaviour can be captured by the diathesis alternation of a verb, which are the relevant syntactic and semantic properties for a verb description? The syntactic structures are relevant for the argument functions, 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 have chosen exactly these three feature levels to describe the verbs by their behaviour.

Assuming that we are provided with a feature description for verb behaviour, how can we obtain a semantic verb classification? I have applied a clustering algorithm which uses the syntactico-semantic descriptions of the German verbs as empirical verb properties and learns to induce a semantic classification from this input data. But 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 technique should be based on the linguistic background of the task. Therefore, I have focused on the following sub-goals of the clustering task: I 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 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.

2. Clustering Methodology

The clustering methodology brings together the concept of a German semantic verb classification, empirical data for a verb description at the syntax-semantic interface, a clustering technique, and methods for the evaluation of the clustering experiments. The clustering results are interpreted with respect to the empirical relation between verb meaning and verb behaviour, the development of a methodology for natural language clustering, and the acquisition of semantic verb classes.

<u>German Verb Classes</u>

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. 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*. Because of the meaning-behaviour relationship at the syntax-semantic interface, the verbs grouped in one class show a certain agreement in their behaviour. The class size is between 2 and 7, with an average of 3.9 verbs per class. Eight verbs are ambiguous with respect to class membership. The classes include both high and low frequency verbs: the corpus frequencies of the verbs range from 8 to 71,604. The manual classification represents a gold standard in order to evaluate the reliability and performance of the clustering experiments.

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

Empirical Distributions for German Verbs

I have developed, implemented and trained a statistical grammar model for German which is based on the framework of head-lexicalised probabilistic context-free grammars. The idea originates from Charniak (1995), and this work has used an implementation by Schmid (2000). The statistical grammar model provides empirical lexical information, specialising on but not restricted to the subcategorisation behaviour of verbs.

The German verbs are represented by distributional vectors, with features and feature values in the distribution being acquired from the statistical grammar. The distributional description is based on the hypothesis that 'each language can be described in terms of a distributional structure, i.e. in terms of the occurrence of parts relative to other parts', cf. Harris (1968). The verbs are distributionally described on three levels at the syntax-semantic interface, each of them refining the previous level by additional information. The first level $D1$ encodes a purely syntactic definition of verb subcategorisation, the second level $D2$ encodes a syntactico-semantic definition of subcategorisation with prepositional preferences, and the third level $D3$ 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.

The following table presents three verbs from different verb classes and their ten most frequent frame types with respect to the three levels of verb definition, accompanied by the probability values. The frame types indicate possible arguments in the frames: nominative (n), dative (d) and accusative (a) noun phrases, reflexive pronouns (r), prepositional phrases (p), expletive *es* (x), non-finite clauses (i), finite clauses (s-2 for verb second clauses, s-dass for *dass*-clauses, s-ob for *ob*-clauses, s-w for indirect *wh*-questions), and copula constructions (k). Prepositional phrases in the frames are referred to by case and preposition, such as 'mit$_{Dat}$', and 'für$_{Akk}$'. The selectional preferences in the frames on $D3$ refer to the 15 top level nodes of *GermaNet* (Hamp and Feldweg, 1997; Kunze, 2000): *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'. The preferences have been obtained by frequency propagation through GermaNet on basis of nominal fillers for arguments slots in the grammar. The relevant frame slot for selectional preference refinement is underlined. The core part of the verb description, the subcategorisation frames, has been evaluated on levels $D1$ and $D2$: Schulte im Walde (2002b) describes the induction of a subcategorisation lexicon from the grammar model for a total of 14,229 verbs with a frequency between 1 and 255,676. 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.

