German Clause-Embedding Predicates: an Extraction and Classification Approach

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

Vorgelegt von

Ekaterina Lapshinova-Koltunski

aus Wolgograd

Hauptberichter:	apl. Prof. Dr. phil. habil. Ulrich Heid
Mitberichter:	Prof. Dr. phil. habil. Grzegorz Dogil
Tag der mündlichen Prüfung:	9. Februar 2010

Institut für maschinelle Sprachverarbeitung
Universität Stuttgart

2009

Acknowledgements

The work on this thesis was performed within the PhD program 'Graduiertenkolleg-609' funded by DFG. I was supervised by Ulirich Heid at the Institut für Maschinelle Sprachverarbeitung of the Universität Stuttgart. I am very grateful to him for his interest in my work, his motivation and feedback, and especially for the fact that he could always find time for me, although it was sometimes difficult due to his full-scale schedule at both universities and his activities in research projects. I would like to thank him for giving me the opportunity to develop my own ideas: without him I would have been just rambling from one topic to another all the time.

I am grateful to all of the 'Graduiertenkolleg-609': our professors - Grzegorz Dogil, Jürgen Pafel, Hans Kamp, Klaus von Heusinger, Hinrich Schütze, Achim Stein, Bernd Möbius; as well as my fellow PhD and our students - Olga Anufryk, Natalie Lewandowska, Klaus Rothenhäusler, Dennis Spohr, Henrike Baumotte, Charles Yee, Uta Benner, Simone Heinold, Manuela Korth, Petra Augurzky, Zorica Trpcevska, Regine Brandtner, Dolgor Guntsetseg, Melanie Uth and Manuel Kuntz. Their helpful comments and suggestions gave me a lot of ideas during our annual 'Klausurtagungen' in Kleinwalsertal, Austria, and the entertaining trekking tours in the Alpes brought us all together as friends.

I also thank the administrative staff of the 'Graduiertenkolleg-609', namely, Eva Schmid and Christian Bär, who helped me to avoid a lot of the bureaucratic issues I would have to go through otherwise. The financial support provided by the 'Graduiertenkolleg-609' made my participation in international conferences and this whole research project possible.

I would also like to thank professors and the administrative staff at the IMS, who made sure that my working environment was both efficient and comfortable. I am especially obliged to Antje Roßdeutscher and Kurt Eberle for their advice in semantic issues and to Gertrud Faaß, Sabine Schulte im Walde, Klaus Rothenhäusler, André Blessing, Julia Weidenkaff and Fabienne Fritzinger for their help with technical problems concerning CQP and UNIX. My gratitude is due to those who provided me with comments on the draft of this thesis. I am grateful to all my colleagues at the IMS for creating a friendly and stimulating atmosphere. Special thanks go to Olga Anufryk, Natalie Lewandowska, Klaus Rothenhäusler and Dennis Spohr, who helped me to forget about work for at least half an hour per day, enjoying our lunch or coffee inside or outside the IMS. I also thank Olga Anufryk, Natalie Lewandowska and my both office mates, Hannah Kermes and Oksana Stus, for discussing not only matters concerning work but also problems concerning social issues.

Finally, I would like to thank my family and my friends for all their love and support. I am so lucky to have a lot of friends in Russia, Germany and, actually, all over the world. Our communication helped me greatly, be it sometimes only through

the internet. Another round of special thanks go to my train mates, Isabelle, Ann and Ulrike, who made my almost daily train rides between Freiburg and Stuttgart funny and less stressfull. I am grateful to all my family members, especially my grandma and my late grandpa, who would be happy to share share my Rigorosum with me. I would like to thank my aunt, uncle and cousins, especially Natasha, with whom I chatted almost every day during the last months of my work over the dissertation and who helped me to switch off from the daily routine. Huge thanks to both of my parents, Irina and Sergey, my parents-in-law, Irena and Gerd, as well as Anna and Christian, for their emotional support. Special thanks to my parents for letting me go abroad and live thousands of kilometres away from them. I know they miss me just as much as I miss them. And, last but not least, major thanks to my husband Krzysiek for his patience with me during the countless days that I spent in front of my laptop instead of spending them with him. He never stopped believing in me even if I stopped doing so. I am happy to have him on my side and share the burden and the fun of my work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction: Aims and Motivation

This thesis is concerned with experiments on the automatic extraction and classification of predicates in German. Analysing subcategorisation properties of German verbs, nouns and multiword expressions, as well as their relations, we focus on a number of questions. How can data about subcategorisation properties be extracted from text corpora? How can lexical items be classified according to their subcategorisation properties? Do nominalisations, compound nouns and multiwords have their own subcategorisation properties; which of their properties are inherited from their base lexical units? These questions outline our main aims, which are described in this introductory chapter. Moreover, we are going to present the reasons that were responsible for us to address the above mentioned questions at all.

Aims of the Present Reasearch

The purpose of the present research is to extract and classify German verbs, nouns and multiwords automatically according to their subcategorisation properties. Besides that, we aim at comparing valency properties of morphologically related predicates, such as verbs and their derivatives; therefore, we analyse the phenomenon of "inheritance" in subcategorisation (for instance in the case of deverbal nouns which share their subcategorisation properties with the underlying verbs).

The aims of the present thesis include several aspects, which can be divided into three parts.

Extraction of predicates along with their subcategorisation information We analyse subcategorisation properties of lexical units of German by means of extracting evidence for subcategorisation from text corpora. For this purpose, we elaborate an extraction architecture based on available linguistic (lexical, grammatical) knowledge about the phenomena we extract. We extract verbal, nominal and multiword predicates from text corpora along with their subcategorisation information. The lexical data are created to serve symbolic NLP, especially large symbolic grammars for deep processing, such as HPSG[1] or LFG[2]. For these, detailed linguistic knowledge about lexical data is necessary.

[1] Cf. work in the LinGO project (Copestake et al. 2004).

[2] Cf. the PARGRAM project (Butt et al. 2002).

In this thesis we concentrate on sentential complements only, although our methods can be also applied to the extraction of other complement types. The choice of this complement type is caused by our aim to compare subcategorisation properties of morphologically related words. Sentential complements are allowed by all predicates under analysis – verbal, nominal and multiword ones.

Classification of predicates based on their subcategorisation information We classify the automatically extracted data according to the subcategorisation information extracted along with the predicates. A classification based on the ability of predicates to allow for a certain type of complements enables a systematic description of the subcategorisation behaviour of predicates, as well as of their morpho-syntactic preferences (selectional restrictions of predicates determine their subcategorisation behaviour).

The analysis of "inheritance" relations between related predicates We compare subcategorisation properties of verbs and their nominalisations (the ones occurring freely in corpora, as well as those occurring within a multiword) and analyse the phenomena of "inheritance" and "non-inheritance" in subcategorisation. We analyse these phenomena also for multiword expressions and compound nouns, in terms of their relations with their constituents. Besides that, we automatically classify "inheritance" relations based on the comparison of subcategorisation properties of verbs and their derivatives.

Motivation for the present Research

The reasons that motivate us to address the problems mentioned above originate from both, a linguistic and a NLP background. Thus, the current section describes the importance of our research from the point of view of linguistics (including lexicography) and NLP (including formal grammars and NLP applications).

Linguistic Background

Contribution to linguistics One of the main reasons for our analysis of subcategorisation properties of German predicates is the important role of subcategorisation information for not only sentence structure, but also semantic and pragmatic levels of language description, cf. (Helbig 1992) and (Fischer 1999). Besides that, computational experiments show that surface syntactic and semantic indicators can help us to understand verb semantics, cf. (Schulte im Walde 2000), (Merlo/Stevenson 2001). The description of the obtained subcategorisation properties is necessary for different fields of linguistics. Automatic acquisition of predicates along with their subcategorisation properties provides linguists with up-to-date information about the language. The automatic classification of predicates according to their valency behaviour helps to analyse systematically such phenomena as selectional restrictions of lexical units, which depend on the semantics or contextual parameters of the predicates. Moreover, the automatic analysis of relations between morphologically related predicates

allows us to explain the presence or absence of certain morpho-syntactic features, for example verbal valency features in the respective derived nouns.

Contribution to lexicography Subcategorisation information is not detailed enough in most general dictionaries and, apart from that, dictionaries that list subcategorisation frames often list expected patterns, rather than actual observed ones. Hence, the evidence from corpora obtained with our extraction and classification architecture is necessary for lexicologists and lexicographers who need access to this information. Furthermore, to our knowledge most lexicographic studies on subcategorisation concentrate exclusively on verbal predicates, cf. section 3.1 below. In this thesis we focus also on further predicate types, analysing nouns and multiword expressions, cf. sections 3.2 and 3.3. Additionally, just a few authors, e.g. (Sommerfeldt/Schreiber 1996), consider relations which exist between morphologically related predicates, for instance correspondences between verbal subcategorisation properites and those of their nominalisations. These phenomena have not received enough attention so far.

Contribution to language learning and multilingual processing Our extraction and classification system can also find application in learning and teaching of German as a foreign language. Information on subcategorisation properties of German verbs, nouns and multiword expressions is important for learners of German as a foreign language. Syntactic and semantic features of predicates allow learners to induce their meaning. The knowledge about valency properties of German verbs, nouns and multiword expressions helps learners to produce correct sentences in German. Besides that, the output of our extraction and classification tool provides German teachers with information on regularities and exceptions in the behaviour of German words, which can help them to explain these phenomena systematically.

In multilingual processing, for instance in translation, subcategorisation properties also provide information on sentence structures and morpho-syntactic preferences of words. In the process of translating from German, the data obtained with our tool allows to understand the meaning of predicates. For the translation into German, our tool delivers the necessary information for the production of correct sentences in German.

Natural Language Processing

A dictionary containing structured lexical data with complex information is one of the most important components of many NLP systems.

Manual creation of lexical resources, containing such complex information as predicate-argument structures, is difficult and takes much time and effort. Moreover, manually acquired lexicons usually contain a great number of inaccuracies. Lexical information, automatically retrieved with acquisition tools, can be stored in machine-readable lexicons and updated dynamically, cf. (Schulte im Walde 2006). Besides that, the automatic extraction of subcategorisation data provides statistical information about the behaviour of predicates (like their occurrence with different complement types). This information is important for most NLP applications and

cannot be produced manually for large sets of data. Inaccuracies in manually created lexicons cause problems to symbolic parsing systems.

This calls for a semi-automatic approach to corpus-based lexicon acquisition. Therefore, we decide for a (semi)-automatic precision-oriented approach in extracting and classifying predicates. We apply the term 'semi-automatic', as for the extraction of some predicate types (for example multiwords), we need to use manual sorting procedures which are necessary to avoid noise or to resolve ambiguities, e.g. we manually sort out certain noise-bearers or select a proper candidate if the tools propose several of them.

To contribute to the reduction of manual effort invested into sorting of noisy cases, we focus on high accuracy of our extraction and classification results. High precision is opposed to completeness, compensated by the application of extraction procedures on larger corpora. The obtained information on subcategorisation properties of verbs, nouns and multiword expressions is important for different areas of NLP. In the following we briefly describe the possible contribution of the extracted information for formal grammars, computational lexicons and parsers, as well as Information Extraction (IE).

Role of subcategorisation information for symbolic grammars and parsers Subcategorisation information is contained in most formal grammars. For instance, both HPSG and LFG include information on predicate-argument structures in their lexicons, cf. 3.1.2.1 below. The inaccuracies, which might occur in manually created lexicons usually cause errors in the process of syntactic analysis of sentences. Therefore, detailed linguistic knowledge about lexical data for symbolic grammars should be created (semi-)automatically. A syntactic parser needs information about the number and the nature of arguments of predicates, accordingly a subcategorisation lexicon is a key component for most syntactic parsers, cf. (Poibeau/Messiant 2008). Moreover, such lexicons can also be used to enhance semantic classification of predicates, cf. (Schulte im Walde 2002).

Predicate-argument structure for IE Subcategorisation information is also used for IE purposes. For instance, (Surdeanu et al. 2003) describe an approach to the application of predicate-argument structure data for Information Extraction. The authors claim that some of the most successful IE techniques are built around a set of domain-relevant linguistic patterns based on selected predicates. These patterns are matched against documents for identifying and extracting domain-relevant information. Such patterns can be created either manually or automatically. As already mentioned, automatically acquired data is more accurate, which increases the efficiency of IE systems.

Beside this point, the automatic analysis of the relations between verbs and their derivatives allows to reduce the number of patterns applied, as some derived nouns completely inherit their subcategorisation properties from verbs.

Overview of Chapters

The chapters of this thesis are organised as follows.

Chapter 2 describes main categories of the phenomenon of valency or subcategorisation. In the first part, in section 2.1, we review the relevant literature (both linguistic and NLP) on the description of subcategorisation. Analysing different approaches to the notion of valency and the description of its main types of information, we define the categories that are of particular importance for our thesis. The second part of chapter 2 (section 2.2) focuses on the analysis of syntactic and semantic features of sentential complements, as we concentrate on the extraction and classification of predicates subcategorising for subclauses. We summarise different studies on sentential complements and define their features, which are important for the analysis of the valency behaviour of clause-embedding predicates.

Chapter 3 describes subcategorisation properties of the predicates under analysis – verbs, nouns and multiwords. We analyse the related literature on verbal, nominal and multiword predicates (in sections 3.1, 3.2 and 3.3) according to the valency categories defined in chapter 2. In section 3.4, we analyse the phenomenon of "inheritance" in subcategorisation, which is observed both in the valency of multiwords and compound nouns (described in section 3.4.1), and in the valency of deverbal nouns, described in section 3.4.2. For multiword expressions we analyse the relations between subcategorisation properties of the whole construction and those of its nominal constituents, cf. section 3.4.1.1. For compound nouns, we compare subcategorisation properties of their head and non-head constituents and address the problem of their valency bearers, cf. section 3.4.1.2. "Inheritance" in nominalisations is presented by the valency properties inherited from verbs. We analyse both, cases where valency properties of nominalisations correspond to those of their base verbs, and cases where they differ from the valency of their underlying verbs.

Chapter 4 presents approaches on classification of the predicates described in chapter 3. We analyse the related work on the classification of verbs (in section 4.2.1), nouns (in section 4.2.2.5) and multiwords (in section 4.2.4.4), which can be grouped according to various criteria. In sections 4.2.1.3, 4.2.2.5, 4.2.3.3 and 4.2.4.4, we present different classifications of verbs, nouns, nominal compounds and multiwords based on their subcategorisation features.

Chapter 5 describes the extraction and classification architecture elaborated within the present thesis. We describe the input corpora used in this thesis along with the set of pre-processing procedures applied on these corpora, cf. section 5.1. We then specify the extraction context, which is determined by both, the German word order and our aim to elaborate precision-oriented procedures, cf. section 5.2. The procedures to extract verbal, nominal and multiword predicates and to classify them according to their subcategorisation properties are presented in section 5.3.

Chapter 6 presents the results of extraction and classification performed by our procedures. Section 6.1 describes sample extraction results and their interpretation. In section 6.2, we evaluate the extraction and classification procedures presented in chapter 5 above. We evaluate the procedures of our architecture calculating precision

and recall obtained on the extracted and classified data. We also discuss possible improvements of the procedures, which can increase the accuracy of our results.

Chapter 7 discusses the contribution of this thesis and suggests directions for future work. The main aspects of the contribution include the automatic extraction of valency data for different types of predicates, their automatic classification according to the subcategorisation properties, as well as the automatic analysis of "inheritance" relations between morphologically related predicates.

Appendix We attach examples of the extracted and classified data in the appendix[3]. It contains also the description of annotations applied on the used corpora.

[3]Further lists of extracted data will be available at a site on the computing system of IMS (Universität Stuttgart) after this thesis has been reviewed.

Chapter 2

State of the Art on Subcategorisation in Linguistics and NLP

To establish a theoretical background to this thesis, we analyse a number of studies on different aspects of subcategorisation (or valency) research in linguistics, lexicography and NLP. In this chapter we summarise the main categories, which are related to the notion of valency. As we concentrate in this study on the analysis of predicates that subcategorise for sentential complements, we describe their features in detail. Studies on subcategorisation, which describe the phenomena, we deal with in the present research.

In section 2.1, we recall the main categories of the subcategorisation phenomena, look at the development of the notion in linguistic literature, which is described in section 2.1.1 of this chapter. We go on with the description of different levels and types of valency, as well as its main categories in sections 2.1.2 and 2.1.3. Afterwards, in section 2.2, we concentrate on the description of sentential complements.

2.1 The Notion and its Basic Categories

Subcategorisation or *valency* is often seen as the property of lexical units to determine the occurrence of other elements in a sentence[1]. In other words, valency is the number of elements a word can take. These phenomena are described (depending on the theoretical framework and terminology) as semantic cases, theta-roles, arguments or complements. According to (Herbst 2007), subcategorisation can be represented by complement inventories, which are lists of complements with which a lexical unit can occur, or by valency patterns in which a lexical unit can occur. Both frameworks are compatible with each other.

General and computational linguistics employ different terms for the phenomenon of subcategorisation. For instance, it is called *government* or *Rektion* in traditional grammar, or *complementation* in descriptive grammar. In this thesis, we call lexical units which determine the occurrence of other elements in a sentence, *predicates*, the elements, which are taken or determined by the lexical units – *arguments* or *complements* of the predicates, and the abilty of the lexical units to take a certain number of elements – their *valency*, *subcategorisation* (which is often used in NLP

[1] Cf. (Herbst 2007).

work and comes from generative framework) or *predicate-argument structure*. That means that we use the term valency in a relatively wide sense, which comprises both, the number of elements a word can take and their realisations.

2.1.1 Development of the Notion

In the present section, we briefly recall the most important works on subcategorisation (especially for the description of valency for German) starting from the 1950s.

Most authors state that the notion of valency was first applied in the dependency grammar of Lucien Tesnière (developed in the 50s), although similar concepts were put forward even earlier, in the 30s-40s. Tesnière's model differs from the modern approaches in being verb-concentrated. It is generally accepted that valency theory in linguistics appears within dependency grammar. The main notion for Tesnière is that of *dependency* between elements of the sentence. The term *valency* is used by Tesnière in his work *El'ements de syntaxe structurale* (1959), and the author borrows its meaning from the definition of valency in chemistry (*"number of actants"*). Language elements (at that time verbs) are compared to chemical elements, which have a capacity to be combined with a fixed number of atoms of another element. The etymology of the word "valence" derives from the 15th century, from Latin *valentia* "strength, capacity", and means "extract, preparation", and the chemical meaning referring to the definition "combining power of an element" is recorded from 1884, from German *Valenz*. Tesnière applies this chemical concept for the description of French verbs. He establishes the base for the linguistic theory of valency in dependency grammar, such as the distinction between complements and adjuncts (see section 2.1.3.2), characteristics of verbs as avalent, monovalent, etc. (see section 2.1.3.1), as well as syntactic and semantic functions of complements (see sections 2.1.4 and 2.1.5).

The concept was taken over by many European linguists. However, most of the work on valency is applied for the teaching of German as a foreign language, for the linguistic description of German and was later on further developed for other languages, as for instance for English (see (Emons 1974), (Emons 1978) and (Matthews 1981)) or for French. A great contribution to the description of valency in English is done in (Herbst 1999), (Herbst 2004) or (Herbst/Götz-Votteler 2007) which are the latest of his umerous studies on valency later on. German grammarians and linguists started using the concept of valency formulated by Tesnière already in 70s, e.g. (Heringer 1970), (Helbig 1971), (Brinkmann 1971), (Erben 1972) and (Engel 1977).

The early development of valency theory is still closely related to that of dependency grammar and other syntactic theories with a dependency component (e.g. in (Engel 1977), (Heringer 1970), (Matthews 1981) or (Mel'čuk 1988)). However, the main impact on the development of this theory as such comes to a greater degree from foreign language teaching and the creation of valency dictionaries for language learners, and not from the work within dependency grammar. Since 1969, as Helbig and Schenkel's *Wörterbuch zur Valenz und Distribution deutscher Verben* was published, a number of dictionaries have appeared describing valency for French, German, Italian, Spanish, Latin and Japanese. Some of these dictionaries are intended for linguistic research, others are developed to be used by foreign language learn-

ers, e.g. (Herbst et al. 2004). We outline dictionaries describing subcategorisation properties for English, German and some other languages in section 3.1.1 below.

Helbig also uses the concept of valency and states conditions for a clear and formalised valency concept. The most important one is that the finite verb should be regarded as the structural center of a sentence. esides that, he defines valency as the ability of a verb to fill argument places with complements, whereas both, the number and the kind of the actants is important, cf. (Helbig/Schenkel 1973).

However, most works on valency concentrate on the description of verbal predicates. A few authors only, e.g. (Grebe 1973) or (Herbst 1983), mention subcategorisation properties of other predicates, i.e. nouns, adjectives and adverbs.

For the description of subcategorisation of predicates in German, a significant contribution was done by the linguistic works in the 70's mentioned above. Besides that, in this thesis we refer to the theoretical works, such as (Engel 1988) and (Engel 1991), (Helbig 1992), (Heringer 1996), and the lexicographic works, such as dictionaries of Helbig and Schumacher[2], as well as those of Sommerfeldt and Schreiber[3]. One of the latter works, which is also significant in the decription of subcategorisation background is *Valenz und Dependenz. An International Hanbook of Contemporary Research* by Agel[4], who summarises works on major problems of valency theory.

In the description of subcategorisation in NLP, there exist numerous works on different kinds of predicates, which we refer to in the following sections of the thesis. Most important theoretical issues on valency in NLP are described by H. Somers in (Somers 1987).

2.1.2 Syntactic and Semantic Levels of Valency

Linguists have been discussing the problem of valency description levels since the first works on valency appeared. Is valency a formal phenomenon on the phrase level, is it a conceptual phenomenon on the content level or does it appear first at the communicative level?

For the most part grammarians and linguists call valency a theory of syntax. In this case it describes the valency patterns of predicates as their syntactic patterns into which they enter. At the same time there exist semantic approaches, which give semantic characteristics of the arguments in a given predicate.

The two levels of description mentioned above do not exclude each other and are not isomorphic. This can be illustrated with the German verb *erfahren* ("to find out"), which has three argument places to fill on the logical-semantic level : *erfahren(x,y,z)*. *Experiencer, Source* and *Patient* are the roles which can fill these argument places. These three potential arguments are realised through subject, accusative and prepositional objects on the syntactic level and are differentiated as obligatory or optional complements (cf. examples in figure 2.1).

The realisation of words and their ability to be combined with other words in a sentence are described on the syntactic level. Syntactic valency defines constituents and sentence functions: it determines the number and the kind of arguments, which

[2]Cf. (Helbig/Schenkel 1991), (Schumacher 1986) and (Schumacher 2004).
[3]Cf. (Sommerfeldt/Schreiber 1983) and (Sommerfeldt/Schreiber 1996).
[4]Cf. (Agel 2003).

DE example	**EN translation**	**semantic level**	**syntactic level**
Ich	"I"	*Experiencer*	subject
erfuhr	"found out"	**verbal predicate**	
diese Nachricht	"this news"	*Patiens*	accusative object
(von ihr)	"from her"	*Source*	prepositional object (optional)

Figure 2.1: Valency levels

can be realised in a sentence, their obligatoriness (obligatory and optional complements or adjuncts), and assigns both a morpho-syntactic form (nominative, accusative, their syntactic categories, etc.) and a grammatical role of complements (subject, object, etc.).

The semantic level of valency description is based on the relations between the word meaning of the valency bearer and its potential arguments. Formally it can be expressed with the terms of predicate-argument structures used in predicate logic: *P(x,y,z,...)*, where a *P* is a predicate and *x,y,z* are its srguments. Each argument place is characterised by some semantic attributes, which allow the verb to fill the argument places with certain elements, on the principle of *selection restrictions*. Argument places are filled with semantic roles, e,g, *Agens, Patients, Source, Experiencer*, as shown in figure 2.1 above.

In our study we concentrate on the syntactic level of valency description. However, in order to to explain some phenomena we also analyse semantic factors, such as selectional restrictions of predicates, which allow them to take certain predicate types. Therefore, existing semantic approaches are also relevant to the issues we deal with, i.e. in the description of complement realisation (see section 2.1.3).

2.1.3 Complementation: Quantitative and Qualitative Valency

As mentioned in section 2.1 above, alency theory includes the analysis of sentences, which focuses on the roles certain words play in sentences with respect to the necessity of other elements to occur. This largely coincides with what is often called realisation of predicate-argument structure or complementation, cf. (Herbst 1999). The basic assumption of the valency theory is that verbs occupy the central position in sentences because they determine how many other elements are to occur in order to form a grammatical sentence. As already mentioned in section 2.1, uch elements are called complements and the number of the ones a verb can take constitutes its valency, cf. (Herbst 2004).

2.1.3.1 Quantitative and qualitative valency: valency patterns

Describing predicate argument structure and its realisation in complements, most authors, e.g. (Herbst 1999) and (Jacobs 2003), mention quantitative and qualitative valency on both levels of it description.

At the **syntactic level**, subcategorisation is seen in terms of the complements that a predicate takes, which includes:

- *Quantitative syntactic valency* – the number of complements a predicate can take. Jacobs call this feature *Realisierungsforderungen* ("realisation requirements"). This coincides with the concept of valency or complementation patterns of predicates.
- *Qualitative syntactic valency* – the formal character of these complements, their grammatical functions or syntactic categories, called *Merkmalsforderungen* ("feature requirements") by Jacobs.

At the **semantic level**, subcategorisation is generally seen in terms of the argument description that a predicate takes, which includes:

- quantitative semantic valency – the number of argument positions a predicate can open, or *Relatforderungen* ("relator requirements");
- qualitative semantic valency – the semantic character of these arguments, their selectional restrictions. Jacobs describes this feature as *content requirements for the relators* and subclassify them into *sortale Forderungen* ("sortal rewuirements"), which correspond to sortal or selectional restrictions, and *Rollenforderungen* ("role requirements"), which coincide with the concept od semantic roles.

Thus, the quantitative valency describes on both levels, how many arguments or complements a predicate can take. In this case, we speak about avalent (with a dummy subject), monovalent (with one argument), bivalent (with two arguments), trivalent (with three arguments) or even tetravalent (with four arguments, which is uncommon for German) predicates, cf. examples (2.1a) to (2.1d).

(2.1a) *Es regnet.* ("It rains").

(2.1b) *Sie arbeitet.* ("She works").

(2.1c) *Er öffnet die Tür* ("He opens the door").

(2.1d) *Sie geben mir eine Aufgabe* ("They will give me a task").

In many cases quantitative syntactic and semantic valency coincide, as for instance in (2.1c) where the complement *Er* ("he") represents one argument ('someone who opens the door') and *die Tür* the other ('the object being opened'). the other ('the object being opened'). In the case of an avalent predicate, like in (2.1a), the predicate has no arguments but has a syntactic placeholder *es* ("it"), which is technically the complement of the verb *regnen* ("to rain").

Qualitative and quantitative valency on both levels is related to the concept of *valency patterns* or *subcategorisation frames*, which includes both, the number of arguments and complements a predicate can take, as well as the type or form of their realisation. A valency pattern reflects a specific combination of arguments or complements for a given predicate. A sentence is grammatical if all the arguments of

a given predicate are realised. This coincides with the notions of *completeness* and *coherence* in LFG (**L**exical **F**unctional **G**rammar)[5].

In the following part of the work, we analyse the main issues concerning the structure, types and position of complements (or arguments) with respect to their description in different linguistic and NLP studies. The focus of our research is on sentential complementations, which are described in section 2.2.

2.1.3.2 Complements and adjuncts

Predicates determine the number of other elements that have to occur in a sentence to make it grammatical. However, since not all elements are dependent on the governing element, a distinction is made between elements, which are part of the valency of the predicates and those, which are not. This was already mentioned by Tesnière who distinguishes between the elements, which are directly involved in the action described by the verb and those whose occurrence has no restriction.

The dependent elements are complements (as was already mentioned in 2.1 above) and the latter elements are referred to as *adjuncts*.

The distinction between complements and adjuncts can be observed not only in the theoretical linguistic work. In formal grammars, e.g. in the theory of LFG, the complement-adjunct description is represented with the terms *subcategorisable* and *non-subcategorisable* grammatical functions. The non-subcategorisable ones, such as ADJ(unct)s and XADJ(uncts), are adjuncts of verbs and not their complements. Subcategorisable or governable functions in LFG are SUBJ, OBJ, XCOMP, COMP, OBL[6].

This distinction is also necessary for NLP, as its absence can result in numerous errors in the NLP processing. Therefore, complements vs. adjuncts are described in several NLP resources used for the parsing of natural language, e.g. COMLEX, which is an NLP lexicon of subcategorisation features for verbs[7], or in FrameNet, which is a lexical resource describing semantic and syntactic valency, cf. (Filllmore 2007)[8]. In FrameNet, peripheral elements, such as *Manner, Location, Means,* etc., provide aspects of the setting, which can modify any frame of the relevant type, i.e. act, state, etc. and are not necessary to the central meaning of a frame. The characterestics of these elements correspond to those of adjuncts in traditional linguistics.

Obligatory vs. optional complements An important aspect of the complement description is obligatoriness. Most authors classify complements into *obligatory* and *optional* ones, and both traditional grammar and modern syntactic theory distinguish between these two complement types. This distinction is required by the structural property of the verb meaning.

Obligatory complements cannot be omitted in a sentence without changing the grammaticality of the sentence, in which the predicate occur or changing the meaning of the predicate. To illustrate this, (Helbig 1992) gives an example, shown in (2.2)

[5] see (Bresnan 1982a) and (Bresnan 2001).

[6] The more detailed description of complement types in LFG is given in section 2.1.4 below.

[7] COMLEX is described in (Meyers *et al.* 1994). We give a more detailed information on this lexcion in in section 3.1.2.2 below.

[8] This lexicon is also described in more details in section 3.1.2.2 below.

In (2.2a), the obligatory complements *Ute* and *den Linguisten* cannot be deleted. Optional complements, such as *Frankenwein* in (2.2b) can be omitted, which, however, changes the meaning of the verb used intransitevely, as in (2.2c).

		obligatory	**optional**
(2.2a)	*Ute besucht*	*den Linguisten.*	–
	("Ute visits")	("the linguist.")	–

		obligatory	**optional**
(2.2b)	*Ute trinkt gern*	–	*Frankenwein.*
	("Ute likes to drink")	–	("Frankenwein.")

		obligatory	**optional**
(2.2c)	*Ute trinkt gern.*	–	–
	("Ute likes to drink.")	–	–

Obligatory complements are dependent on the predicate in form, whereas optional complements demonstrate characteristics of a complement, but are not syntactically required to be expressed at all. They differ from obligatory ones in the fact that their occurrence in a sentence is not dependent on the predicate. Besides that, the predicate does not determine them in form, cf. (Herbst 1999) and (Herbst et al. 2004). T.Herbst states that there also exist contextually optional complements, which are optional only if their referent can be identified from the context.

The concepts of obligatoriness and optionality are used both, in linguistic and NLP studies. For instance, in the above mentioned subcategorisation lexicon COMLEX, obligatoriness or optionality of certain complement types (e.g. *that*-clauses) are described with the features *required* and *optional*. This is illustarted in figure 2.2. The sentential complement introduced by *that* is optional in the first sentence, *They thought (that) he was always late*. The verb *thought* has a sentential complement, which can be optionally introduced by the pronoun *that*. In the second sentence, *He complained that they were coming*, the *that*-clause is obligatory, it is required by the argument structure of the verb *complain*.

1. (vp-frame s :cs (s 2 :that-comp **optional**)
   :gs (:subject 1 :comp 2)
   :ex *"they thought (that) he was always late."*)

2. (vp-frame that-s :cs (s 2 :that-comp **required**)
   :gs (:subject 1 :comp 2)
   :ex *"he complained that they were coming."*)

Figure 2.2: COMLEX verb frames for verbs with a sentence clause

In FrameNet, which describes subcategorisation in terms of frames and frame elements, obligatoriness of complements (as well as their distinction from adjuncts) is expressed by the concept of *coreness*, which includes four possible levels: *core, peripheral, extra-thematic* and *core-unexpressed*. In the description of obligatoriness, FrameNet also operates with the term *Null Instantiation: Definite Null Instantiation (DNI), Idefinite Null Instantiation (INI), Constructional Null Instantiation (CNI)*,

is used for the description of frame elements that do not show up in the sentence, but whose semantic role should be still identified. DNI is used for the missing element which is understood in the linguistic or discourse context, thus coinciding with anaphora. INI is used for the missing complements of transitive verbs, which can be used intransitively, thus, optional complements, and CNI is used for the elements whose omission is licensed by the construction or the structure of the sentence they are used in. The latter coincides with what T. Herbst calls contextually optional complements.

Optional complements vs. adjunts Optional complements shouldn't be confused with adjuncts, which have no specific relation to the meaning of the verb. The deletion of adjuncts does not affect the grammaticality of a sentence. Adjuncts have no restriction on their occurrence. They differ from complements in that their occurrence and number in a sentence is not dependent on the predicate. Moreover, they are not determined in their form by the predicate.

It is obvious that adjuncts are not totally freely addable since their occurrence is subject to general semantic compatibility, but the important distinction criterion with respect to complements is that the occurrence of adjuncts is not in any way dependent on particular lexical items.

The criteria to distinguish optional complements vs. adjuncts are widely discussed in linguistics and NLP studies. One of them is based on the fact that the omission of adjuncts can affect neither grammaticality nor the meaning of the predicate. To prove this, we can apply the deletion test in which adjuncts can be just removed from a sentence. For instance, the adjunct *in London* can be deleted in (2.3) without changing the grammaticality of the sentence or the meaning of the verb *besuchen*.

		obligatory	**optional**	**adjunct**
(2.3)	*Ute besucht*	*den Linguisten*	–	*in London.*
	("Ute visits")	("the linguist")	–	("in London.")

The elimination test was suggested by several grammarians and is described, for instance, in (Helbig/Schenkel 1973). However, this test is discussed controversially. Some authors, e.g. (Somers 1987) critise it for not taking into consideration the possibility of omitting deep subjects in imperatives, passives or infinitives. Most linguistic and lexicographic criteria for the distinction of complements from adjuncts leave obscureness, which is the reason for the problems in the distinction between optional complements and adjuncts in the theories and NLP applications.

In some NLP works, e.g. in COMLEX, further criteria are developed, which are justified on the basis of experimental evidence. The criteria for *complement-hood* are based on the examination of the data as well as statements made in the linguistics. For instance, a complement is obligatory if it makes the realisation of a predicate grammatical, it can only be the subject of the passive, has an argument theta-role (e.g. theme, goal, patient, etc.). Furthermore, the authors mention *rules of thumb*, which are useful for identifying complements and include such criteria as usage with frequent predicates, realisation as typical complements (nominal and prepositional phrases, clauses), etc. Adjuncts are given *adjunct-hood* criteria, e.g. typically prepositional, adverb and subordinate clauses, position in a sentence, etc. Conflicts between

criteria for adjuncts and those for complements are usually resolved in favour of complements.

The statements described in the two sections above show that the distinction between obligatory and optional complements, as well as the one between complements and adjuncts is analysed both, in linguistic and NLP studies, where it finds different terminological description. In table 2.1 we summarise some of the above mentioned descriptions, according to the studies they are used in.

studies	**complements**		**adjuncts**
	obligatory	**optional**	
Tesnière	actants (prime, second, third)		circonstants
Helbig	obligatory actants	optional actants	adjuncts
LFG	subcategorisable		non-subcategorisable
COMLEX	required	optional	adjuncts
FrameNet	core FE		peripheral FEs
		INI and CNI (contextually optional)	

Table 2.1: Complements and adjuncts in different studies

As mentioned in 2.1, we use the terms *complements* or *arguments* in the description of the elements, which predicates subcategorise for. We also apply the terms *optional* and *obligatory* for the description of obligatoriness of complements and adjuncts in the description of elements, which are not dependent on the predicates. As the aim of this thesis is to elaborate on an architecture to extract and classify subcategorisation properties of predicates, which do not include adjuncts (as they are non-subcategorisable elements), we do not analyse them in the following sections.

2.1.4 Complement Realisation on the Syntactic Level

As mentioned above, on the syntactic level, subcategorisation is seen in terms of the complements that a predicate takes, which includes the statements: how many elements a predicate can take and what is their formal character (the grammatical function they carry within a sentence).

The arguments, which constitute the predicate-argument structure of a verb can be realised as subjects and objects, according to their grammatical role in a sentence. Using the concept of valency in a wide sense, we call all the subcategorised elements complements of a predicate. In some linguistic and NLP studies the subject does not belong to the class of complements. The concept of the subject is used along with that of the complement (represented by objects, obliques, etc.). We consider complements of a predicate as the realisation of its predicate-argument structure, which also includes subjects.

For instance, we admit that the German verb *arbeiten* (“to work”) has only one complement (a subject) in its syntactic environment, whereas the verb *fragen* (“to

ask") has several arguments, cf. examples (2.4a) and (2.4b).

	sentence	**complements**
(2.4a)	*Sie arbeitet.*	*sie*
	("She is working.")	("she")

	sentence	**complements**
(2.4b)	*Sie fragte ihren Vater, warum es passierte.*	*sie, ihren Vater, warum es passierte*
	("She asked her Dad why it happened.")	("she, her Dad, why it happened")

In (2.4b), the argument *sie* of the verb *fragen* is its subject, the complement *ihren Vater* is its indirect object, and the complement *warum es passierte* is its direct object. Complement realisations can have a simple form, e.g. the third person pronoun *sie* or a complex form, e.g. a subclause. We describe possible forms or types of complements, as well as their syntactic categories in the following sections.

2.1.4.1 Grammatical functions of complements

As already mentioned, arguments of a predicate can be assigned grammatical functions, e.g. subjects (subj) or objects (obj). Traditionally, objects fall into three classes: direct (dir), indirect (indir) and prepositional (prep). However, the classification of argument grammatical functions varies in different theories in linguistic, lexicographic and NLP literature. In table 2.2, we summarise the diversity of terms used in linguistics and NLP to describe the same phenomena. We follow the traditional classification of grammatical functions into subjects, direct, indirct and prepositional objects.

function/study	**subj**	**dir**	**indir**	**prep**	**further**
traditional grammar and lexicography					
Tesnière	subj nom	dir acc	indir dat		
German grammar, VALBU and ViF	NomE	AccE, GenE	DatE	PrepE	AdvE, etc.
Herbst, VDE	$[C]_a$	$[C]_p$			
formal grammar and NLP					
LFG	SUBJ, COMP	OBJ	OBJ2	OBL_θ	COMP, XCOMP
COMLEX	subj	obj	obj2, obj3, obj4		
FrameNet	External	Object	Dependent		

Table 2.2: Grammatical functions in linguistic and NLP studies

In the relationship between predicate arguments and their grammatical functions, an important condition is the principle that the same grammatical function cannot be assigned to different arguments and different grammatical functions cannot be assigned to the same argument, for example, cf. the principle of Function-Argument-Biuniqueness in (Bresnan 1982a).

In the following we describe some of the approaches mentioned in table 2.2.

Traditional grammar and lexicography Most traditional grammars use functional labels of subject and object, which are described depending on the approach the authors decide for. For instance, Tesnière distinguishes three classes complements according to their syntactic function in a sentence – subjects, objects and indirect objects. For inflected languages he also suggests the classification into nominative (nom), accusative (acc) and dative (dat) complements, which coincides with the case-marking theory, followed by the modern grammar description. For instance, for German, some grammarians, e.g. (Engel 1991) and (Helbig/Buscha 2005), also distinguish nominative, accusative and dative complements. Furthermore, they expand this classification including genitive and prepositional complements. As shown in table 2.2 above, nominative complements correspond to subjects, e.g. *die Studie*, *wir* and *sie* in (2.5). Accusative, e.g. *ihren Vater* in (2.5b) or *die Wirkung dieses Materials* in (2.5a), dative, such a the pronoun *Ihnen* in (2.5b), and genitive complements, e.g. the noun phrase *seiner Hilfe* in (2.5c), are objects.

(2.5a) ***Die Studie*** *untersucht* ***die Wirkung dieses Materials***.
("The study examines the impact of this material").

(2.5b) ***Wir*** *werden es* ***Ihnen*** *mitteilen.*
("We will inform you about this").

(2.5c) ***Sie*** *bedarf* ***seiner Hilfe***.
("She needs **his help**").

In the lexicographic work, for instance, in the dictionaries for German verbs *Verben in Feldern* and *VALBU*[9], he classification includes not only nominative, accusative, dative and genitive complements (called NomE, AccE, DatE, GenE[10]) bbut also prepositional ones and further complement types, such as adverbial, predicative and verbative complements (AdvE, PredE, VerbE, which are not considered in this thesis). The case-based classification for German is significant as it is an inflected language. In this study we analyse sentential complementation, which do not contain any formal case-marking. Therefore, the above mentioned classification is not relevant for this thesis and we do not go into detail in its description.

One of the important characteristics of verbal complements is their ability to function as subjects of active or passive clauses, cf. (Herbst 2004). Therefore, in *VDE*[11], this ability is indicated with indexes for a possible active subject or a possible passive subject. [C]a-complements can occur in an active clause, functioning as the subject of the clause or in a passive clause, taking the form of a phrase or a clause introduced by the preposition by (provided another complement can function as subject). [C]p-complement can occur in an active clause following the verb phrase or as the subject of a passive clause.

Formal grammars and NLP Valency description in formal grammars and NLP work also involves the grammatical functions, e.g. subjects and objects. For instance in

[9] Cf. (Schumacher 1986) and (Schumacher 2004).
[10] E stands for the German word *Ergänzung* ("complement").
[11] Cf. (Herbst et al. 2004).

LFG[12], grammatical functions provide a mapping between the syntactic structures and the predicate-argument structure. The grammar of each language characterises, which syntactic constituent can be mapped onto the argument of a predicate. The characterisations differ from language to language, but the functions remain the same – all languages have subjects, objects, oblique objects and complements of various types, which can be mapped onto the arguments of verbal or other predicates (see (Bresnan 1982a)). Grammatical functions in LFG are classified into two groups – subcategorisable and non-subcategorisable (for adjuncts). Subcategorisable functions are subdivided into semantically unrestricted functions, which include SUBJ (subject), OBJ and OBJ2 (objects), and semantically restricted ones, which contain OBL_θ (prepositional object), COMP and XCOMP (complex objects expressed by complement clauses).

In NLP lexicons, for instance, in COMLEX, the role each constituent plays in a sentence is indicated in the grammatical structure. The grammatical structure consists of a list of grammatical relations (summarised in figure 2.3), each referring to an element from constituent structure.

:Subject
:Head
:Obj
:Obj2
:Obj3
:Obj4
:Omit-subc (for omission of complements)
:Comp
:Prep (for bare prepositions)
:Part (for bare particles)
:Mod (for adverbs in advp-frames)

Figure 2.3: COMLEX grammatical relations

The principal grammatical functions in FrameNet , e.g. *External Argument, Object*, and *Dependent*, are described in (Filllmore 2007). *External* corresponds to the subject. The term subject is not used in the theory of Frame Semantics because in most cases the relevant constituent in its own location is not a subject. The grammatical function *External* is used for the subject function of both, finite and non-finite verbs (cf. (2.6a) and (2.6b) respectively), as well as for dependents of governing or frame-bearing nouns, cf. (2.6c). Thus, in (2.6b) *the general*, which is the *Object* of *persuade*, is the *External Argument* of *to release* and in the sentence (2.6c), *physician* is the *External argument* of the verb *perform* but has the *Genitive* relation to the noun *decision*. In some cases, nominal predicates have their own *External*, e.g. if used with support verbs within a multiword expression as in (2.6d) or when the frame-bearing noun is governed by a control noun, e.g. in (2.6e).

(2.6a) *The **physician** performed the surgery.*

[12]A more detailed description of LFG is given in section 3.1.2 below.

(2.6b) *We persuaded **the general** to release the prisoners.*

(2.6c) *The **physician**'s decision to perform the surgery.*

(2.6d) ***He** made a statement to the press.*

(2.6e) ***My** attempt at an agreement with Path failed.*

Object is assigned to any traditionally accepted object, however, only to verbal predicates. *Dependent* is referred to both complements and adjuncts, e.g. in (2.7a) and (2.7b), as both, the adjunct *in order to finance a concert* and the optional complement *to me*, carry the same grammatical function Dependent. The distinction between complements and adjuncts is expressed in FrameNet with the concept of coreness (as mentioned in section 2.1.3.2) and is not replicated in a grammatical function description. This grammatical function is especially interesting for this study as *Dependent* is also assigned to all sentential complements (both of verbal and nominal predicates), cf. (2.7c) and (2.7d). As mentioned above, frame-bearing nouns are also assigned their own grammatical function of *Genitive* as illustrated in example (2.6c above). The other two grammatical functions, which are specific only for nouns are *Quant*, which is assigned to prenominal determiner when they express a number, e.g. ***three** bottles of wine*, and *Appositive*, which is assigned to post-target appositional Ns and NPs, e.g. *Libel lawyer* **Jonathan Crystal** *represented the plaintiff*.

(2.7a) *Bill sold the house **in order to finance a concert**.*

(2.7b) *Pat spoke **to me**.*

(2.7c) *I believe **that you are the winner**.*

(2.7d) *The fact **that cats have fur**.*

2.1.4.2 Syntactic categories of complements

Traditional theoretical grammar, lexicography and NLP literature also employ formal categories to describe the formal realisation of complements as phrases or clauses.

In (Herbst et al. 2004), the authors describe complements with respect to their formal realisation in terms of phrases and clauses or sentential complements. Like many other linguists, they distinguish between phrases and clauses (sentential complements). Phrases include noun phrases (NPs), adjective phrases (APs) and prepositional phrases (PPs), whereas clauses (SC) include infinitive clauses, such as *zu*-infinitives ("to"-infinitives) (2.8a) and *dass-*, *w-* and *ob*-clauses (that-, wh-and if-or whether-clauses in English), (2.8b) to (2.8d) in German[13].

Systematic correspondences of syntactic categories to grammatical functions are described in (Schumacher 1986) and (Schumacher 2004). We summarise a part of this description in table 2.3 below.

Most approaches in linguistics and NLP work (e.g. FrameNet, COMLEX, etc.) operate the same terms to describe syntactic categories. In this study, we also follow this terminology. As we concentrate on sentential complements, we do not go into details

[13]Further examples of sentential clauses are described in the section 2.2

in the description of other complement types. The subject of this study are *dass-*, *w-* and *ob*-clauses,therefore, therefore, we outline their characteristics and description in section 2.2 below.

(2.8a) *Sie hat versprochen,* ***zurückzukommen***.
("She promised **to come back**.")

(2.8b) *Sie hat versprochen,* ***dass sie zurück kommt***.
("She promised **that she comes back**.")

(2.8c) *Sie wußte nicht,* ***wann sie zurück kommt***.
("She didn't know **when she comes back**.")

(2.8d) *Sie wußte nicht,* ***ob sie zurück kommt***.
("She didn't know **if she comes back**.")

function	**syntactic category**	**example**
NomE	NP	***Die geplante Flugverbindung*** *wird nicht zustandekommen.*
	SC	***Was ich gesagt habe****, beruhte auf einem Irrtum.*
AccE	NP	*Die Studie untersucht* ***die Wirkung dieses Materials****.*
	SC	*Man hat nicht beachtet,* ***dass es passieren kann****.*
GenE	NP	*Sie bedarf* ***ihrer Hilfe****.*
	SC	*Der Minister wurde beschuldigt,* ***die Angelegenheit verzögert zu haben.***
DatE	NP	*Ich traue diesem Mann nicht.*
	SC	*Ich leihe meine Bücher nur,* ***wem ich will****.*
PrepE	preposition + NP in Acc	*Ich verlasse mich* ***auf deine Hilfe***
	NP in Dat	*Die Medien beschäftigen sich* ***mit diesem Problem****.*
	adjective	*Der Ausschuss hielt keinen der Bewerber* ***für qualifiziert****.*
	SC	*Er hat nicht (****daran****) gedacht,* ***was auf ihn zukommen könnte****.*

Table 2.3: Correspondences of grammatical functions to syntactic categories

2.1.5 Complement Description on the Semantic Level

On the semantic level, valency is seen in terms of argument positions opened by predicates (cf. section 2.1.3.1 above), as well as semantic features of predicates, e.g. sortal or selectional restrictions.

The semantic description of complements is already present in Tesnière's work. The author mention semantic functions – those that perform the action, those that undergo the action and those to whose benefit the action takes place. Variety of approaches in the valency description produces the variety of different descriptions of complement semantics. In (Götz-Votteler 2007), he authors summarise four approaches, which describe complements by means of semantic roles, semantic components, semantic categories and verb-specific description.

Theoretical approaches to valency widely use the concept of *semantic roles*, which are often called *thematic roles* or *theta-roles*. For the first time, *semantic roles* were

mentioned by Charles Fillmore in (Fillmore 1968). The author describes semantic valency as a set of *semantic roles* associated with a word in a given meaning. These roles characterise a central component of the semantic structure of any phrase or clause that can be built around that word in that meaning, see (Fillmore 2003). Case roles defined by Fillmore include *Agent, Instrument, Stimulus, Patient, Theme, Experiencer, Content, Beneficiary, Source, Goal* and *Path*. The inventory of case roles used by other authors, e.g. (Allerton 1982), and other theories, e.g. LFG[14], are similar to Fillmore's description in most cases.

FrameNet also operates with the categories defined by Fillmore[15]. Each frame provides its set of semantic roles. The verbs belonging to a particular frame share the same collection of frame-relevant semantic roles. The "general-purpose" semantic roles (as *Agent, Patient, Theme, Instrument, Goal*, and so on) are replaced by "frame-specific" role names (e.g. *Speaker, Addressee, Message* and *Topic* for "speaking verbs").

The idea of semantic role is similar to Wierzbicka's theory of Natural Semantic Metalanguage (NSM), which allow to analyse words from ordinary language by means of script-like explications based on reductive paraphrase (plainly speaking, to break words down into combinations of simpler words) using a small collection of semantic primes, see (Wierzbicka 1972). The semantic primes are believed to be atomic, primitive meanings present in all human languages: *substantives (I, YOU, SOMEONE, PEOPLE, SOMETHING), mental predicates (THINK, KNOW, WANT, FEEL, SEE, HEAR), actions, events and movement (DO, HAPPEN, MOVE), time (WHEN/TIME, NOW, BEFORE, etc.), space (WHERE/PLACE, HERE, ABOVE)*, etc.

Similarly, in (Schumacher 2004) the author defines semantic categories for every argument, depending on what it refers to, e.g. person, animal, object or force. Semantic categories depend on the usage of an argument within the semantic range of a certain category. Another option of describing complements at the semantic level is the verb-specific description of participants, which is also employed by (Schumacher 2004). For instance, the first argument of the verb *sich verletzen* ("to injure oneself") is characterised as "somebody who gets injured by sth: person/animal". This method is mostly used in lexicographic frameworks, such as *VALBU* or *VDE*. Besides that, in (Schumacher 1986), he author describes parallels between the selectional restrictions (e.g. *fact* or *event*) of verbs and the complements they take.

In NLP lexicons, e.g. in VerbNet (VN), which is the largest online verb lexicon currently available for English[16], semantic categories are combined with semantic roles. VerbNet provides detailed syntactic-semantic descriptions of Levin classes organised into a refined taxonomy. Each verb class in VN is completely described by thematic roles, selectional restrictions on the arguments, frames consisting of a syntactic description and semantic predicates with a temporal function. Semantic restrictions (such as *abstract, animate, human, organisation*, etc.) serve to constrain the types of thematic roles allowed by the arguments. They do not depend on the meaning of the verb, but can be regarded as properties of the arguments. Each frame is associated

[14]The ordering of semantic roles to a universal hierarchy is described in later works on LFG, e.g. in (Zaenen/Engdahl 1994).

[15]Semantic roles in FrameNet are described in (Baker *et al.* 1998) and (Johnson *et al.* 2002).

[16]Cf. (Kipper-Schuler 2005).

with explicit semantic information. Thematic roles applied in VN are *Actor, Agent, Asset, Attribute, Beneficiary, Cause*, etc.[17]

Although we extract predicates in terms of their syntactic valency in this thesis, the analysis of semantic features of both, predicates and their complements, are important to explain some phenomena under analysis (e.g. the preference of different ppredicates for certain subclause type). Due to that, of particular importance are the correspondences between selectional or sortal restrictions of the predicates under analysis and their sentential complements, which are described in section 2.2.2.3.

2.2 Sytax and Semantics of Sentential Complements

We summarise semantic and syntactic approaches in the analysis of subclauses in the following part of this thesis. Sentential complements have been a research topic in different linguistic theories. Most studies concentrate on the problem of what predicates require what clausal complements, i.e. for what sentential complement they subcategorise for. For instance, (Grimshaw 1979) claims that predicates are endowed not only with syntactic subcategorisation but also with semantic selectional restrictions, which means that they can select for the same type of semantic complement but the syntactic realisation of that complement can vary.

On the syntactic level, sentences are traditionally subdivided into declaratives and interrogatives. Declaratives are introduced by the conjunction *dass* ("that"), and interrogatives are introduced by *w*-words ("wh-"words) or the conjunction *ob* ("if" or "whether"). On the semantic level, there are statements, which correspond to declarative sentences and questions, which correspond to interrogatives.

2.2.1 Sentential Complements on the Syntactic Level

In the following sections, we analyse syntactic features of sentential complements, their forms, possible positions in a sentence and grammatical functions.

2.2.1.1 Forms of sentential complements

Forms of sentential complements differ in the type of the introductory word, e.g. *dass-,w-* or *ob*, in the form the main verb takes, e.g. infinitive clauses, where the main verb has an infinite form, vs. other clauses, where the verb has a finite form or in the verb position in the subclause, e.g. *dass, w-* and *ob*-clauses or infintives, where the verb occupies the final position in the sentence vs. *Verbzweitsätze V2,* where the finite verb occupies the position after the subject. *Verberst* V1, *Verbzweit* V2 or *Verbletzt* (VL) as shown in table 2.4. V1 is used in direct questions, V2 is usually the form of a main clause and VL is used in embedded sentence starting with certain introductory words.

In table 2.5 below, we give a list of different forms of sentential complements based on the classification given in VALBU, cf. (Schumacher 2004). As we concentrate on the extraction of predicates, which take *dass-*, *w-* and *ob*-clauses, in the

[17]Lists of the thematic roles, selectional and syntactic restrictions, predicates, and frame types are available under http://verbs.colorado.edu/verb-index/reference.php

Type	1st position	2nd position	final position
V1	Kommt	sie heute zurück?	
V2	Sie	kommt heute zurück	
VL	dass sie heute zurück		kommt

Table 2.4: Verb position in a German sentence

following, we give a more detailed analysis on these three types of sentential complements.

form	**explanation**	**VALBU terminology**
dass-clause	subclause introduced by *dass*	*dass*-clause
ob-clause	indirect question introduced by *ob*	*ob*-question
w-clause	indirect question introduced by *w*-words	*w*-question
zu-infinitive	infinitive with *zu*	Inf -
ifinitive	infinitive without *zu*	Inf +
V2	subclause in form of a main clause	*HPTS*
direct questions or statements	direct speech	*DIRR*

Table 2.5: Forms of sentential complements in German.

Subclause types: *dass*-clauses Subclauses, introduced by the conjunction *dass* belong to declarative clauses, which we call (following most linguistics and NLP studies) *dass*-clauses. If a subclause starts with the conjunction *dass*, he main verb moves to the final position in the clause, cf. VL in table 2.4 above and sentence (2.9) below. This subclause type cannot occur with every German predicate, which is explained both, by their semantics and the semantics of the predicates, cf. section 2.2.2.3 below.

(2.9) *Sie verspricht, dass sie zurückkommen wird.*
("She promises that she will come back").

Subclause types: *w*- and *ob*-clauses Sublcauses introduced by the conjunction *ob* or a *w*-word belong to interrogative subclauses and are called *w*- or *ob*-clauses. Although *w*- and *ob*-clauses are analysed together in most studies, some authors, e.g. (Schumacher 2004), claim that these two clause types should be distinguished from each other. The reason for this is that not all the verbs subcategorising for a *w*-clause can also have an *ob*-clause and vice versa, cf. (2.10a) and (2.10b).

(2.10a) *Er hat gewusst, warum sie nicht kommen konnte.*
("He knew why she couldn't come.")

(2.10b) *Er hat gewusst, *ob sie nicht kommen konnte.*
("He knew *if she couldn't come.")

We also admit that there are some differences between *w-* and *ob*-clauses. However, analysing the ability of predicates to take different types of subclauses, we mostly consider *w-* and *ob*-clauses within one category – interrogative clauses –, which are compared to declarative clauses introduced by *dass*. Therefore, we often use the term *w-/ob*-clauses in this thesis.

W-clauses are introduced by *w*-words. The term *w*-word refers to a special group of words, most of which begin with the letter *w-*, like *warum* ("why"), *wo* ("where"), *wann* ("wann"), etc. English grammar operates the term *wh-word*. In German, these words are often called interrogative pronouns (*Interrogativpronomen, Fragewörter*). The group of items that can introduce dependent interrogative clauses comprise both, noun phrases, such *welchen Zug* in example (2.11a), or adverbs, such as *wie oft* in example (2.11b) below. We list more frequent German words in the section A.1 in the appendix.

(2.11a) *Er fragt, welchen Zug sie nehmen sollen.* ("He asks which train they should take").

(2.11b) *Er fragt, wie oft die Züge fahren.* ("He asks how often trains go").

2.2.1.2 Position in a sentence

Sentential complements occupy certain positions in a sentence, depending on their grammatical functions.

Subclause positions can be described according to the topological field model of (Höhle 1986) shown in table 2.6, mentioned in works on the grammar of German, e.g. (Helbig/Buscha 2005). Subclauses in German can occupy either the *Vorfeld* (VF, pre-field) or the *Nachfeld* (NF, post-field). These topological fields (called *Stellungsfelder* in German) describe the position in a clause determined by left and right sentence brackets (called *linke Satzklammer (LSK)* and *rechte Satzklammer (RSK)* in German). Sentence brackets is the German word order principle that divides a clause or sentence into the mentioned positions. The *Vorfeld* is the position in front of the finite verb, whereas the *Nachfeld* is the position after the infinite verb forms. Another topological field (which is not relevant for sentential complement analysis) is the *Mittelfeld* (MF, middle field), the position between the finite verb and the infinite verb forms or the end of the clause.

VF	LSK	MF	RSK	NF

Table 2.6: Topological field model

Subclause in the Vorfeld A subordinate clause in the *Vorfeld* occupies the position before the main clause which contains the main verb. The main clause following the subclause in the Vorfeld starts with the finite verb as illustrated in table 2.7.

In this study, we extract subclauses in the VF, which are preceded by nouns within the analysis of nominal predicates. In this case, nouns precede subclauses in the VF, as shown in table 2.8.

VF, subclause	LSK	MF, main clause	RSK	NF
Dass die Unternehmenseigentümer keine Dividenden erwarten können, "That enterprise bondholder won't get any dividends"		*kann nicht ausgeschlossen werden.* "shouldn't be excluded"		

Table 2.7: Subclause in the Vorfeld

VF, subclause	LSK	MF, main clause	RSK	NF
Die Vorstellung, dass die Menschheit unbedeutend werden könnte, "The idea that humanity could become insignificant"		*ist Unsinn.* "is nonsense."		

Table 2.8: Subclause preceded by a noun in the Vorfeld

Subclause in the Nachfeld The position in the *Nachfeld* is the most frequent for subordinate clauses. In this work we extract sentential complements in the NF for most predicates under analysis (except nominals, which are extracted along with their subclauses in the VF). A subclause in the NF is placed after the main clause, as illustrated in table 2.9 below.

VF	LSK	MF, main clause	RSK	NF
Es "It"		*kann nicht ausgeschlossen werden,* "shouldn't be excluded,"		*dass die Unternehmenseigentümer keine Dividenden erwarten können.* "that enterprise bondholder won't get any dividends."

Table 2.9: Subclause in the Nachfeld

2.2.1.3 Grammatical functions of subclauses

Sentential complements can have almost all grammatical functions as described in section 2.1.4.1. An exception are, for instance, indirect objects (which are often realised as nominal phrases in dative), which cannot be realised as sentential complements at all. This is mentioned by several authors, e.g. (Schumacher 1986) or (Oppenrieder 2006).

(Schumacher 1986) explains this phenomenon by the semantics of predicates and sentential clauses. The author claims that sentential complements give a specification to predicates, expressing either a *fact* (*Sachverhalt*) or an *event* (*Ereignis*). For instance, the accusative complement of the verb *mitteilen* can be realised as a clause, whereas the accusative complement of the verb *bekommen* can not be realised as a sentential clause because it is impossible for it to become a fact, cf. (2.12a) and (2.12b).

(2.12a) *Er teilt uns **etwas** mit. – Er teilt uns mit, **dass er kommen werde**.*
("He notifies us of **something** – He notifies us **that he will come**")

(2.12b) *Er bekommt **etwas**. – *Er bekommt, **dass...***
("He gets **something**. – *He gets **that...**").

In table 2.10 below we summarise the grammatical functions, which sentential complements can have in a sentence, based on the classification formulated in (Schumacher 2004).

grammatical function	**example**
subj	***Dass die wahren Fans draussen sind**, hinterlässt einen bitteren Nachgeschmack.* "**That the real fans are outside** leaves a bitter aftertaste"
dir	*Du muss uns erzählen, **ob deine Bewerbung erfolgreich war**.* "You must tell, **wether your application was succefull**".
prep	*Er interessiert sich **dafür, was damals passierte.*** "He is interested (to find out) **in what happened those days**".

Table 2.10: Grammatical functions of German sentential complements

Subject clauses Both declarative *dass*- and interrogative *w-/ob*-clauses can be used as subjects in a sentence, as seen in table 2.10 above and in example (2.13a) below. In sentences subject subclauses usually occupy the VF or are moved to the end of a sentence. In the latter case, the subject position in the VF is occupied by the *Korrelat*[18] *es* ("it"), which acts as a placeholder or a marker, see (2.13b). This usage of *Korrelat* is called "dummy", "empty" ot "explitive" because the English *it* or the German *es* do not refer to anything, except for the subclause itself, cf. (Lester 2008).

(2.13a) ***Ob die wahren Fans draussen sind**, lässt sich klären.* ("Whether the real fans are outside can be clarified").

(2.13b) ***Es** lässt sich klären, **ob die wahren Fans draussen sind**.* ("It can be clarified whether the real fans are outside").

Object and prepositional clauses Object clauses can occupy both, the VF and the NF in a sentence. As mentioned above, sentential complements can express direct objects or prepositional complements but no indirect objects.

If a sentential clauses expresses a prepositional object, a *Korrelat* is inserted into the main clause, cf. examples (2.14a) and (2.14b). The prepositional complement, expressed by *für diese Information* ("in this information") in (2.14a) can be replaced by the *w*-clause *was damals passierte* ("what happened those dayd") in (2.14b), whereas the preposition is replaced by the Korrelat *dafür* ("in it").

(2.14a) *Er interessierte sich **für diese Information**.*
("He was interested **in this information**").

(2.14b) *Er interessierte sich **dafür, was damals passierte**.*
("He was interested **in what happened those days**").

[18] Korrelat is a correlative word, a correlate, which is a referential element in the main clause whose function is to refer to a subclause.

The presence of a *Korrelat* with a sentential complement depends on the verb and can be obligatory, facultative or impossible. In table 2.11, we summarise Korrelats according to grammatical functions of sentential complements, following the classification described in (Schumacher 2004).

grammatical function	**Korrelat**
subject	*es*
object	*es*
prepositional	*da(r)*+preposition, e.g. *darauf, davon*

Table 2.11: Korrelat with sentential complements.

The Korrelat *es* depepends on the position of the subclause and is only used if the subclause occupies the NF and not the VF, cf. (2.15a) and (2.15b).

(2.15a) *Alle geht* ***es*** *an,* **dass die Umwelt zerstört wird**.
("**It** concerns everyone **that the environment is being destroyed**").
vs.

(2.15b) **Dass die Umwelt zerstört wird**, *geht alle an.*
("**That the environment is being destroyed** concerns everyone").

In this study we extract predicates with sentential complements, which have all the three grammatical functions illustrated in table 2.10. However, we cannot automatically identify grammatical functions. Therefore, we do not take into account grammatical functions in the classification of predicates according to their subcategorisation features, cf. section 4.

2.2.2 Semantics of Sentential Complements

Declarative *dass*-clauses and interrogative *w-* and *ob*-clauses are different in their semantics. They are compatible with different predicates depending on the selectional restrictions of these predicates. As in this study, we analyse predicates, which occur with both, declarative and interrogative sentential complements, the relations between the semantics of these predicates and the one of their complements are important for the interpretation of their subcategorisation behaviour. Therefore, in the following sections we summarise approaches on the semantic interpretation of declarative and interrogative clauses (sections 2.2.2.1 and 2.2.2.2), aas well as studies on the problem of compatibility of sentence semantics and the one on the predicatesm which take sentential clauses (section 2.2.2.3).

2.2.2.1 Semantics of declarative clauses

From the semantic point of view *dass*-sentences denote statements or propositions, which are in most cases closed and express definite truth values. The truth values can be either true or false but are always definite.

Some views on *dass*-clauses are derived from more philosophic theories, e.g. the singular term theory, supported by (Schiffer 1996) and (Bealer 1998) or relational theory of attitudes. According to the former, *dass*-clauses are singular terms and the complementiser dass is a term-forming operator that turns meaning ful sentences of the language into complex singular terms. The latter understands *dass*-clauses as singular terms whose semantic value is the intentional content of the expressed relation and takes the objects of attitudes to be propositions. (Moffett 2002) states that *dass*-clauses denote not only propositions as the singular term theory suggests, but they sometimes can occur as non-argument modifiers, e.g. if embedded by adjectival or nominal predicates.

According to the traditional view in linguistics, for instance in (Bausewein 1990) or (Moltmann 2003), a *dass*-clause embedded under an attitude verbal predicate expresses a certain kind of object, a proposition, which acts as an argument of this attitude predicate.

There exist different views on what propositions are. Some authors, e.g. in the type theory, call them functions of possible worlds or situations into the truth values, others state that they are complexes of the meanings of constituents and some authors consider them to be primitive entities. We follow the view that proposition is, on the one hand, the meaning of a sentence, and on the other hand, an object of a propositional attitude. Sentential clauses have a functional category, which is represented semantically as a proposition.

We agree that, if seen semantically, *dass*-clauses denote propositions, which have a closed character and are defined in their truth values. They can be either true or false, depending on the interpretation situation. The fact that they can be false is seen from such sentences as in (2.16).

(2.16) *Die Kirche glaubt, dass die Erde das Zentrum des Universums ist.* ("The Church believes that the Earth is the center of the Universe").

2.2.2.2 Semantics of interrogative clauses

Interrogative clauses, which are introduced by *w*-words and the conjunction *ob*, have a much more complex internal structure than *dass*-clauses.

Interrogative *w*-clauses, which are often referred to as 'question clauses', are usually interpreted as a set of answers on to the question they express. For instance, (Karttunen 1977), who applies the framework for linguistic description developed by Richard Montague in (Montague 1974) in his analysis, claims that every question denotates a set of true answers or propositions. The proposition in this set is expressed not by possible answers but by their true and complete answers to the question. If a sentence is φ and $\|\varphi\|$ is the proposition this sentence expresses, the interrogative *ob*-clause denotes the singleton $\{\|\varphi\|\}$ or the negation $\{\ \|\varphi\|\ \}$, depending on whether φ is true or not.

The translation of *what John reads*[19] denotes a set, which contains for each thing that John reads, the proposition that he reads it. If John happens to read only 'New York Times' and 'Playboy', then the indirect question *what John reads* denotes a set containing only the two propositions expressed by *John reads New York Times* and

[19]The author analyses direct and indirect questions.

John reads Playboy. If John doesn't read at all, this indirect question denotes an empty set, cf. (Karttunen 1977).

This concept is based on Hamblin's idea about the treatment of questions[20], in which the author states that every question denotes a set of propositions. For instance, the direct question *Is it raining?* denotes the set of prepositions expressed by *it is raining* and *it is not raining*.

Some authors, e.g. (Hintikka 1967), analyse interrogative sentences contextually. According to this approach, interrogative clauses are not assigned any meaning as such and interpret indirect questions replacing interrogative questions with the corresponding declarative sentences. For instance, the sentence in (2.17a) is equivalent to the sentence in (2.17b). If analysed in this way, an indirect question denotes a kind of function, which takes intensions of question embedding verbs as arguments.

(2.17a.) *John remembers whether it is raining*.

(2.17b.) *If it is raining then John remembers that it is raining, and if it is not raining then John remembers that it is not raining*.

According to the contextual approach, *w*-clauses in general are ambiguous between a universal and an existential reading in the interrogative quantifier. That means that the sentence *John remembers who came* is equivalent to the sentence *Someone came and John remembers that he came*.

However, this kind of analysis cannot be applied to all predicates, which is admitted by several authors, e.g. (Karttunen 1977) or (Bäuerle/Zimmermann 1991). Not all predicates take *dass*-clauses as complements and for some of them, e.g. those that only allow for interrogatives, such paraphrase means something different. For instance, example (2.18a) does not have the same meaning as the corresponding sentences in (2.18b). There are two senses of *wonder* involved here. In (2.24a), *wonder* means "wish to know", in (2.18b) "be amazed at". In the first sense *wonder* allows only for interrogatives, in the second sense only declaratives. We analyse the interaction between the semantics of predicates and the clauses they subcategorise for in section 2.2.2.3 below.

(2.18a.) *John wonders whether it is raining*.

(2.18b.) *If it is raining then John wonders that it is raining, and if it is not raining then John wonders that it is not raining*.

Some authors differentiate between alternative *ob*-clauses and *w*-questions. For instance, (Karttunen 1977) uses the term *search questions* for the latter because semantically these questions involve a search for a suitable value for a variable. Alternative questions, which are introducted with the conjunction *ob*, can be considered as syntactically 'degenerate' alternative questions. The author claims that questions like *whether Mary cooks* come to be semantically equivalent questions like *whether Mary cooks or Mary doesn't cook* although they are syntactically generated by different rules.

[20] Cf. (Hamblin 1973).

(Karttunen 1977) states the difference in the semantics of these two interrogative clauses. For instance, example in (2.19a) denotes the set containing all true propositions expressed by sentences of the form "x dates Mary", and (2.19b) on the other hand, picks out the set containing all true propositions expressed by sentences of the form "x dates Mary" and "x doesn't date Mary". In other words, (2.19b) denotes a set, which contains for each person who dates Mary the proposition that he dates Mary, and for each person who doesn't date Mary the proposition that he doesn't date Mary.