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

$D1$ for *beginnen* 'to begin' defines 'np' and 'n' as the most probable frame types. Even by splitting the 'np' probability over the different PP types in $D2$, a number of prominent PPs are left, the time indicating um_{Akk} and nach_{Dat}, mit_{Dat} referring to the begun event, an_{Dat} as date and in_{Dat} as place indicator. It is obvious that not all PPs are argument PPs, but also adjunct PPs represent a part of the verb behaviour. $D3$ illustrates that typical selectional preferences for beginner roles are *Situation, Zustand, Zeit, Sache*. $D3$ has the potential to indicate verb alternation behaviour, e.g. 'na(Situation)' refers to the same role for the direct object in a transitive frame as 'n(Situation)' in an intransitive frame. *essen* 'to eat' as an object drop verb shows strong preferences for both intransitive and transitive usage. As desired, the argument roles are strongly determined by *Lebewesen* for both 'n' and 'na' and *Nahrung* for 'na'. *fahren* 'to drive' chooses typical manner of motion frames ('n', 'np', 'na') with the refining PPs being directional (in_{Akk}, zu_{Dat}, nach_{Dat}) or referring to a means of motion (mit_{Dat}, in_{Dat}, auf_{Dat}). The selectional preferences represent a correct alternation behaviour: *Lebewesen* in the object drop case for 'n' and 'na', *Sache* in the inchoative/causative case for 'n' and 'na'.

Clustering and Evaluation Techniques

The clustering of the German verbs was 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 (i) that verbs are represented by distributional vectors, and (ii) that 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 was varied according to how much pre-processing we invested. k-Means is sensitive to the input, and the resulting cluster shape should match the idea of verb classes. I therefore tried random cluster input and hierarchically pre-processed cluster input (with amalgamations single-linkage, complete-linkage, average distance between verbs, distance between cluster centroids, Ward's method) 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 have been developed with reference to a small-scale German verb classification and discussed and tested on the acquisition of a large-scale German verb classification.

A clustering evaluation demands an independent and reliable measure for the assessment and comparison of clustering experiments and results. In theory, the clustering researcher has acquired an intuition for the clustering evaluation, but in practise the mass of data on the one hand and the subtle details of data representation and clustering algorithms on the other hand make an intuitive judgement impossible. There is no absolute scheme with which to measure clusterings, but a variety of evaluation measures from diverse areas such as theoretical statistics, machine vision and web-page clustering are generally applicable. Based on a series of general evaluation demands, general clustering demands and specific linguistic clustering demands, I compared a number of measures against each other and according to the demands, and determined three measures as the most suitable for the linguistic clustering task: a pair-wise precision and recall measure which has been used in adjective clustering before (Hatzivassiloglou and McKeown, 1993) and provides an easy to understand percentage, an adjusted pair-wise precision measure which introduces a scaling factor based on the size of clusters and comes closest to the linguistic demands on a clustering result (Schulte im Walde and Brew, 2002), and the adjusted Rand index which is a measure of agreement vs. disagreement between object pairs in clusterings that is corrected for chance (Hubert and Arabie, 1985) and provides the most appropriate reference to a null model. The measures compared the results of clustering experiments against the manual verb classification as gold standard.

Clustering Examples

For illustrative purposes, I present representative parts of the cluster analysis as based on the following parameters: the clustering input is obtained from a hierarchical analysis on the German verbs (Ward's amalgamation method), the number of clusters being the number of manual classes (43); similarity measure is performed by the skew divergence, a variant of the Kullback-Leibler divergence. The cluster analysis is based on the verb description on $D3$, with selectional roles for 'n', 'na', 'nd', 'nad', 'ns-dass'. I compare the respective clusters with their pendants under $D1$ and $D2$. For each cluster, the verbs which belong to the same gold standard class are presented in one line, accompanied by the class label.