(2.19a) *Who dates Marry.*

(2.19b) *Whether he0 dates Marry.*

We admit that there exist differences in the semantics of *w*-and *ob*-clauses. We follow the view that *w*-clauses denote a set of answers on the questions they express, whereas *ob*-clauses express *yes/no*-questions, which ask about the truth value of the proposition. However, we describe *w*-and *ob*-clauses in one category, i.e. interrogative clauses, as in our work, the analysis of these subclause types is related to the analysis of subcategorisation properties of predicates, which take declarative vs. interrogative clauses. Most predicates, which license *w*-clauses can also take *ob*-clauses and vice versa. Besides that, our approach has a more syntactic character and several authors, e.g. (Karttunen 1977) also mention that *w*- and *ob*-questions belong to one syntactic category.

2.2.2.3 Sentential complements vs. predicates: their semantics

As mentioned in 2.2.2.1 and 2.2.2.2 above, *dass*-clauses are closed propositions defined in their truth values, *ob*-clauses are open about the "positive" or "negative" value of the proposition and *w*-clauses are open sets of propositions. They can become closed propositions only after the with *w*- introduced gaps can be filled.

The compatibility of the declarative and interrogative clauses with different kinds of predicates depends on the semantics of both subclauses and that of their predicates (their selectional or sortal restrictions), which is admitted in a number of works on the semantics of subclauses, e.g. (Karttunen 1977), (Bäuerle/Zimmermann 1991), (Schwabe 2004), (Fischer 2005) and (Oppenrieder 2006). This coincides with the statement of (Schumacher 1986) who describes the relation between the semantics of verbal predicates and their choice for the complement types, as illustrated in examples (2.12a) and (2.12b) in section 2.2.1.3 above. (Vendler 1967) also admits that the choice for complements depends on the semantics of predicates, e.g. 'factive' predicates, in contrast to 'non-factive' ones, presuppose the truth of the subordinate clause. This assumption coincides with the classification of (Kiparsky/Kiparsky 1970) who argues that there are three classes of verb: factives, such as *know, find out, discover, notice, realise, remember, know, result*, half-factives, e.g. *tell, anticipate* and nonfactives, e.g. *think, believe, assume*.

Therefore, the semantic characteristics of sentential complements are very important to explain the subcategorisation behaviour of predicates that select them. Sentential complements can be combined with a variety of matrix verbs, which subcategorise for a certain subclause type depending on the verbal semantics. Several

authors (e.g. (Lewis 1970) and (Tichý 1978)) consider semantics of sentential complements as a part of lexical semantics of the verbs, which allow for them. For instance, Tichý claims that the difference between (2.20a) and (2.20b) is explained by the difference in meaning between the verbs *asserts* and *asks*.

(2.20a) *Tom asserts that Bill walks.*

(2.20b) *Tom asks whether Bill walks.*

In (Schwabe 2004), the author distinguishes between matrix predicates that have propositional arguments and those that have situational arguments. Predicates of the former group are the German verbs *glauben* ("to believe"), *wissen* ("to know") and *hoffen* ("to hope"). Predicates of the latter group are the German verbs *bedauern* ("to regret"), *wollen* ("to want") and *zeigen* '("to indicate").

However, this view is criticised by other authors, e.g. (Groenendijk/Stokhof 1984) or (Bäuerle/Zimmermann 1991), who admit that the lexicon can not be endless, whereas question sequences can be. The difference between sentences with declarative and interrogative complements cannot be explained just by the meaning of the verbs. Thus, (Groenendijk/Stokhof 1984) in examples (2.21a) and (2.21b) show that if *wissen* ("to know") had the same meaning in both sentences (which follows from Tichy's statement), then (2.27a) and (2.27b) would have the same meaning. However, it is not the case if Bill does not sleep and Tom knows that, which means that (2.21b) is true and (2.21a) false.

(2.21a) *Tom weiss, dass Bill schläft*. ("Tom knows that Bill is sleeping").

(2.21b) *Tom weiss, ob Bill schläft*. ("Tom knows whether Bill is sleeping").

This means that the semantics of some predicates allows them to subcategorise not only for one subclause type. For instance, according to (Fischer 2005), predicates that take sentential complements fall into three classes: those that subcategorise only for statements, those that allow only for questions and those that take both subclause types. This is also mentioned by (Karttunen 1977), who claims that some verbs, e.g. *sagen* ("to say"), allow not only for a set of propositions (interrogative clauses) but also for propositions themselves (declarative clauses), as seen in examples (2.22a) and (2.22b).

(2.22a.) *Erwin sagt, dass die Tagesschau pünktlich beginnt*.
("Erwin knows that the Tagesschau starts on time").

(2.22b.) *Erwin sagt, ob die Tagesschau pünktlich beginnt*.
("Erwin knows if the Tagesschau starts on time").

The explanation for the ability to take both, declarative and interrogative clauses could lie in the ambiguity of predicates – in one reading, they take *dass*-clauses in other readings they show preferences for *w-/ob*-clauses. This distinction is present in some lexical resources. For instance, the electronic dictionary *ELDIT*[21] provides different valency patterns for predicates, depending on their meaning. Thus, the

[21] We describe this dictionary in section 3.1.1.2 below.

verb *(sich) entscheiden* ("to decide"), in its meaning as 'resolving of a problem with one of the possible solutions which are available', can take *dass*-clauses. However, in its second reading *sich entscheiden* as 'to choose an option after a long consideration' this noun prefers *w-/ob*-clauses.

Some verbs allow for both interrogative and subordinate clauses, if they occur under certain contextual conditions only, e.g. with the changed modality (embedded under a modal verb), polarity (used with negation), mood (declarative vs. interrogative context) or others (e.g. the usage of a Korrelat). These parameters can change the truth values of the predicates and thus, allow them to take other types of complements.

Several authors, e.g. (Egli 1974), (Bäuerle/Zimmermann 1991), (Fischer 2005) and (Schwabe/Fittler 2009a) state the role of contextual conditions, which influence the occurrence of both, interrogative and declarative clauses with verbs. For instance, (Egli 1974) introduces the category of mood for German questions: *es ist so, dass* ("it is so that") which is equivalent to *ja* ("yes") and *es ist nicht so, dass* ("it is not so that") which is equivalent to *nein* ("no"). This shows that contextual properties should be taken into account while interpreting the semantics of sentences with declarative and interrogative subclauses. Semantics is correct for (2.23a) and (2.23b) but does not work in (2.23c), cf. (Bäuerle/Zimmermann 1991).

(2.23a) *Kommt Urs? Ja* $\equiv$ *Es ist so, dass Urs kommt*.
("Will Urs come? Yes $\equiv$ It is so that Urs will come").

(2.23b) *Kommt Urs? Nein* $\equiv$ *Es ist nicht so, dass Urs kommt*.
("Will Urs come? No $\equiv$ It is not so that Urs will come").

(2.23c) *Kommt Urs? Nein* $\neq$ *Es ist nicht so, dass Urs kommt*.
("Will Urs come? No $\neq$ It is not so that Urs will not come").

In (Fischer 2005), the author states contextual parameters for the German verb *zweifeln* ("to doubt"). It allows for *dass* and *ob*-clauses, but never appears with a *w*-clause in corpora. The author shows that *zweifeln* mostly subcategorises for an *ob*-clause when used positively (there is only one case of negated *zweifeln* out of 367 used with an *ob*-complement). Negated *zweifeln* can take only *dass*-clauses.

Contextual conditions of subclause-taking verbs are systematially described in (Schwabe/Fittler 2009a). The authors present semantic conditions determining question-embedding of German verbs. Besides that, in (Schwabe/Fittler 2009b)hey discusses particular logical consistency conditions satisfied by German proposition-embedding predicates. The authors analyse negative contexts as well as the usage of Korrelat. For instance, the verb *nachdenken* tends to take *w-/ob*-clauses with the Korrelat *darüber* only, cf. examples (2.24a) vs. (2.24b) and (2.24c).

(2.24a) *Er denkt, dass sie kommen.* ("He thinks that they will come").

(2.24b) *Er denkt darüber, wer kommt.* ("He thinks about who will come").

(2.24c) *Er denkt darüber, ob sie kommen.* ("He thinks about if they will come").

The electronic dictionary *EDLIT* also contains a number of contextual restrictions for sentential complements. For instance, the verb *sich entscheiden* allows *w-/ob*-clauses either without any Korrelat or with the Korrelat *darüber* ("for it"), but not with the Korrelat *dagegen* ("against it"), whereas *zu*-infinitives are allowed under both Korrelats (*darüber* and *dagegen*).

The contextual conditions for the choice of subclauses (or other complements) are included in the morpho-syntactic restrictions of argument in COMLEX, the features of which are described in section 3.1.2.2 below. (Ehrich 1991), , who analyse nominal predicates, points out that negated nominal predicates cannot express events as it is impossible to find time or place where the "non-occurrence" of the event can take place, cf. section 4.2.2 below.

Summarising different studies on the relations between subclauses and their predicates, we admit that both, the semantics of the matrix predicates and the semantics of sentential complements influence the compatibility of predicates with declarative vs. interrogative subclause types, and thus, have an impact on the subcategorisation behaviour of predicates. We follow the view that there exist three types of predicates: those which take declaratives only, those which take interrogatives only and those which allow for both subclause types, cf. our classification in section 4. However, we also agree that some predicates take both subclause types under certain conditions only, e.g. if the modality, polarity or mood is changed or if they are used with a Korrelat.

2.3 Summary: types of valency information related to this study

[Types of valency information related to this study] We summarise the main types of subcategorisation information analysed in the described studies above (e.g. valency patterns (or frames), grammatical functions (GF), case, syntactic categories (SCs), semantic roles (SemR) and selectional restrictions (SelR) in table 2.12.

	valency patterns & their realisation			
GF	subj	indir	dir	prep
case	nominative	dative & genitive	accusative	prep
SC	NP, pronoun, subclause	NP, pronoun	NP, pronoun, subclause	PP, subclause with a Korrelat
SemR	Agent, Patient, Experiencer, Theme, Instrument, Goal, etc.			
SR	human, animal, fact, event, etc.			

Table 2.12: Main types of subcategorisation information

As our aim is to automatically extract and classify different types of predicates according to their subcategorisation properties, we consider those aspects of valency description only, which are relevant to the elaborated architecture. We concentrate on the extraction of sentential complements only, thus, the description of further syntactic categories, e.g. NPs, as well as the analysis of indirect object are not relevant

here. As subclauses do not have any case markers, we do not consider this aspect within our procedures.

Our extraction and classification architecture is based on the comparison of subcategorisation properties of predicates. As we concentrate only on the predicates, which occur with subclauses (described in section 2.2 above) in our corpora, we restrict their valency patterns (VP)[22] to this, which include sentential complements. Besides that, for the purpose of classification we are interested in the type of subclauses these predicates can licence (declarative *dass*- or interrogative *w-/ob*-clauses, cf section 2.2.1). These clauses have all the three grammatical functions sentential complements can take in a sentence, cf. section 2.2.1.3. Therefore, analysing the related work on the description of valency of verbal, nominal and multiword predicates, we are interested in the following types of valency information: valency patterns, syntactic categories of the described predicates, as well as their grammatical function.

The decision of a predicate to allow for one of the mentioned subclause types is based on the sortal or selectional restrictions of verbs, which are presented on the semantic level of valency description[23], cf. section 2.2.2.3 above. In this study, we analyse predicates, which are compatible with clauses defined semantically as statements and/or questions, as described in 2.2.2. Therefore, of particular importance for this thesis is the description of selectional restrictions and partially semantic roles (we define them as SR in one category) in the related work on subcategorisation description. Thus, analysing subcategorisation of verbs, nouns and multiwords, we operate with categories defined on both, semantic and syntactic levels of valency description, as illustrated in table 2.13.

valency level & information type		**their relevance for the present study**
syntactic	**VP**	predicates whose valency patterns include subclauses
	SC	sentential complements, declarative *dass* and interrogative *w-/ob*-clauses
	GF	subject, object and prepositional clauses (although they are not defined within our architecture)
semantic	**VP**	predicates whose valency patterns include statements and/or questions
	SR	selectional restrictions of predicates, which allow for statements and/or questions

Table 2.13: Types of valency information and their relevance to our study

In this chapter we summarise the main categories of the phenomenon of valency, which are related to this study. However, we do not analyse the related work on subcategorisation of verbal, nominal or multiword predicates in detail. This kind of analysis follows in chapter 3 below.

[22]The concept of valency patterns belongs both to the syntactic and semantic levels of subcategorisation description, cf. section 2.1.3.1

[23]We also analyse the presentation of semantic roles (SemR), as this kind of information sometimes replaces the information on selectional restrictions.

Chapter 3

Subcategorisation of Verbs, Nouns and Multiwords

Subcategorisation features of different predicates have been an important research object in linguistics and lexicography. (Agel 2000) notes that there are more than 3000 publications dealing with valency theoretical problems. We do not aim at reviewing all the issues analysed in those studies. We only want to discuss the aspects relevant for the present research, cf. section 2.3 above. Concentrating on different types of predicates in our study, we refer to a number of studies on valency of these predicates, including linguistic theoretical work, subcategorisation dictionaries, as well as works on acquisition tools. Table 3.1 contains a list of works we will study in this chapter.

predicates	**linguistics**	**lexicography**	**NLP**
verbs	(Tesnière 1980), (Engel 1988), (Engel 1994), (Engel 1996), (Agel 2000), (Agel 2003), (Götz-Votteler 2007)	(Helbig/Schenkel 1969), (Engel/Schumacher 1976), (Schumacher 1986), (Herbst et al. 2004), (Schumacher 2004), *ELDIT*, *ADNW*, *DAFLES*	HPSG (Pollard/Sag 1994), LFG in (Kaplan/Bresnan 1982), (Bresnan 2001) and (Dalrymple 2001), COMLEX, (Merlo/Stevenson 2001)
nouns	(Teubert 1979), (Teubert 2003), (Ehrich 1991), (Agel 2003), (Schierholz 2005),	(Sommerfeldt/Schreiber 1983), (Sommerfeldt/Schreiber 1996), (Herbst et al. 2004)	(Crouch *et al* 2006), (Gurevich et al. 2007), NOMLEX, etc.
multiwords	(Krenn/Erbach 1994)	(Fellbaum et al 2006), (Heid 2006)	(Bartsch 2004), (Storrer 2007), (Lapshinova/Heid 2007)

Table 3.1: Predicate types and studies on them

Most authors concentrate on the analysis of verbal predicates. There are studies, which also describe subcategorisation properties of nominal predicates, for example, (Teubert 1979) or (Schierholz 2005), and multiword predicates, for instance, (Krenn/Erbach 1994) or (Fellbaum et al 2006), which are mostly discussed in connection with verbal predicates, i.e. when the properties of verbs are compared with the properties of their derivatives.

3.1 Subcategorisation of Verbs

Knowledge about verbs is important for most systems of automatic acquisition of lexical information. As admitted by (Tesnière 1959), (Helbig/Schenkel 1973), (Engel 1994) or (Engel 1996) and (Götz-Votteler 2007), verbs play a central role for the structure and the meaning of sentence and discourse and they are the primary source of relational information in a sentence – the predicate-argument structure that relates an action or state to its participants (i.e., who did what to whom).

Traditionally, verbal subcategorisation is described as the potential of verbs to choose their complements. For example, the verb *abwarten* ("to wait") in sentence (3.1a) can chose a direct object expressed by a *wh*-clause in addition to the obligatory subject expressed by the pronoun *wir*.

(3.1a) *Wir werden abwarten, wer sich bewirbt.*
("We will wait (for), who applies".) ('*Frankfurter Rundschau*')

(3.1b) **Wir werden ihm abwarten, wann er nach seinem schweren Unfall wieder auflaufen kann.*("We will wait him, when he can take to the field again after the accident") ('*Frankfurter Rundschau*')

(3.1c) *Wir werden ihm sagen, wer sich bewirbt*
("We will tell him, who applies".) ('*Frankfurter Rundschau*')

However, the verb *abwarten* cannot take an indirect object expressed by the pronoun in dative *ihm*, as in (3.1b). For a different verb, e.g. *sagen* ("to say"), the combination of these two complements (direct and indirect objects) is acceptable (cf. (3.1b) and (3.1b)). The analysis of different grammatical functions of complements and their syntactic categories is given in the section 2.1.3 above.

The current section describes the related work on subcategorisation features of verbal predicates. Most aspects of verbal valency analysed in linguistic work are already described in chapter 2. Therefore, we do not repeat these works and proceed with the description of valency categories (based on the list presented in table 2.13 in section 2.3), such as valency patterns, syntactic categories of complements, their grammatical functions, their semantic roles, and selectional restrictions of predicates in lexicography and NLP studies.

3.1.1 Verbal Predicates in Lexicographic Work

In this section, we analyse the description of the subcategorisation properties of verbs in two printed valency dictionaries – *VALBU*, cf. (Schumacher 2004) and *VDE*, cf. (Herbst et al. 2004), as well as several electronic dictionaries.

3.1.1.1 Verbal predicates in printed dictionaries

The valency dictionary *VALBU* is a didactically-oriented valency dictionary of German verbs and is based on the theoretical principles of the dictionary *Verben in Feldern* (*ViF*, cf. (Schumacher 1986)), as well as the grammar of (Engel 1988). This dictionary contains semantic and syntactic descriptions of verbs and their specific environment, as well as some information on morphology, word formation, the ability

of verbs to build passive forms, their phraseology and stylistics, and numerous usage examples. However, it does not provide any information on valency of nouns or multiwords.

The Valency Dictionary of English, (*VDE*, (Herbst et al. 2004)), provides valency description for different types of predicates in English, concentrating especially on verbal ones. The description of subcategorisation properties comprises statements about the quantitative valency of the lexical units established, an inventory of their complements, as well as systematic information on the semantic and collocational properties of the complements.

Valency patterns Valency patterns in *VALBU* are represented in terms of complementation patterns which represent the predicate-argument structure of the described verbs. Entries contain information on obligatory and optional complemets, as well as adjuncts. For instance, the verb *erfahren* ("to find out") in *VALBU* has a complementation or valency pattern containing two obligatory complements: NomE (subject) and AkkE (direct object), and one optional complement PräpE (prepositional object), as seen in example (3.2).

(3.2)

verb	**complementation pattern**
erfahren	NomE AkkE (PräpE)

VDE also contains a list of complementation patterns identified on the basis of the COBUILD/Birmingham corpus of English, cf. figure 3.1. Every entry includes information on the number of possible patterns (D1 to D4), on the possible usage as an avalent predicate (General:0), and besides, minimum and maximum of valency depending on the voice of predicate (Active: 1/3 and Passive:1/3). The letter *T* indicates the possibilty to be used as a trivalent predicate. Degree of obligatoriness of complements is indicated with *obl* (obligatory) or *cont* (contextually obligatory), ehich is not indicated in figure 3.1

discuss *verb*

	Active: 1/3	Passive: 1/3	General: 0
I	$[N]_A$/[by N]		
II	$[N]_P$	D1	T
	$[V\text{-}ing]_P$	D2	T
	$[wh\text{-}CL]_{P(it)}$	D3	
	$[wh\ to\text{-}INF]_{P(it)}$	D4	
III	[with N]		T

Figure 3.1: VDE entry of the verb *to discuss*

Grammatical functions Grammatical functions in *VALBU* are case-specified (like in a number of works on German grammar, e.g. (Engel 1991) and (Helbig/Buscha 2005)) and are assigned directly to the indicated complements (NomE, AkkE and PräpE in example (3.2)). Every entry contains are detailed specification of grammatical functions. For instance, for prepositional objects, the authors give the information on which preposition can be used with this verbal predicate.

VDE does not contain any information on grammatical functions of verbs. Patterns with prepositional phrases only contain the information about the prepositional usage of the complement.

Syntactic categories Although *VALBU* does not give any explicit information on syntactic categories in entries, it provides the information about subclauses. If a complement can have a sentential form, the entry contains the indicator SE (which stands for *SatzErgänzung* in German).

Syntactic categories of complements are expressed directly in the patterns, cf. figure 3.1. For instance, the verb *erfahren* can have noun phrases (N) or different kind of subclauses as its complements (V-ing, wh-CL, etc.).

Semantic roles and selectional restrictions The representation of semantic roles in *VALBU* is related to selectional restrictions of verbal predicates. Semantic roles are given in form of paraphrases, which are specified according to the verb meaning. For instance, the verb *aussprechen* ("to pronounce") will determine its subject as "someone who expresses or utters something" and its object as "something which is being expressed". Semantic roles in *VDE* are identified in verb entries in the preamble (I, II, III, etc.), and are linked to the their syntactic realisations (syntactic categories under the same Roman figure express the same semantic role), as well as illustrating them examples. Thus, the entries include various kinds of cross-references: the preamble links semantic roles to appropriate example sets and the meaning descriptions link components of the definitions to semantic roles, while associating senses to example sets. Although the dictionary does not contain any indicators for the selectional restrictions of verbs, they can be derived from the examples.

3.1.1.2 Verbal predicates in electronic dictionaries

In spite of the fact, that some new printed valency dictionaries, such as *VALBU, VDE* have appeared recently, more authors think of creating electronic versions of their works. The space, the capacity of computers and internet connection allow for the complete substitution of paper dictionaries with electronic ones.

In (Heid 2006), the author outlines a number of online valency dictionaries, describing their role and pecularities of the given information on valency. Electronic dictionaries aimed at second language learners provide the most comprehensive valency discription. Information on subcategorisation properties is included in *DAFLES* (*Dictionnaire Actif de Français Langue Etrangère ou Seconde*)[1], which is a monolingual learner dictionary of French for Dutch speakers, cf. (Selva et al. 2002). The bilingual *ELDIT* (*Elektronisches Lernerwörterbuch Deutsch/Italienisch*[2], "Electronic Learner's Dictionary for German/Italian") and *DNW* (*Deutsch-Niedersorbisches Wörterbuch*[3], "German-Lower Sorbian Dictionary") describe both languages in detail and contain detailed information on valency properties of lexical units.

Valency patterns The valency specification in all the above mentioned dictionaries provides the information on complementation patterns. For example, in *ELDIT*, the subcategorisation frames of the verb erfahren are given in examples of possible combinations of this verb with other words, cf. example (3.3). Optional complements and adjuncts are given in brackets.

[1] http://www.kuleuven.ac.be/dafles
[2] http://www.eurac.edu/eldit.htm
[3] http://www.dolnoserbski.de/dnw

(3.3)
1. *jemand* **erfährt** *(von jemandem) etwas*
   ("Smb learns sth (from smb)")
2. *jemand* **erfährt** *von einer Sache*
   ("Smb ears about sth")
3. *jemand* **erfährt** *(durch jemanden) etwas*
   ("Smb learns (from smb) sth")
4. *jemand* **erfährt** *(irgendwie) (von jemandem), dass*
   ("Smb learns (somehow) (from smb) that")

DNW gives a similar description of complements, whereas *DAFLES* applies structural set phrases, which give the information on the number and position of complements.

Grammatical functions All the above mentioned electronic dictionaries contain information on grammatical functions. In *ELDIT*, grammatical functions, such as subject,accusative, dative, genitive and prepositional objects, are linked directly to the verbal complements given in complementation patterns. or example, the subject of the verb *erfahren* is *jemand*, its accusative objects are *etwas* and the subclause is introduced by *dass*.

Syntactic categories Syntactic categories are contained in *DNW* and *DAFLES* but are not explicitly present in *ELDIT*. Each complement is illustrated with sentence examples from which the user can deduce the possible syntactic categories. For instance in sentence (3.4), which is an example for valency pattern 4 in example (3.3), the subject can be realised as nominal phrase, e.g. *die Flugsicherung*, the prepositional object – as prepositional phrase, e.g. *vom Piloten*, the accusative object – as declarative *dass*-clause.

(3.4) *Die Flugsicherung hat zu spät vom Piloten* **erfahren***, dass die Boeing Probleme beim Landeanflug hat.* ("The air traffic control found out too late from the pilot that the Boeing have problems during final descent").

ADNW provides a detailed description of complement variants, e.g. if it is expressed by a *dass-, w-, ob*-clause or a *zu*-infintive both for German and Lower Sorbian, and contains information on the absence or presence of Korrelats and lexically restricted complements.

Semantic roles and selectional restrictions The lexical description of complements in form of paraphrases, applied in *ELDIT* and *DNW*, allows for a certain semantic classification. Semantic information is more detailed in *ELDIT*, which contains more specific details about complement selectional restrictions, such as the information if the complement is animated or not, cf. (3.3). Besides that, *ELDIT* has a number of contextual constraints, which allow verbal predicates to take complements under certain contextual conditions only, cf. section 2.2.2.3.

DAFLES includes restrictions for verbal complements, they can denote action, place, result, etc.

3.1.1.3 Summary: verbal predicates in lexicography

In table 3.2 below, we summarise the types of subcategorisation information contained in the dictionaries described above (both, printed and electronic): valency patterns (VP), grammatical functions of complements (GF), their syntactic categories (SC), selectional restrictions (SelR) and semantic roles (SemR).

types of information	*VALBU*	*VDE*	*ELDIT*	*DAFLES*	*DNW*
VP	+	+	+	+	+
GF	+	-	+	+	+
SC		+	-	+	-
SelR	+	-	+	+	+
SemR		+	-	-	-

Table 3.2: Verbal valency in printed and electronic dictionaries

3.1.2 Verbal Predicates in NLP Work

All current symbolic approaches to syntactic analysis in NLP rely on valency data. Subcategorisation is represented in terms of syntactic and semantic features of complements. NLP-oriented grammatical theories suggest that the lexical description of predicates should go hand in hand with a grammatical rule that allows a system to derive a (syntactic or semantic) representation of a sentence, which singles out the predicate and its arguments.

3.1.2.1 Verbal predicates in formal grammars

We analyse the subcategorisation description of verbal predicates in formal grammars, Head-Driven Phrase Structure Grammar (HPSG) and Lexical Functional Grammar (LFG). Both of them are constraint-based.

Head-Driven Phrase Structure Grammar is a non-transformational phrase-structure based theory representing syntactic categories by feature structures. In this formal grammar, all linguistic knowledge is expressed in the feature structure format. The main principle of HPSG is the head-drivenness, which means that there exist head and non-head constituents and head features are shared by them. The stipulation of this sharing belongs to one of the main principles of HPSG, described by Pollard and Sag who state that the head value of any headed phrase is structure-shared, i.e. identical with the head value of the head daughter, cf. (Pollard/Sag 1994). That means that if the head is a lexical constituent and its complements are realised as sisters to its head, and the dependency structure is projective, the syntactic structure assumed by HPSG and Dependency Grammar can be considered only notational variants, both expressing the same linguistic facts (the head and the complements) and bearing the same features.

Lexical Functional Grammar (LFG)[4] has two interrelated syntactic representations: constituent structure (C-structure), which encodes details of surface syntactic

[4]See (Kaplan/Bresnan 1982), (Bresnan 2001) and (Dalrymple 2001) for details.

constituency, and functional structure (F-structure), which expresses abstract syntactic information about predicate-argument-modifier relations and certain morpho-syntactic properties such as tense, aspect and case. C-structure takes the form of phrase structure trees.

The level of F-structure is produced from functional annotations on the nodes of the C-structure and implemented in terms of recursive feature structures (attribute value matrices). For instance, the annotated F-structure for the sentence *John believes that Howard loves Sue* is shown in figure 3.2.

$$\begin{bmatrix} \text{PRED} & \text{'believe } \langle(\uparrow \text{SUBJ}),(\uparrow \text{COMP})\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'John'} \end{bmatrix} \\ \text{COMP} & \begin{bmatrix} \text{PRED} & \text{'love } \langle(\uparrow \text{SUBJ}),(\uparrow \text{OBJ})\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Howard'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'Sue'} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 3.2: F-structure for the sentence *John believes that Howard loves Sue.*

Valency patterns Valency or complementation patterns are represented in lexicon entries in both grammars.

In HPSG the local syntactic feature SUBCAT is the bearer of subcategorisation information. It is used to encode the various dependencies that exist between a lexical head and its complements. SUBCAT takes a list of (partially specified) synsems as its value. In a headed phrase the SUBCAT value of a daughter is the concatenation of the SUBCAT value of the phrase's *subcat* list with the list of the complement daughters (see figure 3.3). In HPSG the flow of subcategorisation information up projection paths is handled by the *Subcategorisation Principle*, which establishes that the SUBCAT value of a phrase is the SUBCAT value of the lexical head minus those specifications already satisfied by some constituent in the phrase, see (Pollard/Sag 1994) for details.

$$\left[\text{dtrs: } \left[\text{head\_dtr: } \left[\ldots\right]\right]\right] \Rightarrow$$
$$\left[\text{subcat: } \boxed{1} \text{ dtrs: } \left[\text{head\_dtr}\left[\text{subcat:}\boxed{1}+\boxed{2}\right] \text{ comp\_dtr } \boxed{2}\right]\right]$$

Figure 3.3: An example of HPSG's Subcategorisation Principle

Figure 3.4 illustrates the HPSG lexicon entry for the verb to *believe*, whose predicate-argument structure contains three elements: subcat <np1,s[comp]>.

An LFG entry also includes complementation patterns, expressed by the predicate-argument structure (PRED), which is the most important attribute of any lexeme. The predicate-argument structure (PRED) of the verb *focused* from the sentence in figure 3.2 shows that the valency pattern of the verb has two elements, cf. figure 3.5.

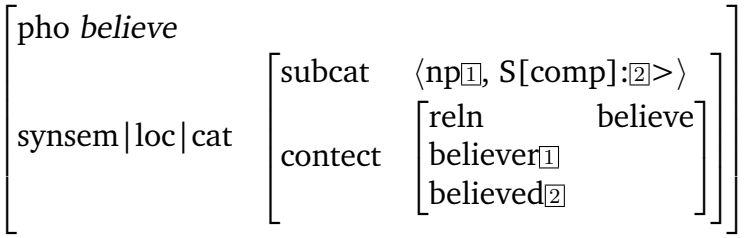


Figure 3.4: HPSG lexicon entry for the verb *to believe*

believe V (↑ PRED)= 'BELIEVE<(↑ SUBJ),(↑ COMP)>'
(↑ TENSE) = PRESENT

Figure 3.5: LFG lexicon entry for the verb *believe*

Grammatical functions In HPSG grammatical functions, such as subject, object, etc. are defined in terms of the order of the corresponding elements on the head's SUBCAT list. A more recent version of HPSG distinguishes between two features SUBJ (for subject) and COMPLS (for further complements), which were conflated in the former feature SUBCAT. Therefore, subject has its own list feature, whereas grammatical functions of further complements are defined as first object, second object, etc.[5] Order on this list corresponds to the traditional grammatical notion of obliqueness of grammatical relations, with more oblique elements occurring further to the left.

Grammatical functions (e.g. SUBJ, OBJ, or COMP) in LFG are the most important attributes of F-structures. LFG treats subcategorisation basically as a functional phenomenon – functors do not subcatgorise for categories but for grammatical functions. In (Bresnan 1982a), grammatical functions are classified according to two main parameters – subcategorisability and semantic restrictedness, cf. section 2.1.4.1 above. (Bresnan 1982a) also includes TOPIC and FOCUS into the category of grammatical functions. Their subcategorisability is subject to parametric variation and distinguishes between subject-oriented and topic-oriented languages. As we study subcategorisation features of verbs, which take subclauses, the grammatical function COMP is of particular importance for this thesis. In LFG grammtical functions are bound to the concepts of biuniqueness (which was already described in section 2.1.4.1 above), completeness and coherence (mentioned in 2.1.3.1 above). A local F-structure is only *complete* if it contains all the governable grammatical functions that it predicate governs, and an F-structure is *coherent* if all the governable grammatical functions that it contains are governed by a local predicate.

Syntactic categories In HPSG lexical dependencies involve category selection, which is achieved in the SUBCAT list specifications. Thus, in figure 3.4 the SUBCAT description for *believe* specifies that the category of its subject is an NP and the category of its object is a subclause.

[5] A similar approach is decribed for COMLEX, cf. table 2.2 in section 2.1.4.1

In LFG complementation patterns contain grammatical functions and not syntactic categories. However, syntactic categories are included into C-structures based on phrase structure rules. For instance, one of the phrase rules for the sentence in figure 3.2 is VP $\rightarrow$ V CP, where VP is a verbal phrase, V is a verb and CP is a complement clause.

Semantic roles and selectional restrictions Semantic roles in HPSG are included into the feature CONTENT, which is, together with SUBCAT, a part of the CAT decription, cf. figure 3.4. RRole assignment is the connection between the constituents of a utterance and the constituents of the topic the utterance is about, cf. believer and believed for the verb *to believe*. The roles present in the situation described by the verb have correspondence to the variables of grammatical functions in SUBCAT. Thus, the subject variable (first element of the SUBCAT list) unifies with the variable filling the believer role and the first object variable (second element of the SUBCAT list) unifies with the variable corresponding to the believed role. Therefore, grammatical functions in SUBCAT are restricted to certain semantic roles in CONTENT.

The assignment or mapping of grammatical functions to semantic roles in LFG also proceeds with selectional restrictions. Semantic roles are contained in the semantic form of predicates, e.g. the semantic form of the verb *to believe* has the semantic form believe(experiencer theme). The semantic role 'experiencer' is restricted to the subject of *to believe*, whereas the semantic role 'theme' is restricted to the object or to a complement clause, as in figure 3.5. LFG has a hierarchy of semantic roles, which includes the following roles in descending order: agent, beneficiary and maleficiary, recipient and experiencer, instrumental, patient and theme, locative, motive.

Besides that, to restrict grammatical functions to semantic roles, LFG includes further features [+/-r] (thematically restricted or unrestricted) and [+/-o] (objective or not) as follows. SUBJ and OBL are unrestricted [-r], whereas OBL_θ is unrestricted [+r], SUBJ is non-objective [-o], OBL is objective [+o] and OBL_θ can be both [+/-o].Semantic roles are associated with the specified grammatical functions according to several lexical mapping prinicples. For instance, according to the Intrinsic Role Classification, [-o] is assigned to the agent, [+o] is assigned to instrumental, patient and theme, locative, motive, whereas [-r] is assigned to all roles accept for the agent.

3.1.2.2 Verbal predicates in NLP-based dictionaries

A number of NLP-based dictionaries, such as FrameNet, WordNet, COMLEX, IMSLex, HagenLex, VALLEX and others provide detailed information on complement structure of verbs. We analyse the description of verbal subcategorisation presented in those dictionaries according to the types of information relevant for our study.

The Berkeley **FrameNet** (see section 2.1.3) is a lexical resource based on frame semantics and supported by corpus evidence. The aim is to combine both, semantic and syntactic valency. It contains about 10,000 lexical units (a pairing of a word with a meaning) and about 800 semantic frames, which are hierarchically related. Usually each sense of a polysemous word belongs to a different semantic frame.

WordNet WordNet is an online lexical reference system in which not only English verbs but also nouns and adjectives are organised into synonym sets, each represent-

ing one underlying lexical concept. Different relations link the synonym sets[6]. The special feature of WordNet is its attempt to organise lexical information in terms of word meanings, rather than word forms.

The **COMLEX** computational lexicon, described in (Grishman *et al.* 1994), provides detailed syntactic information for about 38.000 English words and includes 92 subcategorisation features of verbs.

IMSLex[7], a lexical resource comprising morphological and syntactic information for the German language. The purpose of IMSLes is to link together several lexical resources developed at the Institute for Natural Language Processing (IMS) of the University of Stuttgart.

HaGenLex (HAgen GErmaN LEXicon, cf. (Osswald 2004)) is a semantics-based computational lexicon for German developed at the Intelligent Information and Communication Systems (IICS) group of the FernUniversität in Hagen. HaGenLex contains detailed morpho-syntactic and semantic information. The lexical material of HaGenLex has been manually compiled on the basis of frequency lists and publicly available dictionaries. Verb entries constitute about 30 % of all the HaGenLex entries, the majority of which are nominal entries.

VALLEX 1.0, a valency lexicon for Czech verbs, Version 1.0[8], is a collection of linguistically annotated data and documentation, resulting from an attempt at formal description of valency frames of Czech verbs. It contains about 1400 verbs, which were collected (according to their number of occurrences in a part of the Czech National Corpus[9]). The set of verbs in VALLEX 1.0 is closed under the relation of "aspectual pair", see (Žabokrtský 2005).

Valency patterns For every verb FrameNet delivers a number of valency patterns, which contain the dependents of the verb. Valency patterns contain both, the information on the Frame Element (FE) this dependent belongs to, their syntactic realisations (SC) and grammatical functions (GF). For instance, the verb *believe* has three valency patterns in terms of Frame Elements of the arguments and each of them contains further complement patterns in terms of syntactic categories, cf. table 3.3. Thus, for instance pattern 1 contains 8 patterns of syntactic realisation, pattern 2 – 3 patterns, and pattern 3 only one pattern.

To represent valency patterns of verbs, WordNet includes for each verb synset one or several sentence frames, which specify the subcategorisation features. Examples of sentence frames for the verb *believe* are illustrated in (3.5).

(3.5) *Somebody* ***believes*** *something*
Somebody ***believes*** *somebody*
Somebody ***believes*** CLAUSE

In COMLEX, complements are formally defined by patterns which describe their constituent and grammatical structure (':cs' and ':gs), and examples (':ex'), as illus-

[6] A detailed description of WordNet is given in (Miller et al. 1990).
[7] Cf. (Lezius et al. 2000).
[8] http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0
[9] http://ucnk.ff.cuni.cz

believe				
1.	FE	Cogniser	Content	
	a.	NP	AVP	
		Ext	Dep	
	b.	PP[by]	NP	
		Dep	Dep	
	...	etc.		
2.	FE	Cogniser	Content	Content
	a.	CNI	NP	VPto
		–	Ext	Dep
	b.	CNI	NP	NP
		–	Ext	Obj
	c.	NP	NP	Sfin
		Ext	Obj	Dep
3.	FE	Cogniser	Content	Evidence
		NP	Sfin	NP
		Ext	Dep	Dep

Table 3.3: FrameNet valency pattern for the verb *to believe*

trated in figure 3.6. For example, the first frame 'np' represents a nominal complement, the second ('s') a full sentential complement with an optional *that* complementiser. Subjects are not specified in the argument list.

(vp-frame np	:cs ((np2))
	:gs (:subj 1 :obj2)
	:ex "John believed Howard")
(vp-frame s	:cs ((s 2: that-comp optional))
	:gs (:subj 1 :comp 2)
	:ex "John believed (that) Howard loves Sue")

Figure 3.6: Examples of COMLEX subcategorisation frames for the verb *believe*

Subcategorisation patterns contained in IMSLex are encoded in the TSNLP format, cf. (Estival et al. 1995). In table 3.4, we illustrate TSNLP format of two valency patterns for the verb *glauben* ("to believe"). The patterns can be mapped both, to macros that supply the verb's predicate-argument structure in LFG format, as well as to macros providing for other information relevant for parsing.

HaGenLex is a semantics-based lexicon, its entries include both, semantic and syntactic information, contained in valency patterns, illustrated in figure 3.7.

Each word entry in VALLEX consists of a non-empty sequence of pattern entries, typically corresponding to the individual meanings of the headword lemmas. Each pattern entry contains a description of the valency frame itself and of the frame attributes, such as functors, list of possible morphemic forms and type of complementation.

lemma	valency pattern	example
glauben	(subj(NP_nom), obj(NP_dat))	*John glaubt ihm nicht.* ("John does not believe/trust him")
glauben	(subj(NP_nom), s-comp(C_dass))	*John glaubt, dass Howard Sue liebt.* ("John believes that Howard loves Sue")

Table 3.4: An example of valency frames for the verb *glauben* in TSNLP

action, MENTAL

AGT	ORNT	MCONT
[LEGPER+]	[LEGPER+]	
np/nom	np/dat	*dass*-comp
	optional	optional

Figure 3.7: An example of the valency pattern for the verb *glauben* in HaGenLex

Grammatical functions The information on grammatical functions is available in all the above mentioned lexicons except for WordNet and HaGenLex[10], which are semantic-based systems. In FrameNet, COMLEX and IMSLex[11], grammatical functions are defined directly in subcategorisation patterns, cf. tables 3.3 and 3.4 and figure 3.6. However, the latter does not contain the subject argument on the level of constituent structure. In VALLEX, grammatical functions are defined within the description of the surface-syntax roles of every node, such as Pred – predicate, Sb – subject, Obj – object, Atr – attribute, Adv – adverbial etc. Besides that, VALLEX entries contain the information on complement case, such as nominative, accusative, instrumental, etc.

Syntactic categories Syntactic categories are included into valency patterns in FrameNet, COMLEX and ImsLex, as seen in tables 3.3 and 3.4 and figure 3.6. HaGenLex contains information on syntactic categories (e.g. np-acc, pp, etc.) in the extended version, which is represented in the form of attribute-value matrice. VALLEX and WordNet do not contain any information on syntactic categories of complements. However, WordNet provides information on the possibility of a clause, cf. example (3.5) above.

Semantic roles and selectional restrictions Semantic roles in FrameNet are expressed within frames, as already describe din section 2.1.5. Each frame provides its set of semantic roles. The verbs, which belong to this frame, share the same semantic roles. For instance, the verb *to believe* belongs to the frame 'Awareness', whose core semantic roles are *Cogniser, Content, Expressor and Topic*, and peripheral semantic roles – *Degree, Evidence, Manner, Role* and *Time*[12]. Besides that, frames contain se-

[10] A valency pattern of HaGenLex contains the feature OBJ, which stands for a neutral object, defined in terms of semantic roles, and not grammatical functions

[11] The description of the grammatical functions in FramNet and COMLEX is given in section 2.1.4.1 above. Grammatical functions of IMSLex are based on the LFG description.

[12] See table 2.1 in section 2.1.3.2 for the description of core and peripheral FEs.

mantic and syntactic restrictions for semantic roles. Thus, the semantic role *Cogniser* can be 'the person whose awareness of phenomena is at question', whereas *Content* is 'the object of the *Cogniser*'s awareness'. Valency patterns include restrictions for syntactic categories and grammatical functions of every semantic roles, cf. table 3.3.

As valency frames in WordNet are represented in terms of sentence frames, cf. example (3.5), this lexicon does not explicitly present semantic roles. However, it specifies whether the complements are animated or not, cf. somebody vs. something in example (3.5), and therefore contains selectional restrictions.

COMLEX and IMSLex do not include any information on semantic roles or semantic restrictions of predicates. However, COMLEX contains a number of morpho-syntactic restrictions, which allow predicates to take only certain type of predicates. These restrictions are introduced for encoding verbs, which occur with a given complement structure provided that certain morpho,syntactic conditions are met. The argument structure of these verbs is qualified by specification of some additional parameters. This feature is particularly interesting in the description of verbs, which take sentential complements. For instance, the fact that one can say *she didn't realise whether...* but cannot say **she realised whether...* is captured in COMLEX through the condition 'neg t' (i.e. 'negative = true'), cf. section 2.2.2 for the description of context parameters and their influence on the choice for a subclause).

Both HaGenLex and VALLEX contain semantic roles directly in valency patterns. Moreover, HaGenLex also includes semantic restrictions of arguments, such as *animal, animate, human*, etc. For instance, in the valency pattern in table 3.7, the restriction 'LEGPER' mmeans that both arguments tagged with this feature should be 'juridical or natural persons'. In VALLEX, this kind of information is missing.

3.1.2.3 Summary: verbal predicates in NLP

In table 3.5, we summarise the features of the above described formal grammars and NLP-based dictionaries according to the types of subcategorisation information, relevant for this study: valency patterns (VP), grammatical functions (GF), syntactic categories (SC), selectional restrictions (SelR) and semantic roles (SemR), as described for lexicographic work in table 3.2 in section 3.1.1.3 above.

types	*HPSG*	*LFG*	*FrameNet*	*WordNet*	*COMLEX*	*IMSLex*	*HaGenLex*	*VALLEX*
VP	+	+	+	+	+	+	+	+
GF	+	+	+	-	+	+	-	+
SC	+	+	+	-	+	+	+	-
SelR	+	+	+	+	+	-	+	-
SemR	+	+	+	-	-	-	+	+

Table 3.5: Verbal valency in formal grammars and NLP-based lexicons

3.2 Subcategorisation of Nouns

For NLP it is necessary to have subcategorisation information about different predicate types: verbal, nominal and multiword ones. Although the first works on valency,

starting with Tesnière[13], concentrated only on verbs, the later valency theory studies also describe nouns and adjectives as potential valency-bearers.

The current section describes the related work on subcategorisation features of nominal predicates[14]. Nominal predicates do not posses all the types of subcategorisation information verbal predicates do, e.g. the description of grammatical functions is missing[15]. However, the information on their subcategorisation features is important. Most valent nouns are derived from verbs or adjectives and their subcategorisation features can be deduced from their base verbs or adjectives[16]. We analyse works on the valency of deverbal nouns and its correspondences with the valency of their underlying verbs in section 3.4.2 below.

In the following sections we describe valency categories, such as valency patterns, syntactic categories of complements, their semantic roles or selectional restrictions of predicates, described in linguistic and lexicographic work, as well as NLP studies.

3.2.1 Nominal Predicates in Linguistic and Lexicographic Work

3.2.1.1 Linguistic studies

The descriptions of nominal valency differs significantly among the linguists. Various scholars, studying valency, admit the existence of nominal valency. However, there are also studies, which reject it. For instance, (Mackenzie 1997) claims, there exist functional reasons to propose that nouns and nominalisations should be analysed as avalent predicates. claims there exist functional reasons to propose that nouns and nominalisations should be analysed as avalent predicates. The sceptical view is also supported by Eisenberg, e.g. in (Eisenberg 1994). Nevertheless, every modern valency dictionary, if not systematically, describes at least the existence of subcategorisation properties of nouns. The reason for it is that grammarians, describing subcategorisation phenomena, are mostly interested in regularities. Therefore, they describe nominal valency as a secondary phenomenon, which can be deduced from the verbal one. Lexicographers, contrary to grammarians, describe nominal subcategorisation as idiosyncratic morpho-syntactic and semantic features of nominal predicates, cf. (Teubert 2003).

Valency patterns W. Teubert claims that subcategorisation phenomena, which he believes to belong to both, lexicon and syntax, describe features of valent words including not only verbs but also nouns, see (Teubert 2003). He describes nominal valency within a notion of nominal complex, cf. (Teubert 1979). A nominal complex is a syntactic construction, which houses the realisation of nominal subcategorisation. It consists of the reference group and denominal complements. Teubert mentions that the categorial description of a noun includes not only the indication of case, gender and number, but also nominal complements, as shown in figure 3.8.

[13] Cf. (Tesnière 1959).

[14] In this section we describe three properties of simplex noun predicates only, as compound nominal predicates are analysed within the phenomenon of subcategorisation "inheritance".

[15] Some authors describe grammatical functions of nominalisations based on the grammatical functions of their base verbs.

[16] Cf. (Teubert 2003).

Nom(case, gen, num) + WNom(X1,...,Xn)

Figure 3.8: Subcategorisation pattern of a noun in (Teubert 1979)

Valent nominal predicates in German can have up to three compliments and all of them are optional. (Teubert 1979) admits that most nouns are avalent, which means that they do not open any argument positions. However, such nouns can still have adjuncts.

Syntactic categories Syntactically noun compliments can be expressed by nominal and prepositional phrases or infinitive and sentential clauses.[17]. Nominal phrases can occupy the position both, before and after nominal predicatea, prepositional phrases, infinitives and sentential compliments can be only after a nominal predicate.

Although the number of syntactic categories of complements is limited, the syntactic realisation of nominal arguments in German can remain functionally ambiguous. For instance, the German nominal phrase in genitive can express both, an agentive complement, e.g. *die Ermittlung der Polizei* ("investigation of the police"), and a possessive attribute, e.g. *die Pistole des Polzisten* ("the gun of the policeman"). Moreover, one complement can be realised by different syntactic categories. Both the subclause *wann er ankommt* ("when he arrives") and the prepositional phrase *nach seiner Ankunft* ("about his arrival") express the same complement of the nominal predicate *ihre Frage* ("her question"), cf. *ihre Frage, wann er kommt* ("her question when he arrives") vs. *ihre Frage nach seiner Ankunft* ("her question about his arrival").

Semantic roles and selectional restrictions Most studies mention neither semantic features of nominal predicates, nor selectional restrictions of nominal predicates. (Teubert 2003) describes different noun classes according to semantic features of complements they take (agent, object, etc.). For instance, the noun *Ermittlung* ("investigation") can subcategorise for an agentive complement (*Agentivergänzung*), e.g. *Polizei* ("police"): *die Ermittlung der Polizei* ("investigation of the police"), whereas the nominal predicate *Vorrat* ("reserve/supply") can subcategorise for an objective complement (*Sachergänzung*) only, e.g. *Erdöl* ("oil"): *der Vorrat an Erdöl* ("the supply of oil").

3.2.1.2 Nominal predicates in lexicography

In this section we analyse the description of nominal predicates in two printed dictionaries – the *Wörterbuch zur Valenz und Distribution der Substantive*[18] (we use the abbreviation *WVDS* in the following), cf. (Sommerfeldt/Schreiber 1983), and the

[17]Syntactic categories in German are described in section 2.1.3

[18]Translated as "Dictionary of Valency and Distribution of Nouns".

above mentioned *VDE*, cf. (Herbst et al. 2004), and in the electronic dictionary *EL-DIT*, described in 3.1.1.2.

WVDS was the first dictionary, which described subcategorisation properties of nouns that are represented both, on the semantic and on syntactic levels. The authors analyse valency features of nominals according to a number of classes nouns can be classified to (this classification is described in section 4.2.2 below). They state that nominals, which belong to different semantic groups can also have different subcategorisation features.

Valency patterns In figure 3.9, we give an example of the *WVDS* entry for the noun *Beurteilung* ("apprasal/evaluation"). This entry contains two complementation patterns for *Beurteilung*, indicating the number of complements (two complements in the first pattern and one complement in the second), their syntactic and semantic features.

Beurteilung = 'Einschätzung'
1.1. → (2)
1.2. → Sg, pS (durch)
1.3. → fest: Sg + pS (durch)
(die Beurteilung *der Leistungen durch die Jury*)
2. Sg→ 1.± Anim
(die Beurteilung *der Prüfung/Pferde/Pflanzen/Geräte*)
2. Abstr
(die Beurteilung *der Vorträge/Aussprache*)
p = durch
S→ Hum
(die Beurteilung ... durch *den Lehrer/Kommission/Jury*)
Anm.: Das Substantiv Beurteilung bezeichnet auch das Schriftstück, das die Einschätzung enthält: *Die Beurteilungen müssen bis Montag abgegeben werden.*

Figure 3.9: Entry from (Sommerfeldt/Schreiber 1983) for the noun *Beurteilung*

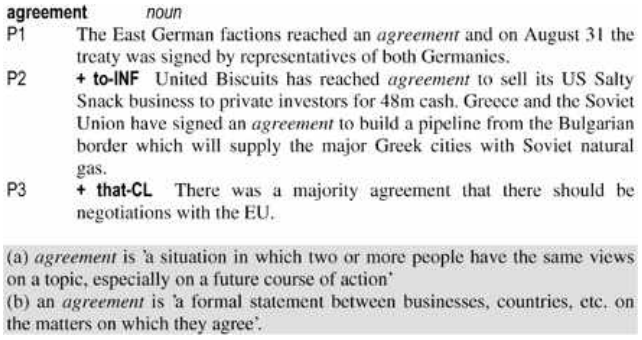
Sommerfeldt and Schreiber agree wwith W. Teubert that most nouns have optional complements. However, they admit that the optionality should be considered from the grammatical point of view. The deletion of an optional complement does not lead to an ungrammatical sentence but can lead to another meaning (cf. (3.6a) and (3.6b)).

(3.6a) *Er ist Vertreter seines Landes in der UNO* ("He is the representative of his country by the UNO").

(3.6b) *Er ist Vertreter* ("He is a representative") = this is his profession.

The Valency Dictionary of English, *VDE* (see (Herbst et al. 2004) and section 3.1.1), provides the valency description not only for English verbs but also for nouns. Noun entries in *VDE* are limited to the pattern-sorted examples block and the meaning description. In figure 3.10, we illustrate examples of valency patterns for the

noun *agreement*, which are indicated with P1, P2, P3 and illustrated by sentence examples, cf. figure 3.10



agreement *noun*

P1 The East German factions reached an *agreement* and on August 31 the treaty was signed by representatives of both Germanies.

P2 **+ to-INF** United Biscuits has reached *agreement* to sell its US Salty Snack business to private investors for 48m cash. Greece and the Soviet Union have signed an *agreement* to build a pipeline from the Bulgarian border which will supply the major Greek cities with Soviet natural gas.

P3 **+ that-CL** There was a majority agreement that there should be negotiations with the EU.

(a) *agreement* is 'a situation in which two or more people have the same views on a topic, especially on a future course of action'
(b) an *agreement* is 'a formal statement between businesses, countries, etc. on the matters on which they agree'.

Figure 3.10: VDE entry of the noun *agreement*

The electronic dictionary *ELDIT* does not give systematic specification for nominal complementation patterns, as it does for verbal predicates. However, it contains lists of possible combinations of nouns, with other words including nominal complements (besides complements it also includes various collocations that contain these nouns). For example, the entry for the noun *Glaube* ("belief") contains the phrase *der Glaube an jemdn./etw.* ("the belief in smb/sth"), which indicates the prepositional complement of this noun. For the noun *Frage* in the meaning "problem", a restriction for a genitive nominal phrase (*Frage*+Genitive) is enclosed directly in the entry. Besides that, the list of combinations of *Frage* with further words contains the expression *es bleibt die Frage, ob,* which allows to state that this noun can have an interrogative sentential complement.

Syntactic categories In *WVDS* nominal compliments can have the following syntactic categories – nominal phrases in in genitive (Sg in figure 3.9), prepositonal pharses (pS in figure 3.9), possessive pronouns, relative adjectives, infinitives and subordinate clauses.

Syntactic categories in *VDE* are given for every pattern, and contain additional information on their types, e.g. kind of clause, cf. *to*-INF or *that*-CL in figure, kind of preposition, e.g. *by* N, and on morpho-syntatic features of complements, e.g. *between* N*pl*/N *and* N.

ELDIT does not provide any information on the syntactic categories of complements of nominal predicates.

Semantic roles and selectional restrictions Semantic information is expressed in *WVDS* in form of selectional restrictions, which specify whether the semantic features of complements, e.g. Hum – human, Anim – animated, Abstr –abstract in figure 3.9.

The authors also state that sentential complements subcategorised by nominal predicates can either describe the "content" of the event notion (mostly embedded by bivalent deverbals), as in (3.7a) or the content of features and states (mostly embedded by adjectivals), as in (3.7b).

(3.7a) *seine Behauptung, dass die Antwort stimme.*
("his thought that the answer is right").

(3.7b) *der Stolz darauf, dass er eine so gute Prüfung abgelegt hat.*
("his pride about that he has such a good result in the exam").

VDE does not contain any information on semantic feature of nominal complements. However, the indication of some morpho-syntatic features restricts the choice for a certain type of complements. For instance, one of the valency patterns of the noun *agreement* contains the phrase introduced by the preposition *among(st)*. The noun, which follows this preposition should have a plural form or belong to nominals describing groups – *among(st)* N$pl/group$. This means that the noun *agreement* can take prepositional phrases introduced by *among* with plural nouns or nouns, which indicate a group, as shown in example (3.8).

(3.8) *After two hours of delibiration, the council president failed to find agreement* ***among the 15 members****.*

Selectional restrictions for animated and non-animated complements in *ELDIT* are expressed with the words *jemand* ("somebody") and *etwas* ("something").

3.2.1.3 Summary: nominals in linguistics and lexicography

In table 3.6, we summarise the above described linguistic and lexicographic work according to the types of subcategorisation information relevant for the analysis of nominal predicates: valency patterns (VP), syntactic categories (SC) and selectional restrictions or semantic roles (SR).

types	*Teubert*	*WVDS*	*VDE*	*ELDIT*
VP	+	+	+	+/-
SC	+	+	+	-
SR	+	+	+	+

Table 3.6: Nominal valency in lingistic and lexicographic work

3.2.2 Nominal Predicates in NLP Work

Some NLP studies also describe subcategorisation of nominal predicates. However, most works in NLP concentrate on the analysis of deverbal nouns, which behave similarly to their underlying verb. In the current section, we describe the presentation of nominal subcategorisation features in formal grammars and some NLP-based dictionaries.

3.2.2.1 Nominal predicates in formal grammars

We analyse the description of nominal valency in two fromal grammars – HPSG and LFG[19].

[19]The main features of these grammars are described in section 3.1.2.1 above.

Valency patterns In the HPSG framework subcategorisation patterns for nouns are also expressed in the feature SUBCAT, cf. the description of SUBCAT for verbal valency in section 3.1.2.1 above. In HPSG subcategorisation features of nouns are applied to build a nominal phrase whose head is a noun: NP → DP N'. To build a nominal phrase a noun or a nominal complex (saturated) needs a determiner, cf. (Pollard/Sag 1994), thus, the SUBCAT of a nominal phrase containing a noun contains the structure for a determiner phrase DP, as shown in figure 3.11. The saturated N' can be either a single noun or a combination of a noun with further elements, e.g. phrases in genitive, such as *Entscheidung der Frau*, or prepositional phrases, such as *Entschediung gegen mich*, etc.

$$\begin{bmatrix} \text{SYNSEM}|\text{LOC}|\text{CAT} & \begin{bmatrix} \text{HEAD} & \textit{noun} \\ \text{SUBCAT} & \langle \text{DP} \rangle \end{bmatrix} \end{bmatrix}$$

Figure 3.11: HPSG entry for a noun

LFG also contains subcategorisation patterns for nouns. However, only subclauses are seen as complements (see figure 3.12), prepositional complements and gentitive nominal phrases are represented as adjuncts, cf. figures 3.13 and 3.14[20].

$$\begin{bmatrix}
\text{PRED} & \text{'Ermittlung } \langle [\boxed{1}\text{:stimmen}] \rangle\text{'} \\
\text{SPEC} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'die'} \\ \text{DET-TYPE} & \text{def} \end{bmatrix} \end{bmatrix} \\
\text{NTYPE} & \begin{bmatrix} \text{NSYN} & \text{common} \end{bmatrix} \\
\text{COMP1 } \boxed{1} & \left\{ \begin{bmatrix}
\text{PRED} & \text{'stimmen } \langle [\boxed{2}\text{:Antwort}] \rangle\text{'} \\
\text{TNS-ASP} & \begin{bmatrix} \text{TENSE pres, MOD indicative} \end{bmatrix} \\
\text{SUBJ} & \boxed{2} \begin{bmatrix}
\text{PRED} & \text{'Antwort'} \\
\text{SPEC} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'die'} \\ \text{DET-TYPE} & \text{def} \end{bmatrix} \end{bmatrix} \\
\text{NTYPE} & \begin{bmatrix} \text{NSYN} & \begin{bmatrix} \text{COMMON} & \text{mass} \end{bmatrix} \\ \text{PERS} & \text{3, NUM sg,} \\ \text{GEND} & \text{fem, CASE nom} \end{bmatrix}
\end{bmatrix} \\
\text{VTYPE} & \text{main, PASSIVE -,} \\
\text{COMP-FORM} & \text{ob, CLAUSE-TYPE int}
\end{bmatrix} \right\} \\
\text{STMT-TYPE} & \text{header, PERS 3, NUM sg, GEND fem, CASE nom}
\end{bmatrix}$$

Figure 3.12: F-structure for the phrase *die Ermittlung, ob die Antwort stimmt.*

Syntactic categories Syntactic categories of nominal complements are expressed in both formal grammars. In HPSG nominal complements can be prepositional or genitive phrases as well as determiners, which are required to build a nominal phrase.

[20] All the three F-structures were generated with the XLE-Web, web-based tool for parsing with LFG, http://decentius.aksis.uib.no/logon/xle.xml

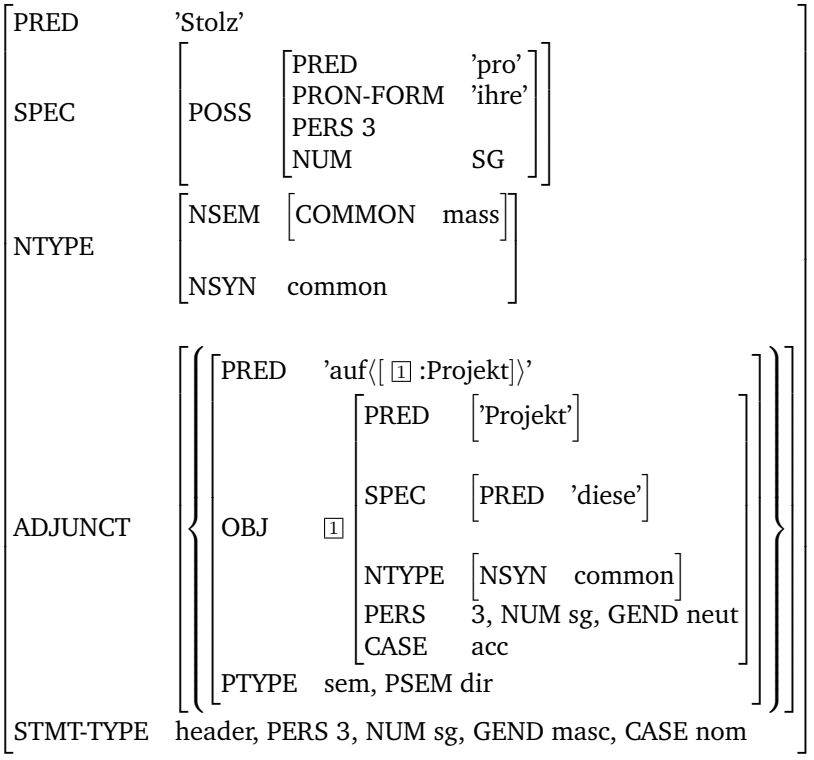


Figure 3.13: F-structure for the phrase *sein Stolz auf dieses Projekt*

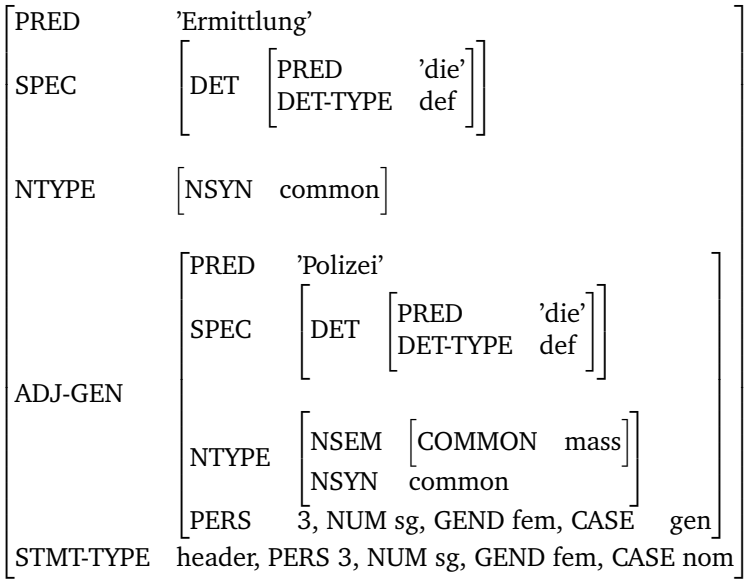


Figure 3.14: F-structure for the phrase *die Ermittlung der Polizei*

As mentioned above, nominal complements in LFG can be realised only as subclauses. The subcategorised *dass-* or *w-*clauses of nominal predicates (*die Frage, warum...* ("the question why"), *die Ermittlung, dass* ("the investigation that"), etc.) are defined as COMP or COMP1. In figure 3.12 the subclause *ob die Antwort stimmt* ("wether the answer is correct") is subcategorised by the noun *Ermittlung* ("investigation"). COMP-FORM gives the information about the subclause form (*dass-, w-* or *ob*-clauses).

Prepositional and genitive phrases are defined as adjuncts ADJUNCT or ADJ-GEN in LFG, cf. figures 3.13 and 3.14 above.

Semantic roles and selectional restrictions Both HPSG and LFG apply semantic restrictions for the description of nominal predicates. For instance, in a HPSG attribute-value matrice for the expression *Entscheidung des Kandidaten* ("decision of the candidate"), the restrictions contain the semantic role 'THEMA' for the noun *Kandidat*, which is restricted into the relation of possession with the noun *Entscheidung*.

In LFG nouns are marked with selectional features, classes of the sort hierarchy – a noun is annotated with the class(es) it belongs to. A specification of selectional features for a noun consists either of one sort or of several sorts, which are connected by logical AND or logical OR. Additionally, the feature COMP-TYPE provides information about subclause type (declarative vs. interrogative).

3.2.2.2 Nominal predicates in NLP-based dictionaries

Subcategorisation of nouns is also described in several NLP-based lexicons. In this section we analyse the description of nominal subcategorisation contained in FrameNet, IMSLex, HaGenLex, COMLEX, NOMLEX[21] and STO[22].

The latter lexicon, STO (SprogTeknologisk Ordbase), is a Danish lexicon resource for language technology applications. NOMLEX is a computational lexicon of nominalisations created on the basis of COMLEX, cf. (Macleod *et al.* 1998b), which describes different types of deverbal nominalisations, including those whose relationships to the base verbs are predictable and those, which are believed to be lexicalised. In NOMLEX the argument structures of deverbal nouns are mapped onto those of their base verbs. However, nominalisations in NOMLEX are still treated as nouns, therefore their argument structures are related but differ from the verbal ones. A more detailed description of the relations between predicate-argument structures of nouns and verbs represented in NOMLEX is given in section 3.4.2 below.

Valency patterns Valency patterns of nominal predicates in FrameNet have the same structure as those of verbal ones. For instance, the noun *belief* (which belongs to the same frame together with the verb *to believe*) has two valency patterns in terms of Frame Elements and each of them contains further complement patterns in terms of syntactic categories, cf. table 3.7. The first pattern contains 9 patterns of syntactic realisation, the second one – 8 patterns (we list just several examples of these patterns).

[21] http://nlp.cs.nyu.edu/nomlex/index.html
[22] http://www.cst.dk/sto/uk/

belief			
1.	FE	Cogniser	Content
	a.	AJP	Sfin
		Dep	Dep
	b.	CNI	PP-ing[about]
		–	Dep
	c.	CNI	Sfin
		–	Dep
	...	etc.	
2.	FE	Cogniser	Topic
	a.	CNI	PP[about]
		–	Dep
	b.	PP[of]	DNI
		Dep	–
	b.	PP[of]	N
		Dep	Dep
	...	etc.	

Table 3.7: FrameNet valency pattern for the noun *belief*

HaGenLex contains about 13.00 lexicon entries for nouns, subcategorisation patterns for which are represented in the same way as verbal valency patterns. Figure 3.15 illustrates an example of a complementation pattern for the noun *Glaube* ("belief"), for instance, in the expression *Der Glaube der Menschen, dass* ("the belief of people that").

MENTAL

ORNT	MCONT
[LEGPER+]	
np/gen	*dass*-comp
optional	optional

Figure 3.15: An example of the valency pattern for the noun *Glaube* in HaGenLex

IMSLex also contains information on nominal valency patterns, which are also encoded in TSNLP format, as shown in table 3.8, which illustrates that the noun *Glaube* ("belief") can have a number of valency patterns.

Nominal complement frames in COMLEX are defined by patterns, which describe their constituent and grammatical structure (':cs' and ':gs'), and examples (':ex') (cf. section 3.1.2.2 for the description of verbal comlements), as illustrated in figure 3.16. The first frame represents a sentential complement introduced by *that*, the second – are *to*-infinitive.

Some nouns also have the attribute ':feature', which describes their semantic features, e.g. *countable, collective, human, time*, etc.

The valency patterns for nouns in NOMLEX are classified into deverbal (VERB-SUBC) and ordinary noun complements (NOUN-SUBC). For instance, the subcategorisation of the noun *announcement* can be represented by several valency patterns.

lemma	valency pattern	example
Glaube	(pp-obj(P_an))	*Viele Menschen haben den Glauben an die Gerechtigkeit verloren.* (“Many people have lost the belief in justice”).
Glaube	(corr(an_acc), s-comp(C_dass))	*Viele Menschen haben den Glauben daran verloren, dass die Gerechtigketi existiert.* (“Many people have lost the belief (in) that justice exists”).
Glaube	(s-comp(C_dass))	*Den Glauben, dass die Gerechtigkeit existiert, haben viele Menschen verloren.* (“The belief that justice exists, many people have lost”).

Table 3.8: Subcategorisation information for the noun *Glaube* in IMSLex

(np-frame noun-that-s	:cs (:head (NOUN1) :post-modifier(s 2 :that-comp required))
	:gs (:head 1 :comp2)
	:ex “the plan that he will go there”)
(np-frame noun-to-inf	:cs (head (NOUN1) :post-modifier(vp 2 :mood to-inf :subj anyone))
	:gs (:subj 1 :comp 2)
	:ex “the plan to go there”)

Figure 3.16: Examples of COMLEX subcategorisation frames for the noun *plan*

We illustrate two of them in table 3.9. The prepositional complement introduced by *of* is deverbal, whereas the prepositional compliment with *about* is non-deverbal.

Pattern 1 (NOUN-SUBC (NOUN-PP:PVAL(“about”)))	
noun	*the announcement about the war*
base verb	–
Pattern 2 (VERB-SUBC(NOM-NP:OBJECT((DET-POSS)(PP-OF))))	
noun	*the announcement of the war*
base verb	*someone announced the war*

Table 3.9: Compliment patterns for the noun *announcement* in NOMLEX

Nominal subcategorisation patterns in STO are encoded with the help of letters and digits, e.g. Dn2GPn-med, where ’Dn’ means the start of the lexical description for a noun, ’2’ - the number of arguments this noun can has. The letter ’G’ stands for genitive, ’Pn’ – for a preposition complement, which is governed by a nominal phrase ’n’, ’med’ – for the preposition *med* ("with/by"). The authors sometimes encode modifiers along with the argument, if they occur frequently and are part of the ’core meaning’ of the noun.

Syntactic categories All the above mentioned lexicons contain information on the syntactic categories of nominal complements directly in valency patterns, cf. tables

3.7, 3.9 and 3.8, and figures 3.16 and 3.15, and the above mentioned pattern in STO. In most cases nominal complements can be realised as prepositional complements, genitive nominal phrases (in German) and sentential complements. Some lexicons, e.g. STO, also include modifiers as nominal complements.

Semantic roles and selectional restrictions Semantic information is contained in FrameNet, NOMLEX, HaGenLex and partially in STO. Nominal predicates in FrameNet have the semantic roles (frame elements). which are included into the frame they belong to (e.g. the noun *belief* belongs to the frame 'Awareness'). Furthermore, frames contain semantic and syntactic restrictions for semantic roles, cf. section 3.1.2.2 above. HaGenLex also incldes information on semantic restrictions mentioned in section 3.1.2.2 above.