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

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

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

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

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

(f) nieseln regnen schneien – *Weather*

(g) dämmern – *Weather*

The weather verbs in cluster (f) strongly agree in their syntactic expression on $D1$ and do not need $D2$ or $D3$ refinements for a successful class constitution. *dämmern* in cluster (g) is ambiguous between a weather verb and expressing a sense of understanding; this ambiguity is idiosyncratically expressed in $D1$ frames already, so *dämmern* is never clustered together with the other weather verbs on $D1-3$. *Manner of Motion, Existence, Position* and *Aspect* verbs are similar in their syntactic frame usage and therefore merged together on $D1$, but adding PP information distinguishes the respective verb classes: *Manner of Motion* verbs primarily demand directional PPs, *Aspect* verbs are distinguished by patient mit_{Dat} and time and location prepositions, and *Existence* and *Position* verbs are distinguished by locative prepositions, with *Position* verbs showing more PP variation. The PP information is essential for successfully distinguishing these verb classes, and the coherence is partly destroyed by $D3$: *Manner of Motion* verbs (from the sub-classes *Locomotion, Rotation, Rush, Vehicle, Flotation*) are captured well by clusters (b) and (c), since they inhibit strong common alternations, but cluster (a) merges the *Existence, Position* and *Aspect* verbs, since verb-idiosyncratic demands on selectional roles destroy the $D2$ class demarcation. Admittedly, the verbs in cluster (a) are close in their semantics, with a common sense of (bringing into vs. being in) existence. *laufen* fits into the cluster with its sense of 'to function'. Cluster (d) contains most verbs of *Quantum Change*, together with one verb of *Production* and *Constitution* each. The semantics of the cluster is therefore rather pure. The verbs in the cluster typically subcategorise a direct object, alternating with a reflexive usage, 'nr' and 'npr' with mostly auf_{Akk} and um_{Akk}. The selectional preferences help to distinguish this cluster: the verbs agree in demanding a thing or situation as subject, and various objects such as attribute, cognitive object, state, structure or thing as object. Without selectional preferences (on $D1$ and $D2$), the change of quantum verbs are not found together with the same degree of purity. There are verbs as in cluster (e), whose properties are correctly stated as similar on $D1-3$, so a common cluster is justified; but the verbs only have coarse common meaning components, in this case *töten* and *unterrichten* agree in an action of one person or institution towards another.

The same cluster analysis has been applied in a large-scale experiment: 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 resulted 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), resulting in a total selection of 883 German verbs. The number of clusters was set to 100, which corresponds to an average of 8.83 verbs per cluster. 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 present examples. The verbs which I think semantically similar are marked in bold font.

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

(b) *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'

(c) ***ahnen*** 'to guess', ***<u>bedauern</u>*** 'to regret', ***<u>befürchten</u>*** 'to fear', ***<u>bezweifeln</u>*** 'to doubt', ***merken*** 'to notice', ***vermuten*** 'to assume', *weißen* 'to whiten', ***wissen*** 'to know'

(d) ***anbieten*** 'to offer', *angebieten* is 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'

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

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

Cluster (a) is an example cluster 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. Cluster (b) is an example cluster where a part of the verbs shows overlap in their meaning aspects (quantum change), but the clusters also contain considerable noise. Clusters (c) to (f) are example clusters where most or all verbs show a strong similarity in their conceptual structures: Cluster (c) 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 (d) 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*. In cluster (e) all verbs refer to a basis, and in cluster (f) the verbs describe the process from arresting to treating a suspect. A number of semantically coherent classes needs 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.

3. Conclusions

I have presented a clustering methodology for German verbs whose results agree with the manual classification in many respects and should prove useful as automatic basis for a large-scale clustering. I did not arbitrarily set the parameters, but tried to find an at least near-optimal compromise between linguistic and practical demands. Without any doubt the cluster analysis needs manual correction and completion, but represents a plausible basis. Key issues of the clustering methodology refer to linguistic aspects on the one hand, and to technical aspects on the other hand.

Linguistic Aspects

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

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. The clustering input plays an important role. k-Means needs similarly-sized clusters in order to achieve a linguistically meaningful classification. 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 is of secondary importance, since the differences in clustering with varying the 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 supports 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.

Bibliography

Naoki Abe and Hang Li. Learning Word Association Norms Using Tree Cut Pair Models. In *Proceedings of the 13th International Conference on Machine Learning*, 1996.

Steven Abney. Partial Parsing via Finite-State Cascades. *Natural Language Engineering*, 2(4): 337–344, 1997.

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

Michael R. Anderberg. *Cluster Analysis for Applications*. Academic Press, San Diego, CA, 1973.

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

Thomas T. Ballmer and Waltraud Brennenstuhl. *Deutsche Verben*. Gunter Narr Verlag, Tübingen, 1986.

Leonard E. Baum. An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities*, III:1–8, 1972.

Franz Beil, Glenn Carroll, Detlef Prescher, Stefan Riezler, and Mats Rooth. Inside-Outside Estimation of a Lexicalized PCFG for German. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, 1999.

Avrim Blum and Pat Langley. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence*, pages 245–271, 1997.

Branimir Boguraev, Ted Briscoe, John Carroll, David Carter, and Claire Grover. The Derivation of a Grammatically-Indexed Lexicon from the Longman Dictionary of Contemporary English. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 193–200, Stanford, CA, 1987.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria, 2002.

Michael R. Brent. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19:203–222, 1993.

Chris Brew and Sabine Schulte im Walde. Spectral Clustering for German Verbs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 117–124, Philadelphia, PA, 2002.

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

Carsten Brockmann and Mirella Lapata. Evaluating and Combining Approaches to Selectional Preference Acquisition. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 2003.

Hadumod Bußmann. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart, Germany, 2nd edition, 1990.

Glenn Carroll. Slam Tools Tutorial. Manuscript, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1997.

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

Eugene Charniak. Parsing with Context-Free Grammars and Word Statistics. Technical Report CS-95-28, Department of Computer Science, Brown University, Providence, Rhode Island, 1995.

Eugene Charniak. A Maximum-Entropy-Inspired Parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, Washington, 2000.

Stanley Chen and Joshua Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Center for Research in Computing Technology, Harvard University, 1998.

Stephen Clark and David Weir. Class-Based Probability Estimation using a Semantic Hierarchy. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Germany, 2000.

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

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:154–163, 1960.

Michael Collins. Three Generative, Lexicalised Models for Statistical Parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, 1997.

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

Ido Dagan, Lillian Lee, and Fernando Pereira. Similarity-Based Models of Word Cooccurrence Probabilities. *Machine Learning*, 34(1-3), 1999. Special Issue on Natural Language Learning.

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

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

David R. Dowty. *Word Meaning and Montague Grammar*. Reidel, Dordrecht, 1979.

Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2000.

Dudenredaktion, editor. *DUDEN – Das Stilwörterbuch*. Number 2 in ‘Duden in zwölf Bänden’. Dudenverlag, Mannheim, 8th edition, 2001.

Judith Eckle. *Linguistic Knowledge for Automatic Lexicon Acquisition from German Text Corpora*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1999.

Jason Eisner and Giorgio Satta. Efficient Parsing for Bilexical Context-Free Grammars and Head Automaton Grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

Stefan Engelberg. Verb Meaning as Event Structure. In Alan K. Melby and Arle R. Lommel, editors, *The Lexicon*, LACUS Forum XXVI, Fullerton, CA, 2000a.

Stefan Engelberg. *Verben, Ereignisse und das Lexikon*. Number 414 in Linguistische Arbeiten. Max Niemeyer Verlag, Tübingen, 2000b. Doctoral Dissertation.

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

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

Ana Fernández, María Antonia Martí, Gloria Vázquez, and Irene Castellón. On the Concept of Diathesis Alternations as Semantic Oppositions. In *Proceedings of the ACL-SIGLEX Workshop on Standardizing Lexical Resources*, Maryland, MD, 1999.

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

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

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

Edward B. Fowlkes and Collin L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78:553–569, 1983.

Pablo Gamallo, Alexandre Agustini, and José Gabriel Pereira Lopes. The Role of Co-Specification for the Acquisition of Selection Restrictions from Unsupervised Corpora. In *Proceedings of Journee Applications, Apprentissage et Acquisition de Connaissance a partir de Textes Electroniques*, Grenoble, France, 2001.

Ralph Grishman, Catherine Macleod, and Adam Meyers. COMLEX Syntax: Building a Computational Lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 268–272, Kyoto, Japan, 1994.

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

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

Vasileios Hatzivassiloglou and Kathleen R. McKeown. Towards the Automatic Identificaton of Adjectival Scales: Clustering Adjectives According to Meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 172–182, Columbus, Ohio, 1993.

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

Gerhard Helbig and Wolfgang Schenkel. *Wörterbuch zur Valenz und Distribution deutscher Verben*. Max Niemeyer Verlag, Tübingen, 1969.

John A. Hertz, Anders Krogh, and Richard G. Palmer. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Boston, MA, 1991.

Donald Hindle. Noun Classification from Predicate-Argument Structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, 1990.