COMLEX provides information on the semantic types of nominal predicates but does not give any information on the semantics of their complements. NOMLEX contains selectional restrictions for the complements of nouns, which are expressed with the features SUBJ-ATTRIBUTE or OBJ-ATTRIBUTE, whose values can be HUMAN, COMMUNICATOR, LOCATION, etc. Moreover, NOMLEX provides information on sortal readings of nouns. For instance, the noun *announcement* illustrated in table 3.9 belongs, according to NOMLEX, to the semantic type 'result'.

IMSLex does not contain any semantic information. Semantic component in IMSLex describes just the semantic type of proper nouns. However, IMSLex includes such morpho-syntactic restrictions as the presence of Korrelat, cf. (corr(an_acc)) in table 3.8 above.

3.2.2.3 Summary: nominals in NLP

In sections 3.2.2.1 and 3.2.2.2 above we analyse different NLP approaches on the description of nominal subcategorisation. We summarise these approaches in table 3.10, indicating which types of subcategorisation information are present in the above described studies (the relevant types of information are definbed in section 3.2.1.3) above.

types	*HPSG*	*LFG*	*FrameNet*	*IMSLex*	*HaGenLex*	*COMLEX*	*NOMLEX*	*STO*
VP	+	+	+	+	+	+	+	+
SC	+	+	+	+	+	+	+	+
SR	+	+	+	-	+	+/-	+	+

Table 3.10: Nominal valency in NLP work

3.3 Subcategorisation of Multiword Expressions

In this thesis, we analyse subcategroisation properties of not only simplex predicates (verbs and nouns) but also complex ones, which include more than one constituent, e.g. compound nouns and noun+verb-multiwords. Commonly, valency properties of multiword and compound predicates are described within the problem of the identification of valency bearer – which constituent defines the valency properties of the

whole construction. Therefore, we describe the properties of compound nouns and multiword expressions within the analysis of the "inheritance" phenomenon in section 3.4 below.

However, the subcategorisation properties of multiword expressions are also analysed within the general description of valency information, which includes the description of their valency patterns, the grammatical functions (if available) and syntactic categories of their complements, as well as their semantic roles and selectional restrictions, cf. criteria for the description of verbs and nouns in sections 3.1 and 3.2 above.

We concentrate on the analysis of multiword expressions consisting of a preposition, a noun and a support verb[23], which are commonly called support verb constructions (SVCs)[24]. However, some of the multiword predicates under analysis do not fall into the category of support verb constructions because of their idiomaticity. There exists a number of terms for the description of the phenomena under analysis. Phraseological research distinguishes between collocations and idioms, sometimes with an intermediate category of partial idioms, cf. (Burger 1998), or with a subclassification of collocations into transparent vs. opaque ones, described by (Grossmann/Tutin 2003). Collocations are assumed to include support verb constructions, cf. (Storrer 2007), (Krenn 2000) and others. As the borderline between these categories is not a clearcut one (cf. the detailed discussion of the state of art about collocations and their "neighbours" in (Bartsch 2004)), we will use the term 'mutiword expression' (MWE) in the following to refer to the targeted class of items in a general way, the terms support verb construction (SVC) and idiom to refer to the common classificatory intuition.

Most linguistic and lexicographic studies do not describe subcategorisation of multiword expressions. This information is important for NLP applications, therefore there exist more approaches on the description of multiword valency among NLP studies. In the current section we summarise the related work on subcategorisation features of multiwords[25]. The related studies are analysed according to the types of subcategorisation information described in 3.1, which are based on the categories listed in table 2.13 in section 2.3.

3.3.1 Multiword Predicates in Linguistic and Lexicographic Work

The following section describes subcategorisation features of multiword presented in linguistic and lexicographic work.

3.3.1.1 Multiwords in linguistic studies

A number of phraseological studies, e.g. (Wojtak 1992), (Keil 1997), (Burger 1998), (Engel 2004) and (Wotjak/Heine) state that multiwords have syntactic functions of

[23]For the analysis of "inheritance" relations between verbs and their derivatived, we also analyse nominalisation+support verb multiwords, which are not taken into account in the general analysis of multiwords in this work (their extraction and classification).

[24]See the definition of support verb construction introduced by (Heringer 1968).

[25]Although we concentrate on the analysis of preposition+noun+verb multiwords, we analyse works, which describe subcategorisation of different multiword types.

predicates. Subcategorisation of multiword expressions is one of the aspects, analysed related to the syntactic idiosyncracy of multiwords.

Valency patterns In (Burger 1998) and (Keil 1997), the authors claim that idioms are able to have argument places, which can or must be filled. They analyse deverbal idioms and state that the whole idiom is a complex verbal lexeme. The authors describe internal and external valency of an idiom. Internal arguments of a verbal idiom are the verb arguments that are constituent parts of the idiom, e.g. the prepositional phrase *auf die Palme* ("at the palm") in the idiom *jmdn auf die Palme bringen* ("to drive sb crazy") in example (3.9) bellow. The words *Ute* and *Jens* are external arguments of the idiom.

(3.9) *Ute bringt Jens auf die Palme.* ("Ute drives Jens crazy").

(Wojtak 1992) describes valency patterns of some multiwords, i.e. those which have somatic constituents or those describing clothes, e.g *Sand in die Augen streuen* ("to throw dust into smb's eyes"). The author also admits that multiwords have external and internal arguments. Thus, the expression *Sand in die Augen streuen* has two external arguments, expressed by *jemand* and *emandem* in the phrase *jemand streut jemandem Sand in die Augen* ("somebody throws dust into smb's eyes"), and two internal arguments, expressed by *Sand* ("sand") and *in die Augen* ("into eyes"), as illustrated in figure 3.17.

external:
Sn – MWE – Sd
internal:
Verb – Sa – Ps

Figure 3.17: Valency patterns of the expression *Sand in die Augen streuen*

(Fellbaum et al 2006) makes use of dependency structures to describe the syntactic form of MWEs, which would allow for a subcategorisation description.

Grammatical functions and syntactic categories Valency patterns of multiword expressions are usually described with the help of syntactic features of multiword arguments. For instance, (Wojtak 1992) and (Engel 2004) indicate the syntactic categories of complements, as well as their case, cf. Sn, Sa, Sd (S stands for *Substantiv* ("noun") and n for nominative, a – accusative, d – dative). In German case marking of complements can serve as indication of their grammatical functions. Multiword expressions can also take prepositional complements (pS) and subclauses (NS).

Semantic roles and selectional restrictions Only a few authors mention semantic features of multiword valency. For example, (Wojtak 1992) and (Wotjak/Heine) provides not only syntactic valency patterns (see figure 3.17) but also semantic ones. In table 3.11, we illustrate an example for which the authors give both, syntactic and semantic valency patterns. Besides that, the author describes additional restrictions for some complements. For instance, the subcategorised subclause expresses 'the subject of illusion' (the expression *jemandem den Zahn ziehen* means "to take away smb's illusion").

MWE	Sn	Sd	(NS, dass)
predicate	Agent	Adressee	Content/Theme

Table 3.11: Valency patterns of the expression *den Zahn ziehen*

3.3.1.2 Multiwords in lexicography

Most dictionaries we know do not contain enough information about the valency properties of multiwords. In the following we summarise valency features of multi words described in some printed dictionaries, e.g. *VALBU*[26] or the *Wörterbuch der Valenz etymologisch verwandter Wörter* (*WVEVW*, "Dictionary of the Valency of Etimologically Related Words")[27], and electronic dictionaries, e.g. *ELDIT*, as well as some lexicographic approaches, e.g. (Krenn 2000) and (Hanks *et al.* 2006).

Valency patterns Although most printed and electronic dictionaries do not explicitly represent valency patterns of multiword expressions, some of them include informal indication of multiword subcategorisation in their entries. For instance, *VALBU* includes multiword expressions at the end of entries describing verbal subcategorisation. For some of them, the authors see the whole construction as a valency bearer. For example, the entry of the verb *treten* ("to kick"), which is often used as a support verb in the meanning "to get", contains a list of SVCs with this verb. Although these SVCs are not provided with the full syntactic and semantic information about their valency features, in some constructions, their subcategorisation can be deduced from the given phrases, cf. example (3.10), in which they are used with further context partners.

(3.10)
Etwas tritt (jemandem) ins Bewusstwein
("Sth becomes aware for someone")

Jemand/etwas tritt in Erscheinung
("Sbm/sth becomes visible/obvious")

WVEVW gives informtaion on multiwords, which are derived from verbs. Their valency patterns are listed together with the verbal patterns.[28]

In *ELDIT* (described in sections 3.1.2.2 and 3.2.2.2), multiwords are listed in the entries for verbal and nominal predicates, which are constituent parts of these multiwords. Subcategorisation patterns are given informally like in *VALBU*.

In (Hanks *et al.* 2006), the author also gives informal indications of complements, e.g. *jemandem* {*Informationen*|*eine Antwort*|...} *erteilen* ("to give information/answer to someone"). The database of support verb constructions of (Krenn 2000) contains a field for the subcategorisation frames of support verb constructions, illustrated by a three-place frame, e.g. (NP_{nom}, (NP_{dat}), NP_{acc}) for *zur Verfügung stellen* ("put at someone's disposal").

[26] *VALBU* was already described above, cf. section 3.1.1.1

[27] Cf. (Sommerfeldt/Schreiber 1996).

[28] We describe the correspondences between morphologically related words in section 3.4.2.2 below.

Grammatical functions and syntactic categories Grammatical functions and syntactic categories of multiword complements are presented formally in (Krenn 2000) only. For instance, NP in the complements NP_{nom}, NP_{dat} and NP_{acc} indicates a nominal phrase, which is case-marked. The case marking serves to indicate grammatical function in inflected languages. In other above mentioned works, these types of information remain informal. We can deduce the grammatical functions of complements in example (3.10) as we know that *jemand* ("smb") is a nominative form and *jemandem* is a dative form. However, for the pronoun *etwas*, the case is ambiguous.

Semantic roles and selectional restrictions The description of complements within a phrase in *VALBU*, *ELDIT* and (Hanks *et al.* 2006) allows us to deduce the restriction 'animated' vs. 'non-animated' for multiword complements. In *WVEVW* semantic roles and selectional restrictions, such as *Täter/Mensch* ("agent/human"), *Thema/Geschehen* ("theme/event") are expressed directly in the valency patterns. Other studies, to our knowledge, do not provide any information on semantic features of multiword complements.

3.3.2 Multiword Predicates in NLP Work

Subcategorisation properties of multiword predicates are also described in NLP studies. Formal grammars mostly only provide mechanisms for those collocations where the subcategorisation properties of the noun are preserved, cf. (Krenn/Erbach 1994). In this case they are treated in the same way as the nominals, cf. section 3.2.2.1 above. Other cases are not convered by formal grammars. Therefore, we concentrate in this section on the description of multiwords in NLP-based dictionaries, such as NOMLEX, FrameNet and SyntLex.

Valency patterns The dictionary of nominalisations NOMLEX (see section 3.2.2.2) contain entries that capture the relationships nominalisations with their base verb. The dictionary also incldues notes about if the verbal arguments maybe found in the noun phrase containing a nominalisation. In NOMLEX2, cf. (Macleod 2002), the authors describe and extension to NOMLEX covering SVCs, e.g. *they took a walk*, *she made a discovery*, etc. NOMLEX2 has 28 entries containing 158 SVCs.[29]. The entries are based on nominalisation entries from NOMLEX, cf. section 3.2.2.2, as well as verb entries from COMLEX, cf. 3.1.2.2. Particular support verbs are represented either by single verbs or by lists of similar verbs. Complementation patterns of SVCs are represented with the help of lexical function labels. The authors apply Mel'čuk's lexical function notation, see (Mel'čuk 1988) and (Mel'čuk 1996). Lexical functions indicate what relationship the nominlaisation has with its support verb. Therefore, we give the detailed description of lexical functions within the analysis of the "inheritance" phenomenon in section 3.4.1.1 below. In figure 3.18 we give a NOMLEX2 entry for the multiword to make accusation. The entry for the noun accusation contains further support verbs, such as *level, hurl, bring*, etc., which are listed along with their complementation patterns.

[29]The project serves as a test to ascertain whether a lexicon of this sort could be created.

```
(NOM    :ORTH accusation
        :VERB accuse
        :NOM-TYPE ((VERB-NOM))
        :NOUN-SUBC ((NOUN-PP :PVAL (about)))
        :VERB-SUBJ ((DET-POSS)
                    (N-N-MOD))
        :SUBJ-ATTRIBUTE ((COMMUNICATOR))
        :OBJ-ATTRIBUTE ((COMMUNICATOR))
        :N-N-MOD-NO-OTHER-OBJ ((SUBJECT))
        :VERB-SUBC ((NOM-NP :OBJECT ((PP :PVAL (against))
                                     (PP-OF)))
                    (NOM-NP-PP :OBJECT ((PP :PVAL (against)))
                                :PVAL (of))
                    (NOM-NP-P-ING-OC :OBJECT ((PP :PVAL (against)))
                                     :PVAL (of)))
        :SUPP-V
        ((SVERB1 :ORTH make
                 :LEX-FUNC ((OPER1 :TRANSP T))
                 :REQUIRED ((NOM-DET :INDEF T))
                 :VERB-SUBC ((NOM-NP :OBJECT ((PP :PVAL (against))))
                             (NOM-NP-PP :OBJECT ((PP :PVAL (against)))
                                        :PVAL (of))
                             (NOM-NP-P-ING-OC :OBJECT ((PP :PVAL (against)))
                                              :PVAL (of)))
```

Figure 3.18: An example of the NOMLEX2 entry for the multiword *to make accusation*

In FrameNet information about multiword expressions is represented in a variety of ways. Examples of MWEs entered as such will include lexicalised noun-noun compounds (*wheel chair, middle of nowhere*, etc.), verb-particle lemmas (*trip up*, etc.), and various kinds of SVCs and idioms (*give the slip (to), cook someone's goose*, etc.).

Support verb constructions have the same valency patterns as their nominal components. According to the FrameNet conception, nouns with event and state readings frequently select support verbs, which permit them to enter into predications. In these case syntactic arguments of support verbs are filled with nouns, whereas nouns evoke frames.

For example, (3.14) reports an act of revenge and not taking. The verb *to take* is annotated as a support verb in the FrameNet entry for the noun, cf. figure 3.19. Therefore, the valency patterns of the SVC *to take revenge* are transferred from the nominal element, its arguments are Frame Elements of the frame 'Revenge' – *Avenger, Injury, Offender* and *Punishment*.

(3.14) *King Menephta [took SUPP] awful* ***revenge*** *on a Libyan army he defeated around 1300 BC.*

More idiomatic multiwords have their own valency patterns in FrameNet. For example, *to give the slip* have one valency pattern in terms of Frame Elements and three frame elements in terms of syntactic categories and their grammatical functions, as illustrated in table 3.12.

SyntLex, a dictionary of collocations for Polish, cf. (Vetulani *et al.* 2008), contains

revenge.n
Frame: Revenge
Definition

COD: retaliation for an injury or wrong.
Supports: *exact, get, have, in,* ***take****, wreak*
Governors: *consider, seek, vow*

Figure 3.19: A FrameNet entry for the noun *revenge*

give the slip	
Evader	Pursuer
CNI	NP
–	Ext
CNI	NP
–	Obj
NP	NP
Ext	Obj

Table 3.12: FrameNet valency pattern for the multiword *to give the slip*

about 16,000 collocations (mostly SVCs), which are semi-automatically extracted from corpora. Every entry contains a nominal predicate along with the list of their support verbs and valency patterns of the SVCs, which result from their combination, cf. figure 3.20. For example, the noun *rozmowa* ("conversation") can be supported by the verbs *nawiązać* ("to connect") and *odbyć* ("to take place"). In both cases, the valency pattern for the resulting SVCs is (Acc)/N1 z (Instr). The entry also provides information on the grammatical case of the noun if used with a certain support verb (Acc – accusative).

rozmowa, f/
  nawiązać (Acc)/N1 z (Instr)
  odbyć (Acc)/N1 z (Instr)

Figure 3.20: SyntLex entry for SVCs which contain the noun *rozmowa*

Syntactic categories and grammatical functions As valency patterns of NOMLEX2 are based on the valency patterns in NOMLEX and COMLEX, it contains the same set of syntactic categories and grammatical functions as in the lexicons of nominalisations and verbs. FrameNet specifies both, syntactic categories and grammatical functions directly in the valency patterns, as seen from table 3.12. In SyntLex grammatical functions are expressed with the case-marking (Polish is an inflected language).

Semantic roles and selectional restrictions Semantic information is contained in NOMLEX and FrameNet, SyntLex describes syntactic valency only.

NOMLEX includes semantic attributes for subjects and objects, e.g. HUMAN, COMMUNICATOR, etc. as seen in figure 3.18 above. Semantic roles in FrameNet are expressed within frames, as already described in sections 2.1.5, 3.1.2.1 and 3.2.2.2 above. Each frame provides its set of semantic roles. The predictaes, which belong to the same frame share the same semantic roles. For instance, the multiword *to give the slip* belongs to the frame 'Evading' whose core semantic roles are *Capture, Evader* and *Pursuer*. Besides, this frame has a number of peripheral semantic roles, e.g. *Area, Degree, Manner*, etc.[30]. Frames also contain semantic and syntactic restrictions for semantic roles. Thus, the semantic role *Evader* belongs to the semantic type 'Sentient'. The *Evader* moves under its own power to *Capture* or contact with the *Pursuer*. Morover, *Evader* is restricted to the grammatical function 'External', cf. table 3.12.

3.3.3 Summary: Multiwords in Linguistics and NLP

In table 3.13 we summarise the above described linguistic and lexicographic work, according to the types of subcategorisation information relevant for the analysis of multiword predicates: valency patterns (VP), grammatical functions (GF), syntactic categories (SC) and selectional restrictions or semantic roles (SR).

types	*Wojtak*	*VALBU*	*WVEVW*	*ELDIT*	*Hanks*	*Krenn*
VP	+	+	+	+	+	+
GF	+	+/-	+	+/-	+/-	+
SC	+	-	+	-	-	+
SR	+	+	+	+	+	-

Table 3.13: Valency of multiwords in lingistic and lexicographic work

The description of valency of multiword predicates id summarised in table 3.14.

types	*NOMLEX*	*FrameNet*	*SyntLex*
VP	+	+	+
GF	+	+	+
SC	+	+	+
SR	+	+	-

Table 3.14: Valency of multiwords in NLP dictionaries

3.4 The Phenomenon of "Inheritance" of Subcategorisation

We define the phenomenon of "inheritance" of subcategorisation as the presence of correspondences between subcategorisation properties of morphologically related

[30] See table 2.1 in section 2.1.3.2 for the description of core and peripheral FEs.

words. This phenomenon is mostly studied within the relationships between verbs and their derivatives, e.g. deverbal nouns, which are morphologically derived from verbs by affixation, and which often share much of their meaning with the base verbs.

In this work, we also analyse "inheritance" relations between nominalisations and verbs, which are described in section 3.4.2 below. However, we are also interested in correspondences between subcategorisation properties of multiword and compound predicates and those of their constituents. For instance, valency features of multiword expressions is determined by their nominal complements in most cases, but in some cases they do not have any correspondences with the valency features of multiword constituents. In section 3.4.1, we describe the problem of identification of valency bearer in multiwords and compound nominals.

3.4.1 "Inheritance" in Multiword and Compound Predicates

The phenomenon of "inheritance" is represented in the current section as the relationship between the subcategorisation properties of multiword and compound predicates and those of their constituent parts. For multiword expressions we summarise different approaches on relations between their nominal and verbal parts and the subcategorisation properties resulting from this relations, cf. section 3.4.1.1. For compound nouns we analyse the related work on the description of their valency properties and the relations between their head and non-head constituents, cf. section 3.4.1.2.

3.4.1.1 Valency of multiwords: nominal or their own?

In this section we study different approaches in linguistics and NLP, which describe relations between valency properties of verbal and nominal constituents of multiwords (mostly SVCs) and their influence on the subcategorisation of the whole construction. We analyse the role of subcategorisation features of nominal constituents, as well as the role of subcategorisation features of verbal constituents.

The role of nominals For support verb constructions, it is assumed that they derive subcategorisation from their nominal component, cf. (Krenn/Erbach 1994), whereas some linguists claim that support verb constructions should be treated as predicates, which have their own valency, e.g. (Heringer 1968) or (Burger 1998). (Krenn/Erbach 1994) suggests that nominals in SVCs determine subcategorisation properties of the whole constructions. This feature is also stated by German grammarians, e.g. (Helbig/Buscha 2005), who admit that the arguments of support verb constructions depend not their nominal components, which are bearers of the lexical meaning. For instance, the multiword *Einfluss nehmen* ("to influence smb/sth") has a prepositional complement introduced by *auf* ("on"). This feature is inhertied from its nominal constituent *Einfluss* that also subcategorises for prepositional complements with *auf*, cf. example (3.15).

(3.15) *Wir nehmen Einfluss auf seine Entwicklung.* ("We take influence on his development").

This approach is also widespread in NLP research. For instance in the formal grammar LFG, the prepositional phrase *auf seine Entwicklung* ("on his development") is treated is an adjunct of the noun *Einfluss*, cf. figure 3.21.

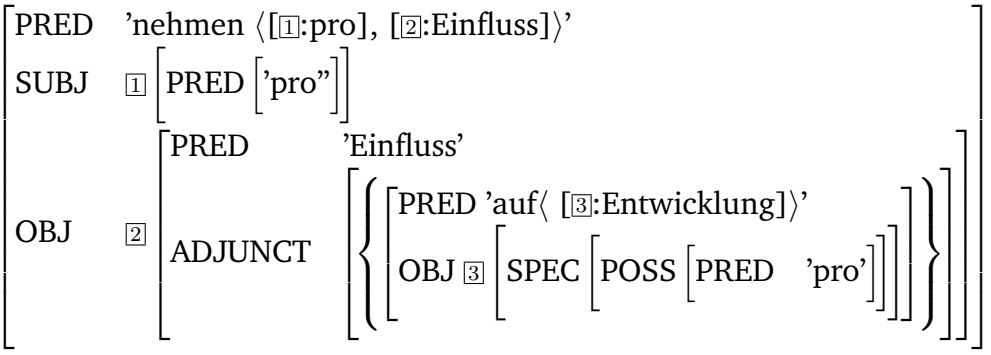


Figure 3.21: F-structure for the sentence *Wir nehmen Einfluss auf seine Entwicklung.*

In the above described computational lexicons NOMLEX and FrameNet, the subcategorisation of support verb constructions is also transferred from nominal predicates, cf. section 3.3.2.

However, some authors note exceptions to this rule. For instance, (Storrer 2007) states that in the multiword *jmdm einen Rat erteilen* ("to give smb advice"), neither *Rat* ("advice") nor *erteilen* ("to give") take a dative, except for further support verb costructions with *erteilen*, such as *jdm. Hausverbot erteilen* ("to ban smb from the house"), *jdm. Vollmacht erteilen* ("to give authority to sb"). However, the verb *raten* ("to advise") does take a dative, which means that subcategorisation properties of deverbal multiword can be, in some cases, inherited from the verb underlying the deverval noun in this multiword. We analyse the related work on the "inheritance" relations between subcategorisation properties of verbs and their nominalisations in section 3.4.2 below.

There exist further cases in which the valency properties can not be explained with those of the nominal or verbal constituent, cf. (Heid 1998) and (Lapshinova/Heid 2007). For example the multiword *zur Sprache bringen* ("to mention/to bring up") takes a *dass*-clause, whereas neither its nominal nor its verbal components do so. These collocations have non-predictable valency and behave similarly to idioms, like *ans Licht kommen* ("to be revealed") or multiwords containing "cranberry"-lexemes: *in Abrede stellen* ("to deny"). The subcategorisation of them is not transferred from the constituent parts, so they have their own subcategorisation properties. This coincides with the view that multiwords have their own subcategorisation properties, cf. (Heringer 1968) and (Burger 1998).

We assume that subcategorisation of support verb constructions, in most cases, is inhertied from their nominal components. However, in some cases SVCs have their own subcategorisation properties.

The role of verbs In SVCs support verbs do not influence the subcategorisation properties of the whole costruction. This approach is supported by most authors, e.g. German grammarians claim that by the transformation of a full verb to a support verb, the lexeme loses not only its lexical meaning but also its valency, cf.

(Helbig/Buscha 2005). However, many authors do not follow the view that subcategorisation patterns of support verbs are completely lost in multiwords. The common view is that support verbs realise their arguments as internal arguments of multiword expressions. For instance, (Wojtak 1992), (Keil 1997) and (Burger 1998) describe internal valency of multiwords. Internal arguments of a deverbal multiword are nominal (or prepositional) complements of the verb, which are constituent parts of the multiword, cf. section 3.3.1.1 above. For instance, in the above mentioned multiword *Sand in die Augen streuen*, cf. figure 3.17, the verb *streuen* subcategorises for an accusative and a prepositional complements, which are expressed by the nominal element *Sand* and the prepositional element *in die Augen*. (Helbig 1984) and (Bergenholtz/Tarp 1994) also mention that nouns can be complements of the verbs in multiword expressions.

In NOMLEX2, which describes subcategorisation of SVCs, the relationship the nominlaisation has with its support verb is indicated by lexical functions, as mentioned in section 3.3.2 above. Lexical function[31] in NOMLEX2 are 'oper(i)', 'func(i)' and 'labor(i,j)'. Oper(i) specifies that the nominalisation occurs as the direct object of the verb. The subscript (i=1,2) indicates whether the subject of the main verb is also the subject of the nominalisation or the object of the nominalisation. In example (3.16a), the combination *pay* and *a visit* would be marked 'oper(1)', as the subject *he* is both of *pay* and *visit*. In example (3.16b) the combination *have* and *visit* is oper(2), since the subject of the main verb is the object of *visit*. 'Func(i)' describes a nominalisation that occurs as the subject of the main verb. The sentence in example (3.16c) is 'func(1), as John accuses (someone), whereas the sentence in (3.16d) is 'func(2)', as 'Someone' accuses John. Labor(i,j) is assigned to a main verb plus nominalisation where the nominalisation occurs as prepositional complement of the main verb. Because of this, there are two subscripts to stand for the main verb subject position (i) and the main verb object position (j). Example (3.16e) represent labor(1,2), as *he* is the subject of the main verb as well as of *interrogate*, *John* is the object of both, the main verb and of interrogate. These functions are included into the entries for multiword as shown in figure 3.18 in section 3.3.2 above.

(3.16a) *He paid a visit to Jane.*

(3.16b) *He had a visit from Jane.*

(3.16c) *The accusation came from John.*

(3.16d) *The accusation names John.*

(3.16e) *He subjected John to an interrogation.*

FrameNet defines support verbs as verbs that combine with state or event nouns to create a multiword predicate, their arguments are filled with the nominal frame, cf. table 3.12 in section 3.3.2 above. The above described lexicon of multiwords for Polish, SyntLex, provides the information on the grammatical function the noun has when used with a certain support verb, which is expressed in the its case. For instance, in figure 3.20 in section 3.3.2 above, the noun *rozmowa* is realised as the

[31] Lexical functions are defined in (Mel'čuk 1988) and (Mel'čuk 1996).

direct object (expressed by the accusative case) of both support verbs listed in the entry.

The above descibed approaches show that although verbs (which build multiword predicates with nouns) do not influence the subcategorisation properties of multiwords, they realise their complements as internal arguments of multiword constructions.

3.4.1.2 Valency of compound nouns: head or non-head?

In the current section we analyse subcategorisation features of German nominal compounds, such as *Journalistenfrage* ("journalist question"), *Motivsuche* ("motive search")[32], etc. Although a compound has a complex form, it acts as a single predicate in a sentence. Therefore, subcategorisation features of nominal compounds in German are described within the features of nominal predicates. However, analysing multiword and compound predicates, we intend not only to describe their subcategorisation features, but also to analyse the relations between their constituent parts.

Nominal compounds are described in a number of linguistic and NLP studies, e.g. (Levi 1978), (Bauer 1983), (Bergsten 1991), (Ortner 1991) in linguistics, and (Bouillon *et al.* 1992), (Johnston/Busa 1999), (Hippisley *et al.* 2005) in NLP.

Most authors, e.g. (Johnston/Busa 1999), admit that compounds have a constituent structure and a compositional reading. The relations between their elements are similar to the relations of sentence parts. (Chomsky 1970) applies the notion of the 'head' and 'X-bar-theory' in wordformation (beforethat it was used for syntax). The authors describe features of a compound in the definition of one of its constituents. The concept of head in the description of compounds is also used by (Zwicky 1985) and (Bauer 1978) or (Bauer 1988). The features of compound heads correspond to the features of phrase heads:

1) the head determines category, gender, flexion class, etc;

2) the head determines argument structure of the construction;

3) the head is also the semantic head of the whole construction;

Thus, subcategorisation properties of compound nominals should also be determined by the head constituent. Non-head are generally seen as modifiers of heads. However, this definition leaves the problem of interpretation of the most compounds. For instance, in the compounds *Bierwurst* and *Rauchwurst*, the non-heads have different relations with the heads. In the first example, it is with what it should be eaten, in the second it is how it has been treated, cf. (Levi 1978). Following this approach does not let to interpret further cases, e.g. *Guinea pig* in English or *Ehrgeiz* ("ambition") in German, as these compounds are not transparent semantically and are mostly lexicalised.

Although compounds have received attention in linguistic and NLP research, a few authors only, e.g. (Grimshaw 1990) and (Johnston/Busa 1999), analyse their

[32]German compounds are written as one word, unlike English nominal compounds that have a form of a nominal phrase.

predicate-argument structure. Authors mostly distinguish between internal or external arguments of compound nouns, which is similar to the analysis of arguments of multiword expressions, cf. section 3.4.1.1 above. Internal arguments are filled with the non-head constituents, whereas external arguments are filled with words or phrases, which occur with compounds. For instance, in *Journalistenfrage an den Präsidenten* ("journalist question to smb"), the nominal compound *Journalistenfrage* has an external argument realised as prepositional complement *an den Präsidenten*.

For English compounds, (Grimshaw 1990) describes a rule for compound whose head takes more than one argument. The rule is based on the hierarchy of semantic roles. The author states that arguments that have less prominent semantic roles must be inside the compound (internal argument), and arguments with more prominent semantic roles – outside (external argument). For instance, both *cookie* and *children* are arguments of the the head noun *baking* in *cookie-baking by children*. The noun *cookie* (which has the semantic role 'theme') is considered to be less prominent than the noun *children* (whose semantic role is 'agent'), cf. *cookie-baking by children* vs.**children-baking of cookies*.

(Johnston/Busa 1999) analysing English and Italian compounds, specify their predicate-argument structures within the qualia structures of the head nouns. Qualia structures provide the 'glue', which links the semantic contributions of modifying nonheads and those of heads. The authors apply the representational framework of the Generative Lexicon (GL) introduced by Pustejovsky[33], simplifying a GL entry to four levels of representations: type structure, argument structure, event structure and qualia structure, as shown in figure 3.22. The latter expresses four aspects of the meaning of the lexical item: FORMAL,CONSTITUTIVE, TELIC, and AGENTIVE. The interpretation of the compound form hunting rifle can sound as follows 'a rifle which is used in its typical capacity (i.e. firing) for the purpose of performing the activity of hunting'. The assignment of a complex structure to an individual quale is coherent with the general interpretation of qualia structure.

$$
\begin{bmatrix}
\textit{hunting rifle} & & \\
\text{TYPESTR} = & \left[\text{ARG1} = \boxed{x}\ \text{rifle}\right] & \\
\text{ARGSTR} = & \begin{bmatrix}\text{D-ARG1} = \boxed{w}\ \text{human}\\ \text{D-ARG2} = \boxed{z}\ \text{prey}\end{bmatrix} & \\
\text{EVENTSTR} = & \begin{bmatrix}\text{D-E1} = \boxed{P}_1\ \text{process}\\ \text{D-E2} = \boxed{P}_2\ \text{process}\end{bmatrix} & \\
\text{QUALIA} = & \begin{bmatrix}\text{FORMAL} = \boxed{x} & \\ \text{TELIC} & \begin{bmatrix}\text{activity lcp}\\ \text{TELIC} = \text{hunt}(\boxed{P}_2, \boxed{w}, \boxed{z})\\ \text{AGENTIVE} = \text{fire}\ (\boxed{P}_1, \boxed{w}, \boxed{x})\end{bmatrix}\end{bmatrix} &
\end{bmatrix}
$$

Figure 3.22: GL entry for compounds

In the above mentioned lexicon STO (see section 3.2.2.2 for details), many en-

[33]For example, in (Pustejovky 1995).

coded nouns are compounds. The authors claim that the non-head constituent of a nominal compound can fill in an argument slot, thereby reducing the number of arguments of a compound noun, cf. (Olsen 2002). For instance, the non-head *brylluppet* in (3.12) fills in the slot for the *af*-complements of the nominal arangement, which can take two complements (*Søsterens* and *af brylluppet*). The resulting compound *brylluparrangement* takes one complement only.

(3.17)

Søsterens arangement af brylluppet (Dn2GPn-af) ("The sister's arrangement of the wedding")	–>	*Søsterens brylluparrangement* (Dn1G) ("The sister's wedding arrangement")

Since nominal compounding is a productive process in Danish, the authors do not intend to lexicalise compound nouns, except for the most frequent ones in corpora. The STO database contains information about elements used for linking of the compound non-heads, therefore every compound can be produced automatically. However, the authors admit that it is not possible to account for the syntactic and semantic behaviour of such a compound automatically.

3.4.1.3 Summary: "inheritance" in multiwords and compounds

In the current section we summarise the above mentioned approaches on the relations between subcategorisation properties of multiword and compound predicates, such as multiwords and compound nouns, and those of their constituent parts.

Multiword expressions The analysis of works on the relations between valency properties of multiwords and those of their constituent parts show that verbs do not have influence on the subcategorisation features of the whole constructions. However, their elements are realised as internal argument of multiwords (nominal/prepositional complements of the verb, which are constituent parts of the multiword). Subcategorisation properties of SVCs are in most cases inherited from their nominal constituents. However, there are limits to the inheritance of subcategorisation. Some SVCs show their own properties, which are not specific for their nominal complements. This allows us to distinguish between multiwords taking over their subcategorisation properties from their nominal complements and those, which possess their own properties. We describe classification of multiwords based on the relations between their subcategorisation properties and those of their nominal constituent is described in section 4.2.4 below. The automatical treatment of multiwords under analysis, their extraction and classification is described in section 5.3.3.4.

Compound nouns To automatically analyse subcategorisation behaviour of compounds, we need to identify the valency bearer in a compound. According to the commonly accepted assumtion, heads of compounds determine their predicate-argument structure. However, there cases that do not correspond with this assumption. For instance, in *Besuchdienste bei älteren Menschen* ("visit services at elder people's"), the prepositional complement *bei älteren Menschen* ("at elder people's") is more likely inherited from the non-head constituent *Besuch* ("visit"), and not from the head *Dienst* ("service"). Most approaches described above, discuss the problem of internal

arguments of compounds. Relations between the constituents of compounds and their external arguments have to our knowledge not received much attention so far. The problem of identification of valency bearer in nominal compounds is described in (Lapshinova/Heid 2008) and (Lapshinova 2008). We believe that this problem is important in the description of subcategorisation properties of nominal compounds and should be taken into account in the process of building or updating NLP lexicons. In section 4.2.3, we describe types of nominal compounds classfied according to the relations between their subcategorisation properties and the subcategorisation properties of their constituent parts. The automatical treatment of these cases is described in section 5.3.3.3

3.4.2 "Inheritance" in Nominalisations

In the current section we describe relations between subcategorisation properties of nominalisations and those of their base verbs. We focus on the following questions: do nominalisations inherit their subcategorisation properties from the underlying verbs or do they possess their own properties, which are not specific for their base verbs?

To answer these questions, we analyse the related work for the description of nominalisations and their base verbs. Nominalisations include not only lexical items occurring freely in context but also those which occur within support verb constructions.

3.4.2.1 Nominalisations and their base verbs

Nominalisations and their argument structure have been a research topic in linguistic, lexicographic and NLP work. Most authors, e.g. (Nunes 1993), (Ehrich/Rapp 2000), (Schierholz 2001), (Meinschaefer 2004), mention correspondences between arguments of nominalisations and those of their underlying verbs, depending on the type of complements and the classes of verbs under analysis.

The view that verbs and nouns seem to share subcategorisation properties is established already in (Chomsky 1970), who illustrated this by the example given in (3.18). The nominalisation *destruction* has the same arguments as its base verb *to destruct*.

(3.18) *The enemy destroyed the city – The enemy's destuction of the city*.

The correspondences between the valency patterns of verbs and their derivatives are studied both on syntactic and semantic levels of valency description. This means that there exist studies, for example, (Grimshaw 1990), (Sommerfeldt/Schreiber 1996), (Schierholz 2001), (Meinschaefer 2004), which analyse correspondences between types and grammatical functions of complements both verbs and nominalisations can take, and there also studies, e.g. (Nunes 1993), (Sommerfeldt/Schreiber 1996), (Ehrich/Rapp 2000), (Meinschaefer 2004), which describe correspondences between semantic roles of complements and their selectional restrictions.

However, only a few authors, e.g. (McCawley 1982), (Harris 1968) and (Lees 1963), mention systematic correspondences between verbs and nominalisations, although deverbal nouns or nominalisations are very common in many Germanic and Romance

languages. Among the recent works in NLP, only a few lexical resources provide systematic correspondences between verbs and nominalisations. (Gurevich et al. 2007) describes the process of mapping the predicate-argument structure of nominalisations onto that of their base verbs, using PARC's text processing system. An earlier example of the description of correspondences between deverbals and verbs is NOMLEX, (Macleod *et al.* 1998b), the above mentioned computational lexicon of nominalisations, which maps noun complements onto the predicate-argument structure of their underlying verbs.

The transformation illustrated in (3.18) cannot be applied for all verb-nominalisation pairs, as shown in example (3.19). The nominalisation *growth* takes the object only, cf. (Wechsler 2008).

(3.19) *John grows tomatoes* – **John's growth of tomatoes*
vs. *the tomatoes' growth*
or *the growth of the tomatoes*.

The cases of non-correspondences between subcategorisation properties of nominalisations and their underlying verbs are described in (Sommerfeldt/Schreiber 1996), (Schierholz 2001), (Meinschaefer 2004), (Wechsler 2008) and others. However, to our knowledge there is no systematic description of correspondences ("inheritance") and non-correspondences ("non-inheritance") between valency properties of verbs and their derivatives.

In the following sections we analyse the related work on the description of both "inheritance" and "non-inheritance" of subcategorisation properties.

3.4.2.2 "Inheritance" of subcategorisation

Mostly, subcategorisation properties of nominalisations are believed to be inherited from their base verbs, cf. examples in (3.20).

(3.20)
- *befürchten, dass* (" to fear that")
  vs. *Befürchtung, dass* ("fear that")
- *erklären, dass/w-/ob* ("to explain that/wh-/if")
  vs. *Erklärung, dass/w-/ob* ("explanation that/wh-/if")

The current section describes linguistic and lexicographic work, as well as NLP studies, which describe correspondences between the predicate-argument structure of verbs and their nominalisations.

"Inheritance" in linguistics and lexicography As mentioned above, "inheritance" of verbal subcategorisation properties by nominalisations is described both in terms of syntactic and semantic valency.

For instance, (Grimshaw 1990) states that on the syntactic level the range of elements, which can occur after nominalisations, is related to the range of elements that can occur after their base verbs, apart from the failure of nouns to take bare noun phrase[34], or (Schierholz 2001) discusses the transfer of subcategorisation properties

[34]in English nominalisations can take only possessives, prepositional phrases and infintive or sentential complements.

of verbal predicates to their nominalisations on the example of prepositional arguments. According to (Schierholz 2001), nominalisations subcategorising a prepositional phrase can build paraphrases with their base verbs, cf. (3.21a) and (3.21b).

(3.21a) *die Abstammung vom Affen* ("the descent from monkeys")
vs. *Jemand stammt vom Affen ab.* ("someone descends from monkeys").

(3.21b) *sein Interesse für Rotwein* ("his interest in red wine")
vs. *Er interessiert sich für Rotwein* ("he is interested in red wine").

Correspondences on the semantic level of valency description are given in, e.g. (Ehrich/Rapp 2000). The authors claim that subcategorisation frames of both, nominalisations and their base verbs are derived from one and the same Lexical Semantic Structure by category specific linking rules. (Nunes 1993) analyses the relations between deverbal nouns and their base verbs, describing the mechanisms of semantic role determination. The author claims that the choice for the role of the nominalisation argument depends on the verb semantics. (Melloni 2007), analysing event and result nominalisations, also states inheritance of verbal predicate-argument structures by nominalisations. In example (3.20a) the nominalisation *cancellation* appears to inherit the argument structure from the corresponding predicate *to cancel*, although only the internal argument (*of all his appointments*) is obligatory and has the argument status. The *by-phrase* corresponding to the external argument (*by the secretary*) is instead optional and can be omitted without affecting the grammaticality of (3.22b). Furthermore, the presence of the aspectual modifier *in a few minutes* proves that the aspectual properties of the base verb are also preserved in deverbal nouns.

(3.22a) *The secretary cancelled all his appointments in a few minutes.*

(3.22b) *The cancellation of all his appointments by the secretary in a few minutes*

(Wechsler 2008)[35] also states that predicate-argument structures of base verbs, both transitive, as in (3.23b), and intransitive, as in (3.23b), are preserved in their nominalisations.

(3.23a) *The letter arrived* vs. *the arrival of the letter.*

(3.23b) *Mary constructed the spaceship* vs. *Mary's construction of the spaceship.*

There exist studies which describe the "inheritance" phenomenon on both levels, e.g. (Meinschaefer 2004) studies syntax and semantics of event-denoting deverbals from the point of view of the theory of argument linking. She shows that nouns and their base verbs have common semantic and syntactic argument structures without any obligatory suppression of arguments, which is evident from nominal constructions that realise all arguments of the base verbs. In *WVEVW*[36], the authors systemise correspondences between verbal and nominalisation subcategorisation on both semantic and syntactic levels.

[35]The author follows the view of (Rappaport 1983).
[36]The description of this dictionary is given in section 3.3.1.2

In the introduction to the *WVDS*, which is described in section 3.2.1.2 above, tstate that nouns deriving from verbs not only take over the meaning structure of their base verbs, but also their valency, cf. (3.24a) and (3.24b).

(3.24a) *Glaube an...* ("belief in...") vs. *glauben an...* ("to believe in...")

(3.24b) *Spiel mit...* ("play with...") vs. *spielen mit...* ("to play with...")

In *WVEVW*, the authors try to systemise the relation between verbal subcategorisation properties and those of deverbal nouns. Valency properties of deverbal nouns are spelled out in the same entries with their base verbs, which allows to find out the similarities and differences in such properties of morphologically related words, as syntactic categories and grammatical functions of complements, as well as their selectional restrictions, cf. the entry for the verb *erklären* ("to explain") and its nominalisation *Erklärung* ("expanation") in figure 3.23.

erklären - Erklärung
Der junge Mann (a) erklärte seiner Freundin (b), dass er am Sonntag keine Zeit habe und nicht kommen könne (c). Die Erklärung des Vorstandes (a) an die Mitglieder (b), dass er für Neuwahlen sei (c), fand einhellige Zustimmung.
1. 'Perspektivierung S/S', 'Bekanntgabe von Fakten', 'offiziell', 'etwas verbal zum Ausdruck bringen'.
2. a – Täter/Mensch, Institution/
V:Sn; S:Sg
b – Adressat/Mensch, Institution/
V:Sd; S:Sp (an)/selten/
c – Thema/Geschehen/
V:Sa/NS (dass, w-)/Inf S:NS (dass, w-)/Inf
3. Der Ministerpräsident erklärte dem Präsidenten seinen Rücktritt. Der Jüngling erklärte dem Mädchen schüchternd seine Liebe. Der Trainer erklärte dem Präsidium, dass kaum noch Aussicht auf den Klassenerhalt bestehe, Die Erklärung des Botschaftlers an die Regierung, dass ein Kompromiß gefunden worden sei, rief Erstaunen hervor.

Figure 3.23: An example of *WVEVW* entry

Arguments of the nominalisation and its base verb are marked with the same letters, which expresses the parallels in the subcategorisation properties of verbs and their deverbals, as shown in example (3.25).

(3.25a) agent/human, institution:
Der junge Mann ("the young man") vs. *der Vorstand* ("the managin board").

(3.25b) adressee/human, institution:
seine Freundin ("his girlfriend") vs. *die Mitglieder* ("the members").

(3.25c) theme/process:
dass er am Sonntag keine Zeit habe und nicht kommen könne ("that he has no time on Sunday and can't come along")
vs. *dass er für die Neuwahlen sei* ("that he is for the snap election").

"Inheritance" in NLP In NLP work, we analyse two studies which describe the "inheritance" of verbal valency properties by their nominalisations systematically. The lexicon of nominalisations NOMLEX[37] gives information on correspondences between complements of nominalisations and their base verbs, describing both their syntactic and semantic features.

The entries of NOMLEX indicate, which verb underlies the given nominalisations, and which complements of the nominalisation are inherited from its base verb[38], cf. table 3.9 in section 3.2.2.2 above. In figure 3.24, we give an entry that shows arguments of the nominalisation *experiment* whose subject can be a possessive, or a noun-noun modifier. Further arguments can be raelised as prepositional phrases only as the underlying verb *to experiment* is instransitive.[39] The subject of the verb (VERB-SUBJ) may appear as noun-noun modifier (N-N-MOD), like in *laboratory experiment* or as possessive determiner, like in *my experiment*. The object can have the form of a prepositional phrase (NOM-PP), e.g. headed by the prepositions *on* or *with*.

The complementation types are prefixed with NOM (NOM-INTRANS, NOM-PP) to indicate that they are treated as nominalisation complements. The ability of nominalisations to absorb the arguments of their base verbs is expressed with the feature VERB-NOM. Although NOMLEX describes the relations of nominal predicate-argument structures to the verbals one, it treats deverbal nouns essentially as nouns.

(NOM	:ORTH	"experiment"		orthography
	:VERB	"experiment"		base verb
	:NOM-TYPE	((VERB-NOM))		nominalisation types
	:VERB-SUBJ	((N-N-MOD)		noun modifier position
		(DET-POSS))		possessive determiner of the noun
	SUBJ-ATTRIBUTE	((COMMUNICATOR))		semantic attribute of the subject
	:VERB-SUBC	((NOM-INTRANS :SUBJECT	((N-N-MOD)	intransitive
			(DET-POSS))	subject features
		:REQUIRED	((SUBJECT)))	overwrite option
		(NOM-PP :SUBJECT	((N-N-MOD)	PP as complement
			(DET-POSS))	subject features
		:PVAL	("on" "with")))	PP values

Figure 3.24: NOMLEX entry for the noun *experiment*

The PARC text processing system, cf. (Gurevich et al. 2007), by contrast does not have a separate level for nominal argument structures. The system is based on semantic description of verbal arguments. Deverbal nouns and their arguments are converted into verb-like event structures. The authors make an attempt to canonise deverbal nouns making them look like verbs for the purposes of knowledge representation. This is important for reasoning systems that rely on semantic information from the input. For example, such a system should be able to answer the question in (3.26a) based on the sentence in (3.26b).

[37] NOMLEX is described in section 3.2.2.2 above.

[38] The information about the predicate-argument structure of the underlying verbs comes from COMLEX described in section 3.1.2.2 above.

[39] For details, see (Macleod *et al.* 1998b).

(3.26a) Input: *The acquisition by US Air of America West last week rocked the financial world.*

(3.26b) Question: *Did US Air acquire America West?*
Answer: *Yes*.

The PARC system identifies arguments of nominalisations and assigns them with appropriate semantic roles, identical to the roles their base verbs have. Then deverbal nouns are rewritten into the realtered verbs with the same argument structure. For instance in figure 3.25, the deverbal noun *death* is converted to a verb-like structure and its possessive or the *of*-phrase are linked to the subject of the verb. The nature of this subject role depends on selectional restrictions of verbs, e.g. the transitive verb *to break* takes a Theme subject, whereas the intransitive *to eat* takes an Agent subject.

Ed's death/the death of Ed
xor(A1,A2)
A1: subconcept(death:6,[death-1,death-2,death-3,
Death-4,death-5,death-6,death-7,end-6])
role(of,death:6,Ed:3)
A2: subconcept(die:6,[die-1,die-2,die-3,fail-4,die-5,
die-6,die-7,die-8,die-9,die-10,die-11])
role(Theme,die:6,Ed:3)
subconcept(Ed:3,[male-2])
alias(Ed:3,[Ed])

Figure 3.25: Mapping rules for the nominalisation *death* in the PARC system

In figure 3.26, we give an example of the PARC system rewriting rules for nominalisations subcategorising for *that*-clauses. The presence of a *that*-clause indicates to the PARC system that a new context must be created, called ctx(disappear), which encompasses the disappearing event. The main participants (Ed, Mary and the stating event) are instantiated in the top-level context, while the disappearing event is only instantiated in the subordinate context.

Mary's statement that Ed disappeared
role(Agent, state:1, Mary:0)
role(Topic, state:1, ctx(disappear:3))
role(Recipient, state:1, implicit_arg:4)
role(Theme, disappear:3, Ed:2)
subconcept(state:1, [state-1,submit-2,express-3])
subconcept(Mary:0, [female-2])
subconcept(Ed:2, [male-2]))
subconcept(disappear:3, [disappear-1,vanish-2,vanish-4,melt-6])

Figure 3.26: Mapping rules for the nominalisation *statement* in the PARC system

In PARC, the expression *Mary's statement that Ed disappeared* is identical to that of *Mary stated that Ed had disappeared*. The authors assume that this correlation pro-

vides and extension of syntactic coverage of the grammar in PARC. The list of nominalisations about which the system knows that they can take *that*-complements can be extended by nominalisations whose underlying verbs can take a *that*-complement. Such nouns are believed to subcategorise for the same complements as their base verbs.

However, this hypothesis is not always valid, cf. example (3.18) above. In the following we analyse the related work on non-correspondences between predicate-argument structure of verbs and their nominalisations.

3.4.2.3 "Non-inheritance" of subcategorisation

In some cases subcategroisation properties of deverbal nouns do not correspond to those of the underlying verb, cf. examples (3.27a) and (3.27b).

(3.28a) *wissen, dass/w-/ob* ("to know that/wh-/if")
vs. *das Wissen, dass/*w-/*ob* ("knowledge that/*wh-/*if")

(3.28b) *vermuten, dass/w-/ob* ("to suppose that/wh-/if")
vs. *die Vermutung, dass/*w-/*if* ("supposition that/*wh-/if*")

The non-correspondence or "non-inheritance" of complements between verbs and their nominalisations is described in a number of works mentioned in section 3.4.2.1 above. However, most authors do not give any systematic analysis of these phenomena.

Authors mention either cases in which nominalisations 'lose" verbal subcategorisation properties, or cases in which the derived predicates seem to have some additional properties, which are not specific for their base verbs. Therefore, we assume that the "non-inheritance" in subcategorisation can be classified into *subcategorisation reduction* (the loss of verbal properties) and *subcategorisation extension* (the existence of other properties not specific for verbs). In the following sections we analyse the related work on "non-inheritance", summarising them according to the two above mentioned *directions* of "non-inheritance" – subcategorisation reduction and subcategorisation extension.

Subcategorisation reduction Describing valency properties of verbs and their nominalisations most authors state that deverbal nouns do not always completely inherit all the complements of the base verbs. Subcategorisation reduction is described both in terms of syntactic properties of complements and their semantic features.

For instance, (Schierholz 2001), who analyses the transfer of subcategorisation properties of verbal predicates to their nominalisations on the example of prepositional complements, shows that sometimes prepositional complements of base verbs do not occur with their nominalisations. The prepositional complement of the verbs *sich ernähren* and *achten auf* in (3.30) can not be found with their nominalisations *Ernährung* and *Achtung*.

(3.29a) *sich ernähren **von*** ("to live/nourish **on** sth")
vs. *die Ernährung *von* ("sustenance/nourishment").

(3.29b) *achten **auf*** ("to pay attention **to**")
vs. *die Achtung *auf* ("attention").

However, non-correspondences are mostly described in terms of valency patterns – the number of arguments or complement both verbs and nominalisations can take. Deverbal nouns are believed to have reduced complementation patterns, as they take over just a part of the arguments of their underlying verbs.

For insatnce, in *WVDS*[40], the authors show that nominalisation have less complements than their base verbs. The verb *belohnen* ("to reward") has three complements, whereas its nominalisation *Belohnung* ("reward") does not necessarily have all of them, cf. examples (3.30a), (3.30b) and (3.30c). However, we know that all their arguments are optional, cf. section 3.2 above. Therefore, the prepositional phrase in (3.30b) is optional.

(3.30a) *Der Kommandeur belohnt den Soldaten mit einem Sonderurlaub.*
("The commander rewards the soldier with a special leave").

(3.30b) *Die Belohnung des Soldaten mit Sonderurlaub*
("The reward of the soldier with a special leave").

(3.30c) *Die Belohnung des Soldaten* ("The reward of the soldier").

Differences in the subcategorisation of verbs and nominalisations are also mentioned by (Camacho/Santana 2004) in their study on argument structure of deverbal nouns in Brazilian Portugese. The authors claim that deverbal nouns preserve just a part of the verbal predicate-argument structure. In (3.31), in the nominal phrase *alguns dos desenhos das cavernas* ("some cavern drawings"), the nominalisation *desenhos* cannot recover the predicate-argument structure of the corresponding verb *desenhar* ("to draw"). Thus, in spite of being a deverbal noun, it works as a prototypical referential noun. Therefore, according to (Camacho/Santana 2004), nominalisations in general are chacterised by valency reduction in comparison to the associated verbs.

(3.31) *em alguns dos desenhos das cavernas principalmente em Altamira... ha uma fidelidade... linear a natureza*
("in some cavern drawings mainly in Altamira... there is a linear fidelity to nature")

(Meinschaefer 2004) describes contrasts in the linking of semantic roles to syntactic types by verbs and nouns. For verbs, agent-like arguments have priority in linking, for event-denoting deverbal nouns, theme-like arguments are linked first, as seen in examples (3.32b) and (3.32c).

(3.32a) *The enemy destroyed the city.*

(3.32b) *the destruction of the city* vs. **the destruction of the enemy*

(3.32c) *the city's destruction* vs. **the enemy's destruction*

[40] Cf. (Sommerfeldt/Schreiber 1983), this dictionary is described in section 3.2.1.2 of this thesis.

This problem is also discussed in (Nunes 1993) who claims that some nominalisations take an undergoer (a patient or a theme) as *of*-PP, whereas others take an actor, cf. examples in (3.33a) and (3.33b).

(3.33a) *Sara knows French* vs. *some knowledge of French*
and not *some knowledge of Sara*.

(3.33b) *the dog barked* vs. *the barking of the dog*.

Nominalisations of some two-argument activity verbs, like *to attack, to investigate* can have both, an actor and an undergoer as a prepositional complement, as shown in (3.34).

(3.34) *Sherlock Holmes investigated the murder.*
vs. *the investigation of Sherlock Holmes into the murder.*
vs. *the investigation of the murder by Sherlock Holmes*.

However, we can paraphrase sentences in (3.32a), (3.33a) and (3.34) into nominalisations with two complements, which express both, an actor and an undergoer, cf. examples (3.35a) and (3.35b). In this case the nominalisation allows for the same number of complements as its base verb does. (Wechsler 2008) gives another example, which shows that this is not valid for all semantic types of nominalisations. For instance the nominalisation *growth*, cf. example (3.19) above, does not inherit all valency properties of the base verb *to grow*. It omits the agent *John* and can take objects only. The author claims that although the grammar itself permits inheritance of the agent, some extra-grammatical factors downgrade this.

(3.35a) *The enemy's destruction of the city.*

(3.35a) *Sara's knowledge of French.*

The difference in the semantic argument-linking is also described in (Pado et al. 2008). The authors construct a model for nominal labelling from verbal training data. They create mappings between arguments of nominalisations and those of their base verbs. For instance, the information from (3.36a) that *Peter* is a COGNIZER can be directly used for the occurrence of *Peter* in (3.36b). The head word of the last word *event* is unseen in (3.36a) and due to its abstract nature, difficult to classify through semantic similarity. The phrase in (3.33b) headed by *event* can be classified as STIMULUS because it is an *about*-PP line in (3.36a). In contrast to that, no direct inferences can be drawn about prenominal genitives or modifiers which do not exist for verbs.

(3.36a) *Peter* (Subj/COGNIZER) *laughs about the joke* (PP-about/STIMULUS).

(3.36b) *Peter's* (prenom-Gen/?) *laughter about the event* (PP-about/?).

In NOMLEX the information about which verbal argument is not inherited by nominalisation is given in the feature :NOM-TYPE, cf. figure 3.24. For nominalisations, which iherit all complements of their base verbs, this feature has the value VERB-NOM. If the value of :NOM-TYPE is OBJECT, the nominalisation cannot take an object itself, cf. (3.37).

(3.37) *Clinton's APPOINTEE to the cabinet.*
vs. *Clinton appointed (the appointee) to the cabinet.*

However, this feature does not solve the problem of ambiguities – the same nominal position can often map into several different verbal arguments. For instance, for the nominalsation *announcement*, both, the verbal subject and the verbal object can be expressed by a possessive (3.38a), a nominal modifier (3.38b) or the of-PP of the nominalisation (3.38c).

(3.38a) **det-poss:**

subj. *his announcement*,

obj. *the product's announcement*

(3.38b) **n-n-mod:**

subj. *The State Department announcement*,

obj. *the product announcement*

(3.38c) **pp-of:**

subj. *The announcement of the White House*,

obj. *The announcement of the product*

Such ambiguities are solved in PARC with the help of selectional restrictions of verbs (known from valency patterns of verbs, e.g. from VerbNet[41]).

Subcategorisation extension Non-correspondences between subcategorisation properties of verbs and their nominalisations can also include additional properties of nominalisations, which are not specific for their base verbs that can be expressed both on syntactic and semantic levels of valency description.

For instance, the authors of *WVDS* show that grammatical forms of nominalisation complements inherited from verbs can change. Indirect and direct objects of a verb can be replaced in nominalisations by prepositional complements, cf. examples (3.39a) and (3.39b).

(3.39a) *Antwort an* ("the answer for") vs. *antworten jmdm* ("to answer smb").

(3.39b) *der Verdacht auf* ("suspicion of")
vs. *verdächtigen jmdn* ("to suspect smb").

In some cases selectional restrictions of nominalisations and their base verbs differ. For instance, the nominalisation *Verdacht* ("suspicion") in (3.39b) in (3.39b) allows for both animate and inanimate content of the prepositional phrase, whereas its base verb *verdächtigen* can take an animate object only, cf. (3.40a) and (3.40b).

[41]The VerbNet project maps PropBank verb types to their corresponding Levin classes of verbs, cf. (Kipper et al. 2000), (Kipper-Schuler 2005) and (Schuler 2005), http://verbs.colorado.edu/verb-index/

(3.40a) *der Verdacht auf den Dieb* (“suspicion of the thief”)
vs. *jemand verdächtigt den Dieb.* (“smb suspects the thief”).

(3.40b) *der Verdacht auf Gelbsucht* (“suspicion of jaundice”)
vs. **jemand verdächtigt (die) Gelbsucht* (*“smb suspects jaundice”)

In *WVEVW* the information on these non-correspondences is included into entries. For instance in figure 3.23 in section 3.4.2.2 above, the nominalisation *Erklärung* can take a prepositional complement, whereas its base verb *erklären* can take direct and indirect objects only.

The difference in syntactic realisation of complements subcategorised by verbs and their nominalisations is also described by (Schierholz 2001), who illustrate cases in which nominalisations subcategorise for prepositional complements, which are not allowed by their base verbs. Many nominal predicates taking prepositional complements are derived from verbs, which take genitive, dative or accusative complements (indirect or direct objects), as shown in (3.41a) and (3.41b).

(3.41a) *Biancas Achtung* ***vor*** *dem Kerl* (“Bianca’s respect for the guy”)
vs. *Bianca achtet* ***den*** *Kerl* (“Bianca respects the guy”).

(3.41b) *ihre Ähnlichkeit* ***mit*** *einer Vorzimmerdame*
(“Her similarity with the receptionist”)
vs. *Sie ähnelt* ***einer*** *Vorzimmerdame* (“She is similar to the receptionist”).

In some cases both, a nominalisation and its related one, take different prepositional complements, as shown in (3.42).

(3.42) *sein Interesse* ***an*** *Rotwein* (“”)
vs. **Er interessiert sich* ***an*** *Rotwein.* (“”)
vs. *Er interessiert sich* ***für*** *Rotwein.* (“”)

Another example of differences in types of complements, nominalisations and their base verbs cases take is given by (Hull/Gomez 2000). The authors notice that the nominalisation *control* subcategorises for a PP with the preposition *over*, e.g. *his control over the business*, while the verb *control* does not.

In NOMLEX, nominal complements, which are not inherited from verbs are marked with the feature :NOUN-SUBC, which means that these comlements are specific for the nominalisation only.

3.4.2.4 Reasons for “non-inheritance”

The above described studies show that subcategorisation properties of nominalisations do not always correspond to those of their base verbs. This subcategorisation behaviour of nominalisations can be explained by several reasons in terms of both syntactic and semantic levels of valency description.

Syntactic features The fact that the same nominal complement type can often map into several different verbal arguments can be explained by nominalisation nature. Argument positions of nominalisations are specific for nouns and are generated by the syntax specific to nouns.

The authors of *WVDS* also explain the "non-inheritance" of subcategorisation properties by syntactic features of nominalisations. For example in German, nominalisations take mostly prepositional or sentential complements, as well as genetive NPs. This explains the occurrence of prepositional complements with the nominalisations, whose base verbs do not allow any, cf. examples (3.39a) and (3.39b) in section 3.4.2.3 above. Besides that, we know for nominal predicates that most their arguments are optional (cf. section 3.2 above). Therefore, nominalisations lose some arguments specific for their underlying verbs, cf. example (3.30b) in section 3.4.2.3 above.

Thus, the ability of nominalisations to realise arguments with certain syntactic types only causes most non-correspondences in their valency patterns with the valency patterns of their base verbs. However, this does not explain cases of "non-inheritance" in which verbs realise their complements with syntactic categories, which are also specific for nominalisations, e.g. prepositional phrases, whereas their derivated nouns do not show these properties, cf. examples (3.29a) and (3.29b) in section 3.4.2.3 above.

Semantic features Semantic features of verbs and their derivatives can also explain the "non-inheritance" relations between nominalisations and their base verbs.

For instance, (Meinschaefer 2004) explains the "non-inheritance" by the contrasts in the linking of semantic roles to syntactic types by verbs and nouns. The authors claim that these contrasts arise because the rules mediating between the predicate-argument structure and syntactic positions of arguments differ slightly for verbs and nouns. As mentioned above, agent-like arguments of verbs have priority in linking, whereas event-denoting nominalisations give priority for theme-like arguments, as illustrated in examples (3.32b) and (3.32c) in section 3.4.2.3 above.

(Nunes 1993) claims that semantic roles of the nominalisation arguments depend on the semantics of the underlying verbs. For instance, if the base verb contains a state predicate (a situation, an event or a process), its nominalisation takes an undergoer (a patient or a theme). Thus, only deverbals derived from activity verbs (describing an action) can take an agent as their argument, cf. (3.33a) and (3.33b) in section 3.4.2.3.

The differences in valency patterns can be also explained by the differences in selectional restrictions of nominalisatins and verbs, cf. examples (3.40a) and (3.40b) in section 3.4.2.3 above. In most cases verbal subcategorisation patterns provide the information on selectional restrictions relevant for different arguments. For example, the verb *to destroy* prefers an animate subject and allows both animate and inanimate objects. In the phrase *the city's destruction,* the possessive phrase *city* generally does not fill the subject role and therefore, is preferably the object. By contrast, in the phrase *Ed's assessment*, *Ed* is much more likely to be the subject of the verb *assess*.

Some authors explain the reasons for the "non-inheritance" diachronically. For instance (Wechsler 2008) claims that the nominalisation *growth*, derived from the originally intrasitive verb *to grow* (illustrated in example (3.19) in section 3.4.2.1 above)

entered the language before the innovation of transitive *to grow* in the specialised 'cultivate' sense. This could probably explain further cases in which the meaning of the nominalisation differs from the underlying verb. For instance, (Schierholz 2001) admits that some nominal predicates can occur as a variant of meaning (*Bedeutungsvariante*), which does not exist for the base verb, cf. examples in (3.43a) and (3.43b).

(3.39a) *der Anschlag **auf** den Präsidenten* ("attempt on the president's life")
vs. **jemand schlägt **auf** den Präsidenten an.* ("smb hits on the president")
vs. *Er soll an diesem Punkt anschlagen wenn er das Ziel erreicht hat* ("He should hit at that point when he achieves the goal").

(3.39b) *das Abkommen **über** die Zusammenarbeit* ("agreemant on the cooperation")
vs. **Jemand kommt "**uber** die Zusammenarbeit ab.* ("smb agrees on coopertaion")
vs. *Er kommt kommt von diesem Weg ab.* ("he deviates from this way").

The nominalisation *Anschlag* ("attack/attampt on smb's life") in (3.43a) is derived from the verb *anschlagen* ("to hit/strike"), but the meaning of the nominalisation *Anschlag auf* differs from the one of the verb *anschlagen*. The verb *abkommen* in (3.43b) underlies the nominalisation *Abkommen über* but does not posess the same meaning.

Ehrich and Rapp (Ehrich/Rapp 2000) claim that the predicate-argument structure of a nominalistaion is not derived from its base verb at all. Both the base verb and its nominalisation derive their subcategorisation frames from Lexical Semantic Structure with the help of linking rules, which control how the elements of the argument structure are syntactically realised. According to the authors, the argument structure of an *ung*-nominalisation consists of all thematic arguments of its Lexical Semantic Structures if the latter does not contain a BECOME term. Otherwise the argument structure consists solely of the effected argument with the lowest rank (beside the referential argument).

3.4.2.5 Summary: "inheritance" in nominalisations

In this section we summarise the above mentioned studies on the "inheritance" relations between subcategorisation properties of verbs and their derivated nouns. We outline the main valency properties of nominalisations, which do not correspond with those of their base verbs. Nominalisations can lose verbal properties ("inheritance" reduction) or gain some further features, which are not specific for their base verbs ("inheritance" extension), cf. table 3.15.

The non-correspondences are listed according to the main types of subcategorisation information, important in this thesis, cf. section 2.3 above. We do not describe correspondences between grammatical functions of verbs and nominalisations, as this information is usually missing for nominalisations. Most authors employ the terms 'subject' and 'object' to describe the verbal complements, which are inherited or lost by the nominalisations. The described features allows us to classify relations between verbs and their nominalisations not only according to the "inheritance" or

	"inheritance" reduction	**"inheritance" extension**
VP	- absence of complements expressed by the noun itself, cf. example (3.37) in section 3.4.2.3 - optionality of nominal complements	- presence of complements which are specific for nouns only, cf. NOUN-SUBC in NOMLEX
SC	- German nominalisations take PP, subclauses and NP in genitive/possessive complements only - nominalisations do not take all subclause types the verb does	some German verbs do not allow for PP complements although their nominalisations do nominalisations take another preposition than its base verb, cf. (3.42) in 3.4.2.3
SR	- different linking rules for semantic roles and their syntactic realisations for verbs and nominalisations - some semantic types of nominalisations do not take agents, e.g. *growth*, cf. (3.19) - nominalisations which express objects/subjects of a verb, cannot take objects/subjects, cf. example (3.37)	- selectional restrictions of nominalisations allow more argument types than those of the verb, cf. (3.40a) and (3.40b)

Table 3.15: "Non-inheritance" of verbal subcategorisation properties

"non-inheritance" features, but also according to the evidence of "inheritance" reduction or extension. We describe classification of "inheritance" relations between nominalisations and their base verbs in section 4.2.5 below. The automatic treatment of "inheritance" relations is described in section 5.3.4.

In table 3.16, we summarise different variants of the "non-inheritance" explanation, which are mentioned in the above described related work. The listed reasons, especially those of semantic character are particularly important for the explanation of the "non-inheritance" cases obtained within the extraction and classification procedures, cf. sections 5.3.4 and 6.1.3.2 below.

	reasons
VP	- nominalisation arguments are specific for nouns - all nominal complements are optional
SC	- nominalisations take PPs, subclauses and NPs in genitive/possessives only
SR	- difference in selectional restrictions of verbs and nominalisations - difference in semantics (meaning) of verbs and nominalisations, cf. (3.39a) and (3.39b)

Table 3.16: Reasons for the "non-inheritance" of verbal properties

Chapter 4

Acquisition and Classification of Predicates

In the following chapter we describe existing studies on automatic acquisition and classification of predicates under analysis. We start with the description of the related work on acquisiton of verbal, nominal and multiword predicates in section 4.1. The next part describes classification approaches for predicates under analysis based on different criteria, including our classification, which is based on subcategorisation properties. The final section describes classification of subcategorisation relations related to this work.

4.1 Acquisition of Predicates

This section describes a number of tools for (semi)-automatic acquisitoon of verbal, nominal and multiword predicates. Creating a computational lexicon by hand is time consuming, prone to errors and requires considerable linguistic expertise. Besides that, manually created lexicons cannot be easily adapted to specific domains. Therefore, there are attempts to automatise lexicon building.

4.1.1 Acquisition Tools for Verbal Predicates

The acquisition of verb subcategorisation from corpora is one of the main issues for most studies, as verbs play an important role in sentences and verbal valency frames provide information for the structural analysis of sentences.

There exist numerous works on induction of verbal subcategorisation. A summary of different approaches to automatical acquisition of verbal valency frames is given in (Schulte im Walde 2009). The author classifies them into five groups according to several dimensions they can be distinguished by: *corpus selection and preparation* (which corpus and what kind of annotation were selected), *frame types* (how many and which types of verb frames are distinguished), *acquisition method* (which computational methods are used), *filtering* (are subcategorisation frames filtered for noise, and what kind of method is used for filtering) and *evaluation* (how is the resulting frame information evaluated).

We do not intend to recall all the existing acquisiton tools and summarise a number of the acquisition methods according to the languages they are applied for. In table 5.2 below we list the works, which describe subcategorisation induction for different languages.

languages	**studies on acquisition**
EN	(Brent 1993), (Ushioda et al 1993), (Manning 1993), (Briscoe/Carroll 1997),(Kinyon/Prolo 2002), (Carroll/Fang 2004), (O'Donovan et al. 2005), etc.
DE	(Schulte im Walde 2002), (Schulte im Walde 2006), (Eckle-Kohler 1999), (Wauschkuhn 1999), etc.
	further languages
FR	(Chesley/Salmon-Alt 2006), (Messiant et al. 2008)
NL	(Spranger/Heid 2003)
CZ	(Sarkar/Zeman 2000)
PT	(de Lima 2002)
IT	(Ienco et al. 2008), (Lenci et al. 2008)
GR	(Maragoudakis et al 2001), (Georgala 2003)
AR	(Bielicky/Smrz 2008)

Table 4.1: Languages and NLP-tools for predicates acquisition

As seen from the table, most mentioned studies concentrate on verbal predicates in English, some of them describe verb subcategorisation induction for German and just a few deal with other languages.

4.1.1.1 Verbal predicates in English

In this section we summarise a number of studies on acquisition of verbal predicates for English.

Acquisition based on finite-state grammar (Brent 1993) focuses on discovering the kinds of syntactic categories semantic arguments of particular verbs can be realised with. The author obtain infinitives, tensed clauses, and NPs. English verbs are identified in a raw corpus as lexical items that appear both, with and without the suffix *-ing*. The verb complements are detected by a finite-state grammar, which define linear patterns, such as *'to V'* for infinitives. This approach is surprisingly successful, but cannot be extended to sufficiently cover all frame types. Moreover, it does not take into account all variation of verbs that can appear in each frame.

A finite-state method is also used by (Manning 1993) who applies a finite-state parser to parse only clauses with auxiliaries, relying on the restricted sentence structure. Subcategorisation extraction in this system is performed by a program that process the output of the stochastic part-of-speech tagger described in (Kupiec 1992). The parser includes a simple NP recogniser (parsing determiners, possessives, adjectives, numbers and compound nouns) and various rules to recognise certain cases that appear frequently.The constituents following the verb are identified as its complements. The described system acquires a dictionary of 4900 subcategorisation

frames for 3104 verbs, an average of 1.6 per verb. This approach can be reliable for a larger set of frame types, but restricts itself to a certain surface pattern, i.e., clauses with auxiliaries.

Acquisition based on probabilistic parser A more complex corpus annotation for verb valency extraction is used in (Briscoe/Carroll 1997) and (Carroll/Fang 2004). For instance, Briscoe and Carroll extract verbs and their complements from parsed corpora, making use of the ranked output analyses of a probabilistic parser trained on a treebank. They allow all patterns which occur according to their grammar. The system of (Briscoe/Carroll 1997) includes several components – a tagger, a lemmitiser, a probabilistic LR parser, a pattern extractor, a pattern classifier and a patternset evaluator – which are applied in sequence to sentences containing specific predicates in order to extract a set of subcategorisation classes for that predicate.

(Carroll/Fang 2004) use the same method described in (Briscoe/Carroll 1997). However, they enhance this with the method used in (Korhonen 2002) and present the automatic acquisition of verb subcategorisation frames from a domain-specific corpus[1], which is lemmitised, tokenised, part-of-speech annotated and syntactically parsed. The authors show that automatically extracted verbal subcategorisation patterns enhance the HPSG parser success rate by 15% in theoretical terms and by 4.5% in practice. This is a promising approach for improving the robustness of deep parsing.

Acquisition tool for subcategorisation evidence (Gahl 1998) presents a tool for extracting evidence for subcategorisation frames from British National Corpus (BNC, (Burnard 2007))[2]. Subcategorisation information is documented in subcorpora, which can be used both to provide evidence for subcategorisation properties of a given lemma, and to determine the frequencies of different syntactic contexts of each lemma. The extraction tool consists of a set of batch files applied with the the Stuttgart CorpusWorkBench (CWB, cf. (Evert 2005)). The tool is used as part of the lexicon-building process in the FrameNet project[3].

Acquisition of verbal predicates from Treebank Some approaches to extract verb valency rely on the information contained in a treebank. In (O'Donovan et al. 2005), verbal frames are derived after performing automatic annotation of the Penn Treebank with LFG structures[4]. The authors present an algorithm for extraction of subcategorisation frames from the Penn-II and Penn-III Treebanks, automatically annotated with LFG f-structures. In contrast to many other approaches, this one does not predefine subcategorisation frames being extracted. The system acquires syntactic-function-based subcategorisation frames (LFG semantic forms) and traditional CFG category-based frames, as well as mixed-function-category-based frames. The algorithm reflects the effects of long-distance dependencies and distinguishes between

[1] The authors use emails about models of mobile phones.
[2] http://www.natcorp.ox.ac.uk/
[3] http://www.icsi.berkeley.edu/~framenet
[4] See the LFG valency description in section 3.1.2.1

active and passive frames, which is particularly important for the accurate assignment of probabilities to semantic forms.

4.1.1.2 Verbal predicates in German and Dutch

Although approaches on the acquisition of verbal predicates in German can base on the above described methods, they differ because the properties of the respective languages, such as morphological marking, word order, etc. differ as well. We describe three works on the acquistion of verbal subcategorisation properties for German.

Statistical methods (Schulte im Walde 2002) presents an automatically induced computational subcategorisation lexicon for German verbs. The lexical entries were obtained by unsupervised learning in a statistical grammar framework of HL-PCFGs, cf. (Carroll/Rooth 1998). The author uses the unsupervised training environment to train on 18.7 million words of a large German newspaper corpora. The author develops a simple methodology to utilise frequency distributions in the lexicalised version of the probabilistic grammar for inducing syntactic verb frame descriptions. Subcategorisation information is extracted for more than 14,000 German verbs, for 38 purely syntactic frame types and a refinement of 178 frame types including prepositional phrase distinctions.

Acquisition based on pattern-grouping Another work on verb valency acquisition, fulfilled in Stuttgart, is described in (Wauschkuhn 1999), who constructs a valency dictionary for 1,044 German verbs with corpus frequency larger than 40. The author extracts a maximum of 2,000 example sentences for each verb from annotated corpus data and creates a context-free grammar for partial parsing. The syntactic analyses provides valency patterns, which are grouped in order to extract the most frequent pattern combinations. The common part of the combinations defines a distribution over 42 subcategorisation frame types for each verb.

Knowledge-based architecture (Eckle-Kohler 1999) develops the system KLAC[5], which is a knowledge-based and, therefore, a context-dependent principle. The author performs a semi-automatic acquisition of subcategorisation information for 6,305 verbs. The used corpora are pre-processed – annotated with lemma and part-of-speech information. The linguistic heuristics are defined in form of regular expression queries through the CWB (mentioned in 4.1.1.1 above) over the usage of 244 frame types including PP definitions. The extracted subcategorisation patterns are manually judged. The extraction steps with CWB are combined with a number of post-filtering steps. This method represents a simulation of a shallow parser with a special-purpose grammar for nearly unambiguous patterns. This system is precision-oriented, which means that some relevant verb candidates remain undetected in corpora.

[5] **K**nowledge-intensive **L**exicon **A**cquisition from **C**orpora.

Chunking-based acquisition for Dutch (Spranger/Heid 2003) present a chunker for Dutch, which is used to extract verb subcategorisation. They make use of robust NLP techniques to extract maximally informative data from the texts, without, however, having to rely on a highly complex grammar. Instead of that a chunk grammar is applied to account for complex structures. It only requires part-of-speech tagging and lemmatisation as an input. The chunking procedure follows the approach described in (Kermes 2003) for German. The chunker relies on part-of-speech and lemma annotations to identify boundaries of chunks and phrases, complex structures are built by embedding simple ones into each other. The chunker in (Spranger/Heid 2003) is based on the above mentioned CWB.

4.1.1.3 Verbal predicates in Romance languages: FR, IT, PT

In this section we summarise acquisition systems for French, which are described in (Messiant et al. 2008) and (Chesley/Salmon-Alt 2006), acquisition tools for Italian presented in (Lenci et al. 2008) and (Ienco et al. 2008), as well as one system for Portugese, described in (de Lima 2002).

Acquisition system for French based on dependency parser The system described in (Chesley/Salmon-Alt 2006) acquires French subcategorisation frames via VISL, a dependency-based parser[6]. The authors test occurrences of 104 frequent verbs from the Frantext[7] online literary database and obtain 27 different subcategorisation frames and 176 verb frame combinations. The syntactic constituents counted as possible frame elements are limited to the following: direct objects, PPs headed by special prepositions, subordinate clauses and small clauses with various heads, infinitive verbs in the case of raising and control verbs, predicative adjectival phrases, and reflexive clitic NPs. The resulting subcategorisation frames can consist of any combination of the above elements.

Multiple-modules acquisition system for French (Messiant et al. 2008) presents a system, which automatically induces large scale lists of subcategorisation frames from a large corpus for French. The system acquires subcategorisation information for 3267 verbs (which occur more than 200 times in the corpus) and 286 subcategorisation frames[8]. The system extracts complements expressed by NPs, infnitve clauses, PPs, subclauses and adjectival phrases. The corpus used in the system is tagged and lemmatised using TreeTagger, cf. (Schmid 1994), and sytactically annotated with the Syntex, a shallow parser specialised in the extraction of lexical dependencies, cf. (Bourigault *et al.* 2005). The acquisition system includes three modules: *extraction of verbs and surrounding phrases, building of subcategorisation frames* (based on morpho-syntactic information and relations between the verb and its arguments) and *filtering* (using statistical filter). The first module takes as input corpora annotated with Syntex and extracts each verb, which is sufficiently frequent (at least 200

[6] Cf. (Bick 2003).

[7] www.frantext.fr/categ.htm

[8] The extracted lexicon is freely available in the web
http://www-lipn.univ-paris13.fr/∼messiant/lexschem.html

occurences). The second module considers the dependencies according to their syntactic category, e.g. NP, and to their grammatical function. The module reconstruct frames with the help of these features. The third module filters the results, which is necessary because of tagging and parsing errors.

Unsupervised automatic acquisition for Italian In (Lenci et al. 2008), the authors report experiments of unsupervised automatic acquisition of Italian and English verb subcategorisation frames from general and domain corpora. This method operates on syntactically shallow-parsed corpora on the basis of a limited number of search heuristics not relying on any previous lexico-syntactic knowledge about subcategorisation frames. The main advantage of such a strategy is that there is no need to presuppose any strict definition of frame structures and to distinguish between the subcategorised arguments and optional adjuncts. The authors also compare verbs, which share similar frames by clustering verbs them using the Minimum Description Length Principle (MDL)[9].

Acquisition tools for Italian based on statistic methods The system described by (Ienco et al. 2008) is based on statistical subcategorisation extraction methods. It is applied on Italian treebank that exploits a rich set of dependency relations. The data set consists of 2,000 Italian sentences from a dependency-based treebank, i.e. the Turin University Treebank (TUT)[10]. The authors apply measures related to the T-Score on the TUT corpus in which annotation of relations include functional-syntactic component. From the training set, they acquire 50 verb with a frequency greater than 5 occurrences and evaluate the corresponding 2,452 subcategorisation frames. Regardless of the small size of the corpus and the rich set of grammatical relations implemented by TUT, the experiments, described in (Ienco et al. 2008), produce satisfactory results.

PCFG-based system for Portugese (de Lima 2002) describes a system to induce lexical information from Portugese text corppora. The author is concerned with both syntactic and morphological information. For this purpose she utilises the HL-PCFG (head-lexicalised probabalistic context-free grammar) framework and emlpoys the LoPar system11 to estimate the parameters of this grammar.

4.1.1.4 Verbal predicates in further languages

The current section describes three acquisition systems for verbal predicates in Czech and Greek.

Acquisition system for Czech based on machine learning (Sarkar/Zeman 2000) present machine learning techniques to identify subcategorisation information for Czech verbs. They use the syntactic dependency definitions in the Prague Dependency Treebank (PDT)[11], to induce subcategorisation frames. A frame is defined as

[9] Cf. (Li/Abe 1998).
[10] Download and more details at http://www.di.unito.it/∼tutreeb
[11] Cf. (Böhmová et al. 2001).

a subset of the annotated verbal dependents contained in the treebank. One of the main features is the argument-adjunct distinction in frames. The authors extract the following argument types: NPs, PPs, clauses, infinitives, reflexive pronouns, passive particles and adverbs. The main aim is to include the obtained subcategorisation frames into a parser for Czech.

Chunking-based statistic acquisition for Greek (Maragoudakis et al 2001) develop a method to obtain verb subcategorisation frames from chunked corpora automatically by using statistic metrics, such as Log Likelihood Statistic and T-score. The authors apply minimum of linguistic resources, such as morphological tagging or phrase chunking, to demonstrate that subcategorisation can be achieved using large corpora without having any general-purpose syntactic parser at all. Moreover, they do not annotate the whole set of training data and increase the performance of the subcategorisation frames learner. The authors estimate that using a free error chunker and eliminating the problem of the conjunction phrases enables the precision higher than 75%.

4.1.2 Acquisition Tools for Nominal Predicates

In this section we summarise existing studies on induction of nominal subcategorisation, which are less numerous than those describing verbal valency acquisition.

Knowledge-based acquisition for German The KLAC system described in section 4.1.1 above is designed to obtain not only subcategorisation patterns for verbs but also those of nominal predicates. The author aims at semi-automatical extraction of a syntactic NLP-lexicon containing full information on subcategorisation properties of different kinds of predicates. Nominal subcategorisation frames acquired by KLAC contain subclauses, infinitives and prepositional complements. Automatic induction of subcategorisation frames is elaborated as a step-by-step abstraction from a concrete word in a concrete sentence to the subcategorisation frame of this word. Automatic extraction is combined with a number of post-filtering steps.

Acquisition based on semantic mapping (Gurevich et al. 2007) attempt to canonicalise deverbal nouns for a knowledge representation within the PARC processing system[12]. Within this system, a text is first parsed with the XLE parser, cf. (Crouch *et al* 2006), then the parse output is rewritten into semantic representations, which are then induced into the abstract knowledge representation. Nominal complements are mapped onto those of their base verbs. To identify deverbal nouns, the authors use WordNet[13] and a list of verbs from the XLE lexicon. The list obtained from WordNet consists of deverbal nouns derived with overt morphology, e.g. *statement* from *state*, words ambiguous between part of speech, e.g. *travel*, or verbs derived from nouns, e.g. *criticise* from *critic*. Nominalisations are obtained along with their base verbs. Afterwards, subcategorisation frames for corresponding verbs are extracted and the system identifies verbal and nominal arguments.

[12]PARC is described in section 3.4.2.2

[13]WordNet is described in section 3.1.2.2 above, cf. (Fellbaum 1998).

Acquisition tool for subcategorisation evidence (Gahl 1998) describes a system to extract subcorpora containing not only verbal subcategorisation frames, but also nominal ones. The system can induce PPs, subclauses, infintives and gerundial complements. This tool relies on subcategorisation information as its input and is not capable of automatically learning subcategorisation frames, e.g. the ones missing in dictionaries or omitted in the input file. The tool facilitates the discovery of evidence for new subcategorisation frames.

Acquisition of semantic roles for nominalisations (Hull/Gomez 2000) describe a computational approach to the semantic interpretation of nominalisations, which also invloves determining semantic roles of nominalisations. The system intends to distinguish between verbal and non-verbal senses of nominalisations, as well as to determine prepositional complements of nominalisations and their semantic roles. The parser used by the semantic interpreter can extract a number of grammatical functions, e.g. subject, object1, object2, and prepositional complements, cf. (Hull/Gomez 2000).

4.1.3 Acquisition Tools for Multiword Predicates

The acquisition of subcategorisation frames for multiword predicates is complicated, as it is mostly difficult to detect multiwords themselves. However, there exist a few studies on morpho-syntactic properties (which include subcategorisation) of multiword extractions, which apply different techniques to achieve the desired results. In this section we describe both, studies that present systems for acquisition of multiwords and studie,s which describe extraction of multiword subcategorisation.

Acquisition of multiwords for German with MI (Breidt 1993) evaluates the usefulness of the statistical approach Mutual Information (MI)[14] for the extraction of verb-noun collocations from German text corpora. The author tests how much can be done with an untagged corpus and what might be gained by lemmatising, part-of-speech-tagging or even superficial parsing. She uses two untagged corpora – the 'Mannheimer Korpus I' (MK1, 2,7 Mio. words) and the 'Bonner Zeitungskorpus' (BZK, 3.7 Mio. words). MI is a function used for the statistical characterisation of collocations. It compares the joint probability of the occurrence of two words within a predefined distance. The author concentrates on support verb constructions. The chosen support verbs belong to the most frequent verbs in the corpus. The author extracts multiwords for 16 verbs, e.g. *bleiben* ("to stay"), *bringen* ("to bring"), *erfahren* ("to find out"), etc. The evaluation of extraction results shows that acquisition from lemmatised, tagged and parsed corpora is more effective due to some properties of the German language, e.g. strong inflection or variable word order.

Acquisition of multiwords for Dutch using statistic methods The system described in (Bouma/Villada 2002) aims at extracting collocational prepositional phrases, e.g. *ten koste van* ("at the expense of"), *in het kader van* ("in the framework of"), etc.

[14] Cf. (Church/Hanks 1989).

To find candidate strings, they extract all instances of the relevant pattern from corpora annotated with part-of-speech tags. Beside the above mentioned MI-test, the authors apply Log-Likelihood score and Paerson's χ^2 test[15]. The authors provide information on a number of idiosyncratic syntactic properties (archaic prepositional and nominal forms and inflection, absence of a determiner, restricted possibilities for modification, restricted functionality as complement). However, they do not provide any systematic description of these properties.

Acquisition of morpho-syntactic properties for German collocations Tools for extraction of collocations along with their morphosyntactic properties are described in (Ritz/Heid 2006). The authors use part-of-speech-tagged and partially parsed German corpora. The extraction tools elaborated by (Ritz/Heid 2006) can identify both collocation candidates and their morpho-syntactic properties and preferences. Therefore, the extraction procedures include two steps: pattern matching and feature determination.

Acquisition of morpho-syntactic properties for Polish collocations An approach for semi-automatic acquisition of morpho-syntactic properties of verb-noun collocations for Polish is presented in (Vetulani *et al.* 2008). This corpus-based acquisition allows to enlarge the verb-noun collocation dictionary for Polish, which is a part of the full lexicon grammar for Polish (SyntLex). The acquisition process includes three phases: the dictionary-based acquisition of collocation lexicon, which is called Baseic Resource (BR), the feasibility study for corpus-based enlargement of the BR and the corpus-based lexicon enlargement. The dictionary-based phase includes manual extraction of collocations from existing dictionaries by a linguist-lexicographer. The second phase is based on the corpus based acquisition of new collocations described in (Vetulani *et al.* 2006). The third part involves semi-automatic transformation of the corpus.

4.1.4 Summary: Related Work on Predicate Acquisition

Analysing the different approaches and methods of the above described studies, we summarise those of them, which are important for the elaboration of our extraction architecture. Table 4.2 presents the list of these features based on the dimensions for acquisition of verbal predicates mentioned in section 4.1.1 above. We apply these features in the extraction of all predicate types under analysis.

For our extraction and classification procedures, we use pre-processed corpora as specified in section 5.1.1. The annotation procedures applied for the corpora under analysis are described in section 5.1.2. As our aim is to compare subcategorisation properties of different predicate types, we decide for the extraction of valency patterns specific for all predicate types under analysis, e.g. subclauses. We intend to achieve high precision in our extraction results and therefore, we apply context-based methods, similar to those described for KLAC, cf. (Eckle-Kohler 1999). Our experiments with dependency-based methods show that their application would not

[15] See (Manning/Schütze 1999) for more details.

dimensions	analysed studies	present study
corpora	- pre-processed: annotated with lemma, token, part-of-speech information, KLAC, (Breidt 1993), (Carroll/Fang 2004) and others - parsed corpora and treebanks, (O'Donovan et al. 2005), (Messiant et al. 2008), etc. - non-annotated, (Maragoudakis et al 2001)	pre-processed
frames	- based on restricted patterns, (Manning 1993), (Gahl 1998), (Schulte im Walde 2002), etc. - is not restricted to patterns, (Lenci et al. 2008), (Messiant et al. 2008), etc.	restricted frames with subclauses
method	- context or knowledge-based methods, (Manning 1993), (Breidt 1993), KLAC, etc. - dependency-based, (O'Donovan et al. 2005), (Chesley/Salmon-Alt 2006) and others - statistical methods, (Schulte im Walde 2002), (Ienco et al. 2008), etc.	context or knowledge-based extraction
filtering	- application of filtering procedures, KLAC and (Messiant et al. 2008)	application of filtering procedures
evaluation	- evaluation of the results, (Breidt 1993), (Briscoe/Carroll 1997) and (Ienco et al. 2008)	precision and recall

Table 4.2: Features of acquisition systems relevant for the present study

increase the precision of our results, cf. section 5.1.3 below[16]. To increase the accuracy of acquisition procedures, we apply a set of filtering procedures, presented in 5.3.1.4. Results of the acquisition with our system, as well as their evaluation are analysed in chapter 6.

4.2 Classification of Predicates

The present section describes related studies on classification of verbs, nouns and multiwords, which are summarised according to the used classification criteria. Furthermore, we present our own classification, which is based on subcategorisation properties in sections 4.2.1.3, 4.2.2.5, 4.2.3.3 and 4.2.4.4.

4.2.1 Classification of Verbal Predicates

Most studies on classification of verbs concentrate on verb classes at the syntactic-semantics interface, cf. (Schulte im Walde 2009). Verb classes are typically constructed on the basis of features describing verb behaviour, particularly with respect

[16]Cf. the work on the syntax-based identification of collocations in (Seretan 2008).

to the choice of their complements, cf. (Pinker 1989) and (Levin 1993) among others. From a practical point of view, such verb classes have successfully been applied in NLP. For instance, the English verb classification by (Levin 1993) is widely used in NLP applications such as word sense disambiguation, machine translation, document classification, and subcategorisation acquisition.

In table 4.3 we summarise a number of studies on verb classification sorting them according to the used classification criteria. Aspectual features are inheritent temporal properties of verbs, semantic aspects are based on verb meaning and syntactic ones are based on verb behaviour or verbal subcategorisation properties.

criteria	**studies**	**classes**
aspectual	(Vendler 1967)	states, achievenments, accomplishments, activities
	(Kiparsky/Kiparsky 1970)	factive, half-factive, non-factive
	(Brent 1991), (Siegel 1998)	states, events
semantic only	(Schulte im Walde 2002) and (Schulte im Walde 2006)	manner of motion, emotion, desire, propositional attitude, communication, observation, description, etc.
syntactic and semantic	(Levin 1993)	put, remove, send, give verbs, etc.
	(Merlo/Stevenson 2001)	unergative, unaccusative, object-drop
	(Grimshaw 1990)	transitive agentive, ditransitive, unergative, etc.
	(Klotz 2007)	according to the meaning - communication, opinion, fact finding, etc. according to complementation - seven classes
	(Bäuerle/Zimmermann 1991) and (Fischer 2005)	those which subcategorise for both declaratives and interrogatives, only interrogatives, only declaratives

Table 4.3: Verb classification according to different aspects

4.2.1.1 Aspectual criteria

As seen from table 4.3, aspectual verb classes based on inherent temporal properties are described by (Vendler 1967), (Brent 1991) and (Siegel 1998). (Vendler 1967) proposes four basic classes - states, achievements, accomplishments and activities. **States** are non-dynamic and temporally unbounded, e.g. *love, know believe, have, be sick, be dead*. **Achievements** code instantaneous changes, usually changes of state but also changes in activities. They have an inherent terminal point, for example, *pop, explode, collapse, shatter*. **Accomplishments** (or **processes**) are temporally extended changes of state leading to terminal point, e.g. *melt, freeze, dry (the intransitive versions), recover from illness, learn*. **Activities** or **actions** are dynamic and temporally unbounded, such as *march, walk, roll, swim, think, rain, read, eat*.

(Vendler 1967)also describes verbs, which can be classified according to the semantics of the complements they take. For instance, the verbs *to assert* and *to believe in* take propositional complements, the verbs *to know* and *to regret* take

factive complements, while the verbs *to hear* and *to continue* take complements referring to events (actions and processes). The author identifies the distinction between predicates, which take complements referring to facts and other predicates with the distinction between 'non-factive' and 'factive' predicates. This classification coincides with the verb classes described by (Kiparsky/Kiparsky 1970), cf. section 2.2.2.3 above.

4.2.1.2 Semantic and syntactic criteria

(Schulte im Walde 2006) describes a classification of verbs based on semantic features. The author classifies German verbs into 43 concise semantic verb classes. The verb class labels refer to the common semantic properties of the verbs in a class at a general conceptual level. Idiosyncratic lexical semantic properties of the verbs are left underspecified. The classification is based on semantic intuition. However, verbs grouped in one class share certain aspects in their syntactic behaviour. The classification target are semantic verb classes such as **manner of motion**, **emotion**, **desire**, **propositional attitude**, **communication**, **observation**, **description**, etc. Some of them are subdivided into subclasses, e.g. verbs of emotion have three subgroups: origin, expression, objection.

Most studies describe verb classes based both on their semantic and syntactic features. For example, (Levin 1993) presents a classification based on the representation of verb meaning and its association with the syntactic expression of verb arguments. Classificatory disctinctions include expression of verbal arguments and their morpholigical properties, e.g. derived adjectives and nouns. In this cross-classification, over 3,000 verbs are grouped into semantically coherent classes with similar meanings. 'Diathesis alternations' include syntactic alternations that verbs are subject to. For instance, transitivity alternations, such as middle, causative-inchoative, instrument alternation, or alternations involving arguments within a verbal phrase, such as dative shift or double object constructions and others, cf. examples (4.1a) and (4.1b).

(4.1a) *David broke the window with a hammer.*
vs. *The hammer broke the window.* (intermediary instrument)

(4.1b) *Doug ate the ice cream with a spoon.*
vs. **The spoon ate the ice cream.* (enabling/facilitating instrument)

Levin's classification includes 48 verb classes sorted according to their meaning. These groups are usually subdivided into five subclasses. Examples of verb classes include **put verbs, pocket verbs, remove verbs, banish verbs, send verbs, give verbs**, etc. The classification proposed by (Levin 1993) is used by many scholars and in many NLP applications. For instance, VerbNet which is mentioned in section 3.4.2.3 contains syntactic and semantic information based on Levin's classification.

In the above mentioned FrameNet (cf. sections 2.1.3 and 3.1.2.2), verbs are grouped according to the conceptual structures or frames that underlie them and their combinatorial patterns. This means that the grouped verbs are semantically

similar but have different alternations. Verbs sharing the same alternation are represented in two different semantic frames[17]. This differs this classification from the one presented by (Levin 1993).

(Merlo/Stevenson 2001) elaborate an automatic classification of verbs, focusing on their *transitivity, causativity, animacy*, and *syntactic features*. The authors distinguish between three verb classes – **unergative, unaccusative**, and **object-drop** verbs – defined according to their argument structure. Semantic roles assigned by verbs to their arguments represent relational semantics at the syntactic level. For instance, the degree of animacy of subject roles is estimated as the ratio of occurrences of pronouns to all subjects for each verb. This is based on the assumption that unaccusatives occur less frequently with an animate subject. The knowledge of argument structure in this classification captures fundamental participant/event relations, which is crucial in parsing and generation, machine translation and in information extraction. Unergative verbs describe manner of motion and include verbs as *jump, rush, march, leap, oat, race*, etc. Unaccusative verbs indicate change of state, e.g. *explode, dissolve, crack, harden*, etc. Object-drop verbs have unexpressed object alternation, for instance, *play, paint, kick, carve, reap, wash, dance* and others.

(Grimshaw 1990) combines aspectual and semantic criteria based on selectional restctions of verbs. The author describes semantic roles, which are typically assigned to arguments of the verbs of different classes: **transitive agentive** – (Agent(Theme)), **ditransitive** – (Agent(Goal(Theme))), **unergative** – (Agent), **psychological state** – (Exp(Theme)), **psychological causative** – (Exp(Theme)), **agentive psychological causative** – (Agent(Exp)).

(Klotz 2007) classifies verbs according to both, their complementation behaviour and semantic properties. The author describes seven complementation classes, which are outlined in table 4.4. This classification is based on the three complement types: that-CL (*that*-clause), N to-INF (infinitive complement) and N V-ing (gerund complement).

class	that-CL	N to-INF	N V-ing
class1	x	x	x
class2	x	x	
class3	x		x
class4		x	x
class5	x		
class6		x	
class7			x

Table 4.4: Complementation classes according to (Klotz 2007)

The author semantically classifies verbs of complementation classes 2 and 5 (those that can take only *that*-clauses vs. those allowing for both *that*-clauses and infinitive complements). The author analyses 112 verbs, which are classified into 9 semantic groups: **communication verbs, opinion verbs, fact finding verbs, fact demonstrating verbs, fact manipulating verbs, fact establishing verbs, emotion verbs, imagination verbs** and **unclassified**, such as *respect, vote* and *wonder*. The tested

[17] See (Baker 2000)

data show that both complementation classes co-occur with all semantic groups.

(Bäuerle/Zimmermann 1991) and (Fischer 2005) classify verbs that allow for sentential complements into three classes: those that license only declaratives, those that allow for interrogatives and those that embed both types of sentences. For instance, (Fischer 2005) distinguishes verbs like *wissen* ("to know"), which license three complement types: *dass, w-* and *ob*-clauses, verbs like *sich fargen* ("to ask oneself"), which allow only for *w-* and *ob*-clauses and verbs like *glauben* ("to believe") subcategorising only for *dass*-clauses. This classification is also based on both syntactic (subclause type) and semantic (selectional restrictions of verbs) features of verbs.

4.2.1.3 Verb classes related to this study

We aim at classifying verbal predicates according to their subcategorisation properties. Although most above mentioned classification approaches involve verbal valency features, e.g. (Schulte im Walde 2006), (Levin 1993), (Merlo/Stevenson 2001) and (Grimshaw 1990), not all of them include criteria based on both syntactic features of verbs and their selectional restrictions, which are important for the present analysis, cf. section 2.2.2.3. As we analyse verbal predicates, which allow for declarative *dass* and/or interrogative *w-/ob*-clauses, our classification is based on the classes distinguished by (Bäuerle/Zimmermann 1991) and (Fischer 2005). We distinguish three classes depending on the relationship between verbs and the types of subclauses they subcategorise for.

The first class, **V1**, includes verbs that license both declarative *dass* and interrogative *w-/ob*-clauses. Verbs that belong to the **V2** class, allow for interrogative *w-/ob*-clauses only. The third class, **V3**, consists of verbs which take declarative *dass*-clauses only, cf. table 4.5.

class	subcategorisation features	DE example	EN translation
V1	interrogatives and declaratives	*äußern* *entscheiden*	to express to decide
V2	interrogatives only	*abstimmen* *abfragen*	to vote/agree to request/ask
V3	declaratives only	*berichtigen* *sich etw. einbilden*	to correct to imagine sth

Table 4.5: Classification of verbal predicates related to the present study

As we analyse interrogative *w-* and *ob*-clauses in one category, cf. section 2.2.1.1 above.

Additionally, the V1 class includes verbal predicates, which take *dass* and *w-/ob*-clauses under certain contextual parameters only, cf. section 2.2.2.3. For instance, the verb *denken* subcategorises for *w-/ob*-clauses provided it is used with the Korrelat *darüber*. However, we do not classify this verb into the V3 class (verbs taking *dass*-clauses only). We claim that this would cause the incompleteness of the subcategorisation properties of such predicates. Therefore, we categorise such verbs as V1 types without indication of the contextual parameters under which they occur with certain subclause types. Contextual parameters are taken into account for the ex-

planation of "non-inheritance"-phenomena between subcategorisation properties of verbs and their derivatives.

We present the procedures to classify verbal predicates automatically according to the types given in table 4.5 in section 5.3.3.1 below.

4.2.2 Classification of Nominal Predicates

Nominal predicates can also be classified according to semantic, syntactic and morphological criteria. In this section we summarise several approaches on the classification of nouns. In table 4.6, we list classes of nominal predicates according to the classification criteria applied in approaches.

criteria	**studies**	**classes**
aspectual	(Grimshaw 1990), (Ehrich 1991)	facts (propositions) and events
apsectual and functional	(Sommerfeldt/Schreiber 1983)	Nomina agentis, relations, actions, processes, states, features
	(Levi 1978)	act, product, agent, patient
	FrameNet	nouns that denote events, relational nouns, artifact nouns, etc.
semantic	(Krifka 1991)	count and mass nouns
	(Sommerfeldt/Schreiber 1983)	concrete and abstract nouns
subcategorisation: quantitative restrictions	(Sommerfeldt/Schreiber 1983)	avalent, with one, two, three or four arguments
	(Teubert 2003)	agentive vs. object complements
morphological	traditional theories	simplex and compound nouns
derivation	traditional theories	bare nouns and nominalisations
nominalisations		
derivation morphological	traditional theories	deverbal and deadjectival
	(Ehrich 1991) and others	infinitive and derivative

Table 4.6: Classes of nominal predicates according to different criteria

4.2.2.1 Aspectual and functional criteria

As most valent nouns are derived from verbs, many authors concentrate on the analysis of deverbal nominal predicates only, classifying them according to their aspectual features into **states** and **events**, cf. (Grimshaw 1990) and (Ehrich 1991).

(Ehrich 1991) points out that events are place or time entities, which occur in certain places and at certain time, whereas facts do not have any place or time features, they are statements about the world. The author claims that negated nominal predicates cannot express events, as it is impossible to find time or place where the "non-occurrence" of the event can take place, cf. examples in (4.2a) and (4.2b). Therefore, Ehrich distinguishes two basic semantic categories of nominalisations: **nominalisations of propositions**, which allow for negations and **event nominalisations**, which do not allow for negations.

(4.2a) *Hans hat die Ankunft des Zuges gefilmt.*
("Hans has taken a video of the arrival of the train").

(4.2b) **Hans hat die Nicht-Ankunft des Zuges gefilmt.*
*("Hans has taken a video of the non-arrival of the train").

Some authors mention further classes based on aspectual and functional features of nouns. (Sommerfeldt/Schreiber 1983) mention the subdivision into the following classes: **Nomina agentis** (*Täterbezeichnungen*), such as *Lehrer* ("teacher"), *Besucher* ("visitor"), etc.; **relations** (*Beziehungsbezeichnungen,*) like *Vater* ("father"), *Freund* ("friend") and also *Präsident* ("president"). The class of abstarct nouns consists of **actions** (*Tätigkeitbezeichnungen*), such as *Spielen* ("play") in *Spielen der Kinder* ("play of kids") or *Einsteigen* ("getting in") in *Einsteigen der Fahrgäste* ("getting in of the passangers"); **processes** (*Vorgangsbezeichnungen*), e.g. *Beginn* ("start"), *Wachstum* ("growth"), etc.; **states** (*Zusatandsbezeichnungen*), e.g. *Aufenthalt* ("stay") or *Verzweiflung* ("desperation"), etc.; **features** (*Eigenschaftsbezeichnungen*) like *Ähnlichkeit* ("similarity") or *Länge* ("length") and others.

(Levi 1978), who analyses complex nominals, distinguishes four classes: **acts** (*enemy invasion, birth control, dream analysis*, etc.), **products** (*oil imports, editorial comment*, etc.), **agents** (*sound synthesizer, financial analyst*, etc.) and **patients** (*student inventions, designer creations, city trainees* and others).

In FrameNet nominals are described as **nouns that denote events** (*withdrawl, replacement*), **relational nouns** (*brother, girlfriend*), **artifact nouns** (*house, vest*) and others. Event and relational nouns are frame-evoking, which means that they posess subcategorisation features.

4.2.2.2 Semantic criteria

Semantic criteria are related to aspectual and functional criteria but depend even more on the meaning of nominalisations. (Sommerfeldt/Schreiber 1983) combine **actions**, **processes**, **states** and **features** into the class of **abstract** nouns, whereas **Nomina agentis** and **relations** are joined into the class of **concrete** nouns,

(Krifka 1991) also mentions similar classes of nouns, *Individualnomina* ("count nouns") and *Massennomina* ("mass nouns"). The latter can be subclassified into *Stoffnomina* ("material nouns") and *Kollektivnomina* ("collective nouns").

4.2.2.3 Morphological and derivation criteria

Most approaches in linguistics and NLP divide nouns into simplex and complex or compound ones, although some authors, e.g. (Levi 1978) also distinguishes between compound and complex nominals, arguing that the latter represent a bigger group of collocations, which includes not only compounds (noun-noun collocations in English) but also adjective-noun collocations, see section 3.4.1.2.

Morover, nouns can be classified according to their derivational origin. There are **bare nouns** (not derived) and **derived ones** (nominalisations), which can be derived either from verbs (**deverbals**) or from adjectives (**deadjectivals**).

According to (Ehrich 1991), nominalisations can be classified into **infintive** and **derivative**[18]. German infinitive nominalisations have a form similar to verbal infinitives, ending with the suffix *-en*. Derivative nominalisations can be generated with the help of different elements: *-ung, -tion, -er, -t, -t, θ, Ge-...-e* and *-erei*.

[18]Cf. **gerundive** vs. **derived** in (Chomsky 1970).

4.2.2.4 Criteria according to the subcategorisation properties

(Sommerfeldt/Schreiber 1983) also give a syntactic classification of nominal predicates, which depends on the number of arguments predicates can open. Thus, the authors distinguish between nouns without any arguments, such as *Donnern* ("thundering") or *Regnen* ("raining"), nouns with one argument, e.g. *das Fallen des Laubes* ("fall of the leaves"), nouns with two arguments like *der Stolz des Sportlers auf den Sieg* ("the proud of the sportman about the win") and nouns with four arguments, such as *die Lieferung der neuen Waren an die Verkaufsstelle durch den Großhandel* ("the delivery of the new products to the shops by the non-retail seller").

(Teubert 2003) also classifies nominal predicates according to their subcategorisation features. But this classification is based on the type of the complements they allow for. The author distinguishes two classes: those that subcategorise for *Agentivergänzung*, "agentive complement", e.g. *Ermittlung der Polizei* ("investigation of the police") and those that subcategorise for *Sachergänzung*, "object complement", such as *Vorrat an Erdöl* ("supply for oil").

4.2.2.5 Nominal classes related to this study

In this study we follow classifications based on the subcategorisation properties of nominal predicates. However, this classification is based on the type of the complements they allow for. For instance, the features of *-ung* and other types of nominalisations, the difference between *facts* and *events*, etc. We also distinguish between simple and compound nominal predicates as the latter possess their own subcategoirsation features, which should be taken into account.

We classify nominalisations that occur freely (not within a multiword) in a sentence into three groups according to their subcategorisation features. As mentioned in section 4.2.1 above, our classification is based on the type of sentential clauses predicates can take. Nominal predicates are classified according to the same criteria used for the classification of verbal predicates and rely on the relationship between nouns and their subcatergorisation properties.

The first class, **N1**, includes nouns that license both interrogative *dass* and declarative *w-/ob*-clauses. Nominals that belong to the **N2** class, allow for interrogative *w-/ob*-clauses only. The third class, **N3**, cconsists of nominal predicates, which take declarative *dass*-clauses only, cf. table 4.7.

class	**complementation features**	**DE example**	**EN translation**
N1	interrogatives and declaratives	*Entscheidung* *Meingung*	decision opinion
N2	interrogatives only	*Befragug* *Überlegung*	question consideration
N3	declaratives only	*Ankündigung* *Bestätigung*	announcement confirmation

Table 4.7: Classification of nominal predicates related to this study

We describe the automatic classification of nouns according to the types given in

table 4.7 in section 5.3.3.2 below.

4.2.3 Classification of Compound Nominals

Compound cominals can also be subclassified into further groups according to morphological or derivational criteria or according to the relations between their head and non-head constituents. In this section, we analyse related work on classification of compound nominals. In table 4.8, we list classes of compound predicates according to the classification criteria applied in approaches.

criteria	**studies**	**classes**
derivational	(Levi 1978), (Maxwell 1995)	verbal and non-verbal
relations between head and non-head	(Marchand 1969) and others	endocentric and exocentric
	(Bisetto/Scalise 2005)	subordinate, coordinate and attributive
	(Levi 1978)	subjective, objective and multi-modifier

Table 4.8: Classes of compound nominal predicates according to different criteria

4.2.3.1 Derivational criteria

According to their derivational origin, compounds can be divided into two classes: **deverbal** and **non-deverbal** compounds. The head of deverbal compounds is derived from a verb (possibly also an adjective), and the non-head is an argument of the base verb. The head of non-deverbal compounds is not derived from a predicate. In some cases the head of such compounds is derived from a predicate but the non-head is not its argument. This approach is supported by (Levi 1978) who calls the processes by which compounds are derived 'predicate nominalisation' and 'predicate deletion'. Predicate nominalisation is the process of derivation of deverbal compounds, whereas predicate deletion is the process by which non-deverbal compounds are derived.

(Maxwell 1995) distinguishes three main subclasses: compounds with a deverbal head (where the non-head functions as an argument of the deverbal base of the head), compounds with a simplex head (where the non-head functions as a classifying argument) and compounds with a deverbal or simplex head (where the non-head functions as a modifier). This approach, which is also called syntactic in linguistic literature, is controversial. Some authors argue that there is no formal structural distinction between deverbal and non-deverbal compounds. We admit that the relations, which are inherent in the classes of compounds differentiated by their derivational structure, usually vary, and therefore, this distinction can contribute to automatic reavealing of their features, e.g. their subcategorisation properties.

As we assume that subcategorisation behaviour of nominal compounds depends in some cases on the derivation of their constituent parts, we distinguish between compounds, which have deverbal head, compounds that have a deverbal non-head, compounds whose both constituents are deverbal and compounds whose both constituents are non-deverbal, cf. section 6.1.1.2.

4.2.3.2 Relations between heads and non-heads as criteria

(Marchand 1969) subdivides nominal compounds into two big classes: **endocentric**

and **exocentric**, cf. table 4.9. Endocentric compounds are also called semantically transparent because the meaning of such a compound can be derived from the meaning of its elements. For example, *laser printer* is transparent - "a printer that uses a laser". An endocentric compound is often a hyponym of its head. *Desktop computer* is endocentric because it is a kind of computer.

(Marchand 1969) admits that all compounds can be explained on the basis of the syntactic relations underlying the corresponding sentences. This assumption is supported by other authors. For instance, (Bisetto/Scalise 2005) propose a classification based on the assumption that two constituents are linked by the grammatical relation, which is not overtly expressed (cf. *apron string* vs. *string of the apron*). The grammatical relations between these two constituents are the relations that hold in syntactic constructions: subordination, coordination and attribution. Therefore, they distinguish between **subordinate, coordinate** and **attributive compounds**. Subordinate compounds have a complement (subordinate) relation between the two constituents. In the compound *taxi driver*, *taxi* is the complement of the deverbal head. Coordinate compounds are represented by those formations in English whose constituents are tied by the conjunction. They are potentially recursive even in Romance languages, e.g. in Italian *poeta pittore regista*, which mean "poet-painter-director". Attributive compounds are formed either by a noun and an adjective, as in *blue cheese*, where the adjective expresses a property and is in a modifier relation to the noun, or by two nouns, where the non-head is used metaphorically expressing an attribute of the head, e.g. *snail mail* or *sword fish*.

(Levi 1978) also suggests a classification of compounds, which is based on the syntactic source of their prenominal modifier. Thus, there are **subjective, objective** and **multi-modifier compounds**. The prenominal modifier of subjective compounds is derived from the underlying subject of the nominalised verb, e.g. in *manager attempts, cell decomposition, faculty decision*, etc. The prenominal modifier of an objective compound is the direct object of the base verb, e.g. *birth control, heart massage, draft dodger* and others. In multi-modifier compounds, both, the subject and the object emerge as prenominal modifiers, cf. *industrial water pollution, student monetary demands*, etc.

4.2.3.3 Classes of compound nominals related to this study

Our classification is based on the relations between subcategorisation properties of heads and non-heads of compound nominals, cf. section 3.4.1.2 above. As we assume that not only the head can determine subcategorisation features of a compound, we distinguish between three classes of nominal compounds[19]: **C1, C2** and **C3**. The **C1**-compounds share their subcategorisation features with the head, the **C2**-compounds share their subcategorisation features with the non-head. The **C3**-compounds are subclassified into two further groups: **C3-1** and **C3-2**. The **C3-1**-compounds include those that share their subcategorisation properties both, with the head and the non-head, and the **C3-2**-class includes compounds, which share their subcategorisation properties with neither the head nor the non-head constituents, cf. table 4.10.

The semi-automatic approach to classify compounds according to the types given in table 4.10 is described in section 5.3.3.3 below.

[19] Cf. (Lapshinova/Heid 2008).

type & description		example
Endocentric Compounds		
a)	a noun is determined by the stem form of another noun	*rainbow*
b)	two nouns may form a group of notionally co-ordinated members, either as an additive group or an appositional group	*fighter-bomber, slave girl*
c)	nouns with "all" or "self" as the rst word	
d)	old genitive groups	*craftsman, bullseye*
e)	Compounds falling under the semantic denominator "appurtenance to a group or solidarity circle"	*landsman, kinsman*
f)	Occupations	*postman*
g)	Verbs plus a common noun	*writing-table*
h)	A verbal item determining a noun	*whetstone, rattlesnake*
i)	An adjective plus a noun	*blackbird, sweetmeat*
j)	Compounds such as	*'he-goat', 'she-dog'*
k)	A predicate plus object	*house-keeping*
l)	Compounds with deverbal nouns as second words	*earthquake, strong- hold*
m)	Compounds where the second word is an agent noun, and the first word is the object or an adverbial complement	*householder, all-seer, self-seeker*
Exocentric Compounds		
a)	Compounds denoting an agent who or which performs what is indicated by the predicate/object nexus of the formal basis	*pickpocket*
b)	Agent nouns from verbal phrases whose second constituent is an adverbial complement	*runabout*
c)	Impersonal deverbal nouns	*blackout*
d)	Formations denoting one who or that which is characterised by what is expressed in the compound	

Table 4.9: Classification of compounds according to (Marchand 1969)

type	valency		DE example	EN translation
	head	non-head		
C1	+	-	*Journalisten**frage**, w-* vs. *Frage, w-*	journalist question wh- question wh-
C2	-	+	***Auswahl**verfahren, w-* vs. *Auswahl, w-*	selection process wh- selection w-
C3-1	+	+	***Wettstreit**, w-* vs. *Wette, w-* or *Streit, w-*	bet battle (competition) wh- bet wh- battle (argument) wh-
C3-2	-	-	***Ehrgeiz**, dass* vs. **Ehre, dass* or **Geiz, dass*	ambition that *honour that *avarice that

Table 4.10: Subcategorisation-based classification of compounds

4.2.4 Multiwords and their Classification

The current section presents a number of studies on classification of multiword expressions according to different criteria, for example, according to their morphological, derivational, syntactic or semantic features. In table 4.11, we list classes of multiwords according to the these classification criteria.

criteria	**studies**	**classes**
compositionality or lexicalisation grade	(Breidt 1993)	idioms, SVCs and collocations in the narrow sense
	(Bauer 1983)	lexicalised and institutionalised phrases
	(Sag *et al.* 2001)	fixed, semi-fixed and syntactically-flexible
	(Storrer 2007)	idioms, lexicalised SVCs, lexicalised multiword compounds, phrasal verbs, and polylexical technical terms
	(Winhart 2002)	lexicalised and non-lexicalised SVCs
structural and syntactic	(Persson 1975)	three classes according to the constituents
	(Winhart 2002)	SVCs with a nominal phrase and with a prepositional phrase
semantic	(Persson 1975)	according to selectional restrictions
	FrameNet	according to their semantic contribution of multiwords
aspectual	(Zifonun *et al.* 1997), (Storrer 2007), (Hanks *et al.* 2006)	causative, inchoative, durative or passive

Table 4.11: Classes of multiword expressions according to classification criteria

4.2.4.1 Criteria of compositionality or lexicalisation grade

(Breidt 1993) classifies multiwords into **verbal phrasemes (idioms)**, such as *to take a fancy*, **support verb constructions**, e.g. *to take into consideration* and **collocations in the narrow sense** (a combination of the support verb with a concrete or non-predicative noun), like *to take a seat*[20]. The author claims that the difference between these types is very gradual and it is difficult to define the criteria for their distinction.

(Bauer 1983) classifies multiword expressions broadly into **lexicalised phrases** and **institutionalised phrases**. Lexicalised phrases have idiosyncratic syntax or semantics or contain 'cranberry lexemes', i.e. words, which do not occur in isolation. Institutionalised phrases are syntactically and semantically compositional, but occur with high frequency in a given context.

In the description of a research for LinGO project, cf. (Sag *et al.* 2001), the authors cite the classification suggested by (Bauer 1983) and propose a further subclassification of lexicalised phrases into **fixed**, **semi-fixed** and **syntactically-flexible expressions**. According to their semantic decomposability, semi-fixed expression are subdivided into **decomposable idioms** such as *spill the beans* and **non-decomposable**

[20]These types are distinguished for German by Brundage (1992), Polenz (1989), Danlos (1992) and Hausmann (1989) who are cited in (Breidt 1993).

idioms such as *kick the bucket*. The authors include compounds and proper names into the class of semi-fixed expressions as well. The class of syntactically-flexible expressions contains **verb-particle constructions** in English, e.g. *write up, look up*, etc. They can be either semantically idiosyncratic, such as *brush up on*, or compositional such as *break up* in *the meteorite broke up in the earth's atmosphere*. Decomposable idioms, such as *let the cat out of the bag* and *sweep under the rug*, tend to be syntactically flexible to some degree. Support verb constructions (which authors call light verb constructions) also belong to this class. They are highly idiosyncratic and thus, it is difficult to predict, which light verb combines with a given noun.

(Villada Moirón 2005) uses the term **fixed expression** to describe multiword expressions and classify them into **idioms**, **collocations**, **metaphors**, **support verb constructions**, **phrasal verbs**, **institutionalised phrases**, **sayings**, **proverbs** and **formulaic expressions**.

(Storrer 2007) mentions the following types of multiword expreessions: **idiomatic expressions, lexicalised support verb constructions, lexicalised multiword compounds, phrasal verbs**, and **polylexical technical terms**.

(Winhart 2002), who concentrates on SVCs in German, distinguishes between lexicalised and non-lexicalised support verb constructions, following the classification given e.g. by Kuhn[21], cf. table 4.12.

type	**DE example**	**EN translation**
lexicalised SVCacc without an article	*Gefahr laufen* *Kenntnis nehmen* *Anwendung finden*	to run into danger to take note to apply
lexicalised SVCacc with an article	*den Vorzug geben* *eine Ausnahme bilden*	to give preference to find an exception
non-lexicalised SVCacc	*Anklage erheben* *eine Beobachtung machen*	to bring in an action to make an observation
lexicalised SVCprep without an article	*in Verwahrung nehmen* *in Vergessenheit geraten*	to take into custody to fall into oblivion
lexicalised SVCprep	*zur Aufführung bringen* *zur Sprache bringen*	to perform sth to bring up
non-lexicalised SVCprep	*zum Abschluss bringen* *auf eine Idee bringen* *unter dem Einfluss stehen*	to bring to a close to give smb an idea to be under smb's influence

Table 4.12: Classes of support verb constructions based on their lexicalisation

4.2.4.2 Structural and syntactic criteria

Multiwords can be also classified according to their structure and syntactic features. For instance, (Persson 1975) classifies multiwords into three groups according to the elements, which constitute the expression:

1) transitive verb + preposition (+ article) + noun, e.g. *zum Schweigen bringen* ("to bring into silence/to make (sb) silent"), *in Bewegung bringen* ("to set into movement/to make sb move");

[21] Cited in (Winhart 2002).

2) intransitive verb + preposition (+ article) + noun, e.g. *zum Erliegen kommen* ("to come to a standstill"), *in Bewegung kommen* ("to come into movement/to start to move");

3) transitive verb (+ article) + noun in accusative, e.g. *Ausdruck finden* ("to find expression"), *eine Erklärung finden* ("to find an explanation").

(Winhart 2002) group German SVCs into two categories: **SVCs with a nominal phrase** and **SVCs with a prepositional phrase**. The first group includes SVCs whose support verbs, such as *machen* ("to make"), *geben* ("to give") or *bekommen* ("to become"), are combined with an NP, e.g. *den/einen Vorschlag machen* ("to make a proposal") or *die Einwilligung geben* ("to give thje consent"), etc. The second group is represented by multiwords, which contain support verbs combined with PPs, like *zum Ausdruck bringen* ("to bring into expression/to express"), *in Frage kommen* ("to come into consideration") or *ins Gerede geraten* '("to get into gossiping").

4.2.4.3 Semantic and aspectual criteria

Semantic criteria of multiword classification are based on either the semantic restrictions and aspects of their elements, cf. (Persson 1975) and (Hanks *et al.* 2006), or their semantic contribution as a whole unit, cf. FrameNet.

(Persson 1975) mentions semantic restrictions of multiwords. Support verbs and prepositions, e.g. *zum* and *bringen* in example (4.3c), cannot be combined with certain nominalisations, e.g. *Anfang* and *Beginn* in (4.3a) or *Geschehen* or *Stattfinden* in (4.3b). These nominalisations belong to the category of events. The nominalisation *Ausbruch*, in contrast, can be combined with both the preposition *zum* and the verb *bringen*.

(4.3a) **zum Anfang kommen, *zum Beginn kommen* (predicate = event)

(4.3b) **zum Geschehen bringen, * zum Stattfinden bringen*

(4.3c) *zum Ausbruch bringen* (predicate = process)

For SVCs containing the preposition *in*, the author also makes distinction between three groups of meaning:

a. space movement: *in Bewegung bringen, in Schwung bringen*

b. emotional movement: *in Rage bringen, in Wut bringen*

c. relations: *in Einklang bringen, in Übereinstimmung bringen*

In FrameNet multiwords are classified with regard to their semantic contribution. The classification is based on the relations to the subcategorisation of the frame-evoking element, which is in most cases nominal predicate and the influence of the support verb onto the subcategorisation of the whole construction, cf. table 4.13. This classification is informal and not encoded into the database.

Multiwords can also be classified according to their aspectual meaning (*Aktionsart*), which is contributed by the constituent verb. There are **causative**, e.g. (*in*

Auftrag geben "to place (an order)", *in Gang setzen* "to set in motion"), **inchoative**, such as (*in Kraft treten* "to become operative", *in Vergessenheit geraten* "to sink into oblivion"), **durative**, e.g. (*in Bewegung bleiben* "to keep moving"), or **passive** like (*zur Aufführung kommen* "to be performed", *Anwendung finden* "to be used"), cf. (Zifonun *et al.* 1997), (Storrer 2007) and (Hanks *et al.* 2006).

Some authors, e.g. (Kamber 2006), apply a set of various criteria in the classification of multiword. For instance, the criteria in (Kamber 2006) used for classification of SVCs combine semantic, word formation and syntactic features:

A the usage of a support verb;

B the derivation of the noun - deverbal vs. non-deverbal;

C the verb semantics: verbs of movement vs. state verbs;

D the presence of a prepositional phrase in the SVC;

types	explanation	examples
Plain Vanilla	the support adds virtually nothing to the frame-evoking element	*make a statement*
Aspectual	the support changes the temporal focus of the event portrayed by the frame-evoking noun	*start* in *start an operation*, this also covers things like *get/go/fall into a (foul) mood* vs. the vanilla support structure *to be in (foul) mood*
Point-of-view	the support changes the profiled point-of-view of the frame-evoking noun	*undego* in *undergo a physical exam* (the patient's point of view) vs. *give a physical exam* (the doctor's point of view)
Registral	the different support verbs appeal to different formal registers	*make a complaint* vs. *register a complaint*; *take revenge* vs. *exact/wreak revenge/vengeance*
Causative	the support adds another participant and the idea of causation to the basic scene	*bring into play* vs. *come into play* or *give a headache* vs. *have a headache*
	normally only the causee is tagged as a frame element evoked by the target	the object of *bring, give*
	additionally the subject of the support verb is tagged when it fills a frame element role that is also part of the basic frame	

Table 4.13: Classes of multiwords in FrameNet

4.2.4.4 Multiword classes related to this study

We classify multword expressions with respect to their subcategorisation properties. Our classification is based on the relationship between subcategorisation properties of muliwords and those of their nominal component, cf. section 3.4.1.1: **M1** and partly **M2** share it, whereas **M3** and **M4** do not, cf. table 4.14. The **M1**-class is represented by multiwords, which subcategorise for the same subclause type that is also subcategorised by their nominal component. The **M2**-class includes multiwords,

which subcategorise for the same subclause type as their nominal component under certain contextual conditions. For example in affirmative context, the nominal *Erfahrung* does not allow for an *ob*-clause, whereas it takes an *ob*-clause in interrogative contexts, cf. examples in (4.4a) and (4.4b).

(4.4a) affirm. *er hat (die) Erfahrung, daß/*ob/w-*
("he has (the) experience that/*if/wh-")

(4.4b) interr. *haben Sie (eine) Erfahrung, *daß/ob/w- ?*
("do you have (any) experience *that/if/wh- ?")

In the **M3** multiwords, neither their nominal nor their verbal component subcategorise for a sentential complement, whereas the multiword itself does. We also group in this class multiwords whose noun takes a subclause, but in a massively different subcategorisation frame: *Beweis* ("proof") takes a *für*-PP or a sentential complement with a(n optional) Korrelat (*dafür*), whereas *unter Beweis stellen* ("to provide evidence for"), which also has a sentence complement, can never take the correlate nor a *für*-PP. These cases are also semantically transparent, i.e. do not qualify for the status of idioms.

The **M4**-class includes multiwords, which can subcategorise for sublcauses even though its nominal constituent does not, and which are commonly seen as idioms, either because they contain 'cranberry' lexemes, cf. section 4.2.4.1 above[22] or because they are non-compositional.

type	feature	DE example	EN translation
M1	"inheritance"	*zur Bedingung machen, dass* vs. *Bedingung, dass*	to condition that condition that
M2	"inheritance" + "switching" of truth values	*in Erfahrung bringen, w-/ob* vs. *Erfahrung, dass/w-/ob*	to find out wh-/if the experience that/wh-/if
M3	"non-inheritance"	*zum Ausdruck bringen, dass* vs. **Ausdruck, dass*	to express that expression that
M4	"non-inheritance" cranberry lexeme non-compositional	 *in Abrede stellen, dass* vs. **Abrede* *ins Auge fallen, dass*	 to deny that accord to catch sb's eye

Table 4.14: MWE classes based on their subcategorisation properties

The semi-automatic approach to classify multiwords according to the types given in table 4.14 is described in section 5.3.3.4 below.

4.2.5 "Inheritance" Relations and their Types

In this section we describe classification of subcategorisation relations between verbs and their nominalisations (both within a SVC and those occurring freely in corpora), based on the ability of derived nouns to inherit verbal valecy properties. no approaches, which systematically describe types of "inheritance" relations, except for

[22]'Cranberry' lexemes in German are described in e.g. (Richter/Sailer 2002) and (Trawiński *et al.* 2008).

the semi-automatic classification described in (Lapshinova 2009). Our classification is based on the description of "inheritance" and "non-inheritance" relations between deverbal nouns and their base verbs presented in section 3.4.2 above. Some nominalisations inherit verbal predicates ("inheritance" cases), some of themlose a part of the verbal subcategorisation features ("inheritance" reduction) and some of them gain subcategorisation properties, which are not specific for their verbs ("inheritance" extension), cf. section 3.4.2 above. Therefore, we distinguish between three classes of subcategorisation relations: **R1**, **R2** and **R3**, cf. table 4.15.

type	feature	DE example	EN translation
R1	"inheritance"	*beweisen, w-/ob –* *Beweis, w-/ob –* *unter Beweis stellen, w-/ob*	to prove wh-/if– proof wh-/if – to put under proof (to prove) wh-/if
R2	"non-inheritance": subcategorisation reduction	*vermuten, dass/w-/ob –* *Vermutung, dass/*w-/*ob –* *zur Vermutung führen, dass*	to assume that/wh-/if – assumption that/*wh-/*if – to bring to the assumption that
R3	"non-inheritance": subcategorisation extension	*überlegen, w-/ob –* *Überlegung, dass/w-/ob- –* *zur Überlegung kommen, dass/w-/ob*	to consider wh-/if – consideration that/wh-/if – to come to consideration that/wh-/if

Table 4.15: Subcategorisation "inheritance" types

The **R1** relations include cases where a nominalisation inherits subcategorisation properties of the corresponding verb. For instance, the verb *entscheiden* (" to decide") allows for both declaratives and interrogatives. So does its nominalisation *Entscheidung* ("decision") and the multiwords containing its nominalisations, such as *zur Entscheidung kommen/gelangen/stellen* ("to come/.../put to decision") or *vor die Entscheidung stellen* ("to put to the decision").

The **R2** relations are observed in verb-nominalistaion pairs, whose nominalisations inherit verbal subcategorisation properties, but some of these properties get lost. This type can be subdivided into two further subclasses: cases where interrogatives get lost and cases, where declaratives get lost. In most R2 cases nominalisations do not take over interrogative complements, which are allowed by their underlying verbs. For instance, the verb *ankündigen* ("to announce") can take both interrogative and declarative complement sentences: *ankündigen, dass* ("to announce that") or *ankündigen, w-* ("to announce wh-"). However, its nominalisation *Ankündigung* allows only for *dass*-clauses: *Ankündigung, dass* ("announcement that"). Theoretically the nominalisation can also lose *dass*-clauses, but such cases are rare or do not exist.

The **R3** relations are specific for those verb-nominalisation pairs where the nominalisation has subcategorisation properties its base verb does not have. These cases can also be subclassified into two groups. The first group include cases in which nominalisations not only inherit verbal subcategorisation properties, but also possess additional subcategorisation features. For example, the verb *überlegen* ("to consider") takes in most cases the interrogative *w*-clause, whereas its nominalisation *Überlegung* ("consideration") allows not only for interrogative but also for declarative complements: *Überlegung, dass* ("consideration that").

As most verbs can take both declarative and interrogative clauses (sometimes under certain contextual conditions only), we assume that relations of type **R3** are

hypothetical.

The procedures to automatically classify verb-nominalisation pairs into the classes listed in table 4.15 are described in section 5.3.4.3 below.

Chapter 5

Extraction and Classification Architecture

The following chapter describes the extraction and classification architecture elaborated within the present thesis. First, in section 5.1, we give detailed information on the corpora used in the study as well as about a set of pre-processing tools involved. We go on with the description of extraction context, cf. section 5.2, which is in most cases determined by the German word order and the precision oriented results. Then, in section 5.3, we represent queries for extracting predicates in different contexts applied in this research and explain the symbolic procedures to classify predicates according to their subcategorisation properties.

5.1 Input: Corpora and their Annotation

Predicates under analysis are extracted from different corpora in German, which are analysed with pre-processing toos. In the following we characterise the corpora used for this study and give an overview of the pre-processing tools applied for the corpora annotation.

5.1.1 Corpora Specification

To extract and classify predicates according to their subcategorisation properties, we use newspaper and web corpora from Germany, Austria and Switzerland, which comprise written texts in German dated from 1988 until 2005, a total of ca. 1563M tokens. In table 5.1, we outline the corpora used in this research, specifying some of them with the information on their size (the number of tokens contained) and time period (if this information is available).

The corpora from Germany include extracts (1988-2001) from German newspapers, such as *die tageszeitung*, *Frankfurter Rundschau*, *Frankfurter Allgemeine Zeitung*, *Stuttgarter Zeitung*, *DIE ZEIT* and *Handelsblatt*. We also use the European Language News Corpus ('ELNC'), which includes online news from 1997. The data in 'ELNC' originates from German news and AFP and NZZ services. A part of this corpus originates from Swiss mass media. Other texts from Switzerland are contained in the Swiss part of DEREKO, which is referred to as DEREKO-CH. It contains data

Corpora	**abbreviation**	**issues**	**size in tokens**
die Tageszeitung	('taz')	1988-1994	111,3M
Frankfurter Rundschau	('FR')	1992-1993	40,6M
Frankfurter Allgemeine Zeitung	('FAZ')	1997-1998	70,2M
Stuttgarter Zeitung	('StZ')	1991-1993	36,2M
DIE ZEIT	('ZEIT')	1995-2001	52,1M
Handelsblatt	('HB')	1986-1988	35,7M
European Lang. News Corpora	('ELNC')	1997	103,8M
'Gutenberg' Literatur Archive	('DE Lit.')	2005	137,3M
BUNDESTAG	('BT')		5,7M
a part of German Web-Corpora	('DeWaC')		286M
Austrian news corpora DEREKO-AT	('AT')	1991-2000	499,7M
Swiss news corpora DEREKO-CH	('CH')	1996-2001	183,9M
TOTAL			ca. **1563M**

Table 5.1: Corpora used in the study

from *Züricher Tagesanzeiger* and *St. Galler Tagblatt*, dated 1996-2001. The Austrian part of DEREKO includes newspaper texts from *Salzburger Nachrichten*, *Oberösterreichische Nachrichten*, *die Presse*, *Kleine Zeitung*, *Tiroler Tageszeitung* and *Vorarlberger Nachrichten*, all dated between 1991 and 2000.

Both, the Swiss and the Austrian parts of the DEREKO corpora are part of the German reference corpus DeReKo and have been made available to us by the Institut für deutsche Sprache, Mannheim, in a cooperative project with the University of Tübingen.

All the corpora described above contain written texts from newspapers. As a rule, newspaper corpora are considered to be non-balanced and non-representative, as they do not contain all conceivable constructions. However, these features are not essential in our study as the phenomena under analysis are rare and somtimes, sheer corpus size has more priority. Besides that, we concentarte on the extraction of the embedded declarative and interrogative sentences, which usually express reported speech. We assume that newspaper texts tend to contain a considerable amount of the sentences of this type.

Nevertheless, we complement the newspaper corpora with other corpora types. For several extraction procedures, e.g. for extraction of multiword expressions and compound nouns, we apply the corpus that consists of written texts from the minutes of the Bundestag (the parliament of Germany) debates. Another alternative is the corpus of literary texts, 'Gutenberg' Achive[1], which includes novels, stories, novellas and poems of over 550 authors.

To achieve a substantial coverage for certain predicate types, we apply web corpora for German (local version of DeWaC), which have been tokenised and tagged by A. Kilgariff and M. Baroni[2], and have been made available for the search with the

[1]The edition of 2005.

[2]Cf. (Baroni/Kilgariff 2006).

query langauge we apply in this study.

As mentioned above, some of the specified corpora are applied only for particular tasks. So we do not apply the whole set of the specified corpora for all our extraction procedures. Their application depends on the features and annotations the corpora possess. The main set of corpora, which are used in almost all experiments and tests, consists of 'FR', 'FAZ', 'taz', 'StZ' and 'ZEIT', a total of 310,4M tokens. Other corpora are used for extraction procedures that involve rare phenomena. For instance, for extraction of multiword constructions of the form PP+N+SV, we apply the extraction queries on almost all the corpora specified in table 5.1.

5.1.2 Corpus Pre-processing Tools

As mentioned in section 4.1.4 above, we operate on pre-processed corpora as extraction from annotated corpora is more effective, due to some properties of the German language, e.g. strong inflection or variable word order, cf. (Breidt 1993).

All the corpora used in the present thesis are sentence-tokenised, part-of-speech-tagged and lemmatised. We use (Schmid 1994)'s TreeTagger and lemmatiser, as well as the STTS tagset[3] for these annotations. A part of the corpora are also chunked. To assess the need for chunking, we use YAC, a recursive chunker for German, cf. (Kermes 2003). Regular expressions for data extraction rely on the IMS Corpus Workbench (CWB, cf. (Evert 2005)). An overview of the used tools is given in table 5.2.

processes	**tools**
tokenising	(Schmid 2000)
pos-tagging and lemmatisation	Tree Tagger (Schmid 1994) and (Schmid 1999)
morphological annotation	Morphology tool SMOR (Schmid *et al* 2004)
chunking	YAC-Chunker (Kermes 2003)
Corpus query tools	CWB (Evert 2005)

Table 5.2: Corpus annotation tools

In the following we illustrate some of the pre-processing steps performed by the above mentioned tools. As an example we take a sentence, which contains the nominal predicate *Ankündigung* ("announcement"), cf. (5.1).

(5.1) *Allein die Ankündigung, dass er komme, hatte den Börsenkurs vergangene Woche in die Höhe getrieben.* ("Alone the announcement that he would come had boosted the course in the previous week").

Token annotations include positonal attributes represented in table 5.3. Each agreement feature has the form ccc:g:nn:ddd with, whereas ccc=case, which can be nominative (Nom), genitive(Gen), dative (Dat) or accusative (Akk); g=gender, which varies between masculine (M), feminine (F) and neutral (N); nn=number, which can be either singular (Sg.) or plural (Pl); and ddd=determination, which can be definite (Def), indefinite (Ind) and Nil.

[3] Cf.
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html.

annotation	meaning
word	word forms ("plain text")
pos	part-of-speech tag (STTS tagset)
lemma	base forms (lemmatised forms)
alemma	ambiguous lemmatisation
agr	noun agreement features

Table 5.3: Token annotation

Part-of-speech-tagging and lemmatisation include the process of marking up the words in a corpus as corresponding to a particular part of speech and the process of determining the lemma for a given word. In table 5.4, we show sentence (4.1) annotated with part-of-speech and lemma information. Every word in the sentence has a triple form, which contains the information on the word form itself, its part of speech (ADV for adverb, NN for noun, etc) and its lemma (haben for *hatte*, etc.).

preprocessed sentence	interpretation		
	word	pos	lemma
<Allein/ADV/allein	*Allein*	adverb	allein
die/ART/d	*die*	article	d
Ankündigung/NN/Ankündigung	*Ankündigung*	noun	Ankündigung
,/$,/,	,	comma	,
dass/KOUS/dass	*dass*	conjunction	dass
er/PPER/er	*er*	personal pronoun	er
komme/VVFIN/kommen	*komme*	finite full verb	kommen
,/$,/,	,	comma	,
hatte/VAFIN/haben	*hatte*	finite auxiliary verb	haben
den/ART/d	*den*	article	d
Börsenkurs/NN/Börsenkurs	*Börsenkurs*	noun	Börsenkurs
vergangene/ADJA/vergangen	*vergangene*	adjective	vergangen
Woche/NN/Woche	*Woche*	noun	Woche
in/APPR/in	*in*	preposition	in
die/ART/d	*die*	article	d
Höhe/NN/Höhe	*Höhe*	noun	Höhe
getrieben/VVPP/treiben	*getrieben*	full verb participle	treiben
./$./.	.	$.	.

Table 5.4: An example of a pos-tagged and lemmitised sentence

The STTS tagset (Stuttgart-Tübingen Tagset) includes 54 tags. 48 of them are sheer part-of-speech-tags, and the other 6 are used for foreign-language material (FM), non-heads of compounds (TRUNC), non-words and punctuation characters ($, or $.). We give the whole list of the STTS tags in the appendix B.1.