John E. Hopcroft and Jeffrey Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Boston, MA, 1979.

Albert S. Hornby. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, 4th edition, 1985.

Frank Höppner, Frank Klawonn, and Rudolf Kruse. *Fuzzy-Clusteranalyse*. Vieweg, Braunschweig, 1997.

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

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

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

Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3), 1999.

Raymond A. Jarvis and Edward A. Patrick. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Transactions on Computers*, 22(11):1025–1034, 1973.

Eric Joanis. Automatic Verb Classification using a General Feature Space. Master's thesis, Department of Computer Science, University of Toronto, 2002.

George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant Features and the Subset Selection Problem. In *Proceedings of the 11th International Conference on Machine Learning*, pages 121–129, San Francisco, CA, 1994.

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

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

Maurice G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30:81–93, 1993.

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

Ron Kohavi and George John. The Wrapper Approach. In Huan Liu and Hiroshi Motoda, editors, *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Kluwer Academic Publishers, Dordrecht, 1998.

Daphne Koller and Mehran Sahami. Toward Optimal Feature Selection. In *Proceedings of the 13th International Conference on Machine Learning*, pages 1316–1321, Nagoya, Japan, 1997.

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

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

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

Pat Langley. Selection of Relevant Features in Machine Learning. In *Proceedings of the AAAI Fall Symposium on Relevance*, New Orleans, LA, 1994.

Maria Lapata. Acquiring Lexical Generalizations from Corpora: A Case Study for Diathesis Alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 397–404, 1999.

Maria Lapata and Chris Brew. Using Subcategorization to Resolve Verb Class Ambiguity. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Maryland, MD, 1999.

Karim Lari and Steve J. Young. The Estimation of Stochastic Context-Free Grammars using the Inside-Outside Algorithm. *Computer Speech and Language*, 4:35–56, 1990.

Lillian Lee. *Similarity-Based Approaches to Natural Language Processing*. PhD thesis, Harvard University, Cambridge, MA, 1997.

Lillian Lee. On the Effectiveness of the Skew Divergence for Statistical Language Analysis. *Artificial Intelligence and Statistics*, pages 65–72, 2001.

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

Wolfgang Lezius, Stefanie Dipper, and Arne Fitschen. IMSLex – Representing Morphological and Syntactical Information in a Relational Database. In *Proceedings of the 9th EURALEX International Congress*, pages 133–139, Stuttgart, Germany, 2000.

Wolfgang Lezius, Arne Fitschen, and Ulrich Heid. Datenbankbasierte Pflege und Verwaltung morphologischer Information im IMSLex. In Jost Gippert and Peter Olivier, editors, *Multilinguale Corpora – Codierung, Strukturierung, Analyse*, pages 122–131. Prag, 1999.

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

David M. Magerman. Statistical Decision-Tree Models for Parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, 1995.

Christopher D. Manning. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242, 1993.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

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

Paola Merlo and Suzanne Stevenson. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3):373–408, 2001.

Paola Merlo, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. A Multilingual Paradigm for Automatic Verb Classification. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, 2002.

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

Hermann Ney, Ute Essen, and Reinhard Kneser. On Structuring Probabilistic Dependences in Stochastic Language Modelling. *Computer, Speech and Language*, 8:1–38, 1994.

Akira Oishi and Yuji Matsumoto. Detecting the Organization of Semantic Subclasses of Japanese Verbs. *International Journal of Corpus Linguistics*, 2(1):65–89, 1997.

Barbara H. Partee, Alice ter Meulen, and Robert E. Wall. *Mathematical Methods in Linguistics*. Kluwer Academic Publishers, Dordrecht, 1993.

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

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

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

James Pustejovsky. The Syntax of Event Structure. *Cognition*, 41:47–81, 1991.

James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995.

William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

Malka Rappaport Hovav and Beth Levin. Building Verb Meanings. In Miriam Butt and Wilhelm Geuder, editors, *The Projection of Arguments – Lexical and Compositional Factors*, pages 97–134. CSLI Publications, Stanford, CA, 1998.

Philip Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, 1993.

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

Francesc Ribas. An Experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 769–774, 1994.