A YAC chunk is a continuous part of an intra-clausal constituent including recursion, pre-head as well as post-head modifiers but not PP-attachment or sentential elements. The YAC-annotated syntactic structures include adverbial phrases (advp), adjectival phrases (ap), noun phrases (np), prepositional phrases (pp), verbal complexes (vc), single verbs (v) and clauses (cl). We list the chunk element in XML-format in section B.2 in the appendix. Additionally, there is also information on feature attributes specifying certain properties of chunks, e.g. their head constituents, np_h for the head of a noun phrase or vc_h for the head of a verbal complex. Further properties of syntactic structures are also give in section B.2 of the appendix. In table 5.5, we illustrate the chunked sentence given in (4.1).

With the help of the morphological analyser SMOR, we can obtain morphological analysis of every word in a corpus. For instance, the definite article *den*, if analysed by SMOR, has the forms shown in figure 5.1. The tool delivers information on case, gender, number etc. This word form can represent, e.g. a definite article, which is accusative masculine singular or dative plural. That information enables us to restrict the search on the basis of certain features. For instance for extraction of direct objects, the query should include the agreement for accusative case, for extraction of indirect objects – agreement for dative, etc.

```
die<+REL><Subst><Masc><Acc><Sg><St>
die<+DEM><Subst><Masc><Acc><Sg><St>
die<+ART><Def><NoGend><Dat><Pl><St>
die<+ART><Def><Masc><Acc><Sg><St>
```

Figure 5.1: Morphological analysis of the article *den*

The corpora, which are linguistcally annotated with the help of the above mentioned tools, can be manipulated by the IMS CWB. The latter includes a set of tools for encoding, indexing, compression, decoding, and frequency distributions, a global "registry" that holds information about corpora (name, attributes, data path) and a corpus query processor (CQP), which enables fast corpus search. In CQP the results of a query are the list of corpus intervals, which match the given query. They can be presented with and without the corresponding corpus annotations. In addition to that, the CQP supports the alphabetical and frequency-based sorting of query results.

The CQP language is represented by regular expression syntax. A regular expression is a concise descriptions of a set of character strings, which are called words in formal language theory. Certain sets of words with a relatively simple structure can be represented in this way. Regular expressions match the words they describe. The language of regular expressions over attribute expressions includes parantheses for marking embedded expressions, concatenation, disjunction, unspecified corpus position, Kleene star, and Kleene plus. In section B.2 of the appendix, we give an overview of the CQP regular expressions syntax.

An example of a simple query in the CQP language is the search for single words or a sequence of words. For instance, to find all occurrences of the noun *Ankündigung* ("announcement"), which is followed by a comma and the conjunction dass in the corpus 'ZEIT', these words should be typed in double quotes at the CQP prompt:

chunked sentence	**interpretation**
<<s>	sentence start
<advp>	start of ADVP
<advp_h allein>*Allein*</advp_h>	ADVP's head
</advp>	the end of ADVP
<np>	start of NP
<np_h Ankündigung>*die Ankündigung*</np_h>	head of NP
</np>	end of NP
,	
<cl>	start of CL
<cl_h dass>*dass*	start of CL'S head
<np>	start of NP
<np_h er>*er*</np_h>	NP's head
</np>	end of NP
<v><vc>	start of VC
<vc_h kommen>*komme*</vc_h>	VC's head
</vc></v>	end of VC
</cl_h>	end of CL's clause
</cl>	end of CL
,	
<v><vc>	start of VC
<vc_h >*hatte*</vc_h>	VC's head
</vc></v>	end of VC
<np>	start of NP
<np_h Börsenkurs>*den Börsenkurs*	start of NP's head
<ac><ap>	start of AP
<ap_h vergangen>*vergangene*</ap_h>	AP's head
</ap></ac>	end of AP
Woche	
</np_h>	end of NP's head
</np>	end of NP
<pp>	start of PP
<pp_h in:Höhe>	start of PP's head
in	
<np>	start of NP
<np_h Höhe>*die Höhe*</np_h>	NP's head
</np>	end of NP
</pp_h>	end of PP's head
</pp>	end of PP
<v><vc>	start of VC
<vc_h treiben>*getrieben*</vc_h>	VC's head
</vc></v>	end of VC
.	
</s>>	sentence end

Table 5.5: An example of a chunked sentence

ZEIT> "Ankündigung" "," "dass";

An extract from the resulting output is shown in figure 5.2. The output consists of a list of individual lines. The corpus intervals, which match the query are marked by angle brackets <...>, whereas the text on both sides is the context.

8083803:	Oskar über die	<Ankündigung , dass>	der Bundespräs
8283690:	h dafür mit der	<Ankündigung , dass>	ein Abschleppw
11000210:	ine deutlichere	<Ankündigung , dass>	die Mehrwertst
19660783:	er historischen	<Ankündigung , dass>	der 25. Dezemb
20535861:	n , und auf die	<Ankündigung , dass>	Wolfgang Cleme
23196265:	wenigstens eine	<Ankündigung , dass>	auch die Bunde
24300132:	rofitiert . Die	<Ankündigung , dass>	nächstes Jahr

Figure 5.2: An example of the CQP output

In this example the searched predicate is lexically specified. However, we are looking for general tendencies over the language. Therefore, our queries for extracting predicates of different types contain lexically underspecified blocks, which are formulated in the form of regular expressions. To restrict the query search for a predicate type, we also apply lexical constraints that are integrated into the queries. The building elements of the regular expressions queries are based on the annotations mentioned above. An overview of the necessary annotations and the query language is given in appendix B.

5.1.3 Experiments with Parsed Corpora

With the help of the above described pre-processing procedures and the CQP queries, we can treat corpora with flat architecture procedures. Alternatively, subcategorisation features can be acquired from parsed corpora, i.e. with the help of dependency-based deep processing. We assume that the application of parsing based techniques can increase the extraction numbers, thus, contributing to a higher recall[4].

Our aim is to extract subcategorisation properties with high precision. However, the preliminary analysis of the data extracted from parsed corpora shows that parsing-based search has both advantages and disadvantages. Although extractions from parsed corpora increase the number of matches and therefore, the recall, at the same time it delivers a great number of non-relevant matches and thus, reduces the precision of the extraction. The noise is caused by the main rules of phrase structure grammars, in which the verb is the central unit of the sentence, which means that sentential complements always depend on the main verb.

Elaborating an extracting procedures from parsed corpora we need to apply a cascaded set of filtering and specification procedures, as otherwise, in parsed sentences the verb will always be identified as the valency bearer. For instance, if we extract

[4]Cf. the results of the experiment on the extraction of verb+object collocations described in (Heid *et al.* 2008). Deep parsing-based processing is compared to chunking-based processing. The results show that the parsing-based method delivers results of up to 70% higher recall than the chunking based one.

sentence (5.1) from parsed corpora by means of simple search (without filtering and specification procedures), the verb complex *hatte...getrieben* and not the noun *Ankündigung* will be identified as valency bearer of the subclause *dass er komme*, which is not correct. In this case the *dass*-clause is subcategorised by the nominal predicate. In some cases it can also be subcategorised by noun-verb multiword expressions. To automatically identify and classify potential valency bearer from parsed corpora, we need to elaborate an architecture which include a number of filtering procedures to exclude typical noise-causing elements, e.g. headless or adverbial relatives antecedents, posessive NPs, etc., as described in section 5.3.1.4. Besides that, morpho-syntactic and lexical specifications are required to identify other than verbal predicate types, cf. section 5.3.2.

The application of filtering and specification procedures in the extraction from parsed corpora improves the extraction results which allows us to achieve both higher recall and precision.

5.2 Extraction Context

As we aim to achieve high precision result in extracting unknown subcategorisation properties of verbs, nouns and multiwords, we elaborate restricted contexts for this purpose. The contexts from which we extract and classify predicates of different types are specified according to the distributional and morpho-syntactic features of predicates under analysis.

5.2.1 Contexts for the Extraction of Verbal Predicates

In the extraction of verbal predicates we use both, active and passive forms of verbs. Sentences containing verbs in passive voice comprise about 6-15%[5] of all corpus text, cf. (Heid/Weller 2008). Extracting active verbal forms, we give preference to Verbletzt, the German verb-final clauses (VL, cf. section 2.2.1), which comprises about 20-25 % of all corpus text, cf. (Balabanov 2007).

5.2.1.1 Verb-final sentences as the most "convenient" context

Structure and features of verb-final sentences The concept of the verb-final (VL) sentence structure is based on the topological field model of (Höhle 1986), mentioned in section 2.2.1 (cf. table 5.6). The Vorfeld (VF, the pre-field) is either occupied by one constituent or remains empty. The left sentence bracket (LSK) contains either a finite verb or a clause introductory element (such as a conjunction, an interrogative or a relative pronoun). The Mittelfeld (MF, the middle field) can include any constituent and the number of constituents is not limited. The right sentence bracket (RSK) contains all the infinite verb forms and also the finite verb. The Nachfeld (NF, the post-field) is filled with sentential complements as well as adverbial and relative clauses.

[5]The amount varies considerably between different corpora types.

VF	LSK	MF	RSK	NF

Table 5.6: Topological field model

In a VL sentence, the verbal complex occupies the position before the comma, and are followed by the conjunction *dass* or *ob*, or by a *w*-word, which introduce a subordinate *dass, ob* or *w*-clause, as shown in table 5.7.

VF	LSK	MF	RSK	NF
main clause				**subclause**
			verb	
	conj *Wenn* If"	**const** *sie* "they"	**v.complex** *erfahren*, "find out"	**sent.complement** *dass John Miller große Mengen Alkohol kauft...* "that John Miller buys much alcohol..."
rel/int *die* "who"		**const** *genau* "exactly"	**v.complex** *wussten*, "knew"	**sent.complement** *worauf es ankommt.* "what it depends on".

Table 5.7: Subclauses after verbal predicates in VL

The verbal complex, which occpupies the LSK can have several forms. In table 5.8, we outline different forms of the German verb *fragen* ("to ask") that can be used in a VL sentence. It can contain up to three constituents, depending on the tense and mood of the main clause. Examples for all the three verb forms are in in 3rd person singular. The verb complexes outlined below include only cases with one full verb *fragen* ("to ask"). The cases like *fragen hörte* ("heard asking") are not included into this table. On the basis of these form, we elaborate queries for verbal predicates extraction, cf. section 5.3.2.1.

Reasons for using verb-final sentences There are several reasons for the application of the VL context in the extraction of verbal predicates. The main reason is the regularity of the sequence of constituents. For this construction, we know which sentence position constituents can occupy. Verbal predicates tend to preceed the sentential complement and that allows us to extract both of them from text corpora, cf. table 5.7.

The other reason is the the higher precision of the expected extraction results. In most cases the sentential complement, which follows the verbal complex, is mostly subcategorised by this verb. Only if the verbal complex is preceeded by other predicates, e.g. nominal or adjectival predicates or the verb construes a multiword predicate with them, the subclause can be subcategorised by these nouns, adjectives or multiwords.

5.2.1.2 Sentences with verbs in passive voice

Another convenient context for extraction of verbal predicates are sentences containing verbs in passive voice. To extract verbal predicates in passive, we consider all the

active			
tense	**infinite forms**		**finite form**
present, past	–		*fragt/frage/fragte/frägte*
perfect/plusquamperfect	*gefragt*		*hat/habe/hatte/hätte*
future I	*fragen*		*wird/werde*
future perfect	*fragen*	*haben*	
with modal verbs	*fragen*		*soll/solle/sollte*
	gefragt	*haben*	
	fragen	*müssen*	*wird/werde*
	gefragt	*haben müssen*	
with *zu*-infinitive	*zu fragen*		*hat/habe/hatte/hätte*
	zu fragen	*gehabt*	
	zu fragen	*haben*	*wird/werde*
	zu fragen	*gehabt haben*	
passive			
tense	**infinite forms**		**finite form**
all tenses	*gefragt*	*werden*	*wird/werde*
		worden sein	
	gefragt	*worden*	*ist/sei/war/wäre*
	gefragt		*wird/werde/wurde/würde*
with modal verbs	*gefragt*	*werden*	*soll/solle/sollte*
	gefragt	*worden sein*	
	gefragt	*werden müssen*	*wird/werde*
	gefragt	*worden sein müssen*	
zu-infinitives	*zu fragen*		*ist/sei/war/wäre*
	zu fragen	*gewesen*	
	zu erwarten	*sein*	*wird/werde*
	zu erwarten	*gewesen sein*	

Table 5.8: The forms of verb complex in VL

three topological field models. However, the passive forms in the VL topological field are described in section 5.2.1.1. Therefore, we only specify the v1 and v2 sentence models in the following .

Structure and features In a passive sentence verbal predicates have a complex form and can occupy both the LSK and the RSK. The first part of the verbal predicate, which is either an auxiliary or a modal verb, fills the LSK. The RSK contains either the full verb in form of a participle or the combination of it with another auxiliary, e.g. *werden* ("to get/be") or *sein* ("to be"), which constitute the second part of the verbal complex.

	VF	LSK	MF	RSK	NF
	main clause				**subclause**
	const	**verb**	**const**	**verb**	**sent.complement**
v1	*(Es)* "It"	*wird* "will be" *Kann* "Can"	*dann* "then" *(es) dann* "it then"	*gefragt* "asked" *gefragt werden* "be asked"	*warum es passierte.* "why it happened". *warum es passierte?* "why it happened?"
v2	*Es* "It" *Der Teilnehmer* "The participant"	*muss* "must" *wird* "is"	*dann* "then" *dann* "then"	*gefragt werden* "be asked" *gefragt* "asked"	*warum es passierte.* "why it happened" *warum es passierte.* "why it happened"

Table 5.9: Subclause after verbal predicates in passive

The Vorfeld is occupied by the Korrelat *es*[6] or by a nominal phrase. The MF can be filled by various constituents and the NF contains the subcategorised sentential complement. The verbal complex can have different forms, depending on the tense and mood of the construction. In table 5.10, we summarise the forms of the verb complex if used in passive in V1 and V2 (the verb is given in the form of the 3rd person singular)[7], cf. table 5.9.

Reasons for using passive clauses Passive clauses are used as context for extraction of verbal predicates for the same reason as the VL clauses. In this sentence construction we also have a regular sequence of constituents, whose position is highly predictable. The subclause, which follows the verb, is typically subcategorised by the verb, except for the cases in which the verb is preceeded by some predicative material, cf. section 5.2.1.3. The regularity of this highly-predictable context enables us to achieve higher accuracy in the extraction.

5.2.1.3 Problems related with further extraction contexts

As mentioned above the VL constructions and the sentences containing verbs in passive serve as the best extraction context for verbal predicates. The search in these context types deliver results with high precision. However, these context types comprise about 30-40% of all corpora, which means that we acquire less than half of

[6] Cf. section 2.2.1.3

[7] Cf. (Eckle-Kohler 1999) for more detailed description.

	LSK	**RSK**	
tense/mood	**finite form**	**participle and infinite forms**	
present/past both moods	*wird/werde/wurde/würde*	*gefragt*	–
indicative mood: perfect plusquamperfect future I future II	 *wird* *war* *wird* *wird*	 *gefragt*	 *werden* *worden* *werden* *worden sein*
conjunctive mood	*werde* *sei/wäre*	*gefragt*	*werden/worden sein* *worden*
with modal verbs, all moods and tenses	*soll/solle/sollte* *wird/werde/hat/habe/* *hatte/hätte*	*gefragt*	*werden/worden sein* *werden müssen/* *worden sein müssen*
with *zu*-infinitive and with *sein* all moods, tenses	*ist/sei/war/wäre* *wird/werde*	 *zu fragen*	– *gewesen* *sein* *gewesen sein*

Table 5.10: The forms of verb complex in passive, in v1 and v2 sentence models

the data contained in corpora. To achieve more substantial extraction results, further sentences models, e.g. Verberst and Verbzweit (v1 and v2 in table 5.11) can be included.

In the v1 sentences, the Vorfeld is empty and the first position in the sentence is occupied by the finite verb, which can be a full, a modal or an auxiliary verb (verb1 in table 5.11). If the sentences starts with a modal or an auxiliary verb, the full verb infinitive or participle precedes the subcategorised clause in the Nachfeld. In the v2 sentences, the Vorfeld is not empty and can be occupied by various constituents - subject, objects, place or time descriptions, etc. The finite verb occupies the second position (verb1). If the finite verb is modal or auxiliar, the position before the Nachfeld is occupied by the full verb infinitive or participle, cf. table 5.11.

These sentence models are more common in corpora. However, they deliver a large number of non-relevant cases, thus, reducing the accuracy of our extraction procedures. Most problems are caused by non-verbal constituents, which occupy various positions in v1 or v2 and can be predicates themselves. The position of the valency bearer in these cases is not as predictable as in the VL and passive sentences. For instance, the MF in 1a can contain a subclause-taking noun, e.g. *Grund* (“reason”). In this case the noun *Grund* and not the verb *erklären* subcategorises for the *w*-clause in the NF, cf. 1a in table 5.12.

To eliminate such cases, we use automatic linguistic fliters based on lexical knowledge, which is not annotated in corpora. For instance, we integrate lexical constraints in the queries to exclude the occurrence of subclause-taking nominal predicates in the searched sentences. In the VL sentences, we know that subclause-taking nouns tend to immediately precede the verb in RSK. This allows us to prevent their occurrence in the sentence. However, in the v2 sentences, these nouns can occupy different po-

		VF	LSK	MF	RSK	NF
		main clause				**subclause**
		const	**verb1**	**const**	**verb2**	**sent.complement**
v1	1a		*Erklärt* “Explains”	*er ihr,* “he to her”		*warum es passierte?* “why it happened”?
	1b		*Kann* “Can”	*er ihr* “he to her”	*erklären,* “explain”	*warum es passierte?* “why it happened?”
	1c		*Hat* “Has”	*er ihr* “he to her”	*erklärt,* “explained”	*warum es passierte?* “why it happened?”
v2	2a	*Er* “He”	*kann* “can”	*ihr* “to her”	*erklären,* “explain”	*warum es passierte.* “why it happened”
	2b	*Er* “He”	*erklärt* “explains”	*ihr,* “to her”		*warum es passierte.* “why it happened”
	2c	*Er* “He”	*hat* “has”	*ihr* “to her”	*erklärt,* “explained”	*warum es passierte.* “why it happened”
	2d	*Gestern* “Yesterday”	*hat* “has”	*er ihr* “he to her”	*erklärt,* “explained”	*warum es passierte.* “why it happened”

Table 5.11: Subclause after verbal predicates in V1 and V2

		VF	LSK	MF	RSK	NF
		main clause				**subclause**
		const	**verb1**	**const**	**verb2**	**sent.complement**
	1a		*Erklärt* “Explains”	*er den Grund,* “he the reason”		*warum es passierte?* “why it happened”?
	1b		*Erklärt* “Explains”	*er es ihr und fragt,* “he it to her or asks”		*warum es passierte?* “why it happened”?
v2		*Er* “He”	*erklärt* “explains”	*ihr den Grund,* “to her the reason”		*warum es passierte.* “why it happened”

Table 5.12: Noise in extraction of verbal predicates in V1 and V2

sitions, which means that to achieve a higher precision in theis extraction context, all the possible positions of nominal occurrence should be calculated and included into the query. Sometimes this can reduce the recall as nouns, which do not take subclauses can also be excluded. In some cases, this can exclude the occurrence of nouns, which do not function as valency-bearer in the particular sentences to extract but can serve as predicates in other cases. Their exclusion can considerably reduce the recall. In (Eckle-Kohler 1998), the author shows with the help of automated linguistic tests that the usage of lexical-syntactic filters can increase the precision in v2 sentences, but the number of extracted true positives declines.

Table 5.13 illustrates one of such ambiguous examples. Our system knows that the noun *Begründung* (“explanation, statement”) takes a subordinate *w*-clause. However, the valency bearer in the given sentence is the verb *erklären* (“to explain”) and not the noun *Begründung*.

The use of lexical constraints in the VL and passive contexts allows us to maximise both, the precision and the recall in the extraction of verbal predicates. The

VF	LSK	MF	RSK	NF
main clause				**subclause**
const	**verb1**	**const**	**verb2**	**sent.complement**
Er "He"	*erklärt* "explains"	*mit einer plausiblen Begrüdung* "with a reasonable statement"		*warum es passierte* "why it happened"

Table 5.13: False negatives after exclusion of nouns in front of a subclause

application of other contexts, e.g. v1 or v2, could deliver more substantial results. However, for these contexts more constraints should be specified, which is time-and labour-consuming.

5.2.2 Context for the Extraction of Nominal Predicates

We extract nominal predicates from the context in which subclauses are unambiguously subcategorised by the nouns. The noun and its sentential complement take the Vorfeld (VF) position in a v2 sentence.

5.2.2.1 Structure and features of the Vorfeld

It is known[8] that the Vorfeld in German is restricted to contain only one syntactic constituent, e.g. a nominal phrase (NP). The NP can have different forms, from a simple noun to a complex nominal phrase, which can contain a number of further constituents, such as a determiner, an adjective, another NP or a PP. The subclause-taking NP can be also embedded in a prepositional phrase, e.g. in (5.2), which means that the NP can be preceded by a preposition or a combination of the prepositon with a definite article, e.g. *zu+der=zur*. We summarise different forms of a full NP, followed by a subcategorised subclause in table 5.14 below.

(5.2) *Mit **der Erkenntnis, dass** auch das Heilige handgefertigt ist, läßt sich danach der eigentliche Rundgang gut beginnen.*
("With the knowledge that the saint is also hand-made the actual tour can be started").

As seen from the table, the NP taking the VF position in a sentence can contain a noun preceded by a determinative, optional adjectives and adverbs or an NP in genitive and followed by and NP in genitive or a PP. The subclause subcategorised by the NP head noun occupies the position between the NP and the main verb of the sentence, which follows the subclause immediately after the comma, cf. table 5.15.

5.2.2.2 Reasons for the use of the Vorfeld

The Vorfeld position of the NP and its sentential complements is the most unambiguous context for the extraction of nominal predicates. According to German grammarians, e.g. (Zifonun *et al.* 1997) or (Helbig/Buscha 2005), if a noun followed by

[8] Cf. works on German grammar, e.g. (Helbig/Buscha 2005).

Vorfeld with a subclause				
before the noun		**noun**	**after the noun**	**subclause**
- preposition preposition +def.article	- determinative, determinative + optional adjectives and adverbs, nominal phrase in genitive	noun	- nominal phrase in genitive, prepositional phrase,	subclause
examples				
- *in* *ins*	- *die, eine interessante,* *die ihnen unbekannte,* *die objektiv gebildete,* *des Managers beste*	*Vorstellung*	- *eines Managers,* *von dem Manager,* *im Parlament*	*dass...*

Table 5.14: Forms of a full NP in the Vorfeld

VF		**MF**	
main clause **1st part**	**subordinate clause**	**main clause** **2nd part**	
noun phrase	**subcategorised clause**	**verb**	**rest**
Die Vorstellung, "The idea"	*dass die Menschheit unbedeutend werden könnte,* "that mankind could become insignificant"	*ist* "is"	*Unsinn.* "nonsense."

Table 5.15: A noun in the VF subcategorising for a sentential complement.

a subclause takes the Vorfeld position, this subclause can only be subcategorised by the noun, cf. figure 5.3.

IF:	an NP followed by a subclause occupies the VF;
THEN:	this subclause can only be subcategorised by the NP

Figure 5.3: The rule for the subcategorisation of a noun in the VF

Therefore, this context type proves to deliver high precision results, ca. 89-99%[9]

As mentioned in section 3.4.1.2 above, some nominal predicates have a compound structure. Compounds subcategorising for sentential complements are also extracted from the Vorfeld context. An example of a sentence, which contains a compound noun in the VF, is shown in (5.3).

(5.3) *Aber all die **Erklärungsversuche**, warum der Teufel sich an die Frau Doktor heranmacht, sind auf der Glatze gedrehte Locken.*
("But all the **expanation-attempts** (attempts to explain why the devil chats up the female doctor are as futile as giving a bald man a comb").

5.2.2.3 Alternative contexts for the extraction of nominals

The extraction of nominal predicates in the Vorfeld delivers high precision results. At the same time we know that the Vorfeld sentences comprise about 5% of all corpus text. Thus, the acquired list of nominal predicates might be incomplete. To obtain more data for nominal predicates further extraction contexts could be used, e.g. extraposed subclauses with nominal predicates in the MF. Although these context types might raise the recall of the obtained data, their application might also scale down the precision as the the relations between the predicate and the subcategorised clause is less evident. We illustrate this in the following examples.

(5.4a) *Er blieb uns eine **Erklärung** schuldig, **warum** diese so lange auf sich warten ließ.* ("He ows us an explanation, why it kept us waiting so long").

(5.4b) *Viele Hochschulen haben mit *__Erklärungen__ deutlich gemacht, **dass** die Hochschulen auch weiterhin offen für Studierende aus dem Ausland sein müssen.* ("Many universities have made it clear with expanations that universities must keep on being open for foreign students")

In (5.4a) the noun *Erklärung* ("explanation") subcategorises for the *w*-clause in the NF. However, the subclause can be also subcategorised by other elements in the

[9]The percentage varies across different subclause types. The lower result for *w*-clauses is explained by the occurrence of the sentences, which are introduced by *wohin, wo* and have a "place"-meaning, e.g. *Im französischen Exil, *wohin er nach seiner Freilassung 1980 emigrierte, nahm er schnell seine politischen Aktivitäten wieder auf* ("In the French exile, where he emigrated after his release in 1980, he took up again his political activities").

MF, e.g. verbs or multiword expressions, as in (5.4b), where the expression *deutlich machen* and not the noun *Erklärungen* subcategorises for the *dass*-clause.

To increase the precision of these extraction contexts, further constraints based on the linguistic knowledge must be included into the queries (e.g. the information on the subcategorisation properties of verbs or adjectives, etc.). In the Vorfeld context, almost no restrictions are needed to achieve the required accuracy in the acquisition of subcategorisation information.

5.2.3 Contexts for the Extraction of Multiword Predicates

Multiword expressions are extracted from the same contexts used for verbal predicates. Thus, we use the VL and passive constructions to extract multiwords, which contain a prepositional phrase, a noun and a support verb, cf. section 3.3 for the definition of the multiwords under analysis. In the VL sentence, the support verb occupies LSK, whereas the preposition and the noun of the multiword tend to immediately precede the verb. The subclause following the verb after a comma is subcategorised by the whole multiword expression, cf. table 5.16.

VF	LSK	MF			RSK	NF
	main clause					**subclause**
	conj	**sent.elements**	**prep**	**noun**	**verb**	
	Weil	*in den Vorwürfen*	*zum*	*Ausdruck*	*kommt,*	*dass unabhängige Software-Häuser in Zukunft wieder bevorzugt Java-Programme schreiben.*
	Because	in the reproaches	to the	expression	comes	that independent software houses will prefer to write Java programmes in the future.

Table 5.16: A multiword predicate in a VL sentence

The passive clause context also shows high predictability. The prepositional and the nominal elements are situated in the MF, taking the position between the verbal parts of the passive verb, cf. table 5.17.

For both contexts we know that the preposition and the noun of the multiword is located at the end of the MF, and is immediately followed by the verb in the RSK. Moreover, no further constituents are allowed between the prepositional and the nominal part of the multiword, except for an article (the possible forms of the searched multiwords are given in table 5.18). Further potential valency bearer of the subclauses in the NF, e.g. nominal predicates, can occur left to the multiword only in the applied contexts. We illustrate this in the following examples. Both, in the VL, cf. (5.5a), and in a passive sentence, cf. (5.5b), the declarative subclause in the NF is subcategorised by the nominalisation *Befürchtung* (“fear”), which is left to the wrongly extracted multiword. These cases can be avoided if a lexical-syntactic filter is included into the query. The filter contains lexical constraints using the knowledge about nouns taking declarative clauses and eliminates all those multiword candidates, which are preceded by the nouns subcategorising for dass-clauses, cf. the description of constraints for verbal predicates in section 5.2.1.3 above.

	VF	LSK	MF		RSK	NF
	main clause					**subclause**
	const	**verb**	**prep**	**noun**	**verb**	
V1		*Kann*	*zum*	*Ausdruck*	*gebracht werden*	*dass unabhängige Software-Häuser in Zukunft wieder bevorzugt Java-Programme schreiben?*
		Can	to the	expression	be brought	that independent software houses will prefer to write Java programmes in the future?
V2	*Es*	*muss*	*zum*	*Ausdruck*	*gebracht werden*	*dass unabhängige Software-Häuser in Zukunft wieder bevorzugt Java-Programme schreiben.*
	It	must	to the	expression	be brought	that independent software houses will prefer to write Java programmes in the future.

Table 5.17: A multiword predicate in a passive construction

PP consituents			**examples**
preposition *in*	no article -	noun *Erfahrung*	*in Efahrung bringen*
preposition *in*	definite article *den*	noun *Blick*	*in den Blick geraten*
combination of a preposition & an article *in+das=ins*		noun *Grübeln*	*ins Grübeln kommen*

Table 5.18: Forms of the PP in the searched multiword expressions

(5.5a) *Weil in den Vorwürfen die Befürchtung **__zum Ausdruck kommt__*, dass unabhängige Software-Häuser in Zukunft wieder bevorzugt Java-Programme schreiben.*
("Because in the reproaches they express the fear that independent software houses will prefer to write Java programmes in the future").

(5.5b) *In den Vorwürfen* __wird__ *die Befürchtung **__zum Ausdruck gebracht__*, dass unabhängige Software-Häuser in Zukunft wieder bevorzugt Java-Programme schreiben.*
("In the reproaches they express the fear that independent software houses will prefer to write Java programmes in the future").

Therefore, we expect that the accuracy of the automatic extraction of multiword candidates in the VL and passive sentences is higher than in other possible contexts, e.g. in v2. In (5.6), we illustrate the examples of error extraction from v2. The valency bearer in these cases is also the nominal predicate, which subcategorises for a subclause. To eliminate the multiword candidates whose context partners are such nominals, we can also use the above mentioned lexical-syntactic filters. However, the number of the positions of subclause-taking nouns, which cause the noise, is higher and thus, less predictable as in the previous cases, cf. (5.6a) to (5.6c).

(5.6a) *In den Vorwürfen* __kommt__ *die Befürchtung **__zum Ausdruck__*, dass unabhängige Software-Häuser in Zukunft wieder bevorzugt Java-Programme schreiben.*
("In the reproaches they express the fear that independent software houses will prefer to write Java programmes in the future")

(5.6b) *Es* __kommt zum Ausdruck__ *die Befürchtung, dass unabhängige Software-Häuser in Zukunft wieder bevorzugt Java-Programme schreiben.*
("In the reproaches they express the fear that independent software houses will prefer to write Java programmes in the future")

(5.6c) __Zum Ausdruck kommt__ *in den Vorwürfen die Befürchtung, dass unabhängige Software-Häuser in Zukunft wieder bevorzugt Java-Programme schreiben.*
("In the reproaches they express the fear that independent software houses will prefer to write Java programmes in the future")

The calculation of the possible positions of subclause-taking nouns in the v2 is more time-and labour-intensive, whereas in the VL and the passive sentence their occurrence is definitely predictable, which allows us to save time and effort in achievement of high precision results.

5.3 Extraction and Classification Procedures

In the following part we describe extraction and classification procedures, used in our analysis. As mentioned above, in this study, we apply the CWB query system. We elaborate a cascaded architecture to extract and classify the predicates described in section 4 above. The architecture is based on symbolic procedures, which procede from the general to the specific, (Lapshinova 2007). First, we apply CQP-queries for extracting all types of predicates in general contexts (for instance verb final sentences and passive constructions). Then we specify CQP-queries to extract different kinds of

predicates: verbal, nominal and multiword, which can be further subclassified into more specific subtypes according to their subcategorisation features, as described in section 4.

An overview of the extraction and classification steps used in this thesis is given in figure 5.4 below. Our algorithm consists of a sequence of procedures to identify and extract different types of predicates, such as verbs, nouns and multiword expressions, as well as classify them according to their subcategorisation features.

1 apply general queries to extract sentences containing predicates
- 1.1 use the VL and passive context for verbal and multiword predicates
- 1.2 for nominal predicates:
- 1.2.1 use the VF context for nominal predicates
- 1.2.2 continue with the step 3

2 apply specific queries to identify predicates
- 2.1 use specific queries for verbal predicates
- 2.2 use specific queries for multiword predicates

3 classify predicates
- 3.1 classify verbal predicates: V1, V2, V3
- 3.2 classify nominal predicates:
- 3.2.1 according to their subcategorisation structure: N1, N2, N3
- 3.2.2 according to their morphological structure: simplex vs. compound

4 compare relations between morphologically related predicates
- 4.1 identify and classify *ung*-nominalisations: (Nung1), (Nung2), (Nung3)
- 4.2 identify, extract and classify base verbs: (Vbase1), (Vbase2), (Vbase3)
- 4.3 classify relations between nominalisations outside and inside a multiword and their base verbs: R1, R2, R3

5 additional procedures
- 5.1 subclassify compound nouns (according to the relations with their head and non-head constituents): C1, C2, C3-1 and C3-2
- 5.2 subclassify multiwords according to the relations with their nominal constituent: M1, M2, M3, M4

Figure 5.4: Cascade of steps to extract and classify predicates

5.3.1 Predicate Extraction: General Queries

The first part of the extraction procedures is represented by general queries, which aim at the extracton of verbal and multiword predicates. As mentioned above, we use restricted contexts allowing for a high-precision extraction, which is in this case, the VL constructions,as well as sentences with verbs in passive voice, cf. 5.2.1. Nominal predicates are extracted in the Vorfeld, cf. 5.2.2.

5.3.1.1 General query 1: VL constructions

To extract both verbal and multiword predicates in the VL sentence model, in which the predicates we aim to explore occupy the final position in the main clause, followed by a comma and the introductory element of the subordinate clause. For example, the sentences in (5.7) are extracted in the VL context. The final position of the main clause is occupied by a finite verb, which either subcategorises for the following subclause, as in (5.7a), or is a part of a multiword predicate, subcategorising for the sentential complement, as in (5.7b).

(5.7a) *...weil nicht mehr die Parlamentarier selbst künftig darüber* **entscheiden sollen***, wieviel Geld sie bekommen.* ("...because not even the parlament members themselves must decide how much money they will get").

(5.7b) *Weil Clinton und sein Anwälte* **in Erfahrung bringen wollen***, was Lewinski zu sagen hat.* ("Because Clinton and his lawyers want to find out, what Lewinski has to say").

In table 5.5, we give a scheme of the general query[10] for the predicate extraction in VL, illustrating it with the sentence in (5.2). The query contains blocks to restrict the extraction to the sentences, which contain verb-final clauses. A verb-final clause starts with a conjunction, a relative or an interrogative pronoun (see line 2). These elements do not necessarily occupy the sentence start and can be preceeded by other sentence constituents, as seen from line 1. Line 3 specifies the extraction of the sentence parts filling the middle field (cf. section 5.2.1) and contains the constraints that eliminate the occurrence of finite verbs and punctuation in this position. Lines from 4a to 4d include elements for the extraction of the verbal complex. The finite verb is in the final position of the verbal complex. It can be preceded by up to three other verbal forms, cf. table 5.8. The verbal complex is located at the end of the main clause and is followed by a comma and the introductory elements of the subcategorised subclause (lines 5 and 6). The subclause can be subcategorised by the verb, the elements preceding the verb or by their combination (multiword expression). It can start either with a *w*-word (line 6a) or with the conjunctions *dass* or *ob* (lines 6b and 6c). The subclause contains optional words (line 7) and is followed by a finite verb (line 8) because it also has the form of the VL. Line 9 is used for punctuation. The explanations for the STTS tagset and the CQP query language are included in the appendix B.

[10]The macros with queries applied in this work are available on request.

	Query building blocks	**comments**	**matching sentence**
	MACRO vl(0)	the start of the macro	
1.		optional elements	...
2.	[pos="KOU.*|PREL.*|PW.*"]	the start of the VL clause conjunction, relat. or interrogat. pronoun	*weil*
3.	[pos!="V.*FIN"&word!=",|-"]*	optional words, no finite verbs or punctuation	*nicht mehr die Parlamentarier selbst künftig darüber*
4			
4a.	<vc>	the start of the verbal complex	
4b.	[pos= "V.*"]0,3	elements of the verbal complex	*entscheiden*
4c.	[pos= "V.FIN"]	finite verb	*sollen*
4d.	</vc>	the end of the verbal complex	
5.	","	comma	,
6.		the subclause start:	
6a.	[(pos="PW.*")]|	*w*-word	*wieviel*
6b.	[word = "dass"]|	*dass*-conjunction	
6c.	[word = "ob"]	*ob*-conjunction	
7.	[pos!="V.*FIN"]*	optional words, no finite verbs	*Geld sie*
8.	[pos="V.FIN*"]	finte verb	*bekommen*
9.	[pos="$."]	the subclause and sentence end	.
	;	the end of the macro	

Figure 5.5: Query for subclause-taking predicates in VL

5.3.1.2 General query 2: passive constructions

Another context for general extraction is represented by passive constructions, cf. section 5.2.1.2. In a passive sentence the verbal complex occupies both, the right and the left sentence brackets in the main clause, and NF is occupied by the subclause, which immediately follows the verb. The valency bearer is still ambiguous. The subclause can be subcategorised by the verb itself or the multiword expression, whose prepositional and nominal elements precede the verb in this context, cf. section 5.2.3. Examples of verbal and multiword predicates in passive are shown in (5.8a) and (5.8b) below.

(5.8a) *Dort* **wird** *bis zum 13. November* **gezeigt**, *was Reinhart Stoll in den Bereichen Malerei und Grafik geschaffen hat.*
("Till November, 13 there will be shown what Reinhart Stoll managed in the areas of painting and drawing").

(5.8b) *Es* **darf** *innerhalb der CDU nicht* **in Frage gestellt werden**, *dass die Republikaner 'eine antidemokratische und autoritäre Partei' seien.*
("It's not allowed to put into a question (to doubt) in the CDU that the Republicans 'are an antidemocratic and an authoritative party' ").

	Query building blocks	**comments**	**matching sentence**
	MACRO passiv(0)	the start of macro	
1a. 1b.	(<s>| [pos = "KON"])	sentence start or conj	
2a. 2b. 2c. 2d.	([lemma = "es"]| (<pp> []* </pp>| [pos = "ADV"])*| (<np> [lemma!=$nounlist]* </np>)*	the Korrelat *es* optional PP or an adverb optional NP, no nouns taking a subclause	*Es*
3.	[pos= "V(M|A)FIN"]	finite aux. or modal verb	*darf*
4.	[pos!="V.*FIN"&word!=",|-"]*	optional words, no finite verbs	*innerhalb der CDU in Frage*
5. 6. 7.	[pos= "VV(PP|IZU)"] [pos= "VA(PP|INF)"]? [pos= "V(A|M)INF"]?	full verb participle or *zu*-infinitive	*gestellt* *werden*
8.	","	comma	,
9. 9a. 9b. 9c. 10. 11. 12.	 [(pos="PW.*")]| [word = "dass"]| [word = "ob"] [pos!="V.*FIN"]* [pos="V.FIN*"] [pos="$."]	the subclause start: *w*-word *dass*-conjunction *ob*-conjunction subclause non-verbal part finte verb of the subclause the subclause and sentence end	 *dass* *die Republikaner ... Partei* *seien* .
	;	the end of the macro	

Table 5.19: Query for subclause-taking predicates in passive

In table 5.19 we show the general query to extract predicates in passive constructions, both, in the v1 and in the v2 sentence constructions, as was described in table 5.9 in section 5.2.1.2. The v1 passive sentence can start either with the Korrelat *es* (line 2a) or directly with a finite auxiliary or modal verb (line 3). The operator | is used to specify the optionality of other constituents in front of the final verb (lines 2a to 2d).

The v2 passive sentence can start either with an adverb or an adverbial prepositional phrase (lines 2b and 2c), or with a nominal phrase, which should not contain any subclause-taking nouns (line 2d). The list of subclause-taking-nouns is obtained from another context, described in section 5.3.1.3 below. The list is defined in form of the variable $nounlist:

```
> define $nounlist< "nounlist.txt"
```

The verbal complex of both v1 and v2 models starts with a finite auxiliary or modal verb (line 3) and is followed by optional elements, which should not contain any finite verbs or punctuation (line 4). The full verb of the verbal complex, which is either a participle or a *zu*-infinitive (line 5), precedes one or two further verbal elements: infintive or participle forms of auxiliaries or modals (lines 6 and 7). The subcategorised subclause follows the verbal complex after the comma. The subclause can start with the conjunctions *dass* or *ob* or with a *w*-word (line 9) followed by optional words (line 10) and a finite verb in the end (line 11).

5.3.1.3 General query 3: Vorfeld

As mentioned in section 5.2.2 above, we use the Vorfeld constructions to extract nominal predicates subcategorising for sentential complements. In this case we search at the sentence beginning for nouns or noun phrases followed by a subclause. The main verb of such a sentence occupies the position after the subclause, cf. example (5.9).

(5.9) *Allein die Ankündigung, dass er komme, hatte den Börsenkurs vergangene Woche in die Höhe getrieben.* ("Alone the announcement that he would come had boosted the course in the previous week").

The scheme of the Vorfeld query to extract sentences as in (5.9) is given in table 5.20. Line 1 contains a constraint, which imposes the query to start the search at the beginning of the sentence. The nominal predicate under analysis (line 4) can be preceded by some prenominal material, e.g. by adverbs (line 2). The nominal predicate we aim to extract can also be a part of a prepositional phrase, thus, we include constraints for a preposition or a combination of a proposition and an article (cf. table 5.14 in section 5.2.2 above) before the nominal phrase in line 3. Line 5 contains a constraint for a Korrelat (e.g. *darüber, dafür*, etc.), which is optional. The subcategorised subclause follows the nominal phrase or the Korrelat immediately after the comma and is introduced by the conjunctions *dass, ob* or a *w*-word, cf. lines 7a to 7c. The structure of the subclause remains the same as in previous cases, cf. tables 5.5 and 5.20. The subclause ends with a comma (line 10), which is followed

	Query building blocks	**comments**	**matching sentence**
	MACRO vf(0)	the start of the macro	
1.	<s>	sentence beginning	
2.	[pos!="NN|V.FIN"]{0,3}	prenominal material, no finite verbs or nouns	*Allein*
3.	[pos="APPR.*"]?	optl. preposition or preposition & an article	
4.	(<np> ... </np>)	noun phrase	*die Ankündigung*
5.	[word= "da.*"&pos= "..."]?	optional Korrelat	
6.	","	comma	,
		the subclause start:	
7a.	[(pos="PW.*")|	*w*-word or	
7b.	(word="ob")| |	*ob*-conjunction or	
7c.	(word="dass")]	*dass*-conjunction	*dass*
8.	[pos!="$.|V.FIN"]*	subclause non-verbal part	*er*
9.	[pos="V.FIN"]	subclause fin.verb	*komme*
10.	","	comma	,
11.	[pos="V.FIN"]	finite main verb	*hatte*
12.	[pos!= "V.FIN"]*	rest of the main clause: oprional words	*den Börsenkurs vergangene Woche in die Höhe getrieben.*
13.	[pos="$."]	the subclause and sentence end	.
14.	within s;	rest of the main clause	
	;	the end of the macro	

Table 5.20: Query for subclause-taking nominal predicates in VF

	Query building blocks	**comments**	**matching sentence**
4.			
option 1			
4a.	(<np>	the start of the nom.phrase	
4b.	([pos = “ART|PIAT|PDAT|PPOSAT”]?	optional determiner	*die*
4c.	[pos!= “NN|V.FIN”]{0,4}	up to 4 optional elements, no nouns and finite verbs	
4d.	**@[pos = “NN”])|**	noun	***Ankündigung***
option 2			
4e.	([pos = “ART|PIAT|PDAT|PPOSAT”]?	optional determiner	*die*
4f.	[pos!= “NN|V.FIN”]{0,4}	up to 4 optional elements, no nouns and finite verbs	
4g.	**@[pos = “NN”]**)	noun	***Ankündigung***
4h.	([(pos= “ART|ADJA|CARD|PDAT| PIAT|PPOSAT”)&(word=".*er|.*es")]?	article or adj in gen.	*des*
4i.	([pos = “NN|NE”]& agr contains “.*Gen.*”))|	noun in genitive	*Präsidenten*
option 3			
4j.	([pos = “ART|PIAT|PDAT|PPOSAT”]?	optional determiner	*die*
4k.	[pos!= “NN|V.FIN”]{0,4}	up to 4 optional elements, no nouns and finite verbs	
4l.	**@[pos = “NN”]**)	noun	***Ankündigung***
4m.	([pos= “APPR.*”]&	preposition or preposition & an article	*vom*
	[pos= “ART”]*&	an optional article	
	[pos= “NN|NE”]	a noun	*Präsidenten*
4n.	</np>)	the end of the nom.phrase	

Table 5.21: Specification to extract nominal predicates in the VF

by the finite verb of the main clause (line 11) and the rest of the clause that can contain various elements.

Line 4 in the given VF query (cf. table 5.20) is underspecified and its application in the search for nominal predicates can reduce the accuracy of the extraction results. We specify the nominal phrase in table 5.21, including constraints for different forms of the full nominal phrase, as shown in table 5.14 in section 5.2.2 above.

As stated in table 5.14, the NP in the VF can contain an optional determiner (an article, indefinite, demonstrative or a possessive pronoun specified in lines 4b, 4e and 4j). It can also contain optional adjectives, adverbs and other elements, the number of which we limit to 4 (lines 4c, 4f and 4k) to avoid the noise. The constraint for the nominal predicate to extract is contained in lines 4d, 4g and 4l). Option 2 includes constraints for a NP in genetive (lines 4h and 4i), whereas option 3 includes constraints for a PP (4m and 4n) which can occur within the searched NP in the VF.

The above described general queries are outlined in the scheme, illustrated in figure 5.6 below. This scheme represents a segment of the whole architecture for the extraction and classification of predicates.

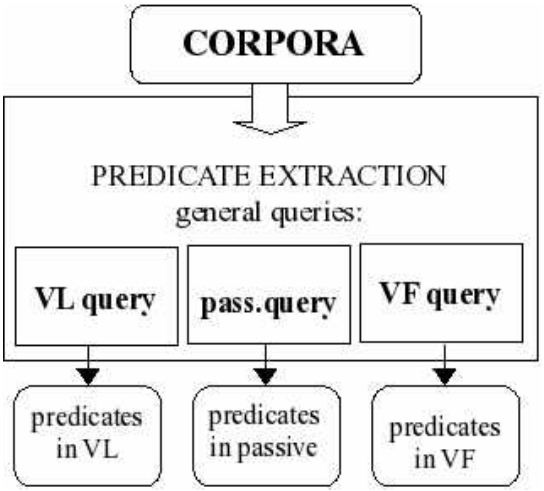


Figure 5.6: General queries in the extraction and classification architecture

5.3.1.4 Filtering procedures for general queries

In some cases further restrictions should be applied to increase the accuracy of the described extraction procedures. These restrictions can either be included as building blocks directly into the main queries, or can be developed as separate queries and applied within a cascaded set of procedures. In the following, we describe cases for which we need to apply additional filtering procedures.

Headless relatives and adverbial relative clauses introduced by certain *w*-words The preliminary extraction tests show that headless relatives as well as adverbial relative clauses can decrease the precision of the obtained predicates, which subcategorise for *w*-clauses. The form of the subcategorised *w*-clauses does not differ from that of the headless and adverbial relatives making their identification problematic. Our extaction tests show that most clauses that start with the *w*-words *wobei,*

wodurch, womit and *wonach* are adverbial relative clauses. We exclude these noisy cases from extraction integrating constraints, which bans the extraction of subclauses starting with the given *w*-words, cf. table 5.22.

	Query building blocks	**comments**
1.	[(pos="PW.*")&	introductory *w*-word
2.	(word!= "wobei|wodurch|womit|wonach")]	no *wobei, wodurch, womit, wonach*

Table 5.22: Query constraint which excludes noisy *w*-words

The described filtering procedure is applied to all the three general queries described above. However, for the extraction of nominal predicates in the VF, we modify it excluding the constraint for the *w*-word *wonach*. Our tests show that 100% of nominal predicates taking subclauses introduced by *wonach* in the VF prove to subcategorise for *dass*-clauses[11], cf. (5.10a) and (5.10b). Thus, the candidates for nominal predicates whose context partners are *wonach*-clauses are not eliminated by the query and extracted along with other *w*-clauses. The nouns with *wonach*-clauses are saved and sorted out as predicates subcategorising *dass*-clauses within the predicates subclassification procedures.

(5.10a) *Gerüchte, wonach der Täter den Gerstensaft selbst gesoffen hätte, bestätigten sich nicht.* ("The rumours according to which the offender boozed the amber nectar himself proved false").

(5.10b) *Gerüchte, dass der Täter den Gerstensaft selbst gesoffen hätte, bestätigten sich nicht.* ("The rumours that the offender boozed the amber nectar himself proved false").

Constructions with antecedents in VL and passive contexts Further constructions, which contribute to the inaccuracy of extraction from VL and passive clauses, are constrauctions with antecendents. The first part of the expression is located before the main verb in the main clause, whereas the second part is the subclause-introducing *w*-word. For instance, the interrogative pronoun *was* ("what") can have the antecedents *das, alles, etwas, nichts*, etc. in the main clause as illustrated in examples (5.11a) and (5.11b). In this case, the *w*-clause is not subcategorised by the verb.

(5.11a) *Wenn du für **irgendetwas** verantwortlich gemacht wirst, **was** du in deinem Amt getan hast* ("If you're blamed for something what you've done in office").

(5.11b) *Wenn ich **alles** aufzähle, **was** wir gemacht haben, verpassen wir heute das Abendessen.* ("When I list **everything what** we have done, we will miss the dinner").

To eliminate the predicate candidates, which occur with antecedent constructions in context, we design a query (shown in table 5.23) containing constraints blocking the occurrence of these constructions in the obtained data.

[11]For this test, we analysed 150 nouns extracted with *wonach*-clauses from 'taz', cf. evaluation results in section 6.2.2.1

	Query building blocks	examples	
1.	[word="das|etwas|alles|irgend.*|nichts"]	***irgendetwas***	***alles***
2.	[pos!="$."& pos!="NN"] {0,4}	*verantwortlich*	-
3.	<vc>...</vc>	*gemacht wirst*	*aufzähle*
","	,	,	
4.	[word="was"]	***was***	***was***

Table 5.23: Constraints to exclude noisy cases with antecendents

Adverbial relative clauses introduced by the pronoun *wo* The preliminary analysis of the extracted predicates, which take subclauses introduced by the interrogative pronoun *wo* ("where"), shows that about 60% of these cases prove to be adverbial relative clauses, thus, false positives in our extraction results. As subclauses introduced by *wo* comprise about 22,6% of all *w*-clauses extracted in the VL context, they cause roughly 13% of the inaccuracy in the extraction results[12].

Therefore, we filter out the predicate candidates, which occur with subclauses introduced by the pronoun *wo* to increase the precision of our extraction procedures. However, these cases should not be completely ignored, as about 40% of them prove to be true positives and their elimination would reduce the recall. A filtering procedure is necessery to exclude the occurrence of the false positives from the obtained data. For this purpose, further constraints should be included into the filtering query.

False positives among the analysed *wo*-clauses are represented by adverbial relative clauses. The clauses of this type have often antecedents in the main clause, for instance, adverbs like *da, dahin* or *dort* ("there"), cf. example (5.12).

(5.12) *Weil er sich seit Jahrhunderten **dort** wohl fühlt, **wo** sie siedeln.* ("Because for hundred of years he feels good **there where** they live").

To eliminate the occurrence of these cases, we apply a query that contain constraints to ban the simultaneous occurrence of *da, dahin* or *dort* in the main clause and *wo* in the subordinate clause, as shown in table 5.24.

	Query building blocks	example
1.	[word="da|dahin|dort"]	***dort***
2.	[pos!="$."& pos!="NN"] {0,4}	*wohl*
3.	<vc>...</vc>	*fühlt*
4.	","	,
5.	[word="wo"]	***wo***

Table 5.24: Constraints to exclude noisy cases with correlative expressions

The elaboration of the filtering query is more complicated for the verbs, followed by *wo*-clauses whose antecendents in the main clause are nouns denoting a location, as illustrated in examples (5.13a) and (5.13b).

[12]We analysed 912 predicates extracted along with *w*-clauses from 'taz'. Sublauses introduced by *wo* comprise 22,6% (206) of all *w*-clauses. More than half of these cases (ca. 60%) prove to be false positives.

(5.13a) *Es wurde **eine ganz neue Welt** *entwickelt, **wo** die Bundesrepublik und eben auch die IG Metall eine führende Rolle spielen werden.* ("EN")

(5.13b) *Ince wird nun wahrscheinlich vor ein Gericht **in seiner Heimat** *gestellt, **wo** ihm die Todesstrafe droht* ("EN")

In this case, the query should include lexical constraints, which bans the occurrence of nominals for which we know that they have a "place"-meaning, cf. line 2 in table 5.25. These nominals can be obtained from another context, a query which is applied to filter out nominals whose context partners in the VF are adverbial relative clauses introduced by *wo*, as shown in example (5.14). The analysis of the nouns extracted along with *wo*-clauses in the VF shows that 98,3% of them are not subcategorised by this noun, cf. evaluation results described in section 6.2.2.1 below.

(5.14) ***Die Orte**, **wo** es den breitesten Urlauberstrom gibt, entsprechen diesem Anspruch gar nicht.*

To filter out such cases from the obtained data, we use the filtering query illustrated in table 5.26. The query contains constraints for the pronoun *wo* that occurs after the NP in VF.

	Query building blocks	**example**
1	...	*in seiner*
2.	[lemma!=$placenounlist]	***Heimat***
3.	[pos!="$."& pos!="NN"] {0,2}	
4.	<vc>...</vc>	*gestellt*
5.	","	,
6.	[word="wo"]	***wo***

Table 5.25: Lexical constraints to exclude "place"-nominals in the main clause

	Query building blocks	**example**
1.	(<np> ... </np>)	***Die Orte***
3.	","	,
4.	[(pos="PW.*")&(word="wo")]	***wo***

Table 5.26: Query to filter out nouns with occurring with *wo*-clauses in the VF

Possessive NPs as valency bearer In some cases, subclauses in VF are subcategorised not by the head noun, but by the embedded NP in genitive, cf. forms of the full NP listed in table 5.14 in section 5.2.2 above. For instance, the sentential complement in (5.15a) is subcategorised by the non-head NP in genitive (*des Problems*) and not by the head noun of the whole NP (*Beherrschung*). The VF query matches head nouns as candidates for nominal predicates. This error-match decreases the accuracy of our extraction results.

However, we shouldn't exclude NPs containing dependent gentitive NPs from the list of candidates, as this elimination would increase the number of false negative candidates. On the one hand, we would automatically lose such true positive predicate

candidates as, for instance, the nominal predicate *Antwort* ("answer") in (5.15b). On the other hand, the NPs containing embedded NP in genitive should not be classified as false positives, as they contain nominal valency bearer, which the VF query fails to identify.

(5.15a) *Die *Beantwortung der **Frage**, **warum** Menschen sich gegenseitig helfen sollen, sei für die Sozialwissenschaften heutzutage schwierig.* ("To answer the question why people should help each other is nowadays very difficult for social scientists").

(5.15b) *Die **Antwort** des Parlamentariers, **dass** man als Regierung eben Opfer bringen müsse, half nicht mehr.* ("The answer of the parlament member that one should make sacrifices as parlament didn't help anymore").

However, we can paraphrase constructions with genitive NPs into compounds.[13] The problem of the predicate identification in this case can be eliminated with the same procedures applied for compound nominal predicates, such as *Ursachenforschung* or *Journalistenfrage*. With the help of regular expression, we convert nominal expressions containing genitive NPs into compound nouns:

Beantwortung der Frage	→	*Fragebeantwortung*
Forschung der Ursachen	→	*Ursachenforschung*
Antwort des Parlamentariers	→	*Parlamentarierantwort*
Frage des Journalisten	→	*Journalistenfrage*

The converted nominal expressions are saved in the obtained data and treated as compound nominal predicates with the procedures described in sections 5.3.2.4 and 5.3.3.3 below.

Clauses introduced by relative pronouns The query for predicates in VL includes constraints for relative pronouns, such as *der, die, das*, which introduce a relative clause containing searched predicates, see section 5.2.1.1. The relative clause *die wußten* in example (5.16a) contains the verbal predicate *wissen*, which subcategorises for *w*-clauses and was obtained with the VL query (see the query specification in table 5.5 in section 5.3.1).

However, our query can also deliver false positives, relative sentences that do not contain any predicates. For instance, in (5.16b), the ob-clause following the relative clause is not subcategorised by the verb in RSK. Its valency bearer is the nominalisation *Überlegung* ("consideration"), which is modified by the relative clause *die er gemacht hat* ("which he made).

(5.16a) *Da saßen Leute, die **wussten**, **worauf** es ankommt.* ("There were sitting people who knew what it depends on").

(5.16b) *Die **Überlegung**, die er *gemacht hat, **ob** das stimmt.* ("The consideration he made if it is true is interesting = The consideration if it is true which he made is interesting").

[13] In section 3.4.1.2, we mention automatical generation of compounds within the STO lexicon.

To filter out these cases from the obtained data, we modify the query specified in table 5.5 and include the constraints, which block subclause-taking nouns[14] in front of relative clauses.

	Query building blocks	**matching sentence**
1.	[lemma=$nounlist]	*Überlegung*
2.	“,”	,
3.	[pos=“PREL.*”]	*die*
4.	[pos!=“V.*FIN”&word!=“,|-”]*	*er*
5.	<vc>...</vc>	*gemacht hat*
6.	“,”	,
7.	[(pos=“PW.*”)]|[word = “dass”]|[word = “ob”]	*ob*

Table 5.27: Filtering query to exclude subclause-taking nouns

Summary The procedures to filtering out noise-causing cases is integrated into the cascaded architecture. Some of them, e.g. the procedures to eliminate the occurrence of headless relatives and adverbial relative clauses are integrated into the general queries. Others are applied on the subcorpora obtained with general queries. In this case, we filter out the irrelevant cases by removing them from the obtained data.

The segment of the extraction architecture which specifies filtering procedures is shown in figure 5.7.

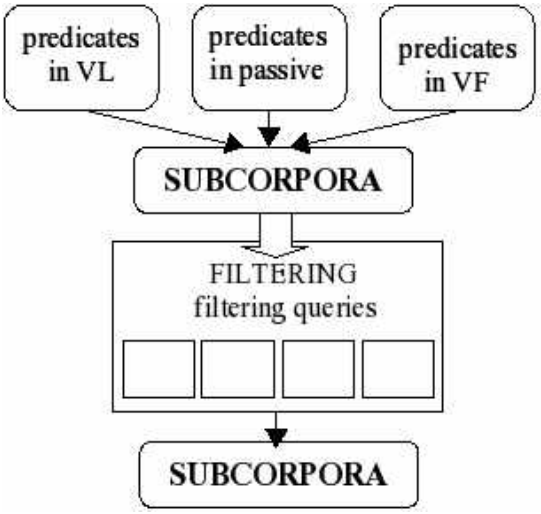


Figure 5.7: Filtering in the extraction and classification architecture

5.3.2 Predicate Classification: Specific Queries

Specific queries are applied on the subcorpus of predicates extracted with the above described general queries. We aim to identify different predicate types and subclassify

[14]We use the list of the nouns obtained with the query described in section 5.3.1.3, which is also applied in the query illustrated in table 5.19 in section 5.3.1 above.

them according to their subcategorisation features. Specific queries are designed of similar elements as the general ones, but contain additional constraints for expected predicate types. For instance, the subclause in the NF extracted both by the general VL query and the general passive query, can be subcategorised not only by the verb in RSK, but by other predicates, which precede RSK, e.g. nominal predicates or multiword expressions[15]. Therefore, to restrict the search to a certain predicate type, we specify the general queries described in sections 5.3.1.1 and 5.3.1.2. For this purpose, we include additional constraints that allow for a more precise definition of the searched predicates.

In the following, we describe the specific queries applied to identify different predicate types (verbs, nouns and multiword expressions), which are classified out of the subcorpora acquired by the general queries described in section 5.3.1 above.

5.3.2.1 Query for verbal predicate extraction

The specific query to extract verbal predicates is based on the general queries for extraction of predicates in the VL and passive context. We only need to include lexical constraints, which exclude the occurrence of other potential valency bearers near the verbal candidate. For example, in (5.17a) it is not the verb *erklären* ("to explain") that subcategorises for the *w*-clause, but the preceding noun *Grund* ("explanation"). To avoid such cases and extract constructions in which the verbal predicates subcategorise for subclauses (as in (5.17b)), we should lexically specify line 3 of the queries in 5.5 and 5.19. We include a constraint that bans the occurrence of subclause-taking nouns mentioned in sections 5.3.1.2 and 5.3.1.4 above. As in previous cases, see queries in tables 5.19 and 5.27, we define the list of the nouns in form of a variable ($nounlist), as illustrated in tables 5.28 and 5.29.

(5.17a) *...weil keine dort den **Grund** *erklären konnte, warum die Treuhand für Singwitz die Flinte ins Korn wirft.*
("...because nobody there could explain the **reason** why the Treuhand throws in the towel for Singwitz").

(5.17b) *...weil keine dort **erklären konnte**, warum die Treuhand für Singwitz die Flinte ins Korn wirft.*
("...because nobody there **could explain** why the Treuhand throws in the towel for Singwitz").

Table 5.28 shows the specified query for extraction of verbal predicates in the VL context. The lexical constraint in line 3b eleminates the extraction of candidates whose context partners are nominal predicates as in (5.17a), which makes the search for verbal predicates as in (5.17b) precise. We also specify the forms of verbal predicates in VL, as given in table 5.8 in section 5.2.1.1 above. For this purpose we modify line 4b of the general query (see table 5.5), in which three constituents of verbal complex remain underspecified.

As the MF of passive clauses can also contain subclause-taking nouns (see example (5.18a)), we should exclude their occurrence in this context type. We insert the same

[15]The subclause can also be subcatgorised by a predicative adjective. As we analyse only verbal, nominal and multiword predicates in this study, we do not consider adjectives.

	Query building blocks	comments	matching words
2.	[pos="KOU.*|PREL.*|PW.*"]	the start of the VL clause conjunction, relat. or interrogat. pronoun	 *weil*
3a.	([pos!="V.*FIN"]&[word!=",|-"]&	optional words, no finite verbs, no punctuation	*keine dort*
3b.	[lemma!=RE(**$imslex_nounlist**)&pos= "NN"])*	no nouns from the list	**den Grund*
4a.	<vc>	start of the verbal complex	
4b.	[pos="V.*(INF|IZU|PP)"]?	infinitive or participle	*erklären*
4c.	[pos="V(A|M)(INF|IZU|PP)"]{0,2}	infinitive or participle of auxiliary or modal verb	
4d.	[pos="V.*FIN"]	final verb	*konnte*
4e.	</vc>	end the verbal complex	

Table 5.28: Specification of the query to extract verbal predicates in VL

lexical constraints to eliminate extraction candidates whose partners are subclause-taking nouns, as we did it for the VL query, cf. tables 5.28 and 5.29. The constituents of the verbal complex are already specified in the general query, so the rest of the query remains unchanged.

(5.18a) *Es konnte dort der **Grund** *erklärt werden, warum die Treuhand für Singwitz die Flinte ins Korn wirft.*
("The **reason** couldn't be explained there why the Treuhand throws in the towel for Singwitz").

(5.18b) *Es konnte dort **erklärt werden**, warum die Treuhand für Singwitz die Flinte ins Korn wirft.*
("There it couldn't be **explained** why the Treuhand throws in the towel for Singwitz").

	Query building blocks	comments	matching words
2.	[pos= "V(M|A)FIN"]	finite aux. or modal verb	*konnte*
3a.	([pos!="V.*FIN"]&[word!=",|-"]&	optional words, no finite verbs, no punctuation	*dort*
3b.	[lemma!=RE(**$imslex_nounlist**)&pos= "NN"])*	no nouns from the list	*der *Grund*
4.	[pos= "VV(PP|IZU)"]	full verb participle or *zu*-infinitive	*erklärt*
5.	[pos= "VA(PP|INF)"]?		*werden*
6.	[pos= "V(A|M)INF"]?		

Table 5.29: Lexical specification of the query to extract verbs in passive

The query containing the building block, which excludes the the occurrence of lemmas from the given list of nouns, aims to search for a sentence with a subordinate

clause that can be subcategorised only by the verb, as other possible predicate types that can appear before the verb are forbidden.

5.3.2.2 Queries for multiword extraction

To extract multiword expressions from the subcorpora obtained by the general queries in tables 5.5 and 5.19, we develop a new query, which is based on the general ones. We concentrate on multiword expressions consisting of a preposition, a noun and a verb (as in (5.19)), and to identify such multiwords, we specify lines 3 and 4 of both general queries (illustrated in tables 5.5 and 5.19 above).

(5.19) *Es fällt mir auch schwer, mich dem Westberliner Schlangestehen anzupassen, unwillkürlich drängle ich, um **in Erfahrung zu bringen**, ob es von der Wurst noch etwas gibt.*
("It is diffcult for me to adapt myself to the West Berlin queueing, I automatically push my way through to **bring into experience** (to find out) if there is enough sausage left")

Line 3 is extended with the constraints for a prepositional phrase, preposition and a noun into line 3. As most V+PP multiwords are support verb constructions, we limit the number of possible verbs to the list of support vebs, lexically specifying the lines of the query, which aims at verb extraction (line 4 in the queries in 5.5 and 5.19). We use the list of support verbs proposed by (Daniels 1963) and (Breidt 1993), and expand it including further verbs that were detected in our preliminary tests.

To include the support verb list into the query, we define it in form of the variable $supportverblist:

```
> define $supportverblist< "supportverblist.txt"
```

In tables 5.30 and 5.31, we illustrate specific queries to extract multiword expressions in the VL (cf. table 5.30) and in passive (cf. table 5.31) contexts.

Constraints for multiword expressions include building blocks for a preposition and an optional article or a combination of a presposition with an article (line 3.2), which can be followed by one optional word (line 3.3) and a noun (line 3.4). We limit the number of words that can occur between the nominal and the verbal elements of a multiword to 3 (line 3.5). The lexical constraint for support verbs has the position before the possible auxiliaries in the VL constructions, see table 5.30. In the passive construction, it can have either a *zu*-infinitive or a participle form, cf. table 5.31.

5.3.2.3 Identification of nominalisations

As mentioned in sections 5.2.2 and 5.3.1.3 above, we extract nominal predicates in Vorfeld, as this context delivers results of higher precision. Nominal predicates, which are represented by different types of nouns (including nominalisations), are identified with the query illustrated in table 5.20 above and do not need further specification. However, for the analysis of the relations between morphologically

	Query building blocks	**comments**	**matching words**
	MACRO mwe_vl(0)	start of the macro	
1.		optional elements	*Es fällt mir auch schwer, mich dem Westberliner Schlangestehen anzupassen, unwillkürlich drängle ich,*
2.	[pos="KOU.*|PREL.*|PW.*"]	conjunction, relative or interrogative pronoun	*um*
3a.	[pos!="V.*FIN"&word!=",|-"]*	optional words, no finite verbs and punctuation	
3b.	(([pos="APPR] [pos= "ART"]?])| ([pos= "APPRART"]))	prep and optional article or prep+article	*in*
3c.	[]?	1 optional word	
3d.	[pos="NN"]	support(ed) noun	*Erfahrung*
3e.	[pos!="NN|PPER"& lemma!="da.*"]{0,3}	up to 3 words, no noun, personal pron., pron.adverb	
4a.	<vc>	start of the verb.complex	
4b.	[lemma=RE($verblist)& pos="VV.*"]	verb from SV list	*zu bringen*
4d.	[pos="V(A|M) (INF|IZU|PP)"]{0,2}	infinitive or participle of auxiliary or modal verb	
4e.	[pos="V(A|M)FIN"]?	optional final auxiliary or modal verb (if the full verb is infinite)	*konnte*
4f.	</vc>	end of the verb.compl.	
5.	","	comma	*,*
6a.	[(pos="PW.*")]|	*w*-word	
6b.	[word = "dass"]|	*dass*-conjunction	
6c.	[word = "ob"]	*ob*-conjunction	*ob*
7.	[pos!="V.*FIN"]*	optional words, no finite verbs	*es von der Wurst noch etwas*
8.	[pos="V.FIN*"]	finte verb	*gibt*
9.	[pos="$."]	the subclause and sentence end	.

Table 5.30: Query to extract multiwords in VL

	Query building blocks	**comments**	**matching sentence**
	MACRO mwe_passive(0)	the start of macro	
1a.	(<s>|	sentence start or	
1b.	[pos = “KON”]|	conj or	
1c.	[lemma = “es”])	the Korrelat *es*	*Es*
1d.	(<pp> []* </pp>|	prep.phrase or	
1e.	[pos = “ADV”])*	an adverb	
2.	[pos= “V(M|A)FIN”]	finite aux. or modal verb	*darf*
3a.	[pos!=“V.*FIN”&word!=“,|-”]*	optional words, no finite verbs and punctuation	
3b.	(([pos=“APPR]	prep and	*in*
	[pos= “ART”]?])|	optional article or	
	([pos= “APPRART”]))	prep+article	
3c.	[]?	1 optional word	
3d.	[pos=“NN”]	support(ed) noun	*Erfahrung*
3e.	[pos!=“NN|PPER”& lemma!=“da.*”]{0,3}	up to 3 words, no noun, personal pron., pron.adverb	
4a.	([pos= “VV(PP|IZU)”]&	participle, *zu*-infinitive	
4b.	[lemma=RE($verblist)])	verb from SV list	*gebracht werden*
4c.	[pos= “VA(PP|INF)”]?		
4d.	[pos= “V(A|M)INF”]?		
5.	“,”	comma	*,*
6a.	[(pos=“PW.*”)]|	*w*-word	
6b.	[word = “dass”]|	*dass*-conjunction	
6c.	[word = “ob”]	*ob*-conjunction	*ob*
7.	[pos!=“V.*FIN”]*	subclause non-verbal part	*es von der Wurst noch etwas*
8.	[pos=“V.FIN*”]	finte verb of the subclause	*gibt*
9.	[pos="$."]	the subclause and sentence end	.

Table 5.31: Query to extract multiwords in passive

related predicates (verbs and their nominalisations), we should identify and sort out deverbal nouns from the list of extracted nominal predicates. For this purpose we use the morphological tool SMOR[16].