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

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. Inducing a Semantically Annotated Lexicon via EM-Based Clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, MD, 1999.

Patrick Saint-Dizier. Constructing Verb Semantic Classes for French: Methods and Evaluation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 1127–1130, Copenhagen, Denmark, 1996.

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

Helmut Schmid. *YAP: Parsing and Disambiguation with Feature-Based Grammars*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1999.

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

Carson T. Schütze. PP Attachment and Argumenthood. In Carson Schütze, Jennifer Ganger, and Kevin Broihier, editors, *Papers on Language Processing and Acquisition*, number 26 in MIT Working Papers in Linguistics, pages 95–152. Cambridge, MA, 1995.

Sabine Schulte im Walde. Clustering Verbs Semantically According to their Alternation Behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 747–753, Saarbrücken, Germany, 2000a.

Sabine Schulte im Walde. The German Statistical Grammar Model: Development, Training and Linguistic Exploitation. Arbeitspapiere des Sonderforschungsbereichs 340 *Linguistic Theory and the Foundations of Computational Linguistics* 162, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2000b.

Sabine Schulte im Walde. Evaluating Verb Subcategorisation Frames learned by a German Statistical Grammar against Manual Definitions in the *Duden* Dictionary. In *Proceedings of the 10th EURALEX International Congress*, pages 187–197, Copenhagen, Denmark, 2002a.

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

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

Sabine Schulte im Walde. Experiments on the Choice of Features for Learning Verb Classes. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 2003b.

Sabine Schulte im Walde and Chris Brew. Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 223–230, Philadelphia, PA, 2002.

Sabine Schulte im Walde, Helmut Schmid, Mats Rooth, Stefan Riezler, and Detlef Prescher. Statistical Grammar Models and Lexicon Acquisition. In Christian Rohrer, Antje Rossdeutscher, and Hans Kamp, editors, *Linguistic Form and its Computation*. CSLI Publications, Stanford, CA, 2001.

Helmut Schumacher. *Verben in Feldern*. de Gruyter, Berlin, 1986.

Shokri Z. Selim and M.A. Ismail. K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(1):81–87, 1984.

Eric V. Siegel and Kathleen R. McKeown. Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights. *Computational Linguistics*, 26(4):595–628, 2000.

Detlef Steinhausen and Klaus Langer. *Clusteranalyse – Einführung in Methoden und Verfahren der automatischen Klassifikation*. de Gruyter, Berlin, 1977.

Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of Similarity Measures on Web-page Clustering. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI 2000): Workshop of Artificial Intelligence for Web Search*, Austin, Texas, 2000.

Leonard Talmy. Lexicalization Patterns: Semantic Structure in Lexical Forms. In Timothy Shopen, editor, *Language Typology and Syntactic Description*, volume III, pages 57–149. Cambridge University Press, 1985.

Vivian Tsang, Suzanne Stevenson, and Paola Merlo. Crosslinguistic Transfer in Automatic Verb Classification. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, 2002.

Zeno Vendler. *Linguistics in Philosophy*, chapter 'Verbs and Time', pages 97–121. Cornell University Press, Ithaca, 1967.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Message Understanding Conference*, pages 45–52, San Francisco, 1995.

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

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

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

Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58:236–244, 1963.

Oliver Wauschkuhn. *Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora*. PhD thesis, Institut für Informatik, Universität Stuttgart, 1999.

Lofti A. Zadeh. Fuzzy Sets. *Information and Control*, 8:338–353, 1965.

Charles T. Zahn. Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers*, 20:68–86, 1971.

Heike Zinsmeister and Ulrich Heid. Collocations of Complex Words: Implications for the Acquisition with a Stochastic Grammar. In *International Workshop on 'Computational Approaches to Collocations'*, Vienna, Austria, 2002.

Heike Zinsmeister and Ulrich Heid. Identifying Predicatively used Adverbs by Means of a Statistical Grammar Model. In *Proceedings of the Corpus Linguistics 2003 Conference*, number 16 in UCREL Technical Papers, Lancaster, UK, 2003a.

Heike Zinsmeister and Ulrich Heid. Significant Triples: Adjective+Noun+Verb Combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research*, Budapest, Hungary, 2003b.