With the help of SMOR, we can automatically analyse the morphological structure of the nouns extracted with the Vorfeld query (cf. table 5.20). An example of the resulting list is shown in figure 5.8 below.

Auskunft	→	Auskunft<+NN><Fem><Nom><Sg>
Bedenken	→	bedenken<V><SUFF><+NN><Fem><Nom><Sg>
Beispiel	→	Beispiel<+NN><Neut><Nom><Sg>
Beweis	→	Beweis<+NN><Masc><Nom><Sg>
Eindruck	→	Eindruck<+NN><Macs><Nom><Sg>
Einsicht	→	Einsicht<+NN><Fem><Nom><Sg>
Entdeckung	→	entdecken<V>ung<SUFF><+NN><Fem><Nom><Sg>
Erfahrung	→	erfahren<V>ung<SUFF><+NN><Fem><Nom><Sg>
Frage	→	fragen<V><SUFF><+NN><Fem><Nom><Sg>
Motiv	→	Motiv<+NN><Neut><Nom><Sg>
Verantwortung	→	verantworten<V>ung<SUFF><+NN><Fem><Nom><Sg>
Vorstellung	→	vor<PREF>stellen<V>ung<SUFF><+NN><Fem><Nom><Sg>

Figure 5.8: List of nominal predicates after the analysis with SMOR

Having the information about morphological features of nominal predicates, we can automatically sort out those, which contain the <V>-feature, thus, being derived from verbs. The list of nouns sorted form the list in figure 5.8 contains the nouns *Bedenken, Entdeckung, Erfahrung, Frage, Verantwortung* and *Vorstellung*. These nominalisations can be sorted further according to their morphological type: e.g *-en* or *-ung*-nominalisations. For example, if we extract *-ung*-nominalisations, we search for the nouns containing <V>ung<SUFF> only. The nominalisations with the suffix *-ung* are especially interesting in the analysis of the subcategorisation relations between verbs and their nominalisations. We can also search for this nominalisation type in corpora using a lexically specified CQP-query.

The analysis of SMOR output shows that the tool can deliver error results. For istance, the morphological analyser does not identify the noun *Beweis* ("evidence/proof") as a deverbal. However, we know that this noun is the nominalisation of the verb *beweisen* ("to prove"). Therefore, we should manually evaluate the output of the SMOR to make sure that all deverbal candidates are automatically identified by our tools.

5.3.2.4 Identification of simplex and compound nominals

The VF query for the extraction of nominal predicates illustrated in table 5.20 delivers not only simplex nominal predicates like *Erklärung* ("explanation"), *Frage* ("question"), *Beweis* ("justification"), but also compound ones, such as *Erklärungsversuch* ("explanation attempt"), *Wahrheitsbeweis* ("truth justification"), *Grundfrage* ("basic question"), etc. For example, this query can deliver both, the sentence in (5.7) and the one in (5.8).

(5.7) *Eine plausible* **Erklärung***, warum die Treuhand für Singwitz die Flinte ins Korn wirft, hat hier niemand.*

[16]Cf. (Schmid *et al* 2004).

("A plausible explanation why the Treuhand throws in the towel for Singwitz, has here nobody = Nobody here has a plausible explanation why...".)

(5.8) *Aber all die* **Erklärungsversuche***, warum der Teufel sich an die Frau Doktor heranmacht, sind auf der Glatze gedrehte Locken.*
("But all the expanation-attempts (attempts to explain) why the devil chats up the doctor are as futile as giving a bald man a comb").

	comments	**sentence with a simplex noun**	**sentence with a compound noun**
1.	sentence beginning		
2.	prenominal material	*Eine plausible*	*Aber all*
3.	optl. preposition or prep/art		
4.	noun phrase	*Erklärung*	*die Erklärungsversuche*
5.	comma	,	,
6.	*w*-word	*warum*	*warum*
7	subclause: non-verbal part	*die Treuhand für Singwitz die Flinte ins Korn*	*der Teufel sich an die Frau Doktor*
10.	finite verb of subclause	*wirft*	*heranmacht*
11.	comma	,	,
12.	finite main verb	*hat*	*sind*
13.	rest of main clause	*hier niemand.*	*auf der Glatze gedrehte Locken.*

Table 5.32: Simplex and compound nominal predicates in the VF

To classify the extracted nominal predicates into simplex and compound ones, we use the morphological tool SMOR, which allows us to obtain morphological features of the extracted nouns. We sort the morphologically analysed predicate candidates into the two groups illustrated in table 5.33.

simplex predicates	**compound predicates**
Beweis	*Beweismittel*
Erklärung	*Erklärungsversuche*
Prognose	*Expertenprognose*
Erfahrung	*Erfahrungsversuch*
Problem	*Forschungsproblem*
Frage	*Journalistenfrage*
Wahrheit	*Wahrheitsbeweis*

Table 5.33: Classification into simplex and compound predicates

Compound nouns can be distinguished from other nominal predicates due to their morphological structure, which has the following form:

wordpart1<NN>wordpart2<+NN>

In figure 5.9 below, we give a few examples of morphologically analysed compound predicates, sorted out from the list of nominal predicates. An overview of the procedures to extract and classify nominal predicates into simplex and compound ones is shown in figure 5.10.

Anhaltspunkt	Anhalt<NN>Punkt<+NN><Masc><Nom><Sg>
Beweis	Beweis<NN>Last<+NN><Fem><Nom><Sg>
Daumenregel	Daumen<NN>Regel<+NN><Fem><Nom><Sg>
Expertenprognose	Experte<NN>Prognose<+NN><Fem><Nom><Sg>
Faustregel	Faust<NN>Regel<+NN><Fem><Nom><Sg>
Forschungsproblem	Forschung<NN>Problem<+NN><Neut><Nom><Sg>
Grundsatz	Grund<NN>Satz<+NN><Masc><Nom><Sg>
Grundziel	Grund<NN>Ziel<+NN><Neut><Nom><Sg>
Legitimationsargumentation	Legitimation<NN>Argumentation<+NN><Fem><Nom><Sg>
Motivsuche	Motiv<NN>Suche<+NN><Fem><Nom><Sg>
Nagelprobe	Nagel<NN>Probe<+NN><Fem><Nom><Sg>
Rundfunk	Rundfunk<NN>Bericht<+NN><Masc><Nom><Sg>
Schlagwort	Schlag<NN>Wort<+NN><Neut><Nom><Sg>
Ursachenforschung	Ursache<NN>Forschung<+NN><Fem><Nom><Sg>

Figure 5.9: A list of sorted compound predicates analysed by SMOR

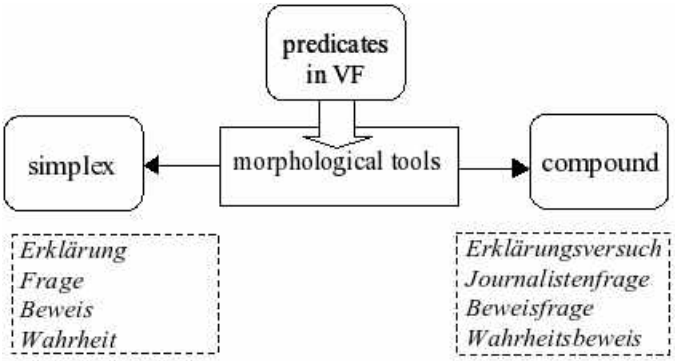


Figure 5.10: Extraction and classification architecture for simplex and compound nouns

5.3.2.5 Summary: specific queries

The specific queries are applied on the subcorpora acquired with the general queries described in section 5.3.1. For the specification of different predicate types, we use lexical and morpho-syntactic constraints which are integrated into the queries. Moreover, for identification of some predicate types, e.g. nominalisations and compound nouns, we use morphological tools. In figure 5.11, we show the segment of the architecture which contains procedures for the above described predicate classification.

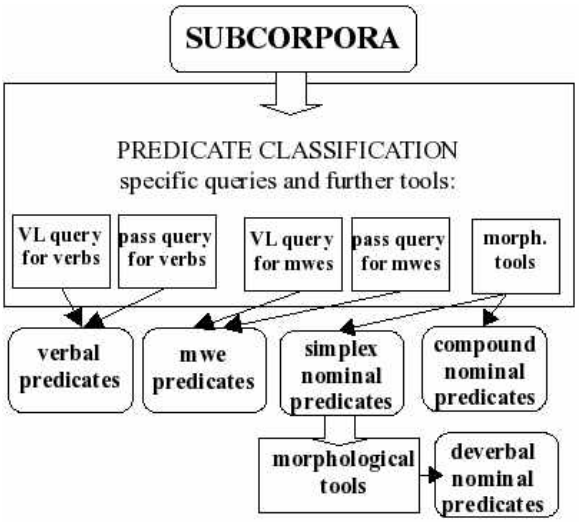


Figure 5.11: Specific queries in the architecture for extraction and classification of predicates

5.3.3 Predicate Subclassification: Further Specification

The predicates obtained with the general queries specified in section 5.3.1 and classified (into verbal nominal and multiword predicate classes) with the specific queries described in section 5.3.2, can be further subclassified according to their subcategorisation properties. The criteria for the subclassification of predicates relies on the absence or presence of declarative *dass*-clauses and interrogative *w-* or *ob*-clauses (the types of subclauses are described in sections 2.2.1.1 and 2.2.1.1 above). We classify each type of predicates (verbal, nominal and multiword) into the three following groups: predicates, which subcategorse for both interrogative and declarative subclauses, predicates, which allow only for interrogatives and predicates taking declarative subclauses only, cf. sections 4.2.1.3, 4.2.2.5 and 4.2.4.4.

In the following part of the thesis, we describe procedures to extract and subclassify verbs, nouns and multiwords according to their subcategorisation properties.

5.3.3.1 Subclassification of verbs

With the help of the queries described in sections 5.3.1 and 5.3.2 above, we obtain lists of verbal predicates subcategorising for *dass, w-* and *ob*-clauses. According to our classification, cf. section 4.2.1.3, verbs can be subdivided into the three following classes: V1, V2 and V3.

V1 contains those verbs, which subcategorise for both declaratives and interrogatives, dass-, w- and *ob*-clauses in our case. The V2 verbs take only interrogatives, both *w*-and *ob*-clauses in our study, whereas V3 includes verbs licensing declaratives only, i.e. *dass*-clauses. To subclassify the extracted lists of verbs, we apply a set of regular expressions formulated in form of shell scripts, which allow us to group these verbs according to the type of the complement clause they subcategorise for, as shown in table 5.34.

predicate class	**declaratives:** *dass*	**interrogatives:** *w-, ob*
V1	+	+
V2	-	+
V3	+	-

Table 5.34: Subcategorisation features specific for different verb classes

In the following table (5.35), we give sample lists of classified verbal predicates, which were extracted from the VL and passive contexts with the queries described in sections 5.3.1 and 5.3.2.

V1	V2	V3
ankündigen, dass/w- *begründen, dass/w-* *bemerken, dass/w-* *entscheiden, dass/w-/ob* *erinnern, dass/w-* *erklären, dass/w-* *festlegen, dass/w-* *mitteilen, dass/w-* *rechnen, dass/w-*	*aufklären, w-* *bestimmen, w-* *befragen, w-* *fragen, w-* *prüfen, w-/ob* *beweisen, w-*	*anordnen, dass* *bedingen, dass* *befürchten, dass* *behaupten, dass* *überzeugen, dass* *vereinbaren, dass* *versichern, dass*

Table 5.35: Sample verbal predicates classified into V1, V2, V3

5.3.3.2 Subclassification of nouns

Nominal predicates[17] extracted with the query specified in tables 5.20 and 5.21, can also be subclassified according to the three classes (N1, N2 and N3) described in section 4.2.2. N1 includes nouns, which subcategorise for both declaratives and interrogatives, N2 contains nouns taking interrogative *w-* and *ob*-clauses only, whereas the N3 nouns can license only *dass*-clauses.

With the help of a set of regular expressions formulated in form of shell scripts, we group nominal predicates extracted in the VF into the three above mentioned classes. The procedures correspond to those applied for the classification of verbal predicates.

[17]Both simplex and compound nouns.

predicate class	**declaratives:** *dass*	**interrogatives:** *w-, ob*
N1	+	+
N2	-	+
N3	+	-

Table 5.36: Subcategorisation features specific for different nominal classes

In table 5.37, we give a list of sample nominal predicates grouped according to the above mentioned classes. Further examples illustrating the three classes are listed in section A.2 in the appendix.

N1	N2	N3
Angst, dass/w-/ob	*Anfrage, w-/ob*	*Aussicht, dass*
Begründung, dass/w-/ob	*Auskunft, w-/ob*	*Bestätigung, dass*
Beweis, dass/w-/ob	*Auswahl, w-/ob*	*Bedingung, dass*
Darstellung, dass/w-	*Befragung, w-/ob*	*Befürchtung, dass*
Entscheidung, dass/w-/ob	*Nachfrage, w-/ob*	*Chance, dass*
Erkenntnis, dass/ob	*Prüfung, w-/ob*	*Drohung, dass*
Erklärung, dass/w-/ob	*Rätsel, w-/ob*	*Erfahrung, dass*
Feststellung, dass/w-/ob	*Überblick, w-/ob*	*Gefahr, dass*

Table 5.37: Sample nominal predicates classified into N1, N2, N3

5.3.3.3 Subclassification of compounds

In section 5.3.2.4 we describe the procedure to distinguish between simplex and compound nominal predicates. Compound nominal predicates can be subclassified into the three groups (N1, N2, N3) mentioned above in the same way as simplex nominals. In this case the classification procedures are the same.

However, compound nouns can also be subclassified according to further features, e.g. to the relations between the subcategorisation properties of compounds and those of their constituent parts. In this case we distinguish three classes of nominal compounds (C1, C2 and C3), which are described and illustrated in table 4.10 in section 4.2.3 above. The C1-compounds share their subcategorisation features with the head of the compound, the C2-compounds share their subcategorisation features with the non-head. The C3-compounds either share their subcategorisation properties with both the head and the non-head C3-1 or share their subcategorisation properties with neither the head nor the non-head constituents C3-2.

To classify the extracted candidates according to the three classes (from C1 to C3), we apply the VF query, which is lexically specified for compound extraction. To insert lexical constraints into the classification query, we use a list of simplex nouns for which we "know" that they subcategorise for a subclause. For this purpose we use the list of subclause-taking nominal predicates, described in sections 5.3.1.2, 5.3.1.4 and 5.3.2.1 above.

To classify compound nominal predicates, we need to undertake modifications of this list. For the extraction of the C1-compounds, the nouns from the "known" list are rewritten into the form (for the CQP query, see appendix B) .+beweis|.+beispiel etc., which means that the first part of the compound, thus its nonhead, allows for subclauses. The modified list is defined as the variable $headnounlist that can now be integrated into the query for compound classification (further examples of the modified list are shown in table 5.12).

To extract the C2-compounds, the subclause-taking nouns are converted into the form Beweis.+|Beispiel.+ (further examples are shown in table 5.13), which means that the first part of the compound, thus its nonhead, allows for subclauses. The list is defined as $nonheadnounlist in the query.

```
".+beweis|.+beispiel|.+chance|.+diskussion|.+einigung|
.+erfahrung|.+fall|.+frage|.+folgerung|.+gefahr|.+hiweis|
.+hoffnung|.+idee|.+information|.+kalkül|.+kritik|.+lösung|
.+lüge|.+mahnung|.+meinung|.+nachricht|.+nachweis|
.+prinzip|.+problem|.+rätsel|.+satz|.+schluß|.+tatsache|
.+trend|.+ursache|.+vermutung|.+warnung|.+zusicherung|
etc."
```

Figure 5.12: $headlist: subclause-taking nouns as the head

```
"Beweis.+|Beispiel.+|Chance.+|Diskussion.+|Einigung.+|
Erfahrung.+|Fall.+|Frage.+|Folgerung.+|Gefahr.+|Hiweis.+|
Hoffnung.+|Idee.+|Information.+|Kalkül.+|Kritik.+|Lösung.+|
Lüge.+|Mahnung.+|Meinung.+|Nachricht.+|Nachweis.+
|Prinzip.+|Problem.+|Rätsel.+|Satz.+|Schluß.+|Tatsache.+|
Trend.+|Ursache.+|Vermutung.+|Warnung.+|Zusicherung|
etc."
```

Figure 5.13: $nonheadlist: subclause-taking nouns as the non-head

We include the above described constraints into the general query for nominal predicate extraction (specified in tables 5.20 and 5.21 in section 5.3.1.3). In table 5.38, we show the query to subclassify compounds of type C1. For this purpose we extend lines 4d, 4g and 4l of the query in table 5.21 and include the variable $headnounlist, which restricts the match to the compounds whose head constituent subcategorises for subclauses.

The queries to extract and classify C2 and C3 types are built up in the same way. For the extraction of C2-compounds, we add a lexical constraint containing the variable $nonheadlist described above. For the extraction of the C3-compounds, we add a constraint, which contains both variables. We assume that compounds whose both parts (head and non-head) are in the list of subclause-taking nouns, belong to the type C3-1, weheras those compounds whose head and non-head constituents are not in the list, belong to the type C3-2. The summary of lexical constraints for different compound types is given in table 5.39.

With the help of the above described constraints, compound nominal predicates are classified as follows:

	Query building blocks	**comments**	**matching words**
4.			
	option 1		
4a.	(<np>	the start of the nom.phrase	
4b.	([pos = "ART|PIAT|PDAT|PPOSAT"]?	optional determiner	*das*
4c.	[pos!= "NN|V.FIN"]{0,4}	up to 4 optional elements, no nouns and finite verbs	
4d.	@[pos = "NN"]&[**lemma=RE($headlist)**])|	noun	*Parade**beispiel***
	option 2		
4e.	([pos = "ART|PIAT|PDAT|PPOSAT"]?	optional determiner	*das*
4f.	[pos!= "NN|V.FIN"]{0,4}	up to 4 optional elements, no nouns and finite verbs	
4g.	@[pos = "NN"]&[**lemma=RE($headlist)**])	noun	*Parade**beispiel***
4h.	([(pos= "ART|ADJA|CARD|PDAT| PIAT|PPOSAT")&(word=".*er|.*es")]?	article or adj in gen.	*des*
4i.	([pos = "NN|NE"]& agr contains ".*Gen.*"))|	noun in genitive	*Präsidenten*
	option 3		
4j.	([pos = "ART|PIAT|PDAT|PPOSAT"]?	optional determiner	*das*
4k.	[pos!= "NN|V.FIN"]{0,4}	up to 4 optional elements, no nouns and finite verbs	
4l.	@[pos = "NN"]&[**lemma=RE($headlist)**])	noun	*Parade**beispiel***
4m.	([pos= "APPR.*"]&	preposition or preposition & an article	*vom*
	[pos= "ART"]*&	an optional article	
	[pos= "NN|NE"]&	a noun	*Präsidenten*
4n.	</np>)	the end of the nom.phrase	

Table 5.38: The classification query to extract the C1 nominal compounds

	Query building blocks	**comments**	**matching words**
C1	([pos = "NN"]&**[lemma=RE($headlist)]**)	compounds whose head is subclause-taking	*Parade**beispiel**, Journalisten**frage***
C2	([pos = "NN"]&**[lemma=RE($nonheadlist)]**)	compounds whose non-head is subclause-taking	***Beweis**mittel, **Erfahrung**swert*
C3-1	([pos = "NN"]&**[lemma=RE($headlist)& lemma=RE($nonheadlist)]**)	compounds whose both constitutents are subclause-taking	***Meinungsstreit**, **Schlußfolgerung***
C3-2	([pos = "NN"]&**[lemma!=RE($headlist)& lemma!=RE($nonheadlist)]**)	compounds whose both constitutents are not subclause-taking	*Ehrgeiz, Schlagzeile, Sehnsucht*

Table 5.39: Query segments with lexical constraints for compound subclassification

- if the head of a compound is in the list of subclause-taking nouns and the non-head is not, the compound belongs to the type C1:
  *Parade**beispiel**, Journalisten**frage***;
- if the non-head of a compound is in the list of subclause-taking nouns, and the head is not, the compound belongs to the type C2:
  ***Beweis**mittel, **Erklärung**sversuch*;
- if both the head and the non-head are in the list of subclause-taking nouns, the compound belongs to the type C3-1: ***Meinungsstreit, Schlußfolgerung***.
- if neither the head nor the non-head are in the list of subclause-taking nouns, the compound belongs to the type C3-2: *Ehrgeiz, Wortspiel*.

We illustrate the subclassification procedures for compounds in figure 5.14.

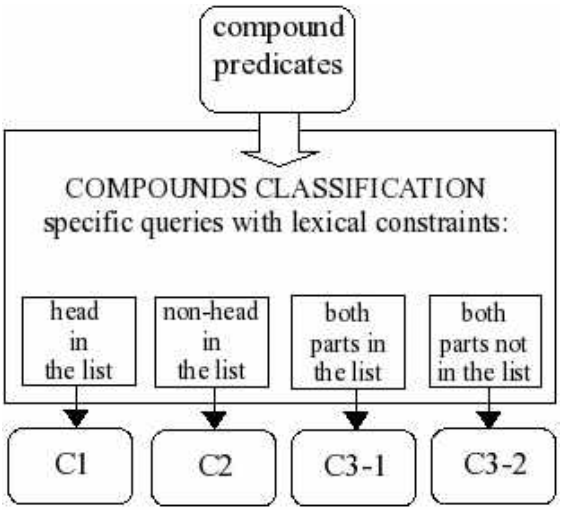


Figure 5.14: Extraction and subclassification of compounds

5.3.3.4 Subclassification of multiwords

To automatically classify the multiword candidates according to M1 to M4 (cf. section 4.2.4), we compare the results of the queries in tables 5.30 and 5.31 (Context 1) with data obtained from another context, namely nominal predicates in the VF, illustrated in table 5.20 (Context 2).

Consequently, cases fulfilling both context tests belong to types M1 of our classification. This means that the nominal component of the multiword extracted in Context 1, appears in Context 2 with the same complement clause. Cases where *dass* and *ob/w*-clauses are switched between the two Contexts (e.g. the nominal element of the multiword extracted in Context 2 subcategorises for a *dass*-clause only under certain contextual parameters, whereas the multiword itself always occurs in Context 1 with *dass*-clauses) belong to type M2. Cases with different subcategorisation in Context 1 and 2 belong to types M3 and M4. The M3 multiwords subcategorise for

a sentential complement, even though neither their nominal nor their verbal component do so[18]. They are semantically trasparent and do not qualify as idioms. The M4 multiword expressions, on the other hand, are either non-compoisitional (idiomatic) or contain "cranberry" lexemes, e.g. *Abrede, Betracht*, lexical items that appear only within an idiom.

A completely automatic distinction between these two types is difficult and not intended here (cf. e.g. (Fazly/Stevenson 2006) for work on the automatic separation of idioms and (rather compositional) collocations). To identify the M3 multiwords, we can apply a lexicon of nouns, which occur freely in corpora but do not subcategorise for sentential complements. To single out some of the M4-cases, we can obviously identify "cranberry" lexical items (see works on "cranberry" lexemes, e.g. (Richter/Sailer 2002) and (Trawiński *et al.* 2008))[19], the rest of them can be identified as such. For the purpose of further discussion, we group the cases M1 and M2 together (cases with "inheritance" of subcategorisation from the nominal part), as well as M3 and M4 ("non-inheritance" cases).

	Context 1 (MWE)		**Context 2 (noun)**	
	dass	*w-/ob*	*dass*	*w-/ob*
M1 *in* **Aussicht** *stellen*	+	-	+	-
	-	+	-	+
	+	+	+	+
M2 *in* **Erfahrung** *bringen*	+	+	+	-
	+	+	-	+
	+	-	+	-
	-	+	-	+
M3 *zu* **Protokoll** *geben*	+	+	-	-
	+	-	-	-
	-	+	-	-
M4 *in* **Abrede** *stellen*	+	+	-	-
	+	-	-	-
	-	+	-	-

Table 5.40: Classification of multiword expressions

Examples of the four types of multiwords are given in table 5.40. The multiword expression *in Aussicht stellen*, extracted in Context 1, subcategorises for *dass*-clauses only. So does its nominal component *Aussicht* when extracted in Context 2. This multiword expression belongs to type M1. The expression *in Erfahrung bringen* can take both declarative (*dass*-clauses) and interrogative complements (*w-/ob*-clauses) in Context 2, whereas its nominal component takes declaratives only. Therefore, this multiword expression is of type M2. The multiword expression *in Abrede stellen* belongs to type M3 because its nominal component *Abrede* is a "cranberry" lexeme and does not appear outside a multiword, thus, cannot take any complements.

[18]As we study multiword expressions, which consist of a prepositional phrase and a support verbs, we assume that the verbal constituent does not subcategorise for the sentential complement of the construction.

[19]A list of "cranberry" words for German is available at http://multiword.sourceforge.net

In table 5.41, we give further examples of multiwords, which were extracted and classified with the procedure described above.

type	DE
M1	*ins/zu(m) Bewußtsein kommen, dass/w-* "to recollect that/wh-" *zur Bedingung machen, dass* "to make it a condition that" *ins Grübeln kommen/geraten, w-/ob* "to brood wh-/if"
M2	*zur Rede stellen, w-/ob* "to tackle about wh-/if" *in Erfahrung bringen, w-* "to find out wh-"
M3	*zu Ansatzpunkten gelangen, ob* "to get to starting points if" *zur Kenntnis geben/gelangen/nehmen, dass* "to make aware/to notice that" *zu Protokoll geben, dass/ob* "to give on record that/if"
M4	*in Abrede stellen, dass* "to deny that" *zum Vorschein kommen, dass/w-* "to appear that" *sich/jmdm in den Kopf setzen, dass* "to set into one/sb's head that" *darüber in die Haare geraten, w-* "to be at loggerheads wh-"

Table 5.41: Examples of classified multiword expressions

5.3.3.5 Summary: subclassification of predicates

In figure 5.15, we summarise the procedures to subclassify the predicates obtained with the general queries described in 5.3.1 and specified with queries described in 5.3.2. Verbal, nominal and mulwtiword predicates are subdivided into further groups according to their subcategorisation properties. We distinguish three classes for verbal and nominal predicates and four classes for compound nominal predicates and multiwords.

5.3.4 Classification of Subcategorisation Relations

As described in 4.2.5, we classify subcategorisation relations between morphologically related predicates: verbs and their nominalisations, into the three classes (R1, R2 and R3), described in section 4.2.5.

To classify predicate relations according to R1, R2 and R3, we need to identify nominalisations and their base verbs (for this purpose we use morphological

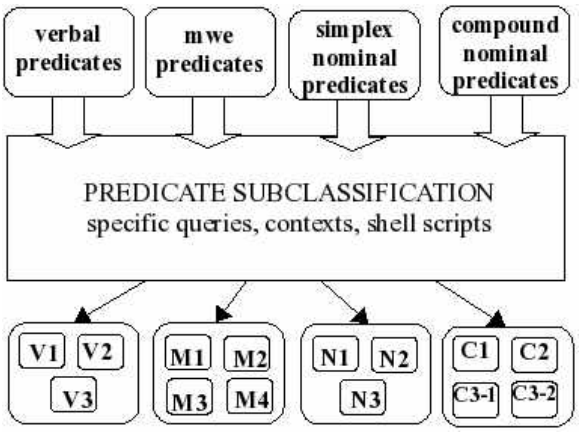


Figure 5.15: Subclassification of predicates in the extraction and classification architecture

tools), extract them along with their subcategorisation features from corpora (to do it, we use specific CQP queries) and classify them accordingly (with the help of shell scripts). In this study we limit nominalisations to those containing the suffix *-ung*.

5.3.4.1 Identification and classification of *ung*-nominalisations

The procedures to identify *-ung*-nominalisations are described in section 5.3.2.3 above. We use morphological tools to sort them out of the list of nominal predicates obtained from the Vorfeld. As a result, we get a list of nominalisations formed with the suffix *-ung*, which subcategorise for *dass, w-* and *ob*-clauses.

To obtain the subcategorisation properties for *ung*-nominalisation from corpora, we apply the VF query specified in tables 5.20 and 5.21. To restrict the search for the *ung*-nominalisations, we include the lexical constraint in form of a variable. The variable $unglist (lines 1d, 2c and 3c), which refers to the list of *ung*-nominalisations obtained with the morphological tools, lexically limits the match of the query.

The identified and extracted nominalisations are classified according to the type of sentential complement they allow for. The classification procedures do not differ from those applied for the subclassification of nominal predicates in general, which are described in section 5.3.3.2 above. Thus, *-ung*-nominalisations are classified into three groups: Nung1, Nung2, Nung3, cf. table 5.43. A list of sample nominalisations grouped according to the three classes is given in table 5.44. Further examples of the three classes of nominalisations are given in section A.6 of the appendix.

5.3.4.2 Identification and classification of base verbs

With the help of morpholgical tools, we get a list of base verbs underlying the extracted and classified *-ung*-nominalisations, cf. table 5.8. The list of the resulting base verbs is given in table 5.9.

The generated list of base verbs is integrated into the specific queries for extraction of verbal predicates in VL and passive, described in tables 5.28 and 5.29. We

	Query building blocks	**comments**	**matching words**
option 1			
1a.	(<np>	the start of the nom.phrase	
1b.	([pos = "ART|PIAT|PDAT|PPOSAT"]?	optional determiner	*die*
1c.	[pos!= "NN|V.FIN"]{0,4}	up to 4 optional elements, no nouns and finite verbs	
1d.	@[pos = "NN"]&[**lemma=RE($unglist)**])|	noun	***Entscheidung***
option 2			
2a.	([pos = "ART|PIAT|PDAT|PPOSAT"]?	optional determiner	*die*
2b.	[pos!= "NN|V.FIN"]{0,4}	up to 4 optional elements, no nouns and finite verbs	
2c.	@[pos = "NN"]&[**lemma=RE($unglist)**])	noun	***Entscheidung***
2d.	([(pos= "ART|ADJA|CARD|PDAT| PIAT|PPOSAT")&(word=".*er|.*es")]?	article or adj in gen.	*des*
2e.	([pos = "NN|NE"]& agr contains ".*Gen.*")))|	noun in genitive	*Präsidenten*
option 3			
3a.	([pos = "ART|PIAT|PDAT|PPOSAT"]?	optional determiner	*die*
3b.	[pos!= "NN|V.FIN"]{0,4}	up to 4 optional elements, no nouns and finite verbs	
3c.	@[pos = "NN"]&[**lemma=RE($unglist)**])	noun	***Entscheidung***
3d.	([pos= "APPR.*"]&	preposition or preposition & an article	*vom*
	[pos= "ART"]*&	an optional article	
	[pos= "NN|NE"]))	a noun	*Präsidenten*
3e.	</np>)	the end of the nom.phrase	

Table 5.42: Lexical constrains to obtain *ung*-nominalisations from corpora

predicate class	**declaratives:** *dass*	**interrogatives:** *w-, ob*
Nung1	+	+
Nung2	-	+
Nung3	+	-

Table 5.43: Subcategorisation features of the three classes of *-ung*-nominalisations

Nung1	Nung2	Nung3
Begründung, dass/w-/ob	*Abklärung, w-/ob*	*Ankündigung, dass*
Darstellung, dass/w-/ob	*Auseinandersetzung, w-/ob*	*Befürchtung, dass*
Entscheidung, dass/w-/ob	*Befragung, w-/ob*	*Erfahrung, dass*
Erklärung, dass/w-/ob	*Klärung, w-/ob*	*Hoffnung, dass*
Feststellung, dass/w-/ob	*Nachforschung, w-/ob*	*Meldung, dass*
Meinung, dass/w-/ob	*Prüfung, w-/ob*	*Überzeugung, dass*
Überlegung, dass/w-	*Überprüfung, w-/ob*	*Voraussetzung, dass*

Table 5.44: Sample nominal predicates classified into N1, N2, N3

nominalisations		**morphological analysis**
Ankündigung	→	ankündigen<V>ung<SUFF><+NN><Fem><Nom><Sg>
Bedingung	→	bedingen<V>ung<SUFF><+NN><Fem><Nom><Sg>
Befürchtung	→	befürchten<V>ung<SUFF><+NN><Fem><Nom><Sg>
Erwartung	→	erwarten<V>ung<SUFF><+NN><Fem><Nom><Sg>
Entscheidung	→	entscheiden<V>ung<SUFF><+NN><Fem><Nom><Sg>
Erklärung	→	erklären<V>ung<SUFF><+NN><Fem><Nom><Sg>
Darstellung	→	darstellen<V>ung<SUFF><+NN><Fem><Nom><Sg>
Vermutung	→	vermuten<V>ung<SUFF><+NN><Fem><Nom><Sg>
Vorstellung	→	vorstellen<V>ung<SUFF><+NN><Fem><Nom><Sg>

Table 5.45: Nominalisations after the morphological analysis

nominalisations vs. base verbs		**EN translation**	
Ankündigung	→*ankündigen*	"announcement"	→ "to announce"
Bedingung	→*bedingen*	"condition"	→"to condition"
Befürchtung	→*befürchten*	"fear"	→"to fear"
Erwartung	→*erwarten*	"expectation"	→"to expect"
Entscheidung	→*entscheiden*	"decision"	→"to decide"
Erklärung	→*erklären*	"explanation"	→"to explain"
Darstellung	→*darstellen*	"presentation"	→"to present"
Vermutung	→*vermuten*	"assumption"	→"to assume"
Vorstellung	→*vorstellen*	"idea"	→"to think"

Table 5.46: Nominalisation-verb pairs after the morphological analysis

insert lexical constraints (the generated list of base verbs defined as the variable $baseverbs) into the block for verbal predicates, as shown in lines 3b and 7a in table 5.47.

	Query building blocks	**comments**	**matching words**
option 1: VL context			
1.	[pos="KOU.*|PREL.*|PW.*"]	the start of the VL clause conjunction, relat. or interrogat. pronoun	 *weil*
2a. 2b.	([pos!="V.*FIN"]&[word!=",|-"]& [lemma!=RE($imslex_nounlist)&pos= "NN"])*	optional words, no finite verbs, no punctuation no nouns from the list	*keine dort*
3a. 3b. 3c. 3d. 3e.	<vc> [**lemma=RE($baseverbs)**& pos="VV.*"] [pos="V(A|M) (INF|IZU|PP)"]{0,2} [pos="V(A|M)FIN"]? </vc>	start of the verb.complex verb from the list infinitive or participle of auxiliary or modal verb optional final auxiliary or modal verb (if the full verb is infinite) end of the verb.compl.	 ***entscheiden*** *konnte*
option 2: passive context			
4.	[pos= "V(M|A)FIN"]	finite aux. or modal verb	*konnte*
5a. 6b.	([pos!="V.*FIN"]&[word!=",|-"]& [lemma!=RE($imslex_nounlist)&pos= "NN"])*	optional words, no finite verbs, no punctuation no nouns from the list	*dort*
7a. 7b. 7c.	[pos= "VV(PP|IZU)"& **lemma=RE($baseverbs)**]] [pos= "VA(PP|INF)"]? [pos= "V(A|M)INF"]?	full verb participle or *zu*-infinitive	***entschieden*** *werden*

Table 5.47: Lexical constraints to extract base verbs in VL and passive

The system searches for base verbs subcategorising for *dass, ob* and *w*-clauses. The list of extracted verbs is used for the analysis of subcategorisation relations between base verbs their nominalisations. For this purpose base verbs are also classified into the three groups mentioned above (Vbase1, Vbase2, Vbase2). The classification procedures do not differ from those applied for the subclassification of verbal predicates in general, which are described in section 5.3.3.1 above, cf. table 5.48 below.

5.3.4.3 Classification of relations

We analyse the relations between the subcategorisation properties of the extracted and classified base verbs and those of their nominalisations and classify them, according to the above mentioned classes (R1, R2 and R3), cf. tables 5.49 and 5.50. Table 5.49 describes to which class of relations verbal and nominal classes can be-

predicates class	**declaratives:** *dass*	**interrogatives:** *w-, ob*
Vbase1	+	+
Vbase2	-	+
Vbase3	+	-

Table 5.48: Subcategorisation features of the three classes of base verbs

long. The classification criterion is the absence or the presence of the complement types.

predicates	*dass*	*w-/ob*	**relations**
(Vbase1), (Nung1)	+	+	→ R1
(Vbase2), (Nung2)	-	+	→ R1, R2, R3
(Vbase3), (Nung3)	+	-	→ R1, R2, R3

Table 5.49: Relations between verbs and their nominalisations – part 1

	Nung1	**Nung2**	**Nung3**
Vbase1	Vbase1Nung1	Vbase1Nung2	Vbase1Nung3
Vbase2	Vbase2Nung1	Vbase2Nung2	Vbase2Nung3
Vbase3	Vbase3Nung1	Vbase3Nung2	Vbase3Nung3

Table 5.50: Relations between verbs and their nominalisatons – part 2

We group verb-nominalisation pairs (cf. table 5.50) acccording to the shared or non-shared features and get the following relations:

Vbase1Nung1 nominalisation and its underlying verb subcategorise only for a *dass*-clause.

Vbase2Nung1 the base verb has all three (or two) complement types, but the nominalisation has only a *dass*-clause (the loss of *ob, w*-clauses).

Vbase3Nung1 the base verb has no *dass*-clause, but its nominalisation has a subcategorised *dass*-clause.

Vbase1Nung2 the base verb has only a *dass*-clause (found in corpora), but its nominalisation has all three (or two) complement types.

Vbase2Nung2 the base verb has all three (or two) complement types, so does its nominalisation (V1N1 and V2N2 – similar relations).

Vbase3Nung2 the base verb has no *dass*-clause, but its nominalisation has all three (or two) complement types.

Vbase1Nung3 the base verb has only a *dass*-clause, but its nominalisation doesn't have any *dass*-clause.

Vbase2Nung3 the base verb has all three (or two) complement types (including the *dass*-clause), but the nominalisation has no *dass*-clause.

Vbase3Nung3 the base verb does not have a *dass*-clause, neither does its nominalisation (V1N1 and V3N2 – similar relations).

We group the verb-nominalisation pairs into the three following classes:

R1 subcategorisation properties are inherited from the verb (Vbase1Nung1, Vbase2Nung2, Vbase3Nung3):
entscheiden, dass/ob/w- ("to decide that/if/wh-")
vs. *Entscheidung, dass/ob/w-* ("decision that/if/wh-")

R2 subcategorisation properties are inherited with the loss of clausal complements by the nominalisation:

- loss of *ob/w*-clauses (Vbase2Nung1):
  ankündigen, dass/w- ("to announce that/wh-")
  vs. *Ankündigung, dass* ("announcement that")
- loss of *dass*-clauses (Vbase2Nung3, Vbase1Nung3):
  ermitteln, dass/ob/w- ("to investigate that/if/wh-")
  vs. *Ermittlung (darüber), ob* ("investigation (about) if")

R3 subcategorisation properties are inherited from the verb, but the nominalisation has additional subcategorisation properties of its own (Vbase3Nung1, Vbase1Nung2, Vbase3Nung2).

The analysis of the preliminary tests shows that the existence of the relations of type R3 proves to be hypothetical. Therefore, we specify the R3 relations as a conceptual type.

In table 5.51, we outline the above described classes of relations and the corresponding verb-nominalisation pairs: R1 nominalisations have the same subcategorisation properties as their base verbs, in R2 they lose some properties and in R3 they show some additional properties, which are not specific for their base verbs.

relation	**subcategorisation features**	**predicate pairs**
R1	= (equal)	(Vbase1)(Nung1), (Vbase2)(Nung2), (Vbase3)(Nung3)
R2	- (lost by the nominalisation)	(Vbase1)(Nung2), (Vbase1)(Nung3)
R3	+ (additional)	(Vbase2)(Nung1),(Vbase2)(Nung3), (Vbase3)(Nung1), (Vbase3)(Nung2)

Table 5.51: Classification of subcategorisation relations

5.3.4.4 Summary: classification of subcategorisation relations

The above described automatic classification of subcategorisation relations between verbs and their nominalisations includes identification of nominalisations and exctraction of their subcategorisation properties from corpora, identification of baseverbs and extraction of their subcategorisation features from corpora, and finally the

comparison of their subcategorisation features and classification of their relations which is based on this comparison.

In figure 5.16 below, we illustrate the procedure to extract and classify the subcategorisation relations between verbs and their deverbals.

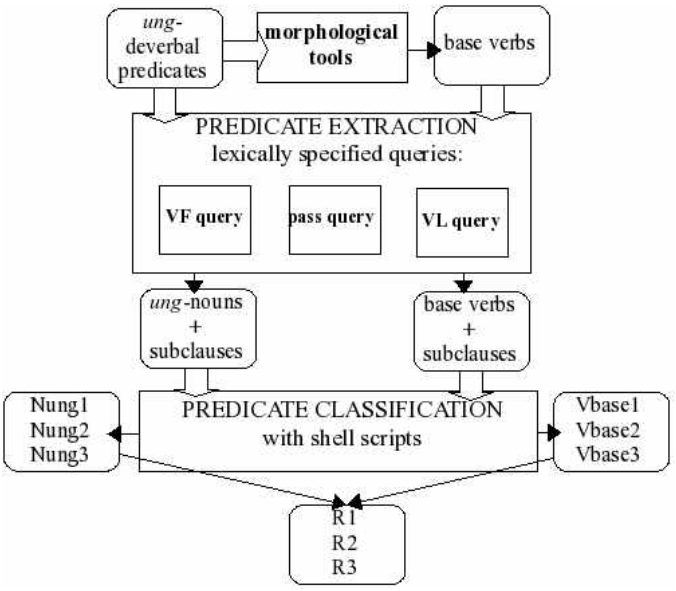


Figure 5.16: Architecture for extraction and classification of subcategorisation relations

5.3.5 Summary of the Procedures to Extract and Classify Predicates

In the the sections above we describe a set of cascaded procedures to extract and classify different types of predicates. As mentioned before, the steps procede from the general to the specific: we start from general queries to extract different predicates along with their subcategorisation information, go on with the specific queries to classify the predicates according to the types under analysis and finally subclassify them according to their subcategorisation features. Our classification is described in section 4 above. As we are also interested in the subcategorisation relations between verbs and their nominalisations, we also obtain the information needed to classify these relations according to the types described in section 5.3.4.3.

To identify and extract predicates, we use both CQP queries and morphological tools, e.g. to obtain compound nouns or *ung*-nominalisations and their base verbs. However, the subcategorisation information for these predicate types is also acquired with the CQP queries, which are lexically specified for this purpose. The classification procedures are elaborated with regular expressions in form of shell scripts.

In figure 5.17 we summarise the above described procedures.

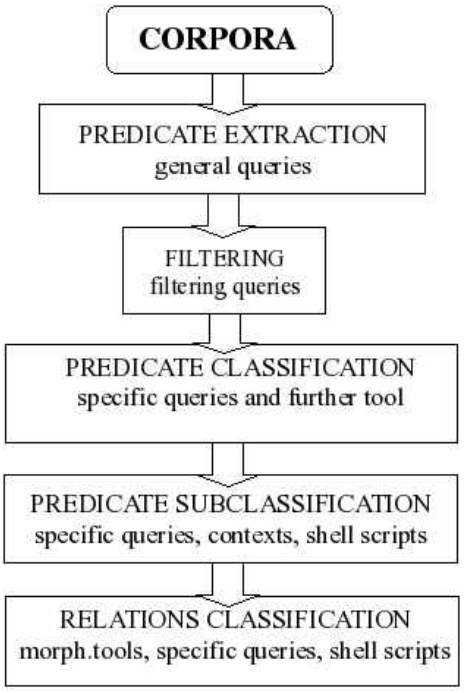


Figure 5.17: Extraction and classification architecture

Chapter 6

Extraction and Classification Results

In the following chapter we present the results and the evaluation of our extraction and classification procedures. In section 6.1 we describe sample extraction results summarising the occurrences of different predicate types and classes in the used corpora and give an interpretation of the obtained figures.

In section 6.2, we evaluate the architecture described in chapter 5 above. The architecture consists of several components that can be applied separately. We evaluate these components, estimating recall and precision obtained on the data.

We also describe the possible applications of the described extraction and classification system for some lexicographic and NLP problems, which are described in section 7.2.4.

6.1 Quantitative Results and their Interpretation

The following section includes sample extraction results for different types of predicates. We report not only on the number of extracted predicates, but also on the proportion of the classified types, as well as on some sample extraction results. Moreover, we interpret the quantitative results obtained for different predicate types.

We start with the analysis of the quantitative results for nominal predicates (both simplex and compound) in section 6.1.1 and go on with the description of the extraction and classification results for multiwords in section 6.1.2. As we are interested in the "inheritance" relations between morphologically related predicates; the quantitative results for verbs are described within the analysis of the extraction and classification figures for subcategorisation relations, cf. section 6.1.3.

6.1.1 Extraction and Classification of Nominal Predicates

In the following sections we present quantitative results for the extracted and classified subclause-taking nouns illustrating them with sample examples. In 6.1.1.1, we describe and interpret the results for nominal predicates (which are not specified according to their morphological structure). The analysis of the occurrence of nominal compounds, as well as calculation of the proportion of different compound nominal types, are described in section 6.1.1.2. The quantitative results for derived nouns are

analysed in section 6.1.3 below, which describes the classiffication of subcategorisation relations between verbs and their derivatives.

6.1.1.1 Nominal predicates in the Vorfeld

Nominal predicates such as *Ankündigung* ("announcement"), *Ansicht* ("view, idea"), *Beweis* ("proof"), *Erfahrung* ("experience"), *Vorstellung* ("idea, vision"), etc. are extracted in the Vorfeld along with their sentential complements, i.e. declarative clauses introduced by the conjunction *dass* or interrogative clauses introduced by a *w*-word or by the conjunction *ob*. We analyse the proportion of nominals taking declarative and interrogative clauses in the Vorfeld. The figures in table 6.1 reveal that subclause-taking nominal predicates show preferences for *dass*-clauses: about 65-67% of the nominal predicates obtained from corpora in VF (ca. 1563 million tokens)[1].

subclause	***dass***	***w-/ob***	**TOTAL**
tokens	40028	19219	59247
in %	67,56	32,44	100,00
context types	10232	5455	15687
in %	65,23	34,77	100,00

Table 6.1: Proportion of *dass* and *w-/ob*-clauses with simplex nouns in the VF

The preference for *dass*-clauses is also obvious in the proportion of the three classes of nominal predicates (N1, N2, N3), described in section 4.2.2.5. The figures in table 6.2 show that predicates, which allow for declarative clauses (type N1 allow for both, declarative and interrogative, e.g. *Angst* (“fear”), and type N3, e.g. *Beispiel* (“example”), allows for declarative clauses only), prove to be the most frequent in our corpora. They make about 75% of all cases extracted from corpora in VF. The number of extracted nouns of type N1 and N2 (nominals subcategorising for interrogative clauses only, e.g. *Motiv, w-/ob* (“reason”), *Zeitpunkt, w-* (“point of time”)) varies between types and tokens. For instance, types of N3 predicates (nouns subcategorising for dass-clauses only) are more frequent than types of N1 and N2 (49,17% vs. 26,24% and 24,59%). However, tokens of the nouns belonging to type N1 (which allow for both *dass* and *w-/ob*-clauses) appear to be more frequent than those of N2 and N3 (61,15% vs. 8,65% and 30,20%), thus, the N1 nominals have more occurrences in the analysed corpora. This means that our system finds more nominals of type N3 than of type N1, but at the same time some of the N1 nominals are generally very frequent in our corpora.

The extraction tests in the Vorfeld performed on corpora of ca. 1563 million tokens show that the predicate *Frage* is the most frequent in our data (10587 tokens). We list the 30 most frequent predicates (sorted according to their occurrence in tokens) in table 6.3. We illustrate the quantitative extraction results for these nominals

[1]The extraction numbers are given for both **tokens** and **types** of predicates. Under **tokens** we understand the number of extracted word forms, whereas **types** indicate the number of query matches for predicates, thus their **context types**. This is especially useful for the study of further context parameters of the extracted predicates.

class	N1	N2	N3	TOTAL
context types	4116	3858	7713	15687
in %	26,24	24,59	49,17	100,00
tokens	36228	5128	17891	59247
in %	61,15	8,65	30,20	100,00

Table 6.2: Proportion of nominal predicate types extracted from corpora

calculating their frequencies. The list does not include any N2 nominals as the number of their tokens is lower than that of the other two types.

Further examples of the predicates classified according to the three classes described in 4.2.2.5 are given in section A.2 in the appendix.

Summary The figures above show that our tools deliver a substantial number of subclause-taking nouns (over 15.000 types and over 59.000 tokens), which can be classified according to the type of the clause they subcategorise for. The obtained results allow us to compare the proportion of declarative and interrogative clauses occurring with nouns in VF, revealing that subclause-taking nouns show preferences for *dass*-clauses. The analysis of the three noun types classified by our tools confirms this tendency. The N1 and N3 nouns prove to be most frequent in our corpora. The N2 nominals, which do not allow for *w*-clauses, e.g. *Motiv, w-/ob*appear to be not so frequent in the analysed corpora.

6.1.1.2 Compounds

As described in section 5.3.2.4 above, we classify nominal predicates into simplex and compound ones with the help of morphological tools. In table 6.4, we analyse the proportion of simplex and compound predicates (lemma types), which occur in Vorfeld in our corpora[2]. The figures show that compound nominal predicates are quite rare (13 % of all nominal predicates in the Vorfeld)[3].

Identified compound nouns extracted along with their sentential complements, are subclassified according to their subcategorisation features, cf. the three classes described in section 4.2.3. In table 6.5 we summarise the frequency of C1 to C3-compounds that occur in the newspaper corpora mentioned above. For this purpose we analyse 628 most frequent compound types in the analysed copora. The figures show that C1 compounds, such as *Agenturbericht* (“agency report”), *EU-Richtlinie* (“EU policy”), *Kanzler-Mitteilung* (“chancellor communication/message”), in which the head constituent assigns the subcategorisation features for the whole compound, are the most frequent in the analysed corpora. They make about 67% of all compounds extracted in the Vorfeld. Compounds of types C2 and C3, however, make over 30% of all compound cases in the Vorfeld, which is an unexpectedly considerable amount. This means that not only the head of a compound can be the valency bearer.

[2]We use corpora of ca. 1563 million tokens.

[3]The tests show that nominal compounds comprise ca. 15% of all nouns in corpora (we automatically analysed 10000 random occurrences of common nouns extracted from our corpora).

predicates	**EN**	**type**	**context types**	**tokens**
Frage	question	N1	839	10587
Tatsache	fact	N1	152	6138
Fall	case	N1	101	2233
Grund	reason	N1	278	1826
Umstand	circomstance	N1	61	1429
Entscheidung	decision	N1	200	1226
Gefahr	danger	N3	32	1021
Hoffnung	hope	N3	127	1017
Chance	chance	N3	34	937
Gerücht	rumour	N1	67	919
Hinweis	advice/clue	N3	181	816
Befürchtung	fear	N3	112	699
Vorwurf	accusation/reproach	N3	131	629
Wahrscheinlichkeit	probability	N3	23	616
Argument	argument/reason	N1	134	554
Erkenntnis	insight/knowledge	N3	71	526
Verdacht	suspect	N1	52	504
Vermutung	guess	N3	79	489
Annahme	assumption	N1	54	475
Bericht	report	N3	106	441
Angst	fear	N1	70	456
Meldung	message	N3	86	436
Vorstellung	idea	N1	43	413
Gedanke	thought	N1	52	386
Einwand	demur/objection	N1	81	384
These	thesis	N1	54	277
Meingung	opinion	N1	81	238
Diskussion	discussion	N1	85	234
Ankündigung	announcement	N3	79	216
Begründung	justification/ground	N1	69	206

Table 6.3: Frequency of the top 30 nominal predicates extracted from corpora

nominal predicates	**simplex**	**compound**	**TOTAL**
types	13650	2037	15687
in %	87,01	12,99	100,00

Table 6.4: The proportion of simplex and compound predicates occurring in VF

In about 8% of all extracted cases the subcategorisation of a compound is determined by its non-head constituent, which is contradictory to the common assumption. For instance, in the predicates *Erklärungsversuch* ("explanation attempt"), *Bedenkzeit* ("consideration time"), *Druckmittel* ("pressure means"), the non-head components *Erklärung-* ("explanation"), *Bedenk(en)-* ("consideration") and *Druck* ("pressure") determine the subcategorisation properties of the compounds. In roughly 24% of the observed cases the subcategorisation of compounds is either determined by both constituents, as in *Denkmodell* ("thinking model"), *Meinungsstreit* ("opinion argument=controversy"), *Zeitplan* ("time plan=schedule"), or by none of them, as in *Ehrgeiz* ("ambition") or *Wortspiel* ("word play") which means that these compounds are idiomatic, in so far that their meaning is lexicalised. Further examples of type C2 and C3-compounds along with sample sentences are given in appendix A.3 below.

compound	**C1**	**C2**	**C3-1**	**C3-2**	**TOTAL**
types	423	53	131	21	628
in %	67,36	8,44	20,86	3,34	100,00

Table 6.5: Occurrence of C1 to C3 types in the Vorfeld

As the morphological tool provides us with information about the part of speech of compound elements, their nominal (NN) or deverbal (V) nature, we test the proportion of compound derivation types, see table 6.6. The data show that compounds whose parts have a deverbal nature, comprise about 46%, which is a considerable amount. We assume that the morphological structure of compounds can also influence their subcategorisation behaviour, in particular in the cases when the non-head is the valency bearer (C2 and partially C3).

type	**example**	**types**	**in %**
NN.NN	*Branchenregel*	341	54,29
NN.V	*Volksbefragung*	188	29,94
V.NN	*Bedenkzeit*	86	13,69
V.V	*Beweisführung*	13	2,07
TOTAL		**628**	**100,00**

Table 6.6: The morphological analysis of compounds: proportion in corpora

Therefore, we analyse the morphological nature of compounds of different types types (C1 to C3), cf. table 6.7.

derivation types	**subcategorisation types**			
	C1	**C2**	**C3-1**	**C3-2**
NN.NN	59,33%	50,94%	38,17%	61,91%
NN.V	28,61%	16,98%	38,93%	33,33%
V.NN	10,87%	20,75%	21,37%	4,76%
V.V	1,18%	11,32%	1,53%	0,00%
TOTAL	100,00%	100,00%	100,00%	100,00%

Table 6.7: The morphological analysis of compounds: subcategorisation types

As seen in table 6.7, over 40% of C1-compounds have a deverbal head and a considerable amount of C2-compounds (over 30%) have a deverbal non-head (V.NN or V.V), which is expected:

C1: *Schlüssel<NN>Frage<V> ('Schlüsselfrage')*
- ("key<NN>question<V> (key issue)")
Experte<NN>Streit<V> (Expertenstreit)
- ("expert<NN>dispute<V> (dispute of experts)")

C2: *beweisen<V>Pflicht<NN> ('Beweispflicht')*
- ("prove<V>duty<NN> (burden of proof)")
Argumentations<V>Kette<NN> ('Argumentationskette')
- ("argumentation<V>chain<NN> (chain of reasoning)")

Besides that, both constituents of most C3-2-compounds have the nominal nature - NN.NN (about 62%), which expected, as most of them are lexicalised, as was mentioned above. Compounds of type C3-1 prove to have both nominal head and non-head, approximately 38% for the NN.NN structure, or have one deverbal constituent, about 39% for the NN.V form and roughly 21% for V.NN:

C3-1: *Ursache<NN>Potential<NN> ('Ursachenpotential')*
- ("reason<NN>capability<NN> (the potential of reasons)")
Wunsch<NN>Traum<NN> ('Wunschtraum')
- ("wish<NN>dream<NN> (great dream)")

C3-2: *Angel<NN>Punkt<NN> ('Angelpunkt')*
- ("tang<NN>point<NN>(pivot)")
Gesicht<NN>Punkt<NN> ('Gesichtspunkt')
- ("face<NN>point<NN>(viewpoint)")

The fact that a great number of non-heads of compounds of types C2 and C3-1 have deverbal nature is not unexpected. Nominalisations which occur as non-head of coumpounds often preserve the subcategorisation properties of their base verbs and thus, influence the unusual subcategorisation behaviour of compound predicates.

The analysis of sample extraction results shows that some predicative constituents of C2-compounds occur more frequently than others. In table 6.8, we display examples of compounds with the frequent deverbal non-head *Beweis* ("proof"). In table 6.9, we show examples of compounds with the frequent non-deverbal head *Weise* ("way").

Interpreting the figures The figures show that about 80% of the C2-compounds contain deverbal nouns as non-heads, e.g. ***Argumentation****skette* ("chain of reasoning"), ***Beweis****last* ("burden of proof"), or ***Erklärung****sversuch* ("expanation"). The choice for the subcategorised sublcause is in this case determined by the deverbal non-head constituent. We assume that there are correspondences between the subcategorisation properties of the deverbal valency-bearers of nominal compounds and those of the verbs, which underlie these deverbal constituents, cf. table 6.10. However, deverbal compound constituents taking over the subcategorisation properties of

non-head	head	tokens	in %
Beweis-	*aufnahme*	2	10,00
Beweis-	*führung*	3	15,00
Beweis-	*lage*	2	10,00
Beweis-	*last*	2	10,00
Beweis-	*mittel*	2	10,00
Beweis-	further heads, e.g. *pflicht*	9	45,00
TOTAL all with the non-head *Beweis-*		20	100%

Table 6.8: Frequent deverbal non-heads of C2-compounds

non-head	head	tokens	in %
Betrachtungs-	*weise*	3	21,00
Denk-	*weise*	3	21,00
Sicht-	*weise*	5	30,00
Verfahrens-	*weise*	1	5,00
Vorgangs-	*weise*	2	10,00
TOTAL with the head *-weise*		14	100%

Table 6.9: Frequent non-deverbal non-heads of C2-compounds

the underlying verbs are also common for other compound types, in which they can be either head (C1 and C3-1) or non-head C3-1 constituents, cf. table 6.11.

We suppose that the unexpected subcategorisation behaviour of C2 and C3-1-compounds can be explained by the deverbal nature of their valency bearer. In the majority of such cases, the non-head constituent which determines the subcategorisation of the whole compound is derived from a verb. As most nominalisations inherit subcategorisation properties of their base verbs, this behaviour proves to be expected.

At the same time, about 38% of C3-1-compounds do not contain any deverbal constituents, e.g. *Kritikpunkt* ("criticism point (point of criticism)", cf. table 6.7, and 29% of C2-compounds contain deverbal heads, e.g. *Ziel**setzung*** ("goal setting (aim)") or *Gesetz**entwurf*** ("law draft (draft law)") which is unexpected[4]. The ex-

[4]Verbs which underly the deverbal heads of these compounds, do not take any subclauses either.

compound	deverbal noun	underlying verb
***Argumentation**skette, dass* "reasoning chain that"	*Argumentation, dass* "reasoning that"	*argumentieren, dass* "to reason that"
***Beweis**last, dass* "burden of proof that"	*Beweis, dass* "proof that"	*beweisen, dass* "to prove that"
***Erfahrung**ssatz, dass* "empirical judgement that"	*Erfahrung, dass* "experience"	*erfahren, dass* "to experience that"

Table 6.10: C2-compounds containing deverbal non-heads vs. base verbs

compound	deverbal noun	underlying verb
C1:		
*DDR-**Erfahrung**, dass* "GDR experience that" *Architekten**meinung**, dass* "architect's opinion that" *Koalitions**vereinbarung**, dass* "coalition agreement that"	*Erfahrung, dass* "experience that" *Meinung, dass* "opinion that" *Vereinbarung, dass* "agreement that"	*erfahren, dass* "to experience that" *meinen, dass* "to think/mean that" *vereinbaren, dass* "to agree that"
C3-1:		
***Schlussfolgerung**, dass* "conclusion that" ***Streitfrage**, ob* "argue question if..."	*Schluss, dass* *Folgerung, dass* "conclusion that" *Streit, ob* "argument if" *Frage, ob* "question if"	*schliessen, dass* *folgern, dass* "to conclude that" *streiten, ob* "to argue if" *fragen, ob* "to ask if"

Table 6.11: C1 and C3-compounds containing deverbal non-heads vs. base verbs

planation for this phenomenon can be the derivation nature of these predicates. Compounds that have the morphological structure NN.V and belong to C2, are often derived from multiword constructions: *Zielsetzung - Ziel setzen* ("goal setting - to set a goal"), *Gesetz entwurf - Gesetz entwerfen* ("law draft - to draft law"), etc. In this case the verb has a function of a support verb, which means that the subcategorisation of the whole construction (both, in the multiword and the compound) is determined not by the verbal but by the nominal element.

Summary Our results show that the general assumption that the head of a compound acts as its valency bearer has exceptions, cf. the C2 and C3-compound types. There are three types of nominal compound predicates in German based on the subcategorisation relationships between the constituents. The data obtained from corpora allow us to classify the extracted compounds into the three mentioned groups.

6.1.2 Extraction and Classification of Multiword Expressions

In this study we extract multiwords containing a preposition, a noun and a support verb, e.g. *in Erfahrung bringen* ("to find out"), with the procedures described in 5.3.2.2 applied on corpora of 1563 million tokens.

According to the extracted data[5], most multiwords show preferences for dass-clauses (about 83-91% of the extracted multiword predicates), cf. figures in table 6.12.

We assume that our tools fail to detect all the cases available in the analysed corpora (see the calculation of recall and precision in section 6.2 below). However, the procedures described in section 5.3.3.4 prove to classify successfully the detected

[5]We consider the multiword types whose frequency is greater than 1 in our data, cf. section 6.2.2.2

multiword predicates	***dass***	***w-/ob***	**TOTAL**
context types	2603	548	3151
in %	82,61	17,39	100,00
tokens	17805	1666	19471
in %	91,44	8,56	100,00

Table 6.12: The proportion of *dass*- vs. *w-/ob*-clauses with multiwords

multiwords, according to the classes described in section 4.2.4.4. As already mentioned in section 5.3.3.4, a complete automatic distinction between M3 and M4 is difficult and not intended in this thesis. To single out the M3 multiwords, we can include a lexicon of nominals, which do not take subordinate clauses and for the detection of some M4 cases, we can obviously identify "cranberry" lexical items.

Therefore, M3 and M4 cases as well as M1 and M2 multiwords are grouped together for the prupose of further discussion: M1 and M2 comprise cases with the "inheritance" of subcategorisation from the nominal part, whereas M3 and M4 cases are those with the "non-inheritance" of subcategorisation from the nominal part. In table 6.13, we demonstrate the occurrence of the two groups of multiwords in the analysed corpora. The figures show that multiwords of types M1 and M2 are more frequent (both in types and tokens) among the obtained data, which is expectable.

multiword	**M1+M2**	**M3+M4**	**TOTAL**
context types	1701	1452	3151
in %	53,98	46,02	
tokens	11687	7787	19474
in %	60,01	39,99	100,00

Table 6.13: The occurrence of M1+M2 vs. M3+M4 classes

6.1.2.1 Sample results

In table 6.14 we summarise some absolute frequency figures from the extraction excercises from the analysed corpora (a total of 1563 million tokens).

In table 6.14, we analyse the multiword expressions *in Aussicht stellen* ("to put into outlook = announce"), *zur Bedingung machen* ("to make it a condition"), *in Erfahrung bringen* ("to bring into experience = find out"), *zur Entscheidung gelangen/kommen/stellen* ("to come to a decision"), *in Rechnung stellen* ("to put into account = to bring to account"), as well as *zum Ausdruck kommen* ("to come to expression = to be expressed"), *in Abrede stellen* ("to deny") and *in Vergessenheit geraten* ("to fall into oblivion") and *zum Protokoll geben* ("to give to minutes=to put on record").

For the first five multiword expressions we have both, a sufficient number of true multiword cases (columns marked with +SV) and enough non-multiword uses of the nouns, which means that the noun occurs freely in context, i.e. in Vorfeld (columns

MWE types	nominal components of multiwords	*dass*		*w-/ob*	
		-SV	**+SV**	**-SV**	**+SV**
		VF	VL, pass	VF	VL, pass
		tokens	**tokens**	**tokens**	**tokens**
M1+M2	*in* ***Aussicht*** *stellen*	83	235	0	0
	zur ***Bedingung*** *machen*	262	176	2	2
	zur ***Entscheidung*** *gelangen/kommen*	110	54	1116	46
	in ***Rechnung*** *stellen*	28	257	3	12
	in ***Erfahrung*** *bringen*	206	417	0	144
M3+M4	*zum* ***Ausdruck*** *kommen/bringen*	0	1700	0	31
	zu ***Protokoll*** *geben/nehmen*	2	120	2	0
	in ***Abrede*** *stellen*	0	185	0	0
	in ***Vergessenheit*** *geraten*	0	123	0	0

Table 6.14: Multiwords vs. nouns, which occur freely in context

with -SV). With the last four multiword expressions, we have few or no Vorfeld (VF) occurrences. This suggests that they are limited to multiword (within a support word construction) or idiom uses, which means that they belong to types M3 or M4. In fact, we expect that "cranberry" nouns, such as *Abrede* in *in Abrede stellen* or *Vergessenheit* in *in Vergessenheit geraten*, do not occur outside multiwords at all.

The table gives the numbers of occurrences of both, declarative (introduced with *dass*) and interrogative (introduced with *w-/ob*) subclause types which occur within a multiword or freely in corpora. We extract nouns both, within and outside a multiword in different word order models, i.e. VF, VL and passive ('pass' in the table).

In section A.4 in the appendix, we show further examples of different multiword types extracted in VL from our corpora. We list M1 and M2 multiwords and types of sentential complements they subcategorise for (indicated as 'compl' in the table). Multiword expressions of type M3 and M4 are given along with their complements and the indicatation if they contain a "cranberry" lexeme (**c**) or if they are idiomatic (**i**).

6.1.2.2 Interpreting the figures

Judging from the samples in table 6.14, we can assume that our tools detected useful results for the interpretation of subcategorisation properties of multiword expressions. By comparing the columns for the occurrence of nominals inside and outside a multiword (+SV and -SV), we can determine multiword expressions of types M1 or M2 as those appearing significantly under both conditions. A separation into M1 vs. M2, i.e. an identification of the "switching" of truth values for complement clauses characteristic of M2, can be observed with the multiword *in Erfahrung bringen*. This multiword seems to accept *dass*-clauses the same way as the noun does outside the multiword (when used in VF), but the multiword-reading in addition occurs consistently with interrogative subclauses (indirect *w-/ob*-questions), which is not the case with the noun used outside the multiword (in VF).

More idiomatic multiwords (our type M3) are *zum Ausdruck bringen/kommen*

("to express"/"to be expressed") or *zu Protokoll geben/nehmen* ("to put on record")[6]. The nominal components of these multiwords subcategorise for subclauses outside a multiword in a few cases only, except for the cases, in which they have a separate lexicalised meaning.

The nouns *Abrede* and *Vergessenheit* seem to take sentential complements only when used in multiwords, which is expected. The only reading of *Abrede* outside a multiword is that of "oral agreement", which is found in 22% of the occurrences of the lemma, but always without a sentential complement. With *Vergessenheit*, we have, besides the targeted idiom, also *der Vergessenheit anheimfallen* ("fall into oblivion") and *der Vergessenheit entreißen* ("avoid that sth falls into oblivion") which are idiomatic, as well as the expression *nach langer Vergessenheit* ("after a long period of oblivion"), a total of 11% of the data.

Moreover, the table shows what kinds of subclauses the multiwords prefer. For instance, with *in Aussicht stellen, zur Bedingung machen, in Abrede stellen, in Vergessenheit geraten, zu Protokoll geben/nehmen*, *w-/ob*-clauses appear in few cases.

Summary These experiments show that certain multiword expressions have their own subcategorisation properties, which are not inherited from their nominal elements. With respect to subcategorisation, such multiwords behave like idioms, even though their semantics is not fully idiomatic: the syntactic behaviour is not fully parallel to the semantic distinctions that are known from phraseology. The comparison of the observed occurrences of nominals both, with and without support verbs, allows us to broadly classify the MWE candidates in terms of their preferences for *dass-* and *w-/ob*-clauses, and with respect to the "inheritance" hypothesis. We think that the observed classes M1/M2 vs. M3/M4, i.e. with "inheritance" vs. without "inheritance", are stable even despite the theoretically problematic status of null occurrences.

6.1.3 Extraction and Classification of "Inheritance" Relations

In the following section we outline the results of extraction and classification procedures for the "inheritance" of subcategorisation, i.e. relations between verbs and their nominalisations.

For this purpose we analyse the occurrence of *ung*-nominalisations and their classification according to the types described in 4.2.5, as well as the retrospective extraction and classification results for their underlying verbs. Furthermore, we additionally analyse subcategorisation properties of *ung*-nominalisations used within a multiword expression. Besides that, we compare the subcategroisation properties inherited from verbs between nominalisations which occur freely in corpora or those which are embedded into a multiword.

6.1.3.1 Sample results and their interpretation

***Ung*-nominalisations and their classification** For the analysis and classification of relations between verbs and their nominalisations, we need to identify deverbal nouns in the list of nominal predicates extracted in the VF, cf. section 5.3.4.1.

[6]The multiword with the support verb *nehmen* ("to take") is less frequent (30% of our data) than the one with *geben* ("to give").

We concentrate on the analysis of *-ung*-nominalisations. They comprise about 22% of all nominal predicates extracted in Vorfeld[7], cf. table 6.15. The figures in table 6.16 show that *-ung*-nominalisations tend to take *dass*-subclauses in most cases. Furthermore, their preference for declarative complements is even greater, if compared to that of other nominal predicate types or nominal predicates general. *Ung*-nominalisations tend to subcategorise *dass*-clauses in approxiemately 77% of all cases extracted in Vorfeld, whereas other nominal predicates subcategorise for declaratives in roughly 61% of the obtained cases.

nominal predicates	**context types**	**in %**
nom.predicates excluding *ung*-deverbals:	12182	77,66
-ung-deverbals:	3505	22,34
all nominal predicates:	15687	100,00

Table 6.15: Proportion of *-ung*-nominalisations in the VF

nominal predicates		***dass***	***w-/ob***	**TOTAL**
nom.predicates excluding *ung*-deverbals:	context types	7528	4654	12182
	in %	61,20	38,80	100,00
-ung-deverbals:	context types	2704	801	3505
	in %	77,15	22,85	100,00
all nominal predicates:	context types	10232	5455	15687
	in %	65,23	34,77	100,00

Table 6.16: Nominal predicates and their preferences for subclause types

To analyse the proportion of nominal types (from Nung1 to Nung3), the classification of which is described in section 5.3.4.1, we test the 290 most frequent *-ung*-nominalisation lemma types[8]. The figures show that similarly to the subclause-taking nouns in general (cf. section 6.1.1.1), the occurrence of the Nung1 and Nung2 nominalisations varies between types and tokens. Types of Nung3 predicates (*ung*-deverbals, which subcategorise for *dass*-clauses only) are more frequent than the types of N1 and N2 (about 56% vs. about22%). However, if compared in tokens, nominalisations belonging to type Nung1 (allowing for both *dass* and *w-/ob*-clauses) appear to be roughly 9% more frequent than those of Nung3 (53,07% vs. 43,66%). The number of occurrences for *ung*-nominalisations of type Nung2 appears to be very low (they comprise about 3% of the analysed nominalisations). The N1 nominals have more occurrences in the analysed corpora.

Examples of *ung*-nominalisations classified according to Nung1 to Nung3 are given in table 5.44 in section 5.3.4.1. In section A.6 in the appendix, we illustrate the three classes of *ung*-nominalisations with further examples extracted by our tools.

[7]We automatically analysed 15670 context types extracted from a corpus of 1563 million tokens.

[8]All of them are not compound, i.e. we exclude compound nominalisations like *Beweisführung* ("proof conduction"), which were extracted in the VF from corpora of 1563 million tokens.

	Nung1	**Nung2**	**Nung3**	**TOTAL**
context types	65	64	161	290
in %	22,41	22,07	55,52	100,00
tokens	4317	266	3551	8134
in %	53,07	3,27	43,66	100,00

Table 6.17: Proportion of different types of *ung*-nominalisations

Base verbs and their classification The list of base verbs is generated from the list of *ung*-nominalisations, as described in section 5.3.4.2 above. We extract base verbs in VL and in passive[9] and obtain 8034 types and 22459 tokens. The proportion between *dass*- and *w/ob*-clauses which occur with verbs shows that verbs do not have the same preferences for *dass*-clauses as their *ung*-nominalisations. The data show that most extracted base verbs occur with *w-/ob*-clauses, as seen from table 6.18.

base verbs	***dass*-clauses**	***w-/ob*-clauses**	**TOTAL**
context types	773	7261	8034
in %	9,62	90,38	100,00
tokens	1892	20567	22459
in%	8,42	91,58	100,00

Table 6.18: Proportion of *dass*- vs. *w-/ob*-clauses with base verbs

We assume that this phenomenon can be explained by contextual parameters of verbal predicates that occur with *w-/ob*-clauses in our corpora (cf. section 2.2.2.3 above). Such parameters as modality (usage with a modal verb), polarity (usage in negative context), as well as their mood (occurrence in interrogative vs. declarative context) can influence the choice of the subclause. The mentioned context parameters may switch thruth values of verbs, thus, the decision to take *dass*-clauses (which express propositions) or *w-/ob*-clauses, which express questions[10] may depend on the context parameters mentioned above. To test this, we analyse the context parameters of verbal predicates extracted from a corpus of ca. 105M tokens and calculate their occurrences with modal verbs, such as *können* (“can”), *müssen* (“must”), *wollen* (“to want”), a semi-modal verb, e.g. *lassen* (“to let”), *scheinen* (“to seem”) or in a negative context, with a negative polarity item, such as *nicht* (“not”), *niemand* (“nobody”), *nichts* (“nothing”), *kaum* (“hardly/barely”), *nie/niemals* (“never”) and others[11]. The proportion of *dass*- vs. *w-/ob*-clauses with the verbs, which occur in one of these contexts is shown in table 6.19.

The figures show that *w-/ob*-clause-taking verbs comprise about 60% of all the extracted verbs occurring with modals or with negative polarity items, e.g *können* (“can”), *müssen* (“must/have to”), etc.

The quantitative classification results (given in table 6.20) of the base verbs show that most of them (76,30% of all the base verbs extracted from corpora of 1563 mil-

[9]Extraction from corpora of 1563 million tokens.

[10]See section 2.2.2 for details.

[11]We use the list of negative polarity items described in (Fritzinger *et al.* 2010).

context parameters	*dass*-clauses		*w-/ob*-clauses		TOTAL	
	context types	in%	context types	in%	context types	in%
+mod	924	36,46	1610	63,54	2534	100,00
+neg	851	46,66	973	53,34	1824	100,00
TOTAL	1775	40,73	2583	59,27	4358	100,00

Table 6.19: Quantitative results for context parameters of base verbs

lion tokens) prove to subcategorise for both, *dass*-and *w-/ob*-clauses, thus belonging to the Vbase1 class (cf. classification described in 5.3.4), which is expected. The Vbase3 class (verbs subcategorising for *dass*-clause only) appears to be very infrequent (0,09% of of all the extracted base verbs).

base verbs	Vbase1	Vbase2	Vbase3	TOTAL
context types	6130	1897	7	8034
in %	76,30	23,61	0,09	100,00
tokens	17628	4822	9	22459
in %	78,49	21,47	0,04	100,00

Table 6.20: Proportion of different types of base verbs

In table 6.21 we illustrate the three classes of verbs with several examples. Further examples are listed in tables in section A.7 in the appendix.

Vbase1	Vbase2	Vbase3
achten, dass/w-/ob	*abwägen, w-/ob*	*berichtigen, dass*
befuerchten, dass/w-/ob	*erforschen, w-/ob*	*beteuern, dass*
entdecken, dass/w-/ob	*konkretisieren, w-/ob*	*einbilden, dass*
erfahren, dass/w-/ob	*nachforschen, w-/ob*	
klarstellen, dass/w-/ob	*prüfen, w-/ob*	
mitteilen, dass/w-/ob	*untersuchen, w-/ob*	
versichern, dass/w-/ob	*verunsichern, w-/ob*	

Table 6.21: Examples of the three types of base verbs

Subcategorisation relations between verbs and nominalisations The fact that the majority of base verbs belong to Vbase1 has an impact on the proportion of subcategorisation relations between verbs and their nominalisations. Our extraction results show that R3[12] relations are very uncommon in German. To find out the porportion of the occurrences of different relation types (R1, R2 or R3, see section 4.2.5 for details), we analyse the 160 most frequent verb-nominalisation pairs[13] extracted

[12]In the R3 relations, nominalisations have additional subcategorisation properties, which are not specific for their verbs.

[13]Lemma types.

from corpora of 1653M tokens. The figures in table 6.22 show that R1 and R2 relations are most frequent in our data, they comprise almost 92% of the classified relations. This means that nominals inherit their subcategorisation properties from the underlying verbs. However, in about 47% of the analysed cases the nominalisation does not possess one of the verbal subcategorisation properties (either *dass*-or *w-/ob*-clauses).

relations	**R1**	**R2**	**R3**	**TOTAL**
types	72	75	13	160
in %	45,00	46,87	8,13	100,00

Table 6.22: Relations in 160 verb-nominalisation pairs in our data

Besides that, the data show that nominalisations tend to lose *w-/ob*-clauses in most cases (84%, cf. table 6.23), which is, on the one hand, predictable if we know that most nouns show preferences for *dass*-clauses, whereas their base verbs do not. On the other hand, if we follow the general assumption that deverbal nouns take over *all* subcategorisation properties of their underlying verbs, the amount of the R2 relations is unexpected. Our data show that most of the extracted cases do not correspond to the general assumption that nominalisations completely "inherit" the subcategorisation properties of the underlying verbs.

subclause	***dass***	***w-/ob***	**TOTAL**
types	12	63	75
in %	16,00	84,00	100,00

Table 6.23: Subcategorisation properties lost by nominals in R2

This means that in the process of derivation, deverbal nouns tend to lose a number of properties that their base verbs possess. Moreover, the absence of the R3 relations in our data reveals that nominalisations rarely gain new properties additional to the ones they take over from the verbs, which means that this type of relation is conceptual – we assume the existence of this relation type but it is very rare in our data (as already mentioned in section 5.3.4 above).

To illustrate relations of the R1 and R2 relations, we give sample verb-nominalisation pairs in table 6.24. The list of all the analysed 160 verb-nominalisation pairs is given in section A.8 in the appendix. Additionally, in table 6.25 we show quantitative results for some R2 pairs where the nominal does not inherit subcategorisation properties of the base verb ("non-inheritance" is extracted from corpora of 1563M tokens).

Experiments with deverbal multiword expressions Analysing the subcategorisation relations between verbs and their nominalisations, we concentrate on the nominalisations occurring freely in text corpora (in the VF context type). However, we assume that the subcategorisation behaviour of the nominalisations embedded into multiword expressions might differ from that of the nominalizations, which occur freely in context. Therefore, we perform additional extraction tests to compare the properties of *ung*-nominalisations inside and outside a multiword with their base

R1 pairs	
äussern , dass/w-/ob	*Äusserung , dass/w-/ob*
anmerken , dass/w-/ob	*Anmerkung , dass/w-/ob*
bedingen , w-/ob	*Bedingung , w-/ob*
berichtigen , dass	*Berichtigung , dass*
einbilden , dass	*Einbildung , dass*
fordern , dass/w-/ob	*Forderung , dass/w-/ob*
meinen , dass/w-/ob	*Meinung , dass/w-/ob*
mitteilen , dass/w-/ob	*Mitteilung , dass/w-/ob*
prüfen , w-/ob	*Prüfung , w-/ob*
überlegen , dass/w-/ob	*überlegung , dass/w-/ob*
überpruefen , w-/ob	*Überpruefung , w-/ob*
R2 pairs	
ankündigen , dass/w-/ob	*Ankündigung , dass*
auffordern , dass/w-/ob	*Aufforderung , dass*
belehren , dass/w-/ob	*Belehrung , dass*
berücksichtigen , dass/w-/ob	*Berücksichtigung, dass*
drohen , dass/w-/ob	*Drohung , dass*
einschränken , dass/w-/ob	*Einschränkung , dass*
erfahren , dass/w-/ob	*Erfahrung , dass*
klären , dass/w-/ob	*Klärung , w-/ob*
prophezeien , dass/w-/ob	*Prophezeiung , dass*
regeln , dass/w-/ob	*Regung , w-/ob*
sicherstellen , dass/w-/ob	*Sicherstellung , dass*
unterrichten , dass/w-/ob	*Unterrichtung , dass*
unterscheiden , dass/w-/ob	*Unterscheidung , w-/ob*
vereinbaren , dass/w-/ob	*Vereinbarung , dass*
vorstellen , dass/w-/ob	*Vorstelllung , dass*
zustimmen , dass/w-/ob	*Zustimmung , dass*
R3 pairs	
aufklären , w-/ob	*Aufklärung , dass/w-/ob*
beurteilen , w-/ob	*Beurteilung , dass/w-/ob*
einschätzen , w-/ob	*Einschätzung , dass/w-/ob*
erwägen , w-/ob	*Erwägung , dass/w-/ob*
überwachen , w-/ob	*Überwachung , dass/w-/ob*
verifizieren , w-/ob	*Verifizierung , dass/w-/ob*
verlautbaren , w-/ob	*Verlautbarung , dass/w-/ob*
verordnen , w-/ob	*Verordnung , dass/w-/ob*

Table 6.24: Examples of the classified subcategorisation relations

predicates	dass		w-/ob		TOTAL	
	tokens	in%	tokens	in%	tokens	in%
ankündigen	31	30,39	71	69,61	102	100,00
Ankündigung	216	100,00	0,00	0,00	216	100,00
bestätigen	41	10,25	359	89,75	400	100,00
Bestätigung	74	100,00	0,00	0,00	74	100,00
erfahren	42	6,43	611	93,57	653	100,00
Erfahrung	124	100,00	0,00	0,00	124	100,00
vorstellen	38	12,22	273	87,78	311	100,00
Vorstellung	403	100,00	0,00	0,00	403	100,00

Table 6.25: "Non-inheritance" extracted from corpora

verbs, e.g. *erfahren - Erfahrung - Erfahrung haben - Erfahrung machen - in Erfahrung bringen* ("to experience - experience - to bring into experience/to find out"). In table 6.26, we give several verb-nominalisation-multiword combinations, for which we indicate the type of clauses obtained from our corpora (ca. 1563M tokens).

The sample results show that in most cases nominalisations both within and outside a multiword prove to "inherit" their subcategorisation properties from their base verbs. For instance, in *entscheiden-Entscheidung-zu(de)r Entscheidung gelangen/kommen/stellen* ("to decide-decision-to to come to/to bring to decision"), the base verb, its nominalisation and the multiword containing the nominalisations subcategorise for both *dass-* and *w-/ob*-clauses. In *abstimmen-Abstimmung-zur Abstimmung kommen* ("to vote - voting - to come to the vote"), both the verbs and their nominalisations take one of the subclause types and do not allow for the other.

Table 6.27 llustrates the case in which the subcategorisation of the nominalisation occurring both, within a multiword and freely in context, differs from that of the underlying verb. The verb *überzeugen* mostly occurs with the interrogative *w-/ob*-clauses. However, its nominalisation *Überzeugung* occurs only with a *dass*-clause in our data[14]. In table 6.27, we summarise quantitative results for the combination *überzeugen - Überzeugung - zur Überzeugung bringen/gelangen/führen/kommen*.

Even the search for the occurrences of the multiwords *zur Überzeugung bringen/gelangen/führen/kommen* along with *w-/ob*-clauses in Google (we do not test verbs or nominalisations outside a multiword as the search delivers much noise)[15], does not deliver any positive results:

[14]We analyse nominalisations extracted in the VF from corpora of 1563M tokens.
[15]Extraction of November, 11, 2009.

relations	predicates	*dass* in %	*w-/ob* in %
R1	*entscheiden*	5,79	94,21
	Entscheidung	33,50	66,50
	Entscheidung treffen	28,26	71,74
	zu(de)r Entscheidung gelangen	87,50	12,5
	zu(de)r Entschediung kommen	60,87	39,13
	vor die Entscheidung stellen	18,18	81,82
	überlegen	4,95	95,05
	Überlegung	56,10	43,90
	zu(de)r Überlegung führen	57,14	42,86
	ahnen	8,79	91,21
	Ahnung	70,59	29,41
	(keine) Ahnung haben	19,05	80,95
	bedingen	+	-
	Bedingung	100,00	0,00
	zur Bedingung machen	100,00	0,00
	die Bedingung stellen	100,00	0,00
R2	*andeuten*	15,22	84,78
	Andeutung	100,00	0,00
	Andeutung machen	81,25	18,75
	erfahren	8,47	91,53
	Erfahrung	100,00	0,00
	Erfahrung haben	89,80	10,2
	Erfahrung machen	87,91	12,09
	in Erfahrung bringen	44,34	55,66
	befürchten	73,33	26,67
	Befürchtung	100,00	0,00
	zu(der) Befürchtung(en) führen	100,00	0,00
	überzeugen	21,74	78,26
	Überzeugung	100,00	0,00
	der Überzeugung sein	100,00	0,00
	zu(de)r Überzeugung bringen	100,00	0,00
	zu(de)r Überzeugung gelangen	100,00	0,00
	zu(de)r Überzeugung führen	100,00	0,00
	zu(de)r Überzeugung kommen	100,00	0,00

Table 6.26: Verbs vs. their nominalisations extracted from corpora

predicates	*dass*		*w-/ob*		TOTAL	
	context types	in%	context types	in%	context types	in%
überzeugen	10	21,74	36	78,26	46	100,00
Überzeugung	34	100,00	0	0,00	34	100,00
der (einer) Überzeugung sein	23	100,00	0	0,00	23	100,00
zu(de)r Überzeugung bringen/ gelangen/führen/kommen	97	100,00	0	0,00	97	100,00

Table 6.27: *überzeugen* and *Überzeugung* extracted from corpora

Google:	*"der Überzeugung ist, dass"*	-	57.700
	"der Überzeugung ist, w-"	-	0
	"der Überzeugung ist, ob"	-	0
	"zur Überzeugung bringt, dass"	-	1.280
	"zur Überzeugung bringt, w-"	-	0
	"zur Überzeugung bringt, ob"	-	0
	"zur Überzeugung führt, dass"	-	220
	"zur Überzeugung führt, w-"	-	0
	"zur Überzeugung führt, ob"	-	0
	"zur Überzeugung gelangt, dass"	-	184.000
	"zur Überzeugung gelangt, w-"	-	0
	"zur Überzeugung gelangt, ob"	-	0
	"zur Überzeugung kommt, dass"	-	16.000
	"zur Überzeugung kommt, w-"	-	0
	"zur Überzeugung kommt, ob"	-	0

Therefore, the nominalisation *Überzeugung* takes over a part of the subcategorisation properties of the base verb *überzeugen* in both cases - occurring freely in corpora and in a multiword with a support verb. Both cases prove to belong to the R2 subcategorisation relations:

überzeugen, ***dass/w-/ob***	→	*Überzeugung,* ***dass****/*w-/*ob*	R2
überzeugen, ***dass/w-/ob***	→	*zur Überzeugung bringen,* ***dass****/*w-/*ob*	R2
überzeugen, ***dass/w-/ob***	→	*zur Überzeugung führen,* ***dass****/*w-/*ob*	R2
überzeugen, ***dass/w-/ob***	→	*zur Überzeugung gelangen,* ***dass****/*w-/*ob*	R2
überzeugen, ***dass/w-/ob***	→	*zur Überzeugung kommen,* ***dass****/*w-/*ob*	R2

Another example (illustrated in table 6.28) shows that there are also cases where the subcategorisation of the nominalisation, which occurs freely in corpora, differs from that of the verb (relation of type R2), whereas the subcategorisation of the nominalisation within a multiword does not (relation type R1), cf. table 6.28. The nominalisation *Erfahrung* ("experience") "inherits" only the *dass*-clause from the base verb. Interestingly, the multiword *in Erfahrung bringen* ("to bring into experience/to find out") tends to take interrogative clauses in over 55% of all extracted cases. This means that the subcategorisation properties of the multiword are closer to the base verb *erfahren* ("to experience/to find out") than to the nominalisation that is contained in this multiword. Examples (6.1a) to (6.1c) confirm that the multiword can occur with all the three complement types. Further multiword expressions containing the nominalisation *Erfahrung* can also take *w-/ob*-clauses. However, their occurrence with interrogative clauses is less frequent, cf. table 6.28.

predicates	***dass***		***w-/ob***		**TOTAL**	
	context types	**in%**	**context types**	**in%**	**context types**	**in%**
erfahren	15	8,47	162	**91,53**	177	100,00
Erfahrung	33	**100,00**	0	0,00	33	100,00
in Erfahrung bringen	47	44,34	59	**55,66**	106	100,00

Table 6.28: *erfahren* and *Erfahrung* extracted from corpora

(6.1a) *Weiter* ***können*** *die Besucher* ***in Erfahrung bringen, dass*** *John Lennon im März 1957 in Liverpool die Gruppe The Quarrymen gründete.*
("Furthermore, the visitors can find out that in March 1957 John Lennon formed the band 'The Quarrymen' in Liverpool".).

(6.1b) *Es fällt mir auch schwer, mich dem Westberliner Schlangestehen anzupassen; unwillkürlich drängle ich, um* ***in Erfahrung zu bringen, ob*** *es von der Wurst noch etwas gibt oder nicht.*

(6.1c) *Außerdem läßt sich* ***in Erfahrung bringen, warum*** *das eine Fahrrad anders um die Kurve fährt als das andere.*

The subcategorisation relations between the verb *erfahren* and its nominalisation *Erfahrung* depends on the context. If the nominalisation occurs freely in context, the subcategorisation relation belongs to type R2. If the nominalisation occurs within a multiword, it has the R1 relation with its base verb:

erfahren, ***dass/w-/ob***	→	*Erfahrung,* ***dass****/*w-/*ob*	R2
erfahren, ***dass/w-/ob***	→	*in Erfahrung bringen,* ***dass/w-/ob***	R1

We also suppose that the subcategorisation of the multiwords, which consist of a preposition, a noun and a support verb, e.g. *in Erfarung bringen* is even closer to the subcategorisation of the verb than that of multiwords containing a noun and a support verb only, e.g. *Erfahrung machen*, cf. the proportion in table 6.26.

6.1.3.2 Towards the explanation for "non-inheritance"

As seen from the above described examples and frequency data, most extracted relations are of type R2: the nominalisation inherits only a part of the subcategorisation properties of the verb. In most cases nominalisations allow for declaratives only (type Nung3). In table 6.29, we give several examples of Nung3 nominalisations (further examples are listed in section A.6 in the appendix).

The "non-inheritance" behaviour of nominalisations can be explained by several reasons. One of them is the fact that nominal predicates show preferences for declarative *dass*-clauses, cf. figures in tables 6.1 and 6.15, whereas their underlying verbs

Ankündigung	*Entdeckung*	*Meldung*
"announcement"	"discovery"	"message"
Bedingung	*Erfahrung*	*Prophezeiung*
"condition"	"experience"	"prediction"
Bestätigung	*Erwähnung*	*Voraussetzung*
"confirmation"	"mention"	"condition"
Beteuerung	*Hoffnung*	*Vorstellung*
"assertion"	"hope"	"conception/idea"

Table 6.29: Nominalisations that take *dass*-clauses only

allow for all the three subclause types. We suppose that the preference for the declarative complement can either be influenced by the semantics of the predicates (their selectional restrictions) or the context they appear in, cf. table 6.19 above.

Hypothesis 1: Semantics of *ung*-nouns and their subclauses We assume that the phenomenon of the "non-inheritance" of interrogative clauses by *ung*-nominalisations can be explained by the semantic features and selectional restrictions of the nominalisations, as well as the semantics of declarative and interrogative sentences. In section 3.4.2.5, we summarised different attempts from the literature to explain the "non-inheritance" of verbal complements by nominalisations. Many authors admit that one of the important reasons is the semantics of nominalisations and verbs, as well as their selectional restrictions.

As described in section 2.2.2.3, the occurrence of declarative or interrogative subclauses with different predicates depends on the semantics of these predicates. For instance, factive verbs such as *wissen* ("to know") or *bedauern* ("to regret") presuppose that the content of their propositional complement is true, e.g. the sentence *Christof weiß, dass Katja Kreuzworträtsel mag* ("Christof knows that Katja likes crosswords") means *Katja mag Kreuzworträtsel* ("Katja likes crosswords"):

$$\text{factive verb, } \textit{dass } p \Rightarrow p$$

Therefore, the nominalisations derived from these verbs, e.g. *Wissen* ("knowledge"), *Bedauern* ("regret") also express true propositions.

Non-factive verbs and their nominalisations, such as *sich vorstellen* ("to imagine") and *Vorstellung* ("idea/thought") or *träumen* ("to dream") and *Traum* ("dream"), may presuppose that the proposition is false, e.g. the sentence *Christof stellt sich vor, dass es regnet* ("Christof imagines that it rains") can mean *Es regnet nicht* ("It doesn't rain"):

$$\text{non-factive verb, } \textit{dass } p \Rightarrow \text{not } p$$

We assume that some nominalisations function as placeholders or containers for true or false propositions, which are, in fact, expressed by the subcategorised clauses. This explains the subcategorisation behaviour of the *ung*-nominalisations which take *dass*-subclauses in over 77% of the obtained data (cf. table 6.16). If seen semantically, *dass*-clauses are propositions, which can be both, true and false:

ung-nominalisation, *dass* p ⇒ p
ung-nominalisation, *dass* p ⇒ not p

W-clauses are questions that can be interpreted as open sets of propositions, *ob*-clauses behave as *yes/no*-questions about the truth values of the propositions. Thus, both interrogative clause types do not express closed propositions with definite truth values. This means that their semantics is not compatible with the restrictions of those *ung*-nominalisations, which serve as placeholders for propositions. Besides that, *ung*-nominalisations are perfective, and predicates allowing for interrogative clauses or questions cannot be perfective, as the latter express an open set of answers.

The propositional meaning of the subclauses subcategorised by *ung*-nominalisations can be introspectively tested with the help of a deletion test[16]. As we assume that many *ung*-nominalisations perform just as containers for the proposition expressed by the *dass*-clause, we can omit this nominalisation from the sentence:

- **if:** the complement clause can be used without the noun ⇒ the nominalisation presupposes a proposition - **else:** no clear support for a proposition

To prove this, we manually check the sentences containing nominalisations followed by a *dass*-clause in the Vorfeld. For instance, the nominalisation *Vorstellung* ("idea"), which subcategorises for a *dass*-clause in (6.2a), can be omitted from the sentence. Example (6.2b) shows that the sentence in (6.2a) is correct even if used without this nominalisation[17], which means that the nominalisation *Vorstellung* has a propositional reading.

(6.2a) ***Die Vorstellung**, dass nur die Mutter die einzig wichtige und gute Bezugsperson für Kinder ist, ist nicht nur Ausdruck von Altruismus sondern kann auch Ausdruck von Egoismus sein, befördert nicht Väterlichkeit, sondern hemmt sie.*
"**The idea** that only the mother is the only important and good psychological parent for children is not only the expression of altruism but can also be the expression of egoism, it does not carry fartherliness but rather blocks it".
vs.

(6.2b) *Dass nur die Mutter die einzig wichtige und gute Bezugsperson für Kinder ist, ist nicht nur Ausdruck von Altruismus sondern kann auch Ausdruck von Egoismus sein, befördert nicht Väterlichkeit, sondern hemmt sie.*
"That only the mother is the only important and good psychological parent for children is not only the expression of altruism but can also be the expression of egoism, it does not carry fartherliness but rather blocks it".

The semantics of nominalisations, which can subcategorise for *w*-and *ob*-clauses, allows not only the expression of closed definite propositions but also questions with open sets of answers and truth values. We suppose that these nominalisations are ambiguous between propositional and further readings, e.g. an event reading. For instance, the nominalisation *Entscheidung* ("decision") can take both, declarative *dass*-

[16]This test was developed in joint work with A. Roßdeutscher.
[17]Which does not mean that the meaning of the sentence does not change. However, it can still confirm the propositional meaning of the nominalisation.

and interrogative *w-/ob*-clauses, cf. examples (6.3a) to (6.3c). The noun in (6.3a) has a propositional reading. However, both in (6.3b) and in (6.3c), it has an event reading due to the meaning of the content - in both cases, the result of the decision is apparently absent.

(6.3a) *Die Entscheidung,* ***dass*** *der FC Bayern Matthäus haben wolle, sei gefallen.* ("The desicion that the FC Bayern wants to have Matthäus has been made").

(6.3b) *Die Entscheidung,* ***ob*** *Deutschland seine Auslieferung fordert, wird für Montag erwartet.* ("The decision if Germany claims for his delivery is expected to be made on Monday").

(6.3c) *Die Entscheidung,* ***wann*** *genau im neuen Jahr gespielt wird, wurde auf die Rückrundenbesprechung am 23. November in Alzenau vertagt.* ("The decision when exactly it will be played next year, was postponed till November, 23rd for the second half of the campaign meeting in Alzenau").

The ambiguity of this noun can be derived from its underlying verb, whose choice for a clause type (declarative vs. interrogative) depends on its readings (cf. the description of valency patterns in *ELDIT*, mentioned in section 2.2.2.3).

Hypothesis 2: Context parameters We assume that the "non-inheritance" of the verbal subcategorisation properties by nominalisations can also result from the contextual parameters of the data, their polarity, modality or their mood, cf. section 2.2.2.3. These factors cause the change of truth values, which influences the choice for sentential complements (propositions vs. questions), cf. table 6.19 above.

The data show[18] that some underlying verbs show preferences for *w-/ob*-clauses if used with modal verbs or negative polarity items, as mentioned in 6.1.3.1 above.

We analyse the contexts of sample verbs, which occur with *w-/ob*-clauses. The figures in table 6.30 show that although most *w-/ob*-taking verbs appear to occur in a positive context and without any modal verbs ('-mod' and '-neg' in table 6.30), a considerable amount of them are used in contexts with the changed modality or polarity ('+mod' and '+neg' in table 6.30). For instance, the verb *vorstellen* subcategorising for a *w-/ob* clause occurs with modal verbs or negative polarity items in almost 62% of the analysed cases.

verbal predicate	**+mod or +neg**		**-mod or -neg**		**TOTAL**	
	context types	**in % in %**	**context types**	**in % in %**	**context types**	**in % in %**
erklären	50	22,73	170	77,27	220	100,00
vorstellen	84	61,76	52	38,24	136	100,00
wissen	341	35,08	631	64,92	972	100,00

Table 6.30: Modality and polarity of the contexts of verbs if used with *w-/ob*-clauses

Moreover, there are verbs that are inherently negative, e.g. *bezweifeln* ("to doubt"), *bestreiten* ("to deny"), *verhindern* ("to prevent"), etc., and attitude verbs, for instance, *glauben* ("to believe"), *vermuten* ("to suppose") *annehmen* ("to assume"),

[18]For this case study, we analysed verbs extracted in VL from a corpus of ca. 1652 million tokens.

which often function as "negative context"[19]. We tested 182 context types of some potentially negative verbs occurring in our data. In 74% of the extracted cases, they take *w-/ob*-clauses, cf. table 6.31.

verbal predicate	***dass***		***w-/ob***		**TOTAL**	
	context types	**in %**	**context types**	**in %**	**context types**	**in %**
annehmen	9	25,00	27	75,00	36	100,00
bestreiten	3	12,00	22	88,00	25	100,00
glauben	12	23,53	39	76,47	51	100,00
hoffen	9	39,13	14	60,87	23	100,00
schätzen	5	25,00	15	75,00	20	100,00
vermuten	9	33,33	18	66,67	27	100,00
TOTAL	47	25,82	135	74,18	182	100,00

Table 6.31: The proportion of *w-/ob*- vs. *dass*-clauses with negative verbs

Nominal predicates occurring in VF are not modified by modal verbs, therefore they are less influenced. Besides that, if used in VF, they cannot be directly negated by the negative dfinite article *keine*[20]. However, multiword expressions, which consist of a nominal predicate and a support verb, behave syntactically like verbs, and thus, can be both, embedded under a modal verb or negated. Therefore, some nominalisations within a multiword show more preferences for *w-/ob*-subclauses than if they occur freely in context. For instance, the nominalisation *Ahnung* can subcategorise both for *dass* and *w-/ob*-clauses. However, if used in the Vorfeld, it shows preferences for declarative sentences (71% vs. 29%), as seen in table 6.24 above. The multiword *Ahnung haben*, on the contrary, prefers *w-/ob*-clauses (81% vs. 19%), and so does their underlying verb *ahnen*. Our data show that both, the verb and the deverbal multiword, if they subcategorise for an interrogative clause, occur with modal verbs or negative polarity items in a considerable number of cases (*ahnen, w-/ob* in over 50% of the analysed cases, and *Ahnung haben* in 69%)[21].

(Ehrich 1991) claims that some semantic types of nominalisations cannot be negated at all (regardless of context). According to the author, event nominalisations cannot be negated as it is impossible that the "non-occurrence" of the event can take place, cf. section 4.2.2.1. This might explain the R2 relations where verbs can take interrogative clauses if they are used in negated context. However, their nominalisations cannot take interrogative clauses, as they cannot be negated at all.

Summary The extraction and classification results obtained within the analysis of the relations between morphologically related predicates show that there are limits to the correspondences or "inheritance" of subcategorisation (e.g. type R2 and R3 relations). Although the obtained figures confirm the generally accepted assumption that subcategorisation properties of deverbal nouns are in most cases taken over

[19] See (Fritzinger *et al.* 2010) for details.

[20] In German, nouns are negated with the negative indefinite article *kein(e)*.

[21] We suppose that the SVC with the nominalisation *Ahnung* has mostly the negated form *keine Ahnung haben, w-/ob*. If used in a positive context, this SVC need the Korrelat *davon*: *eine Ahnung davon haben*.

from their underlying verbs, there are also cases where the process of "inheritance" is limited to certain complement types (e.g. *dass*-clauses only). These phenomena are influenced both, by the semantics of predicates and the semantics of the subclauses they subcategorise for (as described in 2.2.2.3 above). Some nominalisations differ from their base verbs semantically and thus have other selectional restrictions, which influence the choice for the complement they take. Furthermore, such contextual parameters as the occurrence under modal verbs or in negative constructions, can influence the process of "inheritance" of the verbal subcategorisation properties by their derivatives.

We assume that the phenomena of "inheritance" or "non-inheritance" of subcategorisation features between morphologically related predicates can be explained by both, semantics and contextual parameters of the data.

To find out which of these two factors has a greater influence on the studied phenomena we need to conduct a deeper semantic analysis of the predicates on the one hand, and to explore their context on the other. In this thesis we didn't concentrate on the analysis of these factors, although the presented extraction architecture provides information on contextual parameters of predicates, such as embedding under modal verbs, occurrence in negated contexts and others. Besides that, we can apply an existing semantic classification for base verbs, e.g. the one described in (Schulte im Walde 2006) to achieve a cross-classification, which can provide us with the information on the systematicity between semantic verb classes and their subcategorisation behaviour.

The definition of the main reason for the "non-inheritance" cases described within this study cannot be stipulated in this thesis, as a deeper analysis of semantics and contextual parameters of predicates is required which is not the aim of our research.

6.2 Evaluation of Extraction and Classification Procedures

In this section we evaluate precision and recall of the procedures elaborated within this thesis and described in chapter 5 above. These include the identification of certain predicate types in corpora and their exctraction from corpora along with their subcategorisation properties. We also evaluate the precision of single classification steps.

6.2.1 Precision and Recall: their Application

Precision and recall are general measures in information retrieval. They are based on the comparison of an expected result with the effective result of the evaluated system. These results are considered as a set of items, in our case the predicates to be extracted. Precision is a measure of exactness, whereas recall is a measure of completeness.

Mostly there exists an inverse relationship between precision and recall, which means that one can be increased at the cost of reducing the other. For instance, an information retrieval system (such as a search engine) can often increase its recall by retrieving more documents, at the cost of increasing the number of irrelevant

documents retrieved (decreasing precision). In our case, we can increase recall if we increase the number of corpora to expore, which can automatically cause a decrease in precision. As we aim at a precision-oriented extraction, we increase the precision, excluding noise-causing cases from extraction.

In a similar way a classification system, which decides whether or not to include certain items in one class, can achieve high precision by only classifying items with the exact features, but at the cost of lower recall caused by a number of items that match just several criteria of the feature specification.

Precision In the extraction task of our system, precision can be defined as the number of relevant predicates extracted by our tools divided by the total number of predicates extracted.

Extraction:
Precision = Relevant extracted/All extracted

In the classification task, the precision of a class is the number of true positives (TPs), which means the number of items that are correctly labelled as belonging to this class, divided by the total number of elements labelled as belonging to this class. The total number is the sum of true positives and false positives (FPs) (items incorrectly labelled as belonging to this class).

Classification:
Precision = Correctly classified (TPs)/All classified (TPs + FPs)

A perfect precision score (100,00%) means that every predicate extracted by the system is relevant for our analysis. However, it does not give any information on whether all relevant items were identified and retrieved by our tools.

Similarly, in our classification task a precision score of 100,00% means that every predicate classified into a given class does indeed belong to this class. This does not give any information on the number of items from the given class that were not labelled correctly.

Recall In the extraction task of our system, recall is defined as the number of relevant predicates extracted by the tools, divided by the total number of existing relevant predicates, including those that should have been extracted.

Extraction:
Recall = Relevant extracted/All relevant (extracted and not extracted)

In the classification task, recall is defined as the number of true positives divided by the total number of predicates that actually should be in the class, i.e. the sum of true positives and false negatives. False negatives are items, which were not classified as belonging to the positive class, but should have been.

Classification:
Recall = Correctly classified (TPs)/All belonging to the class (TPs + FNs)

A perfect recall score (100,00%) in extraction means that all predicates we are looking for are retrieved by the extraction system. However, this does not deliver any information about how many irrelevant predicates were also retrieved.

A recall of 100,00% in classification means that every predicate from the given class is classified as belonging to this class. However, it says nothing about how many other predicates were incorrectly also classified as belonging to this class.

6.2.2 Precision and Recall of the Extraction and Classification Architecture

To calculate precision and recall for the described tools, we manually analyse the obtained data. For the precision of the extraction of verbal, nominal and multiword predicates, we manually check a number of sentences[22] for each predicate type under analysis. The figures show that the results vary depending on the clause types, *dass*- vs. *ob* and vs. *w*-clauses[23].

predicates	**nominal**	**multiword**	**verbal**
precision in %	99,00	81,06	96,10

Table 6.32: Precision assessed on predicates subcategorising for *w*-clauses

The lower figures for *w*-subclauses can be explained by their systematic ambiguity. Headless relative clauses and adverbal relative clauses in German have the same form as the clauses under analysis. To eliminate the problematic cases, we use filtering procedures, described in section 5.3.1.4 above. The evaluation of the extracted data shows that the application of these filtering procedures reduces noise and, therefore, increases the precision of our results. For instance, the precision of the preliminary extraction tests for verbs and multiwords subcategorising for *w*-clauses was much lower – 60% and 20% respectively. The analysis of the non-filtered data also shows that the reduction of recall due to elimination of the noisy cases is not high, as show in the following sections.

Extraction of predicates with *dass* and *ob*-clauses is less problematic, therefore, in analysing precision and recall for these cases, we obtain higher scores even not applying additional filtering procedures.

[22]The number of the analysed sentences is different for each predicate type.
[23]The extraction results for *w*-clauses show lower precision because of the ambiguity of their form.

6.2.2.1 Evaluation of identification and extraction of nominals

For nominal predicates extracted in VF from the analysed corpora[24], we evaluate 1078 context types (29972 tokens), which occur with different subclauses[25], whose occurrence in our corpora is greater than 5 tokens. Their evaluation provides the precision shown in table 6.33. The precision of the extraction of nominal predicates (an average of 98-99%) varies between about 95% and 100,00% depending on the type of complement extracted along with the noun. The lower figures for the types of *w*-subclauses, as mentioned above, are due to the systematic ambiguity of *w*-indirect questions.

subclause type	***dass***		***w-***		***ob***		**TOTAL**	
	TP	FP	TP	FP	TP	FP	TP	FP
context types	853	5	124	7	89	0	1066	12
precision in %	99,42		94,66		99,00		98,89	
tokens	20018	59	3161	29	6705	0	29884	88
precision in %	99,71		99,09		100,00		99,71	

Table 6.33: Precision of the extraction of subclause-taking nouns in VF

The results presented in table 6.33 are obtained after the application of filtering procedures. The previous precision calculated for *w*-clause-taking nominals was about 10% lower and estimated only 89%, cf. table 6.34.

subclause type	**precision in %**
dass	98,50
ob	99,00
wh-	89,00

Table 6.34: Precision of noun extraction before application of filtering procedures

Most eliminated cases that deliver noise were posed by the nouns extracted with the *w*-words *wo* ("where") and *wobei* ("whereby/whereas"), e.g. *Umgebung, wo* ("environment where") or *Piaffen, wobei* ("piaffes wherby"), cf. (6.4a) and (6.4b).

(6.4a) *Die exakt vollführten* ****Piaffen, wobei*** *das Pferd auf der Stelle trabt, dienen dem Meister zum Verschnaufen, dem Publikum zum Staunen*
("The exactly performed piaffes, whereby the horse is trotting on the spot, are used by the champion to catch one's breath, to astonish the audience").

(6.4b) *In einer unmittelbaren* ****Umgebung , wo*** *viele allenfalls noch auf der Tastatur ihres Computers herumhämmern, arbeitet er am uralten Werkstoff Eisen, wie der Griechengott Hephaistos oder der altnordische Wieland.*
("In the direct environment, where many people are hammering at the best on

[24] A total of 1563 million tokens.

[25] We separately calculate precision for *w-* and *ob*-clauses, as extraction of *w*-clauses delivers more noise than that of *ob*-clauses.

the keyboards of their computers, he is workig on the acient material iron, like the Greek god Hephaestus or the Old Norse Wayland").

(6.4c) *Die **Entscheidung darüber, wo** die Stadtverwaltung zusätzliche Räume bekommen wird, soll während einer zusätzlichen Parlamentssitzung am 25. Februar fallen.*
("The decision (about) where the city administration can get additional rooms, will be made during the additional parlament meeting on February, 25th").

In table 6.35 below, we illustrate the proportion of the unsorted *wo*- and *wobei*-clauses among the extracted *w*-clauses. The figures show that about 76% of all extracted *w*-clauses are introduced by *wo*[26]. To increase the precision, we exclude the occurrence of *wo* and *wobei* with the nominal predicates under analysis.

subclause types and predicates	***w-***	***wo***		***wobei***	
	TOTAL	**TOTAL**	**TPs**	**TOTAL**	**TPs**
all nominal predicates	1560	1175	20	8	1
***ung*-nominalisations**	103	32	4	0	0

Table 6.35: Proportion of *wo*- and *wobei*-clauses among the extracted *w*-clauses

However, this can decrease recall, as such cases as *Entscheidung darüber, wo* ("decision (about) where") in (6.4c) are also excluded by our tools. The figures in table 6.35 show that only 2% of all *wo*- and *wobei*-clauses are true positives. Therefore, the reduction of recall due to the elimination of these cases is not high. Table 6.36 demonstrates that we can achieve a recall of approximately 95-96% if filtering procedures are applicated. The procedures to eliminate these cases are described in section 5.3.1.4.

predicate	**TPs**	**TNs**	**recall**
all nominal predicates	377	21	96,42%
***ung*-nominalisations**	71	4	94,67%

Table 6.36: Recall after elimination of *wo*- and *wobei*-clauses

The next problem is caused by the nominal predicates, whose context partners are clauses introduced by the *w*-word *wonach*. They comprise about 9% of the extracted *w*-clauses. Manual analysis shows that these cases do not belong to interrogative complements. However, 100% of nominal predicates taking *wonach*-subclauses prove to subcategorise for *dass*-clauses (for this test, we analysed 150 nouns extracted with *wonach*-clauses from 'taz'). Therefore, we assume that *wonach*-clauses can serve as indicators for the ability to take the declarative *dass*-clause, see section 5.3.1.4 above. This can be introspectively tested by a substitution test:

if:	the word *wonach* introducing a subclause (which occurs with a noun in the Vorfeld) can be replaced by the conjunction *dass*
⇒	the noun subcategorises for a *dass*-clause

[26]We analysed 1560 sample nominal predicate types (query matches) extracted from our corpora.

We use the conjunction *dass* instead of the *w*-word *wonach*, cf. (6.5a) and (6.5b). Therefore, we reclassify the predicates extracted with *wonach*-clauses into the class of those, which take *dass*-clauses. Nominals occurring with *wonach*-clauses are analysed as nominals which allow for *dass*-clauses. The application of this substitution procedure increases the precision of the obtained *w*-clauses.

(6.5a) *Die bisher geltende Regel, wonach Zeitungen keine preissensiblen Produkte seien, scheint auf jeden Fall widerlegt.*
("The current regulation whereupon newspapers are not price sensitive products seems to be disproved").

(6.5b) *Die bisher geltende Regel, dass Zeitungen keine preissensiblen Produkte seien, scheint auf jeden Fall widerlegt.*
("The current regulation that newspapers are not price sensitive products seems to be disproved").

Nominal predicates, which are embedded as genitives in a nominal phrase pose another problem in the identification of nouns subcategorising for sentential complements. These cases comprise about 10% of the nominal predicates extracted in the Vorfeld. In this case, our tools identify the head noun as the valency bearer of the subclause, which follows the nominal phrase after a comma. However, in 16% of such cases the subclause is licensed by the noun in genitive, as illustrated in example (6.6a). These cases should be distinguished from those where the head noun is the valency bearer of the subclase, as illustrated in (6.6b). A possible solution for the analysis of such cases is described in section 5.3.1.4 above.

(6.6a) *Zur Beantwortung der **Frage, wie** Salböl am zweckmäßigsten zu gewinnen sei, findet sich im zweiten Buch Mose sogar ein Rezept.*
("For answering the question how to become the holy anointing oil in the most appropriate way, there can be even found a recipe in the second book of Mose").

(6.6b) *Auch eine **Erklärung** der Bundesbank, **dass** unordentliche Märkte unerwünscht seien, sei zu erwarten.*

The filtering procedures not only increase the precision and recall of the extraction steps, but also contribute to the achievement of the higher precision and recall of the predicate classification. For instance, the application of *wonach*-indicators to detect *dass*-clause-taking nouns does not allow the tool to classify nouns, which occur with *wonach*-clauses only into N2[27]. In table 6.37, we demonstrate the influence of the filtering procedures onto precision and recall of our extraction and classification results.

In table 6.38, we illustrate the precision for the classification of the non-filtered nominal predicates. The lower figures for the N2 nouns are caused by the lower precision of the extraction of nominal predicates subcategorising for *w*-clauses, as mentioned above. After the application of the filtering procedures described in section 5.3.1.4, we achieve 100% of precision and recall classifying nominal predicates into N1, N2 and N3.

[27]The N2 nominals can subcategorise for *w-/ob*-clauses only.

filtering procedure	precision in %	recall in %
elimination of *wo/wobei*	+76	-4
substitution of *wonach*	+9	+9
resorting nouns in genitive	+2	+2

Table 6.37: The influence of filtering procedures on precision and recall

noun classes	N1		N2		N3	
	TP	FP	TP	FP	TP	FP
context types	1752	5	292	36	1398	22
precision in %	99,72		89,02		98,45	

Table 6.38: Precision of the classification of non-filtered data

Linguistic information required for the extraction work In our extraction procedures, we make use of the grammatical properties of VL and VF sentences (see section 5.2 for details), built into our queries.

The modelling of verb-final sentences depends to some extent on detailed models of nominal and prepositional phrases, and thus profits from additional preprocessing by means of chunking or partial parsing. As our experiments show, for the Vorfeld cases, which are used in the extraction of nominal predicates, no partial parsing is needed. We compare figures of extraction patterns with and without NP and PP boundaries annotated (absolute frequency indicated under **+chunks** and **–chunks** in table 6.39). The chunked corpora provide slightly less data, but most of the cases found without the use of chunking (**diff** in table 6.39) proved to be true positives (**TP**, absolute figures and percentages).

Source	type	+chunks	–chunks	diff.	TP
taz	Vorfeld+*w*-clause	467	484	17	14 (82,4 %)
taz	Vorfeld+*ob*-clause	752	798	46	44 (95,7 %)
taz	Vorfeld+*dass*-clause	2444	2536	92	92 (100,0 %)
FAZ	Vorfeld+*w*-clause	259	283	15	7 (46,7 %)
FAZ	Vorfeld+*ob*-clause	521	538	17	16 (94,1 %)
FAZ	Vorfeld+*dass*-clause	1763	1694	69	69 (100,0 %)

Table 6.39: Extraction results in the Vorfeld position with and without chunking

Evaluation of identification and classification of compounds We also analyse the morphological sorting procedure for a list of extracted nominal predicates (15687 types) and obtain 2037 types of nominal compounds, cf. figures in table 6.4 above. The evaluation of the sorting procedure shows that we achieve a precision of 94,2% and a recall of 95,7% in the automatic identification of nominal compounds. As the morphological tool provides us with information about the part of speech of com-

pound elements, their nominal (NN) or deverbal (V) nature, we illustrate their evaluation in table 6.40. As seen from the figures in table 6.40 below, our tools can successfully distinguish between simplex and compound nominal predicates.

type&*example*	**proportion**	**precision**	**recall**
NN.NN *Branchenregel*	65,58%	97,03%	95,28%
NN.V *Volksbefragung*	12,99%	80,00%	95,24%
V.NN *Bedenkzeit*	21,43%	93,94%	97,06%
TOTAL	**100,00%**	**94,20%**	**95,70%**

Table 6.40: The morphological analysis of compounds: types, precision and recall

The lower precision for the NN.V compounds is caused by the ambiguous output of the morphological tools. Some nominalisations are identified by the morpholgical analyser as both, simplex and compound deverbal nouns. For instance, the nominalisation *Feststellung* ("statement") can be analysed as a deverbal noun, which contains the prefix *fest* ("firm"), or as a compound consisting of a nominal constituent *Fest* ("celebration/festival") and a deverbal *Stellung* ("position"). This problem is caused by the ambiguity of the word form *fest/Fest*. Besides that, this word form can be also an adjective, which results in the ambiguity of the output of the morphological tool:

forms	**morphological analysis**
compound	Fest<NN>legen<V>ung<SUFF><+NN>
simplex	fest<VPART>legen<V>ung<SUFF><+NN>
simplex	fest<ADJ>legen<V>ung<SUFF><+NN>

Table 6.41: Ambiguous output of the morphological analyser

Thus, the nominalisation *Feststellung* is double-classified by our tools both, as a simplex and as a compound nominal predicate. To eliminate the classification of such cases as compound nouns, we should include a disambiguation procedure for the ambiguous verbal prefixes. However, as the number of such cases is not high in our data, we manually sort them out of the list of nominal compounds.

In the classification of compound nominals according to their subcategorisation features (classes from C1 to C3-2), we can achieve high precision, if the nominal predicates are obtained from Vorfeld with high precision. Moreover, the correct distinction between simplex and compound nominals play an important role, cf. the example in table 6.41.

However, some NN.V compounds are ambiguous between C2 and C3-1, which can cause a decrease in precision and recall of their classification. For instance, the compound *Absichtserklärung* is automatically classified into C3-1, as both the noun *Absicht* ("purpose") and the nominalisation *Erklärung* ("explanation) belong to the subclause-taking nominals. Nevertheless, it can also be a C2-compound if we read it as the nominalisation of the multiword expression *Absicht erklären* ("to declare one's intention"), in which the base verb *erklären* ("to explain") serves as a support

verb. These cases are difficult to be detected automatically, as we need contextual information to identify the valency bearer.

In some cases the precision of the detection of the C2-compounds is increased, as the information is delivered by the morphological analyser. For instance, a part of the extraction noise with the CQP query in the preliminary tests comprise cases, such as *Zeitungsbericht* ("newspaper report"), *Zeitungsmeldung* ("newspaper announcement"), where in a naive string comparison, the non-head *Zeitung* ("newspaper") contains the string *Zeit* ("time"), which, if used as a single word, can subcategorise for a subclause. As it is just a graphical part of the compound constituent, the word *Zeit* is not the valency bearer of these compounds. The application of morphological tools allows us to eliminate the extraction of such cases.

To further improve the morphology-based classification, we should include the extracted compound predicates in the "known" list. Some cases, e.g. *Grundsatzmißbilligung* ("deprecation of principle") or *Rechtsgrundsatz* ("principle of law"), contain compound elements, e.g. *Grundsatz* ("principle"), which contain elements that can subcategorise for sentential complements themselves, such as *Grund, dass* ("reason that").

6.2.2.2 Evaluation of multiword extraction and classification

As mentioned in section 6.1.2, multiword predicates are difficult to identify in text corpora. However, we achieve a precision of 81% in their extraction from the analysed corpora. To calculate the precision of the extraction results, we analyse 834 multiword types whose frequency is greater than 1 in our corpora. The experiments show that a limitation to the higher frequency (e.g. frequency>2 or frequency>3) also increases the precision of our results but, at the same time, excludes a great number of true positives. Even the limitation of the frequency to more than 1 occurrence already causes about 13% decrease in the recall, which means that the analysed phenomena are very rare.

The evaluation of our classification method shows that we are able to successfully classify multiwords according to the types described in section 4.2.4.4 above. In table 6.42, we demonstrate the precision and recall of the classification of multiwords into M1+M2 (those sharing their subcategorisation properties with the nominal component) and M3+M4 (those not sharing their subcategorisation properties with the nominal component). The number of correctly classified types is given in line TP types (true positive types). False positives (FP types in the table) consist of the same noisy cases which are thrown up as noise in the extraction of multiwords. We can avoid them by increasing accuracy of the extrcation procedures. Most false positives are classified into the M3+M4 types by our tool. These classed contain multiword expressions, whose nominal constituents do not subcategorise for sentential complements. Therefore, such expressions, as *vom Audipartner erfahren* ("to find out from Audi partner") or *am Anfang stehen* ("to be at the beginning"), which are extracted by our tools as preposition+noun+support verb multiwords, are also classified as M3 or M4 types.

The noise in the M1 and M2 cases comprise the multiword expressions, which contain subclause-taking nouns, e.g. *Fall* ("case"), *Frage* ("question"), but in the extracted context they do not function as support verb constructions, e.g. *um die*

multiwords	M1+M2	M3+M4
TP	836	590
FP	43	40
TN	0	8
precision	95,11%	93,65%
recall	100,00%	98,66%

Table 6.42: Evaluation of the classification procedures for multiwords

Frage gehen (“to be about the question”), in which the verb *gehen* (“to go”) is a part of the expression *um etwas gehen* (“to be about sth/to have to do with sth”), and thus, is not a semantically weak support verb.

Further noise-causing cases are represented by expressions in which the support verb relates to another nominal predicate, as illustrated in (6.7).

(6.7) *Auch diese Rechts**frage** ist nur für den Fall **gestellt** , dass die erste Rechtsfrage bejaht wird.* (“This question of law is also posed for the case that the first question is affirmed”).

The verb *stellen* (“to pose”) builds a support verb construction with the noun *Frage* or *Rechtsfrage* (“question of law/legal issue”), and not with the noun *Fall* (“case”). To exclude such cases, we would need to include lexical constraints to eliminate the occurrence of other candidates for supported nouns in front of the nominal we are looking for. However, we assume that this would decrease the recall of the extraction results. As such cases are not numerous in our data, we decide to manually sort them out from the obtained list of multiwords.

As the classification of multiwords is carried out on the basis of the results of two contexts (extraction of multiwords in VL and passive, as well as extraction of nominal predicates in the VF), the accuracy of the classification procedures depends on the precision and recall of the extraction and classification of nominal predicates, described in section 6.2.2.1.

The anaysis of the classified data reveals no true negatives (TN types in table 6.42) for the M1+M2 classification groups (those wrongly classified as M3+M4), which means that we are able to obtain a recall of 100,00% in the identification of these multiword types. For the M4 multiwords we find cases, which are classified as M1 or M2 by our tools. For instance, the expressions *auf den Punkt bringen* (“to put in a nutshell”) or *jmd beim Wort nehmen* (“to take smb at his word”) are idiomatic and belong to type M4 of our classification. However, they contain nominals for which the tool knows that they take subclauses, therfore, they are automatically classified as M1 or M2 multiword expressions. We manually sort such cases out from our data, as their automatical elimination is still difficult.

6.2.2.3 Evaluation of procedures for “inheritance” relations

The evaluation of procedures for subcategorisation relations includes the calculation of precision and partially recall for the identification and classification of *ung-*

nominalisations, generation of base verbs and their classification, and classification of the subcategorisation relations between verbs and their nominalisations.

Evaluation of the identification and classification of *ung*-nominalisations As described in section 5.3.2.3, we can sort the *ung*-nominalisations out from the list of nominal predicates extracted in the VF with the help of morphological analysis. The morphological tools deliver the information about morphological features (e.g. their deverbal nature) of nominal predicates, which allow us to identify nouns containing the feature <V>ung<SUFF> only.

The analysis of the output of the sorting procedure shows that we can identify *ung*-nominalisations with the precision and recall of 100%. However, this result is achievable, provided that the sorting procedures are applied on to the list of simplex nominal predicates[28].

The accuracy of the classification of the identified *ung*-nominalisations into the three classes (Nung1 to Nung3) depends on the precision of the extraction of their subcategorisation information from corpora. Like other nominal predicates, *ung*-nominalisations are extracted in VF, which means that the extrcation problems both, for nominals in general and *ung*-nominalisations are similar. Most problematic cases in the extraction are comprised by the nominalisations extracted along with *w*-clauses, cf. figures in table 6.33 in section 6.2.2.1 above.

Provided we apply all the above described filtering procedures (see section 5.3.1.4 and section 6.2.2.1 above), we can achieve both, high precision and recall of the results, cf. table 6.43[29].

nominalisations	**Nung1**	**Nung2**	**Nung3**
TP	65	64	161
FP	0	0	0
TN	0	0	0
precision	100,00%	100,00%	100,00%
recall	100,00%	100,00%	100,00%

Table 6.43: Evaluation of the classification procedures for *ung*-nominalisations

The classification of *ung*-nominalisations, which have a compound form is more problematic. The precision and recall of their identification is described in section 6.2.2.1 above. The accuracy can depend on whether the nominalisation is the head or the non-head of the compound, cf. data in table 6.40 above.

Evaluation of base verbs generation and their classification The generation of base verbs from the list of simplex *ung*-nominalisations proceeds with high accuracy. We are able to identify base verbs with the prcision of 100,00%.

However, if we obtain the base verbs for nominalisations, which have a compound structure, we should first identify the valency bearer. For instance, in the compound

[28]For compound nominal predicates and nominal phrases, we should identify the valency bearer first, which is ambiguous in both cases.

[29]To test it, we evaluated 290 most frequent nominalisations (lemma types) extracted from the analyse corpora.

Entscheidungsfindung, ob ("decision making if") or in the nominal phrase *Entscheidung der Vertretung, dass* ("decision of the representatives that"), we need to know that the valency bearer is the nominalisation *Entscheidung* ("decision") and not *Findung* ("finding") or *Vertretung* ("representation"). These problems are solved within the procedures to classify compound nominals in section 5.3.3.3[30]. Therefore, the precision of the compounds classification can also have influence on the detection of base verbs for *ung*-nominalisations, which have a compound form.

The subcategorisation information for base verbs is extracted from corpora with the queries described in sections 5.3.1 and 5.3.2.1. We achieve a precision of about 96% in the acquisition of verbs. The analysis of the preliminary tests shows that we are able to decrease noise by 20-30% if we apply the filtering procedures described in section 5.3.1.4 above. However, some problematic cases (most of them headless relatives introduced by the *w*-words *wer* ("who") or *was* ("wer")) remain undetected.

The evaluation of our base verb classification shows that our tools can successfully categorise the data extracted from corpora into the Vbase1, Vbase2 and Vbase3 classes described in 5.3.4.2. If we assume that the acquired verbs only occur with the subclauses they were obtained with from the analysed corpora, we achieve the precision of 100,00%, cf. table 6.44. For this purpose we analyse 405 base verb types (200 types per Vbase1 and Vbase2 and 5 types for Vbase3) extracted from our corpora.

nominalisations	**Vbase1**	**Vbase2**	**Vbase3**
TP	200	200	5
FP	0	0	0
TN	0	0	0
precision	100,00%	100,00%	100,00%

Table 6.44: Evaluation of the classification procedures for base verbs

The data show that most verbs occur with all the three complement types, thus, belong to the Vbase1 type, cf. figures in table 6.20 above. However, the verbs that belong to type Vbase2 comprise about 20%, which is a considerable amount. We assume that a part of these verbs can also take *dass*-clauses although they were not found with this complement type by our tools. Therefore, we suppose that we can check their subcategorisation properties in other contexts, e.g, further sentence models (e.g. V1 or V2) or in parsed corpora.

Classification of relations The precision and recall of the classification of nominalisations and their underlying verbs have an impact on the accuracy of the classification of subcategorisation relations. If we assume that we classify the subcategorisation relations between verbs and nominalisations whose subcategorisation properties were obtained with 100% of precision and recall, we can obtain the same level of accuracy in the distinction between the R1, R2 and R3 relations.

We manually analyse the 160 verb-nominalisation pairs, whose classification according to subcategorisation relations is demonstrated in table 6.22 above. We achieve 100% of precision and recall in the classification of these relations.

[30]We suggest to treat nominal phrases with genitives as nominal compounds, cf. *Entshceidung der Vertretung* vs. *Vertretungsentscheidung*, as described in section 5.3.1.4 above.

relations	R1	R2	R3
TP	72	75	13
FP	0	0	0
TN	0	0	0
precision	100,00%	100,00%	100,00%

Table 6.45: Evaluation of the classification procedures for subcategorisation relations

However, the analysis of the R3 relations shows that all verb-nominalisation pairs in this relation type contain verbs, which take *w-/ob*-clauses only (cf. appendix A.8). On the one hand, this can mean that the semantics of these base verbs does not allow the expression of propositions. On the other hand, we admit that this can happen due to the low recall obtained by our tools. To find the explanation for these cases, we need to obtain more data from different types of contexts, which would probably decrease the precision of the extraction and classification results.

6.2.2.4 Summary: precision and recall

The evaluation of the single procedures of our extraction and classification architecture shows that our tools deliver sufficient results both, in the procedures to identify predicates and to subclassify them according to their subcategorisation properties. To achieve higher precision, a set of filtering procedures is required. We describe the several filtering steps, which are integrated into our architecture in section 5.3.1.4 above. To achieve higher recall for the acquisition of some predicate types, e.g. verbs, we need to perform additional extractions from other contexts or other corpora types (e.g. parsed corpora), which can increase the recall but, at the same time, also reduce the precision of our results, which is undesired as we aim at precision-oriented extraction.

Chapter 7

Conclusion

In the following chapter we describe our conclusions and show the contributions done within this thesis. We start summarising the results of our analysis, go on with the conclusion and the description of contributions of this research and in the end, we outline the directions for future work

7.1 Summary

This thesis describes a semi-automatic approach to the analysis of subcategorisation properties of verbal, nominal and multiword predicates in German. We semi-automatically classify predicates according to their subcategorisation properties by means of extracting them from German corpora along with their complements. In this work, we concentrate exclusively on sentential complements, such as *dass, ob* and *w*-clauses, although our methods can be also applied for other complement types. Our aim is not only to extract and classify predicates but also to compare subcategorisation properties of morphologically related predicates, such as verbs and their nominalisations. It is usually assumed that subcategorisation properties of nominalisations are taken over from their underlying verbs. However, our tests show that there exist different types of relations between them. Thus, we review subcategorisation properties of morphologically related words and analyse their correspondences and differences.

For this purpose, we elaborate a set of semi-automatic procedures, which allow us not only to classify extracted units according to their subcategorisation properties, but also to compare the properties of verbs and their nominalisations, which occur both freely in corpora and within a multiword expression. The lexical data are created to serve symbolic NLP, especially large symbolic grammars for deep processing, such as HPSG or LFG, cf. work in the LinGO project (Copestake et al. 2004) and the Pargram project (Butt et al. 2002). HPSG and LFG need detailed linguistic knowledge. Besides that, subcategorisation iformation can be applied in applications for IE, cf. (Surdeanu et al. 2003). Moreover, this information is necessary for linguistic, lexicographic, SLA and translation work.

Our extraction and classification procedures are precision-oriented, which means that we focus on high accuracy of our extraction and classification results. High precision is opposed to completeness, which is compensated by the application of extraction procedures on larger corpora.

7.1.1 Data and Existing Approaches

As mentioned above, our interest targets verbs, nouns and multiword expressions. Most works on subcategorisation concentrate on verbal valency. However, there are several studies analysing valency of further predicates. We summarise a number of studies on different predicate types in table 3.1 in chapter 3 above.

Subcategorisation properties of morphologically related predicates have been analysed in a number of linguistic and NLP studies, e.g. (Grimshaw 1990), (Nunes 1993), (Sommerfeldt/Schreiber 1996), (Schierholz 2001), (Meinschaefer 2004) and others. However, only a few lexical resources provide systematic correspondences between verbs and their nominalisations. For instance, (Macleod *et al.* 1998b) describes a computational lexicon of nominalisations NOMLEX which maps noun roles into the predicate-argument structure of their associated verbs. Another example is the analysis described in (Gurevich et al. 2007), where the authors use the PARC's text processing system for the process of mapping the predicate-argument structure of nominalisations and that of their base verbs.

Our preliminary extraction tests also show that there are both correspondences ("inheritance") and differences ("non-inheritance") in the subcategorisation of morphologically related predicates. In many cases subcategorisation properties of deverbal nominal predicates are inherited from their base verbs (example (7.1)).

(7.1)
- *befürchten, dass* (" to fear that")
  vs. *Befürchtung, dass* ("fear that")
- *erklären, dass/w-* ("to explain that/wh-")
  vs. *Erklärung, dass/w-* ("explanation that/wh-")

But there are also cases where subcategorisation of a nominalisation differs from that of its base verb. Verbal subcategorisation properties can either be reduced, cf. (7.2), or extended, cf. (7.3).

(7.2) "Inheritance" reduction:

- *vermuten, dass/w-* ("to suppose that/wh-")
  vs. *die Vermutung, dass/*w-* ("supposition that/*wh-")
- *wissen, dass/w-/ob* ("to know that/wh-/if")
  vs. *das Wissen, dass/*w-/*ob* ("knowledge that/*wh-/*if")

(7.3) "Inheritance" extension:

- *Antwort an* ("the answer for") vs. *antworten jmdm* ("to answer smb").
- *der Verdacht auf* ("suspicion of")
  vs. *verdächtigen jmdn* ("to suspect smb").

7.1.2 Classification of Predicates and their Relations

Although several studies describe classification of different predicates types, e.g. (Vendler 1967), (Levin 1993), (Fischer 2005), (Schulte im Walde 2006), (Klotz 2007) for verbal predicates, (Sommerfeldt/Schreiber 1983), (Ehrich 1991), (Teubert 2003),

(Storrer 2007), etc. for nominal and multiword predicates, none of them apply a coherent systematic classification for different predicate types. We classify verbs and nouns according to the same criteria – their ability to subcategorise for declarative and interrogative sentential complements into three classes. Class 1 includes verbs and nouns which allow for both declarative and interrogative subclauses (V1 and N1), class 2 consists of verbs and nouns which take interrogatives only (V2 and N2) and class 3 comprises verbs and nouns which subcategorise for declaratives only (V3 and N3), cf. example (7.4) and (7.5).

(7.4) Verb classes:

V1: *antworten, dass/w-/ob* ("to answer that/wh-/if")
bestimmen, dass/w-/ob ("to determine");

V2: *abfragen, w-/ob* ("to inquire wh-/if")
befragen w-/ob ("to interview wh-/if");

V3: *berichtigen, dass* ("to correct")
sich einbilden, dass ("to imagine that").

(7.5) Noun classes:

N1: *Angst, dass/w-/ob* ("fear that/wh-/if")
Beweis, dass/w-/ob ("proof/evidence that/wh-/if");

N2: *Abfrage, w-/ob* ("query wh-/if")
Rätsel ("ridle/mystery");

N3: *Eindruck, dass* ("impression")
Gefühl, dass ("feeling that").

The classification of multiword and compound nominal predicates is based on the "inheritance" relations between their subcategorisation properties and those of their constituent parts. Thus, multiword expressions can be classified into those which inherit their subcategorisation properties from their nominal components[1] (M1 and M2) and those which have their own subcategorisation properties (M3 and M4), see example (7.6).

(7.6) M1: MWEs which inherit their subcategorisation from their nominal component ("inheritance" from the noun):
zur Bedingung machen, dass ("make it a condition that")
vs. *die Bedingung, dass* ("the condition that").

M2: MWEs which inherit their subcategorisation from their nominal component under certain contextual conditions only ("inheritance" under certain contextual conditions):
in Erfahrung bringen, w-/ob ("to find out w-/if")
vs. *er hat (die) Erfahrung, dass/*w-/ob*
("he has (the) experience that/*wh-/if")
vs. *haben Sie (eine) Erfahrung, *dass/w-/ob?*
("do you have (any) experience *that/wh-/if?").

[1] We analyse support verb constructions consisting of a preposition, a noun and a support verb.

M3: MWEs whose nominal components do not take any sentential complements:
zum Ausdruck bringen, dass ("to express that")
vs. **der Ausdruck, dass.*

M4: MWEs which are commonly seen as idioms, either because they contain "cranberry" lexemes or because they are non-compositional:
in Abrede stellen, dass ("to deny that") vs. **die Abrede*[2];
ins Auge fallen, dass ("to catch sb's eye that").

Compound nouns are classified into those which inherit their properties from their heads C1, those which inherit their properties from their non-heads C2 and those which inherit their properties from both their heads and non-heads or from none of them, and thus have their own subcategorisation properties (C3-1 and C3-2), see example (7.7).

(7.7) C1: The subcategorisation of the compound is inherited from its head:
*Journalisten**frage**, w-/ob* ("journalist **question**, wh-/if")
vs. *Frage, w-/ob* ("question wh-/if").

C2 The subcategorisation the compound is inherited from its non-head:
***Auswahl**verfahren, w-/ob* ("**selection** process, wh-/if")
vs. *Auswahl, w-/ob* ("selection wh-/if").

C3-1 The subcategorisation of the compound is determined by both, its head and its non-head:
***Wettstreit**, w-/ob* ("bet battle (competition) wh-/if")
vs. *Wette, w-/ob* ("bet wh-/if")
or *Streit, w-/ob*("battle (argument) wh-/if").

C3-2 The subcategorisation of the compound is determined by neither the head nor the non-head:
***Wortspiel**, dass* ("word play that")
vs. *Wort, *dass* ("word *that")
or *Spiel, *dass* ("play *that").

We classify "inheritance" relations between nominalisations and their based verbs according to the correspondences and differences between subcategorisation properties of verbs and their nominalisations. The R1 relations include verb-nominalisation pairs in which nominalisations inherit all valency properties of the underlying verbs. The cases of "inheritance" reduction in subcategorisation (where nominalisations take over only a part of verbal subcategorisation properties) belong to the R2 relations, whereas the cases of "inheritance" extension in subcategorisation (where nominalisations have additional subcategorisation properties that their base verbs do not possess) to the R3 class. We assume that R3 is a hypothetical class as most verbs can subcategorise for both declarative and interrogative clauses (sometimes under certain contextual conditions only), whereas their nominalisations cannot do that[3], cf. example (7.8).

[2]The only non-SVC reading of *Abrede* is that of 'oral agreement', which is found in 22 % of the occurrences of the lemma, but always without a sentential complement.

[3]Our extraction experiments show that nominalisation show preferences for *dass*-clauses (in 65-67% of the analysed cases).

(7.8) R1: subcategorisation properties are inherited from the verb:
entscheiden, dass/ob/w- (“to decide that/if/wh-”)
vs. *Entscheidung, dass/ob/w-* (“decision that/if/wh-”).

R2: subcategorisation properties are inherited from the verb but in a reduced form:

- *ob/w*-clauses are lost:
(sich) erinnern, dass/w-/ob (“to recollect/remind that/wh-/if”)
vs. *Erinnerung, dass* (“recollection that”);
- *dass*-clauses are lost:
klären, dass/ob/w- (“to clarify that/if/wh-”)
vs. *Klärung, w-/ob* (“clarification wh-/if”).

R3: subcategorisation properties are inherited from the verb in an extended form – nominalisations have additional subcategorisation properties of their own:
aufklären, w-/ob (“to clarify wh-/if”)
vs. *Aufklärung, dass/w-/ob* (“clarification that/wh-/if”).

7.1.3 Methods and Tools

Input and Context We use newspaper and web corpora from Germany, Austria and Switzerland, which comprise written texts in German dated from 1988 until 2005, a total of ca. 1563M tokens[4].

All corpora are pre-processed: sentence-tokenised, tagged for part-of-speech, lemmatised and partially chunked[5]. Extraction queries in the form of regular expressions rely on the Stuttgart CorpusWorkBench (CWB, (Evert 2005)).

As extraction context for verbal and multiword predicates, we chose German Verbletzt (verb-final) clauses (VL) and passive sentence. In VL the subcategorised subclause usually follows the verb, prepositional and nominal constituents of MWEs tend to precede it, cf. table 7.1. The subclause is subcategorised either by the full verb or by the multiword. A regular sequence of elements is also present in passive sentences: the subclause follows the 2nd part of the verb, and prepositional and nominal constituents of a multiword precedes the 2nd part of the verb, cf. table 7.2.

Nominalisations are extracted in Vorfeld (pre-field) construction (VF), when a clause is initially positioned before the finite verb in German declaratives. If a noun in VF is followed by a subclause, this subclause can only be subcategorised by the noun (see Table 7.3).

[4]The corpora from Germany include extracts (1988-2001) from German newspapers, such as *die tageszeitung*, *Frankfurter Rundschau*, *Frankfurter Allgemeine Zeitung*, *Stuttgarter Zeitung*, *DIE ZEIT* and *Handelsblatt*. We also use the European Language News Corpus (’ELNC’), which includes online news from 1997. The data in ’ELNC’ originates from German news, and a part of this corpus originates from Swiss mass media. Other texts from Switzerland are contained in the Swiss part of DEREKO, which is referred to as DEREKO-CH. The Austrian part of DEREKO also includes newspaper texts, all dated between 1991 and 2000. Both, the Swiss and the Austrian parts of the DEREKO corpora, are part of the German reference corpus DeReKo and have been made available to us by the Institut für deutsche Sprache, Mannheim.

[5]For annotations we used the Tokeniser of (Schmid 2000), Tree-Tagger described in (Schmid 1994) and (Schmid 1999) and YAC-Chunker (Kermes 2003).

		main clause	subclause
		verb	
verb:			
DE:	*Wenn sie*	*erfahren*,	*dass John Miller große Mengen Alkohol kauft,...*
EN:	"If they"	"find out"	"that John Miller buys much alcohol..."
MWE:			
DE:	*Wenn sie*	*in Erfahrung bringen*,	*dass John Miller große Mengen Alkohol kauft,...*
EN:	"If they"	"find out"	"that John Miller buys much alcohol..."

Table 7.1: *Dass*-clause subcategorised by a verb or a MWE in VL

		main clause			subclause
		verb 1		**verb 2**	
verb:					
DE:	*Es*	*muss*	*heute*	*gesagt werden,*	*dass der Nikolaus ein Türke ist.*
EN:	"It"	"should be"	"today"	"told"	"that Santa Claus is Turk."
MWE:					
DE:	*Es*	*muss*	*heute*	*zur Sprache gebracht werden,*	*dass der Nikolaus ein Türke ist.*
EN:	"It"	"should be"	"today"	"mentioned"	"that Santa Claus is Turk."

Table 7.2: *Dass*-clause subcategorised by a verb or a MWE in passive sentences

	main clause: 1st part noun phrase	**subclause**	**main clause: 2nd part the rest**
DE:	*Die Erklärungsversuche*,	*warum der Teufel sich an X heranmacht*	*sind auf der Glatze gedrehte Locken.*
EN:	"The explanation attempts",	"why the devil chats up X"	"are as futile as giving a bald man a comb."

Table 7.3: *W*-clause subcategorised by a noun in VF

Extraction and Classification Architecture We automatically extract predicates from text corpora classifying them according to their subcategorisation properties. The architecture is based on symbolic procedures, which proceed from the general to the specific, (Lapshinova 2007). First, we apply CQP-queries for extracting all types of predicates in general contexts (for instance verb final sentences and passive constructions). Then we specify CQP-queries to extract different kinds of predicates: verbal, nominal and multiword, which can be further subclassified into the above described specific subtypes according to their subcategorisation features.

An overview of the extraction and classification steps used in this thesis is given in figure 7.1 below. Our algorithm consists of a sequence of procedures to identify and extract different types of predicates, such as verbs, nouns and multiword expressions, as well as classify them according to their subcategorisation features.

1 apply general queries to extract sentences containing predicates
- 1.1 use the VL and passive context for verbal and multiword predicates
- 1.2 for nominal predicates:
- 1.2.1 use the VF context for nominal predicates
- 1.2.2 continue with the step 3

2 apply specific queries to identify predicates
- 2.1 use specific queries for verbal predicates
- 2.2 use specific queries for multiword predicates

3 classify predicates
- 3.1 classify verbal predicates: V1, V2, V3
- 3.2 classify nominal predicates:
- 3.2.1 according to their subcategorisation structure: N1, N2, N3
- 3.2.2 according to their morphological structure: simplex vs. compound

4 compare relations between morphologically related predicates
- 4.1 identify and classify *ung*-nominalisations: (Nung1), (Nung2), (Nung3)
- 4.2 identify, extract and classify base verbs: (Vbase1), (Vbase2), (Vbase3)
- 4.3 classify relations between nominalisations outside and inside a multiword and their base verbs: R1, R2, R3

5 additional procedures
- 5.1 subclassify compound nouns (according to their relations with the head and the non-head): C1, C2, C3-1, C3-2
- 5.2 subclassify multiwords (according to their relations with the nominal constituent): M1, M2, M3, M4

Figure 7.1: Cascade of steps to extract and classify predicates

7.1.4 Results

Extraction and classification of nominals The obtained quantitative and qualitative results show that our tools deliver a substantial number of subclause-taking

nouns (over 15.000 types and over 59.000 tokens[6].), which can be classified according to the type of the clause they subcategorise for. The obtained results allow us to compare the proportion of declarative and interrogative clauses occurring with nouns in VF, revealing that subclause-taking nouns show preferences for *dass*-clauses, cf. table 7.4.

subclause	***dass***	***w-/ob***	**TOTAL**
tokens	40028	19219	59247
in %	67,56	32,44	100,00
context types	10232	5455	15687
in %	65,23	34,77	100,00

Table 7.4: Proportion of *dass* and *w-/ob*-clauses with simplex nouns in the VF

The analysis of the three noun types classified by our tools confirms this tendency. The N1 and N3 nouns (which allow for *dass*-clauses) prove to be most frequent in our corpora. The N2 nominals, which do not allow for *w*-clauses, appear to be not so frequent in the analysed corpora, cf. table 7.5.

class	**N1**	**N2**	**N3**	**TOTAL**
context types	4116	3858	7713	15687
in %	26,24	24,59	49,17	100,00
tokens	36228	5128	17891	59247
in %	61,15	8,65	30,20	100,00

Table 7.5: Proportion of nominal predicate types in our data

Our extraction and classification results for nominal compounds show that the general assumption that the head of a compound acts as its valency bearer has exceptions, cf. the C2 and C3-compound types. There are three types of nominal compound predicates in German based on the subcategorisation relationships between the constituents. Although most compounds belong to type C1, the C2 and C3 compounds make over 30% of all analysed compound cases, which is an unexpectedly considerable amount, cf. table 7.6[7]. The figures also show that in most cases non-head valency bearers of compounds have devrrbal nature, which is expected.

compound	**C1**	**C2**	**C3-1**	**C3-2**	**TOTAL**
types	423	53	131	21	628
in %	67,36	8,44	20,86	3,34	100,00

Table 7.6: Occurrence of C1 to C3 types in VF

[6]The extraction numbers are given for both **tokens** and **types** of predicates. Under **tokens** we understand the number of extracted word forms, whereas **types** indicate the number of query matches for predicates, thus their **context types**. This is especially useful for the study of further context parameters of the extracted predicates.

[7]For compounds, we present extracted lemma types, thus types in their standard meaning.

Extraction and classification of multiwords Our experiments show that certain multiword expressions have their own subcategorisation properties, which are not "inherited" from their nominal elements. With respect to subcategorisation, such multiwords behave like idioms, even though their semantics is not fully idiomatic: the syntactic behaviour is not fully parallel to the semantic distinctions that are known from phraseology. The comparison of the observed occurrences of nominals both, with and without support verbs, allows us to broadly classify the MWE candidates in terms of their preferences for *dass-* and *w-/ob-*clauses, and with respect to the "inheritance" hypothesis. The results of our classification are shown in table 7.7.

multiword	**M1+M2**	**M3+M4**	**TOTAL**
context types	1701	1452	3151
in %	53,98	46,02	
tokens	11687	7787	19474
in %	60,01	39,99	100,00

Table 7.7: The occurrence of M1+M2 vs. M3+M4 classes

Extraction and classification of "inheritance" relations The extraction and classification results obtained within the analysis of the relations between morphologically related predicates show that there are limits to the correspondences or "inheritance" of subcategorisation (e.g. type R2 and R3 relations). Although the obtained figures[8], cf. table 7.8 confirm the generally accepted assumption that subcategorisation properties of deverbal nouns are in most cases taken over from their underlying verbs, there are also cases where the process of "inheritance" is limited to certain complement types, e.g. nominalisation take over *dass*-clauses only.

relations	**R1**	**R2**	**R3**	**TOTAL**
types	72	75	13	160
in %	45,00	46,87	8,13	100,00

Table 7.8: Relations in the most frequent verb-nominalisation pairs in our data

These phenomena are influenced both, by the semantics of predicates and the semantics of the subclauses they subcategorise for, cf. works of (Karttunen 1977), (Bäuerle/Zimmermann 1991), (Schwabe 2004), (Fischer 2005) and (Oppenrieder 2006). Some nominalisations e.g. *Erfahrung* ("experience") or *Vorstellung* ("idea/concept") have different selectional restrictions as their base verbs *erfahren* ("to experience/to find out") and *vorstellen* ("to imagine"), which influences the choice for the complement they take. Furthermore, such contextual parameters, as the occurrence under modal verbs or in negative constructions, can influence the process of "inheritance" of the verbal subcategorisation properties by their derivatives.

Evaluation The evaluation of the single procedures of our extraction and classification architecture shows that our tools deliver sufficient results both in the procedures

[8]The extraction results for subcategorisation relations are given in types in their standard meaning – lemma types, thus 160 different verb-nominalisation pairs.

to identify predicates (see table 7.9) and to subclassify them according to their subcategorisation properties. Precisiona and recall of the classification results depends on the accuracy of our extraction. To achieve higher precision, a set of filtering procedures is required, e.g. the usage of lexical and syntactic restrictions, etc. To achieve higher recall for the acquisition of some predicate types, e.g. verbs, we need to perform additional extractions from other contexts or other corpora types (e.g. parsed corpora or other sentence models), which can increase the recall but, at the same time, also reduce the precision of our results, which is undesired as we aim at precision-oriented extraction.

predicates	**nominal**	**multiword**	**verbal**
precision in %	99,00	81,06	96,10

Table 7.9: Precision results for predicate extraction

7.1.5 Conclusion

The analysis of extracted and classified predicates shows that verbal, nominal and multiword predicates have their own subcategorisation and contextual properties, which should be considered in lexicon acquisition. In the subcategorisation properties of morphologically related predicates, such as verbs and nominalisations, we observe both correspondences, which we call "inheritance", and differences, called "non-inheritance" (including both "inheritance" reducation and "inheritance" extension) of subcategorisation. Searching for the explanation of the "non-inheritance" cases, we state that differences in the subcategorisation behaviour of morphologically related predicates can either be determined by their selectional restrictions (sortal readings of some nominalisations differ from those of their base verbs) or by contextual parameters in which they occur. The analysis of compound nouns and multiword expressions (to a great extent SVCs) show that in some cases subcategorisation properties of the whole construction does not necessarily coincides with the subcategorisation properties of their constituent parts, cf. multiwords of type M3, e.g. *zum Ausdruck kommen* "to be expressed", and M4, e.g. *in Abrede stellen* "to deny", or compounds of type C3-2, such as *Wortspiel* "wordplay". Besides that, for compounds of type C2, e.g. *Erklärungsversuch* "explanation attempt", and C3-1, e.g. *Schlussfolgerung* "conclusion"[9], the commonly accepted assumption that the head of a compound determines its predicate-argument structure is not always valid.

There is need for tools to identify such cases by means of data extraction from corpora, which can be achieved by a precision-oriented semi-automatic extraction and classification elborated within this research. Our method is based on learning subcategorisation properties of predicates from pre-processed corpora. Pre-processing procedures include tokenising, part-of-speech-tagging, lemmatisation, morphological annotation and partially chunking. To design an extraction and classification

[9]Examples of different predicates types extracted and classified by our system are given in the appendix A. Further lists of extracted data will be available at a site on the computing system of IMS (Universität Stuttgart)

architecture, we compile a cascade of procedures, which include both, general and specific queries. General queries are based on word order restrictions that are determined by the features of German, whereas specific queries contain syntactic and lexical restrictions to obtain the targeted predicate types - verbs, nouns and multiwords. Furthermore, for identification of compounds, ung-nominalisations or base verbs, we apply a number of procedures based on morphological analysis. Procedures to subclassify predicates according to their subcategorisation properties include both, specific queries applied on corpora and automatic comparison of the properties of the extracted items. The same procedures are applied to classify subcategorisation relations between verbs and their nominalisations.

As we intend to serve symbolic NLP, especially large symbolic grammars for deep processing, our extraction and classification procedures are precision-oriented. We aim at increasing accuracy, which is opposed to recall, as inaccuracies cause errors in the process of syntactic analysis of sentences. a possible lack in completeness is compensated by the performance of the tools on larger corpora.

7.2 Contributions of the Present Thesis

Within this thesis we provide innovative analysis of different problems, which are important for the areas of linguistics, lexicography and NLP. To assess the contribution of this thesis to these fields, we refer back to the aims presented in the introductory part of this thesis.

7.2.1 Extraction of Subcategorisation for German Predicates

The analysis of the existing studies and approaches shows that most studies concentrate on verbal predicates. However, the information on valency properties of nouns and multiword expressions, as well as relations between subcategorisation features of morphologically related words is also important for linguistic, lexicographic and NLP work. Therefore, in this thesis we analyse not only verbs, but also nouns (both simple, derived and compound) and multiwords by means of automatically extracting them along with their subcategorisation properties from text corpora. This allows us to obtain up-to-date information about the subcategorisation behaviour of German predicates, such as their preferences to take a certain complement type. Only a few dictionaries provide detailed information on these features. besides that, they mostly describe expected subcategorisation features, rather than actually observed ones. Moreover, the description of the relations between subcategorisation features of morphologically related words is mentioned just in a few studies. Our extraction results show that the general assumptions about the behaviour of nominalisations, as well as compound nouns and nominalisations, are not valid for all cases. Nominalisations do not necessarily inherit their valency properties from their base verbs, the heads of nominal compounds do not always function as their valency bearers, and in some cases, nominal constituents of multiword expressions do not determine their valency.

We are able to automatically identify and extract such cases from text corpora. Our tools can operate on pre-processed text corpora. The information automatically

retrieved with our tools can be stored in machine-readable lexicons and updated dynamically. Besides that, our tools provide statistical information about the occurrences of predicates with different complement types, which is important for most NLP applications.

7.2.2 Classification of Predicates in German

The tools developed within this thesis allow us to semi-automatically classify automatically extracted predicates according to their subcategorisation features. In the existing approaches classifications based on subcategorisation criteria are only available for verbal predicates. We are able to classify not only verbal, but also nominal and multiword predicates according to their valency behaviour. Besides that, we provide systematic classification criteria which can be applied for all types of predicates under analysis, which is not present in other studies.

Our classification is based on the types of complements predicates can subcategorise for. The extracted verbs and nouns are classified into three classes: those allowing for both, declarative and interrogative clauses (N1 and V1), those that can take interrogative clauses only (N2 and V2), and those, which license declaratives only (N3 and V3). We focus on sentential complements only, although our methods can be applied for other complement types. This classification enables the systematic description of the subcategorisation behaviour of predicates, of their preferences for *dass-, w-* and *ob*-clauses, i.e. of their selectional restrictions.This can help us to understand the semantics of verbs and nouns better, and in some cases, their subcategorisation properties (e.g. their ability to take *dass*-clauses only) contribute to their disambiguation, as for some cases subcategorisation behaviour of predicates reveals their sortal readings (e.g. facts vs. events), which helps us sometimes to resolve their ambiguities.

The automatic classification of predicates performed by our tools also contributes to lexicography and NLP. In lexicography, the information on the subcategorisation class a predicate belongs to provides information on its features. For NLP, the possibility to predict the subcategorisation behaviour of predicates (relying on the features of the class it belongs to) allows for lexical inference and reduces errors in both, parsing and language generation tasks.

7.2.3 Detection of "Inheritance" Relations and their Classification

Systematicities between subcategorisation properties of morphologically related predicates have not received much attention so far, although these phenomena are important for all areas of language study. We automatically identify subcategorisation relations between morphologically related predicates from corpora. As there are limits to the "inheritance" of subcategorisation (its reduction or extension), we automatically classify "inheritance" relations into three types, cf example (7.8).

This classification is important for the solution of a number of problems in linguistics, lexicography and NLP. For linguistics, it provides information on systematicities of subcategorisation behaviour of verbs and their nominalisations. In lexicography and partially in NLP (creating and updating of computational lexicons), relation

classes provide information on those cases, for which we do not need to spell out subcategorisation information (nominalisations in the R1 relations have the same properties as their verbs). The information that some nominalisations inherit their properties from verbs is default and can be applied, among others, to optimise IE systems.

The phenomenon of subcategorisation "inheritance" in this thesis also includes the analysis of multiwords and compound nouns. They are classified according to the relations between their valency properties and those of their contituents.

Thus, multiwords (consisting of a preposition, noun and a support verb) are classified into four classes: the M1 multiwords inherit subcategorisation properties of their nominal constituents, the M2 multiwords take over the properties of nominal constituents if used under certain contextual conditions only, the M3 and M4 multiwords do not share their subcategorisation properties with those of their nominal constituents. The difference between M3 and M4 lies in their idiomacity. Although the M3 class includes multiwords which are semantically trasparent, their subcategorisation properties are not inherited from their constituents. For instance, the multiword *zum Ausdruck bringen, dass/w-/ob* ("to be expressed that/wh-/if") subcategorises for sentential complements, whereas its nominal constituent *Ausdruck* ("expression") does not. The M4 class includes idioms and those multiwords, which contain 'cranberry' lexemes. This approach allows us to split multiwords into two groups – M3 and M4 are more idiomatic, whereas M1 and M2 contain SVCs whose meaning can be derived from the meaning of their constituent parts, and their subcategorisation properties are also predictable. This shows that there are correlations between compositionality of multiword expressions and their subcategorisation features. More compositional multiwords, e.g. *in Frage stellen, dass/w-/ob* ("to raise to question dass/wh-/if") or *in Erfahrung bringen, w-/ob* ("to find out wh-/if"), inherit their subcategroisation properties from their nominal constituents, e.g. *Frage, dass/w-/ob* ("question that/wh-/if"). Non-compositional multiwords do not inherit their subcategorisation features from their constituents, cf. examples for M3 and M4[10]. This classification approach is important for both, linguistic and lexicographic work, which are concerned with problems of idiom identification. In NLP studies it contributes to approaches for the automatic identification of multiwords.

The analysis of compound nouns shows that the assumption that compounds inherit their subcategorisation properties from their heads does not hold for all cases. Therefore, we elaborate automatic procedures, which classify compounds into four types: those inheriting their valency properties from their heads (C1, e.g. *Journalistenfrage* ("journalist question")), those inheriting their properties from their non-heads (C2, e.g. *Erklärungsversuch* ("explanation attempt")) and those inheriting their properties from both, heads and non-heads (C3-1, e.g. *Schlussfolgerung* ("conclusion")) or from none of them (C3-2, e.g. *Wortspiel* "wordplay"). Knowing that a compound is of type C1 allows us to treat it compositionally, which saves time and effort for listing the subcategorisation properties of the most frequent compounds in a dictionary. A subcategorisation dictionary containing special notes for types C2 and C3 can have an application in language teaching and multilingual NLP.

[10]In some cases, e.g. in M3, compositionality is not always parallel with the "inheritance" of subcategorisation.

7.2.4 Practical Application

The semi-automatic architecture elaborated within this study can find its application for developing dictionaries and NLP lexicons, as well as in teaching, multilingual processing and some other areas of NLP.

The phenomenon of the "inheritance" of certain subcategorisation properties between morphologically related words allows us to systematically describe this process in lexicon construction. If we know for the classified predicates that they take over or inherit their subcategorisation features from the underlying word, we can limit the need of listing all of them in a dictionary or lexicon.

Compounds in lexicon building The treatment of compounds according to the three types described above, limits the need for listing all compounds in an NLP lexicon or a dictionary. On the basis of our automatic classification, we can decide about which compounds to store in the NLP subcategorisation lexicon. Compounds of type C2, e.g. *Erklärungsversuch* ("explanation attempt") and of type C3, e.g. *Schlussfolgerung* ("conclusion") or *Volskmund* ("common speech/vernacular") should be included. There is, however, no need to store compounds of type C1, such as *Journalistenfrage* ("journalist question"). For compounds of this type, a systematic method of matching the arguments of compounds to the subcategorisation frames of their heads, can be used (cf. work of O. Gurevich on methods for mapping deverbal arguments onto those of their corresponding verbs, (Gurevich et al. 2007)). In this case, the matching system would map the subcategorisation frames of the C1 compounds automatically to those of their head nouns, which are already stored in the lexicon.

The system for extraction and classification proposed in this paper is also relevant for building or updating subcategorisation dictionaries for human users. The automatically extracted compound nominals can be classified and added to entries according to the identified valency-bearer noun (cf. figure 7.2):

- Type C1 compounds are listed in alphabetical order, their subcategorisation indications only contain a reference to the respective item of their head.
- Type C2 compounds are listed in alphabetical order, their subcategorisation indications are spelled out and additionally contain a reference to the non-head.
- Type C3 compounds are listed in alphabetical order and contain their own subcategorisation item.

"Inheritance" relations in lexicon building As mentioned above, most dictionaries do not provide systematic correspondences between verbs and nominalisations, as they are described in this work. The classification of the subcategorisation relations between verbs and their nominalisations limits the need to describe the predicate-argument structure of most nominalisations. We can just rewrite this structure from that of the underlying verb. However, the existing non-correspondences between the subcategorisation of nominalisations and their base verbs should also be taken into account. Thus, it is necessary to classify verbs and their derived nominals into the three relation types (from R1 to R3, see example (7.8)) described above.

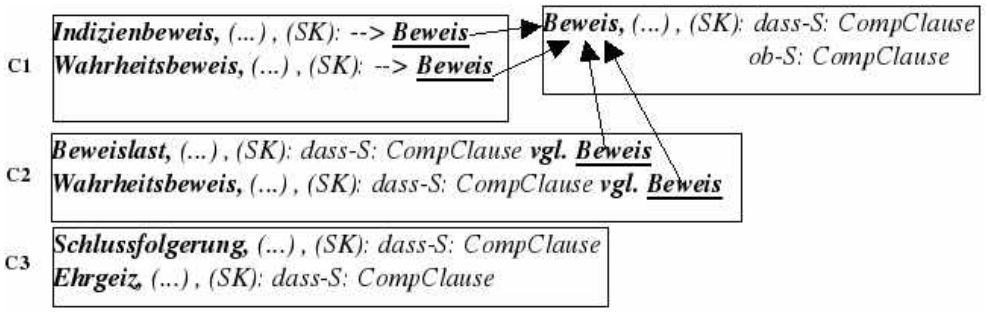


Figure 7.2: Examples of lexicon entries for C1 to C3 compounds

We can use the described classification system for dictionary construction in the following way. List nominalisations with their subcategorisation indications:

1 for R1, e.g. *begründen, dass/w-/ob - Begründung, dass/w-/ob*
("justify that/wh-/if - justification that/wh-/if"), the subcategorisation indications contain references to the subcategorisation of the base verbs.

2 for R2, e.g. *befürchten, dass/w-/ob - Befürchtung, dass*
("to fear that/wh-/if - feat that/wh-/if"), the subcategorisation indications contain references to the subcategorisation of the base verbs and a note about the loss of certain properties.

3 for R3, e.g. *aufklären, w-/ob - Aufklärung, dass/w-/ob*
("to clarify wh-/if - clarification that/wh-/if") the subcategorisation indications contain references to the subcategorisation of the base verbs and a note about additional properties that the verb does not have.

Predicates in further applications Information on subcategorisation properties of predicates, as well as their relations can also be used in multilingual processes, e.g. language teaching or translation, or in applications for IE optimisation.

For instance, if a dictionary contains special notes for types C2 and C3, a user can differentiate these cases from the "inheritance" cases of type C1, as known from the general rule, which is important in both the process of language learning and the translation of compounds without a loss of information. The information about the similarities or non-similarities between the structures of verbs and nominalisations can also be applied in the process of tarnslation, e.g. when an equivalent element of the same word class is missing in the target language. In this case it can be substituted by the morphologically related word and the transfer of its subcategorisation structure allows for the construing of grammatically correct sentences.

As already mentioned above, some successful IE techniques are developed around a set of domain relevant linguistic patterns, based on select predicates. The information on the "inheritance" relations can reduce the number of applied patterns, as some deverbals inherit their properties completely from their base verbs. This means that when searching for a nominalisation or a phrase that contains a nominalisation,

the search engine can also deliver results containing the base verbs, whose predicate-argument structures correspond to those of the nominalisation.

7.3 Directions for Future work

There are various directions for future research, referring to different aspects of this thesis. In the following, we summarise the possible ideas for future work.

Extension of extraction procedures The procedures to extract predicates might be extended in order to include further contexts of predicate occurrences. For instance, we do not take into account active V1 and V2 sentence models, which can be used in the extraction of verbs and multiword expressions. Moreover, in this thesis nominal predicates are extracted only from the Vorfeld. The context for their extraction can be extended to include V1 and V2 sentence models. For this purpose, further refinements of the queries are required. This means that more restrictions should be included to avoid possible noise and to achieve high precision results. Furthermore, extraction procedures might include contexts in which subcategorised subclauses are shifted to the sentence start (topicalised). Besides that, to achieve a higher recall, use could be made of parsing-based extraction procedures, as mentioned in section 5.1.3. The architecture described for flat structures in this thesis, can also be used on parsed corpora. High precision can be achieved with the help of the same set of filtering procedures described in section 5.3.1.4.

Extension of predicate types This thesis only analyses verbal, nominal and multiword predicates. However, the tool can be extended in order to extract and classify adjectives. Additionally, in the analysis of multiwords, we only take into account those containing a preposition, noun and support verbs. However further SVCs may need to be classified, for example those consisting of a noun and a support verb.

Extension of complement types We concentrated on the extraction and classification of predicates, which occur with subclauses. Our method can be also applied to the analysis of further complement types, for instance prepositional clauses or *zu*-infinitives.

The analysis of further parameters For extracted predicates, further morpho-syntactic parameters can be analysed. For instance, the extraction architecture provides information on contextual parameters of predicates, such as embedding under modal verbs, occurrence in negated contexts and others. To explain some features of the subcategorisation behaviour of predicates, a deeper analysis of the mentioned contextual parameters is required.

Extension of predicate classes Our classification of predicates according to their subcategorisation features can also be extended and might include further subclasses according to the morpho-syntactic parameters of predicates. For instance, for verbal predicates subclassification might include classes of verbs allowing for interrogative

or declarative clauses only under certain contextual conditions, if used with a modal verb or in a negated context only, e.g. the verb *zweifeln* (“to doubt”, if negated, can take *dass*-clauses only, cf. (Fischer 2005). Such subclassification might contribute to the explanation of predicate behaviour, as well as to the prediction of the context partners of the analysed predicates.

NLP applications The developed extraction and classification architecture might be used within NLP applications, for instance in lexicon construction or updating as well as optimisation of IE applications. This can prove the usefulness of the presented approach.

Appendix A

Examples of Predicates

A.1 *W*-words in German

German	English
wann	”when”
warum	”why”
was	”what”
was für (ein) + noun	” what (a)”
welche(s/r)	”which”
wer	”who”
wem	”who” (dative)
wen	”who” (accusative)
wessen	”whose”
weshalb	”why”
wie	”how”
wie lange	”how long”
wie oft	”how often”
wieso	”why”
wieviel	”how much”
wo	”where”
wobei	”at what”
wofür	”for what”
wogegen	”against what”
woher	”wherefrom”
wohin	”whereto”
woran	”whereof”
womit	”wherewith”
wonach	”after what”
wozu	”why”

A.2 Examples of Nominal Predicates: N1, N2, N3

This list excludes examples of *ung*-nominalisations which are listed below.

N1	**N2**	**N3**
Angst	*Abfrage*	*Alibi*
Annahme	*Absicht*	*Alternative*
Ansicht	*Abwägen*	*Bedauaern*
Anzeichen	*Anfrage*	*Beispiel*
Argument	*Anlaß*	*Bekenntnis*
Aussage	*Auskunft*	*Beleg*
Beweis		*Bericht*
Diskussion	*Auswahl*	*Beschwerde*
Einwand	*Befragen*	*Bewußsein*
Entscheid	*Beschluß*	*Chance*
Erkenntnis	*Beweggrund*	*Bitte*
Fall	*Bilanziehen*	*Eindruck*
Frage	*Detail*	*Einsicht*
Gedanke	*Eifersüchtelei*	*Furcht*
Glaube	*Einzelheit*	*Garantie*
Grund	*Gleichgültigkeit*	*Gebot*
Hinweis	*Grenzwert*	*Gefahr*
Idee	*Klarheit*	*Gefühl*
Information	*Konflikt*	*Gejammer*
Kritik	*Kontroverse*	*Gewissheit*
Motto	*Motiv*	*Grundsatz*
Nachricht	*Nachdenken*	*Illusion*
Nachweis	*Nagelprobe*	*Indikator*
Plan	*Rätsels*	*Indiz*
Prognose	*Überblick*	*Klage*
Risiko	*Verhältnis*	*Möglichkeit*
Streit	*Zeitpunkt*	*Schluß*

A.3 Examples of compound nouns: C2 and C3

C2	C3-1	C3-2
Argumentationsblock	Denkmodell	Anhaltspunkt
Argumentationskette	Grundsatz	Angelpunkt
Bedenkzeit	Meinugsstreit	Ehrgeiz
Beweisführung	Rätselraten	Gewährsmann
Beweismittel	Rechtsanspruch	Naserümpfen
Denkumweg	Schlüsselkenntnis	Schlagzeile
Druckmittel	Schlussfolgerung	Sehnsucht
Erfahrungssatz	Streitfrage	Volksmund
Erläuterungsbemühen	Verfahrensweise	Wortspiel
Glaubenssatz	Warnsignale	
Hoffnungsschimmer	Werbemythos	
Motivsuche	Wettstreit	
Rechtslage	Wunschtraum	
Schreckgespenst	Wunschvorstellung	
Ursachenpotential	Zeitpunkt	

A.4 Sample Multiwords of Type M1 and M2

multiword	complement	type
(jmdn) zu der Annahme fhren/kommen "to (make smb) assume/believe"	*dass*	M1
zu der Ansicht gelangen/kommen "to take the view that"	*dass*	M1
in Anspruch nehmen "to make use of"	*dass*	M1
zu d Auffassung gelangen/kommen "to take the line"	*dass*	M1
jmdm zur Auflage machen "to impose on sb. as a condition"	*dass*	M1
in Auftrag geben "to comission"	*dass*	M1
in Aussicht stellen "to hold out"	*dass*	M1
zur Bedingung machen "to make it a condition"	*dass*	M1
in Berechnung bringen "to bring into calculation"	*dass*	M1
unter Beweis stellen "to give a proof"	*dass/w-/ob*	M1
ins Bewußtsein bringen/kommen "to dawn"	*dass*	M1
zu(m)/zu dem Bewußtsein kommen "to become conscious"	*dass*	M1
jmdm. zu Bewußtsein kommen	*w-*	M1
jmdm. ins Bewußtsein treten "to sink in"	*dass*	M1
zur Diskussion stehen "to be under consideration"	*w-*	M1
unter Druck setzen "to pout under pressure"	*dass*	M1
zu der Einschätzung gelangen/kommen "to come to the appraisal"	*dass*	M1
jmdn zur Einsicht bringen/gelagen/kommen "to (make sb.) see reason"	*dass*	M1
zur Entscheidung kommen "to reach decision"	*dass*	M1
zum Entschluß kommen "to come to the decision"	*dass*	M1
jmdn. in Erstaunen setzen "to astonish"	*dass*	M1
in Erfahrung bringen "to find out"	*w-*	M2

A.5 Sample Multiwords of Type M3 and M4

multiword	complement	type	note
in Abrede stellen	*dass*	M4	**c**
"to deny"			
zu Ansatzpunkten gelangen	*ob*	M3	
"to come to the starting point"			
jmd. ins Auge fallen	*dass*	M4	**i**
"to strike the eye"			
wie Schuppen von den Augen fallen	*dass*	M4	**i**
"to fall like scales from eyes"			
vor die Augen treten	*dass*	M4	**i**
"to visualise/clarify"			
jmdm/sich vor Augen führen/halten	*dass/w-/ob*	M4	**i**
"to visualise/clarify"			
zum Ausdruck bringen/kommen	*dass/w-/ob*	M3	
"to be expressed"			
an den Beginn stellen	*dass*	M4	**i**
"to put to the start"			
auf die Beine stellen	*dass*	M4	**i**
"to achieve"			
außer Betracht bleiben	*dass*	M4	**c**
"to be let our of the consideration"			
in Betracht kommen/ziehen	*dass*	M4	**c**
"to come/take into consideration"			
jdn ins Bild setzen	*dass/w-/ob*	M4	**i**
"to clue sb. in on sth"			
aus dem Blick geraten	*dass*	M3	
"to pass from view"			
in den Blick geraten	*dass*	M4	**i**
"to com to view"			
im Dunkeln halten	*dass*	M4	**i**
"to keep untold/unknown"			
zu Eigen machen	*dass*	M4	**i**
"to make one's own"			
in Einklang bringen	*dass*	M4	**i**
"to bring in line"			
ins Feld führen	*dass*	M4	**i**
"to invoke"			
jmdm. den Floh ins Ohr setzen	*dass*	M4	**i**
"to put a bug in sb's ear"			
in Gang kommen/setzen	*dass/w-/ob*	M4	**i**
"to get started/start"			
im Gedächtnis (hängen) bleiben	*dass*	M3	
"to stay in mind"			
ins Gewicht fallen	*ob*	M4	**i**
"to carry weight"			

c indicates that the multiword contains a cranberry lexeme;
i indicates that the multiword is idiomatic.

A.6 Examples of *ung*-nominalisations

Nung1	**Nung2**	**Nung3**
Ahnung	*Abklärung*	*Abmachung*
Anmerkung	*Abstimmung*	*Absicherung*
Auffassung	*Abwägung*	*Andeutung*
Aufklärung	*Anleitung*	*Anerkennung*
Aufregung	*Anstrengung*	*Anfechtung*
Auslegung	*Aufführung*	*Anforderung*
Äßerung	*Aufrüstung*	*Ankündigung*
Begründung	*Aufstellung*	*Aufforderung*
Behauptung	*Auseinandersetzung*	*Bedingung*
Bemerkung	*Auslosung*	*Befürchtung*
Beobachtung	*Ausscheidung*	*Begeisterung*
Berechnung	*Auswertung*	*Beglaubigung*
Bestimmung	*Bedingung*	*Behautung*
Betrachtung	*Befragung*	*Bekräftigung*
Beurteilung	*Beratung*	*Bekundung*
Bewertung	*Einstufung*	*Belehrung*
Darstellung	*Erkundigung*	*Bemühung*
Deutung	*Erleuchtung*	*Bescheinigung*
Einigung	*Erörterung*	*Beschuldigung*
Einschätzung	*Fragestellung*	*Bestätigung*
Einstellung	*Klärung*	*Betonung*
Empfehlung	*Konkretisierung*	*Benachrichtigung*
Empörung	*Nachforschung*	*Berücksichtigung*
Entscheidung	*Prüfung*	*Berichtigung*
Entschuldigung	*Präzisierung*	*Bezichtigung*
Erklärung	*Rückmeldung*	*Darlegung*
Ermittlung	*Regung*	*Differenzierung*
Erwägung	*Richtung*	*Drohung*
Erwiderung	*Stellung*	*Einbildung*
Festlegung	*Unterscheidung*	*Einlassung*
Forderung	*Verbesserung*	*Einschränkung*
Formulierung	*Verfassung*	*Einteilung*
Klarstellung	*Verunsicherung*	*Einwendung*
Lösung	*Verwirrung*	*Empfindung*
Meinung	*Weichenstellung*	*Entdekung*
Mitteilung	*Überprüfung*	*Entgegnung*
Mutmassung		*Enthüllung*
Neigung		*Enttäuschung*
Planung		*Erfahrung*
Rechnung		*Eröffnung*

A.7 Examples of base verbs

Vbase1	Vbase2	Vbase3
abmachen	*abfragen*	*berichtigen*
absichern	*abgrenzen*	*beteuern*
achten	*abklären*	*einbilden*
ändern	*abrüsten*	*einlassen*
äussern	*absprechen*	*vermögen*
ahnen	*abstimmen*	
andeuten	*abwägen*	
anerkennen	*anfechten*	
ankündigen	*anfordern*	
anmerken	*anfragen*	
annehmen	*angeben*	
anordnen	*anmassen*	
anregen	*anreden*	
ansehen	*ansagen*	
antworten	*anschauen*	
arbeiten	*ansetzen*	
auffallen	*anspielen*	
auffordern	*ansprechen*	
aufstellen	*anstrengen*	
ausführen	*anweisen*	
auslegen	*auffassen*	
beantworten	*aufführen*	
bedenken	*aufgeben*	
bedeuten	*aufklären*	
befürchten	*auflegen*	
befürworten	*aufstocken*	
begründen	*auseinandersetzen*	
behaupten	*auskommen*	
bekanntgeben	*auslaufen*	
bekräftigen	*auslosen*	
bekunden	*ausnehmen*	
belehren	*ausreden*	
bemerken	*ausschreiben*	
bemühen	*ausstellen*	
beobachten	*bedingen*	
berechnen	*befragen*	
berücksichtigen	*befriedigen*	
bescheinigen	*behandeln*	
beschränken	*beherrschen*	
beschwören	*belohnen*	
besinnen	*benutzen*	
bestätigen	*beraten*	
bestimmen	*beschwichtigen*	
bestimmen	*bestrafen*	
bestreiten	*besuchen*	

A.8 Examples of the R1 relations

base verb	***ung*-nominalisation**
abgrenzen , w-/ob	*Abgrenzung , w-/ob*
abklären , w-/ob	*Abklärung , w-/ob*
abrüsten , w-/ob	*Abrüstung , w-/ob*
abstimmen , w-/ob	*Abstimmung , w-/ob*
abwägen , w-/ob	*Abwägung , w-/ob*
äussern , dass/w-/ob	*Äusserung , dass/w-/ob*
ahnen , dass/w-/ob	*Ahnung , dass/w-/ob*
anmerken , dass/w-/ob	*Anmerkung , dass/w-/ob*
anstrengen , w-/ob	*Anstrengung , w-/ob*
aufführen , w-/ob	*Aufführung , w-/ob*
auseinandersetzen , w-/ob	*Auseinandersetzung , w-/ob*
auslegen , dass/w-/ob	*Auslegung , dass/w-/ob*
auslosen , w-/ob	*Auslosung , w-/ob*
bedingen , w-/ob	*Bedingung , w-/ob*
befragen , w-/ob	*Befragung , w-/ob*
begründen , dass/w-/ob	*Begründung , dass/w-/ob*
behaupten , dass/w-/ob	*Behauptung , dass/w-/ob*
bemerken , dass/w-/ob	*Bemerkung , dass/w-/ob*
beobachten , dass/w-/ob	*Beobachtung , dass/w-/ob*
beraten , w-/ob	*Beratung , w-/ob*
berechnen , dass/w-/ob	*Berechnung , dass/w-/ob*
berichtigen , dass	*Berichtigung , dass*
bestimmen , dass/w-/ob	*Bestimmung , dass/w-/ob*
beteürn , dass	*Beteürung , dass*
betrachten , dass/w-/ob	*Betrachtung , dass/w-/ob*
bewerten , dass/w-/ob	*Bewertung , dass/w-/ob*
darstellen , dass/w-/ob	*Darstellung , dass/w-/ob*
einbilden , dass	*Einbildung , dass*
einlassen , dass	*Einlassung , dass*
einstellen , dass/w-/ob	*Einstellung , dass/w-/ob*
einstufen , w-/ob	*Einstufung , w-/ob*
empfehlen , dass/w-/ob	*Empfehlung , dass/w-/ob*
entfesseln , w-/ob	*Entfesselung , w-/ob*
entlassen , w-/ob	*Entlassung , w-/ob*
entscheiden , dass/w-/ob	*Entscheidung , dass/w-/ob*
entwickeln , dass/w-/ob	*Entwicklung , dass/w-/ob*
erklären , dass/w-/ob	*Erklärung , dass/w-/ob*
erleichtern , w-/ob	*Erleuchtung , w-/ob*
ermitteln , dass/w-/ob	*Ermittlung , dass/w-/ob*
ernähren , w-/ob	*Ernährung , w-/ob*
erörtern , w-/ob	*Erörterung , w-/ob*
erwidern , dass/w-/ob	*Erwiderung , dass/w-/ob*

A.9 Examples of the R2 relations

base verb	***ung*-nominalisation**
abmachen , dass/w-/ob	*Abmachung , dass*
absichern , dass/w-/ob	*Absicherung , dass*
achten , dass/w-/ob	*Achtung , w-/ob*
ändern , dass/w-/ob	*Änderung , w-/ob*
andeuten , dass/w-/ob	*Andeutung , dass*
anerkennen , dass/w-/ob	*Anerkennung , dass*
ankündigen , dass/w-/ob	*Ankündigung , dass*
anordnen , dass/w-/ob	*Anordnung , dass*
anregen , dass/w-/ob	*Anregung , dass*
auffordern , dass/w-/ob	*Aufforderung , dass*
aufstellen , dass/w-/ob	*Aufstellung , w-/ob*
befürchten , dass/w-/ob	*Befürchtung , dass*
behaupten , dass/w-/ob	*Behautung , dass*
bekräftigen , dass/w-/ob	*Bekräftigung , dass*
bekunden , dass/w-/ob	*Bekundung , dass*
belehren , dass/w-/ob	*Belehrung , dass*
berücksichtigen , dass/w-/ob	*Berücksichtigung , dass*
bescheinigen , dass/w-/ob	*Bescheinigung , dass*
beschränken , dass/w-/ob	*Beschränkung , dass*
beschwören , dass/w-/ob	*Beschwörung , dass*
bestätigen , dass/w-/ob	*Bestätigung , dass*
betonen , dass/w-/ob	*Betonung , dass*
bewegen , dass/w-/ob	*Bewegung , w-/ob*
darlegen , dass/w-/ob	*Darlegung , dass*
drohen , dass/w-/ob	*Drohung , dass*
einschränken , dass/w-/ob	*Einschränkung , dass*
einwenden , dass/w-/ob	*Einwendung , dass*
entdecken , dass/w-/ob	*Entdekung , dass*
entgegnen , dass/w-/ob	*Entgegnung , dass*
enthüllen , dass/w-/ob	*Enthüllung , dass*
erfahren , dass/w-/ob	*Erfahrung , dass*
erheben , dass/w-/ob	*Erhebung , w-/ob*
erinnern , dass/w-/ob	*Erinnerung , dass*
erkundigen , dass/w-/ob	*Erkundigung , w-/ob*
eröffnen , dass/w-/ob	*Eröffnung , dass*
erwähnen , dass/w-/ob	*Erwähnung , dass*
erzählen , dass/w-/ob	*Erzählung , dass*
festschreiben , dass/w-/ob	*Festschreibung , dass*
folgen , dass/w-/ob	*Folgerung , dass*
halten , dass/w-/ob	*Haltung , dass*
hoffen , dass/w-/ob	*Hoffung , dass*
kennzeichnen , dass/w-/ob	*Kennzeichnung , dass*
klären , dass/w-/ob	*Klärung , w-/ob*
leugnen , dass/w-/ob	*Leugnung , dass*
mahnen , dass/w-/ob	*Mahnung , dass*

A.10 Examples of the R3 relations

base verb	***ung*-nominalisation**
auffassen , w-/ob	*Auffassung , dass/w-/ob*
aufklären , w-/ob	*Aufklärung , dass/w-/ob*
beurteilen , w-/ob	*Beurteilung , dass/w-/ob*
deuten , w-/ob	*Deutung , dass/w-/ob*
einschätzen , w-/ob	*Einschätzung , dass/w-/ob*
empören , w-/ob	*Empörung , dass/w-/ob*
erwägen , w-/ob	*Erwägung , dass/w-/ob*
spannen , w-/ob	*Spannung , dass/w-/ob*
überwachen , w-/ob	*Überwachung , dass/w-/ob*
verifizieren , w-/ob	*Verifizierung , dass/w-/ob*
verlautbaren , w-/ob	*Verlautbarung , dass/w-/ob*
verordnen , w-/ob	*Verordnung , dass/w-/ob*
verstimmen , w-/ob	*Verstimmung , dass/w-/ob*

Appendix B

Annotations and the CQP language

B.1 STTS tagset

pos	description	examples
ADJA	attributive adjective	*das große Haus*
ADJD	adverbial or predicative adjective	*er fährt/ist schnell*
ADV	adverb	*schon, bald, doch*
APPR	preposition, ambiposition before	*links in der Stadt, ohne mich*
APPRART	preposition with an article	*in+dem=im, zu+der=zur*
APPO	pospsition	*ihm zufolge*
APZR	circomposition	*von jetzt an*
ART	article (definite and indefinite)	*der, die, das, ein, eine*
CARD	numerals	*zwei Manner, im Jahre 1994*
FM	foreign language material	*Er hat das mit 'A big fish' übersetzt*
ITJ	interjection	*mhm, ach, tja*
KOUI	subordinating conjunction with *zu* and infinitive	*um zu leben*
KOUS	subordinating conjunction with a sentence	*weil, wann, dass, ob*
KON	coordinating conjunctions	*und, oder, aber*
KOKOM	comparison without a sentence	*als, wie*
NN	common noun	*Tisch, Herr, das Reisen*
NE	proper nouns	*Hans, Hamburg, HSV*
PDS	substituting demostrative pronoun	*dieser, jener*
PDAT	attributive demonstrative	*jener Mensch*
PIS	substituting indefinite pronoun	*keiner, viele, man, niemand*
PIAT	attributive indefinite pronoun without a determiner	*kein Mensch*
PIDAT	attributive indefinite pronoun with a determiner	*ein wenig Wasser*
PPER	irreflexive personal pronoun	*ich, er, ihm, mich, dir*
PPOSS	substituting possessive	*meins, deiner*
PPOSAT	attributive possessive	*mein Buch, deine Mutter*
PRELS	substiting relative pronoun	*der Hund, der*
PRELAT	attributive relative pronoun	*der Mann, dessen Hund*
PRF	reflexive personal pronoun	*sich, einander, dich, mir*
PWS	substituting interrogative pronoun	*wer, was*
PWAT	attributive interrogative pronoun	*welche Farbe*
PWAV	adverbial interrogative or relative pronoun	*warum, wo, wann*
PAV	pronominal adverb	*dafür, dabei, deswegen*

pos	description	examples
PTKZU	*zu* before an infinitive	*zu gehen*
PTKNEG	negation particle	*nicht*
PTKVZ	separable verb parts	*er kommt an*
PTKANT	answer words	*ja, nein, danke, bitte*
PTKA	adjective or adverb particles	*am schonsten*, *zu schnell*
TRUNC	compound non-head	*An- und Abreise*
VVFIN	finite full verb	*(du) gehst, (wir) kommen*
VVIMP	imperative, full	*Komm!*
VVINF	infinitive, full	*gehen, ankommen*
VVIZU	infinitive with *zu*, full	*anzukommen, loszulassen*
VVPP	perfect participle, full	*gegangen, angekommen*
VAFIN	finite auxiliary verb	*(du) bist, (wir) werden*
VAIMP	imperative, aux	*Sei ruhig!*
VAINF	infinitive, aux	*werden, sein*
VAPP	perfect participle, aux	*gewesen*
VMFIN	finite modal verb	*durfen*
VMINF	infinitive, modal	*wollen*
VMPP	perfect partcicple, modal	*(er hat) gekonnt*
XY	non-words with special characters	D2XW3
$,	comma	,
$.	sentence end punctuation	.?!;:
$(	other sentence punctuation	] ( )

B.2 The CQP regular expressions syntax

Basic Syntax of Regular Expressions

features and examples of basic regular expressions	
letters and digits are matched literally (including all non-ASCII characters)	word → *word*; C3PO → *C3PO*
. matches any single character (“matchall”)	r.ng → *ring, rung, rang, rkng, r3ng,...*
character set: [...] matches any of the characters listed	moderni[sz]e → *modernise, modernize*; [a-c5-9] → *a, b, c, 5, 6, 7, 8, 9*
repetition of the preceding element (character or group)	? (0 or 1); * (0 or more); + (1 or more); n (exactly n); n,m (n ... m); colou?r → *color, colour*; go2,4d → *good, goood, goood*; [A-Z][a-z]+ → “regular” capitalised word VV.* → matches any full verb
grouping with parentheses: (...)	*(bla)+* → *bla, blabla, blablabla,...*; (school)?bus(es)? → *bus, buses, schoolbus*
| separates alternatives (use parentheses to limit scope)	mouse|mice →*mouse, mice*; corp(us|ora) → *corpus, corpora*

Complex Regular Expressions

Complex regular expressions can be used to model (regular) inflection:	
ask(s|ed|ing)?	→ *ask, asks, asked, asking*
(equivalent to the less compact expression ask|asks|asked|asking)	
sa(y(s|ing)?|id)	→ *say, says, saying, said*
[a-z]+i[sz](e[sd]?|ing)	→ any form of a verb with *-ise* or *-ize* suffix

XML Elements Representing Syntactic Structure

<s>	sentences
<pp>	prepositional phrases
<np>	noun phrases
<ap>	adjectival phrases
<advp>	adverbial phrases
<vc>	verbal complexes
<cl>	subclauses
<NN>	common noun
<V>	verb
etc	

Key-value Pairs in XML Start Tags

<s len="..">
<pp f=".." h=".." agr=".." len="..">
<np f=".." h=".." agr=".." len="..">
<ap f=".." h=".." agr=".." len="..">
<advp f=".." len="..">
<vc f=".." len="..">
<cl f=".." h=".." vlem=".." len="..">
len = length of region (in tokens)
f = properties (feature set, see next page)
h = lexical head of phrase (<pp h>: "prep:noun")
agr = nominal agreement features (feature set, partially disambiguated)
vlem = lemma of main verb

Properties of Syntactic Structures (f-Key in Start Tags)

<np f>	**norm** ("normal" NP), **ne** (named entity), **rel** (relative pronoun), **wh** (*wh*-pronoun), **pron** (pronoun), **refl** (reflexive pronoun), **es** (*es*), **sich** (*sich*), **nodet** (no determiner), **quot** (in quotes), **brac** (in parentheses), **numb** (list item), **trunc** (contains truncated nouns), **card** (cardinal number), **date** (date string), **year** (specifies year), **temp** (temporal), **meas** (measure noun), **street** (address), **tel** (telephone number), **news** (news agency)
<pp f>	same as <np f> (features are projected from NP) + **nogen** (no genitive modifier)
<ap f>	**norm** ("normal" AP), **pred** (predicative AP), **invar** (invariant adjective), **vder** (deverbal adjective), **quot** (in quotes), **pp** (contains PP complement), **hypo** (uncertain, AP was conjectured by chunker)
<advp f>	**norm**, **temp** (temporal adverbial), **loc** (locative adverbial), **dirfrom** (directional source), **dirto** (directional path)
<vc f>	**norm**, inf (infinitive), **zu** (*zu*-infinitive)
<cl f>	**rel** (relative clause), **subord** (subordinate clause), **fin** (finite), **inf** (infinitive), **comp** (comparative clause)

Zusammenfassung

Einleitung

Die vorliegende Arbeit beschreibt einen Ansatz zur semi-automatischen Analyse von deutschen Prädikaten. Verben, Nomina und Mehrwortausdrücke (MWAs) werden automatisch aus den Copora extrahiert und nach ihren Valenzeigenschaften klassifiziert. In dieser Arbeit berücksichtigen wir nur satzförmige Komplemente, obwohl diese Methode für die Extraktion weiterer Komplementtypen geeignet ist. Neben der subkategorisierungsbasierten Klassifikation wollen wir auch die Eigenschaften morphologisch verwandter Prädikate (e.g. Verben und ihrer Nominalisierungen) vergleichen. In den meisten Ansätzen wird generell angenommen, dass Nominalisierungen ihre Valenzeigenschaften von den Basisverben übernehmen oder erben. Dennoch zeigen unsere Extraktionsexperimente, dass diese Annahme nicht immer stimmt. Deswegen befaßt sich diese Arbeit mit dem Vergleich der Valenzeigenschaften von Verben und Nominalisierungen und der Analyse ihrer Übereinstimmungen und Unterschiede.

Dafür entwerfen wir ein semi-atomatisches Verfahren zur Extraktion und Klassifikation der Valenzeigenschaften deutscher Prädikate, sowie der Relationen zwischen Valenzeigenschaften von Verben und ihren Nominalisierungen. Die extrahierten Daten können für symbolische NLP-Systeme angewendet werden, besonders für die symbolischen Grammatiktheorien LFG und HPSG[1]. Ausführliche lexikalische Informationen sind für diese Grammatiken sehr wichtig. Außerdem sind Informationen über Subkategorisierung für Linguistik, Lexikographie, sowie multilinguale Ansätze, z.B. Fremdsprachenunterricht oder Übersetzungen, notwendig.

Unser Ziel ist höhere Präzision der Extraktions- und Klassifikationsergebnisse zu erreichen. Somit wird ihre Vollständigkeit vernachlässigt, was wir durch die Anzahl der verwendeten Corpora ausgleichen wollen.

Daten und ihre Beschreibung

Wie bereits angedeutet, werden in dieser Arbeit Verben, Nomina und Mehrwortausdrücke analysiert. Die meisten bisherigen Arbeiten in diesem Forschungsgebiet befassen sich mit Verbvalenz. Dennoch gibt es auch Ansätze, die die Valenz weiterer Prädikate beschreiben. Tabelle 1 zeigt eine Übersicht linguistischer, lexikographischer und NLP-Arbeiten, die sich mit der Subkategorisierung beschäftigen.

[1] vgl. die Arbeiten von (Copestake et al. 2004) und (Butt et al. 2002)

Prädikate	Linguistik	Lexikographie	NLP
Verben	(Tesnière 1980), (Engel 1988), (Engel 1994), (Engel 1996), (Agel 2000), (Agel 2003), (Götz-Votteler 2007)	(Helbig/Schenkel 1969), (Engel/Schumacher 1976), (Schumacher 1986), (Herbst et al. 2004), (Schumacher 2004), *ELDIT*, *ADNW*, *DAFLES*	HPSG (Pollard/Sag 1994), LFG in (Kaplan/Bresnan 1982), (Bresnan 2001) and (Dalrymple 2001), COMLEX, (Merlo/Stevenson 2001)
Nomina	(Teubert 1979), (Teubert 2003), (Ehrich 1991), (Agel 2003), (Schierholz 2005),	(Sommerfeldt/Schreiber 1983), (Sommerfeldt/Schreiber 1996), (Herbst et al. 2004)	(Crouch *et al* 2006), (Gurevich et al. 2007), NOMLEX, etc.
MWA	(Krenn/Erbach 1994)	(Fellbaum et al 2006), (Heid 2006)	(Bartsch 2004), (Storrer 2007), (Lapshinova/Heid 2007)

Tabelle 1: Prädikate in verschiedenen Ansätzen

Auch auf die Subkategorisierungseigenschaften morphologisch verwandter Wörter wird in der linguistischen, lexikographischen und NLP-Literatur eingegangen. Diese Problematik wird in (Grimshaw 1990), (Nunes 1993), (Sommerfeldt/Schreiber 1996), (Schierholz 2001), (Meinschaefer 2004) untersucht. Systematische Beschreibung dieser Phänomene findet man allerdings nur in wenigen lexikalischen Systemen, z.B. in NOMLEX[2] oder im PARC-SYSTEM das in (Gurevich et al. 2007) beschrieben wird.

Unsere Extraktionsergebnisse weisen auf, dass zwischen den Valenzeigenschaften morphologisch verwandter Prädikate sowohl Übereinstimmungen ("inheritance" = Vererbung), als auch Unterschiede ("non-inheritance" = Nichtvererbung) existieren. In den meisten Fällen übernehmen deverbale Nomina ihre Valenzeigenschaften von den Basisverben (Beispiel (1)).

(1) – *befürchten, dass* vs. *Befürchtung, dass*;
– *erklären, dass/w-* vs. *Erklärung, dass/w-*.

Weitere Beispiele zeigen, dass die Subkategorisierungseigenschaften der Nomina sich ebenfalls von den Eigenschaften ihrer zugrundeliegenden Verben unterscheiden können. In manchen Fällen gehen einige Eigenschaften verloren ("inheritance" reduction = Vererbungsreduktion), in anderen Fällen gewinnen Nominalisierungen weitere Eigenschaften, die mit ihren Basisverben nie vorkommen ("inheritance" extension = Vererbungserweiterung), hinzu vgl. (2) und (3).

(2) Reduktion der Subkategorisierungsvererbung:

– *vermuten, dass/w-* vs. *die Vermutung, dass/*w-*;
– *wissen, dass/w-/ob* vs. *das Wissen, dass/*w-/*ob*.

(3) Erweiterung der Subkategorisierugsvererbung:

– *Antwort an* vs. *antworten jmdm*;
– *der Verdacht auf* vs. *verdächtigen jmdn*.

[2]vgl. z.B. (Macleod *et al.* 1998b).

Klassifikation der Prädikate und ihrer Relationen

Umfassende Untersuchungen zur Klassifikation verbaler Prädikate finden sich u.a. in (Vendler 1967), (Levin 1993), (Fischer 2005), (Schulte im Walde 2006), (Klotz 2007), für Substantive und Mehrwortausdrücke seien vor allem (Sommerfeldt/Schreiber 1983), (Ehrich 1991), (Teubert 2003) und (Storrer 2007) genannt. Jedoch bietet keiner dieser Ansätze eine systematische Klassifikation aller Prädikatarten nach ihren Subkategorisierungseigenschaften. In der vorliegenden Arbeit werden Verben und Substantive nach gleichen Kriterien in drei Klassen eingeteilt – nach ihrer Fähigkeit deklarative und interrogative Satzkomplemente zu subkategorisieren. Die Klasse 1 umfasst Verben und Nomina, die beide Typen der Satzkomplemente subkategorisieren können (V1 und N1). Verben und Nomina, die nur interrogative Satzkomplemente (*w-* und *ob*-Nebensätze) subkategorisieren, gehören zur Klasse 2 (V2 und N2), während die Klasse 3 aus Verben und Nomina besteht, die nur mit deklarativen Nebensätzen (*dass*-Sätzen) auftreten (V3 und N3), vgl. (4) und (5).

(4) Verbklassen:

V1: *antworten, dass/w-/ob, bestimmen, dass/w-/ob*;

V2: *abfragen, w-/ob, befragen w-/ob*;

V3: *berichtigen, dass, sich einbilden, dass.*

(5) Nominalklassen:

N1: *Angst, dass/w-/ob, Beweis, dass/w-/ob*;

N2: *Abfrage, w-/ob, Rätsel*;

N3: *Eindruck, dass, Gefühl, dass.*

Die Klassifikation der Mehrwortausdrücke und Nominalkomposita basiert auf den Vererbungsrelationen zwischen den Valenzeigenschaften von MWAs und Komposita und den Valenzeigenschaften ihrer Konstituenten. Darüberhinaus unterscheiden wir vier Klassen der Mehrwortausdrücke, die in zwei Gruppen zusammengefaßt werden können: MWAs, die ihre Valenzeigenschaften von ihren Nominalkonstituenten[3] erben (M1 and M2), und MWAs, die ihre eigenen Valenzeigenschaften haben (die nicht von ihren Nominalkonstituenten übernommen werden) (M3 and M4). In (6) werden diese vier Klassen mit Beispielen illustriert.

(6) M1: MWAs, die ihre Valenzeigenschaften von ihren Nominalkonstituenten übernehmen (Vererbung von dem Nomen):
zur Bedingung machen, dass vs. *die Bedingung, dass.*

M2: MWAs, die ihre Valenzeigenschaften von ihren Nominalkonstituenten nur in bestimmten Kontexten übernehmen (Vererbung in bestimmten Kontexten):
in Erfahrung bringen, w-/ob
vs. *er hat (die) Erfahrung, dass/*w-/ob*
vs. *haben Sie (eine) Erfahrung, *dass/w-/ob?*

[3]Wir analysieren Funktionsverbgefüge (FVGs), die aus einer Präposition, einem Nomen und einem Funktionsverb bestehen.

M3: MWAs, deren Nominalkonstituente keine Satzkomplemente subkategorisieren:
zum Ausdruck bringen, dass vs. **der Ausdruck, dass.*

M4: MWAs, die generell als Idiome definiert sind, entweder weil sie "Cranberry"-Lexeme enthalten oder weil sie nicht-kompositionell sind:
in Abrede stellen, dass vs. **die Abrede*[4];
ins Auge fallen, dass.

Nominalkomposita werden in drei Klassen unterteilt. Die erste Klasse besteht aus Komposita, die ihre Valenzeigenschaften von ihren Kopfkonstituenten übernehmen C1. Die Komposita, die ihre Valenzeigenschaften von ihren Nicht-Köpfen erben, gehören zur Klasse C2. Die dritte Klasse besteht aus zwei Unterklassen. Die Klasse C3-1 umfaßt Komposita, die ihre Valenzeigenschaften sowohl von ihren Köpfen, als auch von ihren Nicht-Köpfen übernehmen können, und die Klasse C3-2 besteht aus Komposita, die ihre Valenzeigenschaften von keinen ihrer Konstituenten erben können. In diesem Fall besitzen Nominalkomposita ihre eigenen Eigenschaften und sind in den meisten Fällen lexikalisiert, vgl. Beispiele in (7).

(7) C1: Komposita, die ihre Subkategorisierungseigenschaften von ihrem Kopf übernehmen:
*Journalisten**frage**, w-/ob* vs. *Frage, w-/ob*.

C2: Komposita, die ihre Subkategorisierungseigenschaften von ihrem Nicht-Kopf übernehmen:
***Auswahl**verfahren, w-/ob* vs. *Auswahl, w-/ob*.

C3-1: Komposita, die ihre Subkategorisierungseigenschaften sowie von ihrem Kopf als auch ihrem Nicht-Kopf übernehmen können:
***Wettstreit**, w-/ob* vs. *Wette, w-/ob* oder *Streit, w-/ob*.

C3-2: Komposita, die ihre Subkategorisierungseigenschaften weder von ihrem Kopf noch von ihrem Nicht-Kopf erben können:
***Wortspiel**, dass* vs. *Wort, *dass* oder *Spiel, *dass*.

Vererbungsrelationen zwischen Nominalisierungen und ihren Basisverben werden nach den Übereinstimmungen oder Unterschieden zwischen ihren Valenzeigenschaften klassifiziert. Nominalisierungen in den R1-Relationen erben alle Subkategorisierungseigenschaften ihrer Basisverben. Verb-Nominalisierung-Paare, deren Nominalisierungen ein Teil der Valenzeigenschaften ihrer Verben verlieren (Reduktion der Subkategorisierungsvererbung), gehören zu den R2-Relationen, während die Verb-Nominalisierung-Paare, in denen Nominalisierungen mehr Subkategorisierungseigenschaften aufweisen (Erweiterung der Subkategoriserungsvererbung) als ihre Basisverben, zu den R3-Relationen gehören. Wir nehmen an, dass R3 eine hypothetische Klasse ist. Verben können meistens sowohl deklarative, als auch interrogative Satzkomplemente subkategorisieren (manchmal unter bestimmten Kontextbedin-

[4]Der einzige mögliche Gebrauch von *Abrede* ausserhalb der FVGS ist mit der Bedeutung 'mündliche Vereinbarung'. Dieser umfasst 22% aller Lemma von *Abrede*, die in unseren Daten gefunden wurden, dennoch ohne Satzkomplemente.

gungen). Nominalisierungen dagegen erlauben meistens nur deklarative Satzkomplemente[5], vgl. Beispiele in (8).

(8) R1: Subkategorisierungseigenschaften von Nominalisierungen werden vom Basisverb geerbt:
entscheiden, dass/ob/w- vs. *Entscheidung, dass/ob/w-*.

R2: Subkategorisierungseigenschaften von Nominalisierungen werden vom Basisverb in einer reduzierten Form geerbt:

- Nominalisierungen verlieren *ob/w*-Nebensätze:
(sich) erinnern, dass/w-/ob vs. *Erinnerung, dass*;
- Nominalisierungen verlieren *dass*-Nebensätze:
klären, dass/ob/w- vs. *Klärung, w-/ob*.

R3: Subkategorisierungseigenschaften von Nominalisierungen werden vom Basisverb in einer erweiterten Form geerbt – Nominalisierungen haben zusätzliche Eigenschaften, die ihre Basisverben nicht aufweisen können:
aufklären, w-/ob vs. *Aufklärung, dass/w-/ob*.

Methoden und Verwendete Tools

Input und Kontext Für die Untersuchung benutzen wir Zeitungs- und Web-Corpora aus Deutschland, Österreich und der Schweiz, die schriftliche Texte auf Deutsch aus den Jahren von 1988 bis 2005 umfassen und insgesamt ca. 1563 Millionen Tokens enthalten. [6]

Alle Corpora sind annotiert mit den folgenden Informationen: Satz-Tokens, Wortart-Tags, Lemmas und teilweise Chunks.[7] Das Corpus wird mit Hilfe von Queries abgefragt, die in Form von regulären Ausdrücken formuliert sind. Die Syntax der regulären Ausdrücke basiert auf der Stuttgarter CorpusWorkBench (CWB, cf. (Evert 2005)).

Für die Extraktion von Verben und Mehrwortausdrücken verwenden wir deutsche Verbletzt-Sätze (VL) und Sätze mit Verben im Passiv. In den VL-Sätzen folgen die subkategorisierten Nebensätze dem finiten Verb, während die präpositionalen und nominalen Konstituenten der Mehrwortausdrücke sich unmittelbar vor dem Verb befinden. Nebensätze werden entweder von den Vollverben oder von den Mehwortausdrücken subkategorisiert, vgl. Beispiele in Tabelle 2. Passive Sätze haben auch eine reguläre Form: die Nebensätze folgen dem zweiten Teil des Verbes, die präpositionellen und nominalen Konstituenten der Mehrwortausdrücke stehen unmittelbar vor dem zweiten Verbteil, vgl. Beispiele in Tabelle 3.

[5]Die Extraktionsergebnisse weisen auf, dass Nominalisierungen in den meisten Fällen einen *dass*-Nebensatz (65-67% der untersuchten Fällen) subkategorisieren.

[6]Corpora aus Deutschland enthalten Ausschnitte (1988-2001) aus deutschen Zeitungen – *die tageszeitung*, *Frankfurter Rundschau*, *Frankfurter Allgemeine Zeitung*, *Stuttgarter Zeitung*, *DIE ZEIT* und *Handelsblatt*. Wir verwenden auch den Corpus 'ELNC' (European Language News Corpus), der Online-Nachrichten aus dem Jahr 1997 umfasst. Einige Teile der Artikel von 'ELNC' stammen aus Schweizer Medienquellen. Weitere Texte aus der Schweiz sind im Corpus DEREKO-CH enthalten. DEREKO-AT umfasst Texte aus österreichischen Zeitungen. DEREKO-CH und DEREKO-AT sind Teile des Referenzcorpus DeReKo, der dem IMS vom IDS in Mannheim zur Vefügung gestellt wurde.

[7]Für Annotationen benutzen wir den Tokenisierer von (Schmid 2000), den Tree-Tagger beschrieben in (Schmid 1994) und (Schmid 1999) und den YAC-Chunker von (Kermes 2003).

	Hauptsatz		Nebensatz
Verb:	*Wenn sie*	*erfahren,*	*dass John Miller große Mengen Alkohol kauft...*
MWE:	*Wenn sie in Erfahrung*	*bringen,*	*dass John Miller große Mengen Alkohol kauft...*

Tabelle 2: *Dass*-Nebensätze mit Verben oder MWAs in VL-Sätzen

	Hauptsatz				Nebensatz
		Verb 1		**Verb 2**	
Verb:	*Es*	*muss*	*heute*	*gesagt werden,*	*dass der Nikolaus ein Türke ist.*
MWE:	*Es*	*muss*	*heute zur Sprache*	*gebracht werden,*	*dass der Nikolaus ein Türke ist.*

Tabelle 3: *Dass*-Nebensatz mit Verben oder MWAs in passiven Sätzen

Nomina (darunter auch Komposita, Nominalisierungen) werden im Vorfeld (VF) extrahiert. Das Vorfeld ist ein topologisches Feld vor dem finiten Verb in deutschen Deklarativsätzen. Wenn neben einem Nomen ein Nebensatz im Vorfeld vorkommt, dann kann dieser Nebensatz nur von diesem Nomen subkategorisiert werden, vgl. Tabelle 4.

Hauptsatz 1	Nebensatz	Hauptsatz 2
Nominalphrase		
Die Erklärungsversuche,	*warum der Teufel sich an X heranmacht,*	*sind auf der Glatze gedrehte Locken.*
Die Erklärung,	*warum der Teufel sich an X heranmacht,*	*sind auf der Glatze gedrehte Locken.*

Tabelle 4: *W*-Satz mit einem Nomen im VF

Extraktions- und Klassifikationsverfahren Wir extrahieren Prädikate automatisch aus den Textcorpora und klassifizieren sie nach ihren Subkategorisierungseigenschaften.

Die Extraktions- und Klassifikationsarchitektur basiert auf symbolischen Verfahren, die sowohl aus allgemein formulierten als auch aus spezifizierten Queries bestehen, vgl. (Lapshinova 2007). Die Suche fängt mit der Extraktion verschiedener Prädikattypen in allgemeinen Kontexten an (z.B. Extraktion der VL- und passiven Sätzen, die verbale Prädikate sowie Mehrwortausdrücke enthalten). Weiter wird die Suche für die Extraktion konkreter Pädikatarten (Verben, Nomina und Mehrwortausdrücke) verfeinert. Wir klassifizieren die extrahierten verbalen und nominalen Prädikate, sowie die Mehrwortausdrücke in die oben genannten Unterklassen.

Abbildung 1 zeigt eine Übersicht der verwendeten Extraktions- und Klassifikationsschritte. Unser Algorithmus ist eine Abfolge von Verfahren zur Identifikation und zur Extraktion von Verben, Nomina (einschließlich Komposita und deverbaler Nomina), sowie ihrer Klassifikation nach Valenzeigenschaften.

1 Allgemeine Queries: Extraktion von Sätzen, die verschiedene Pädikate enthalten
- 1.1 Suche nach VL- und Passivsätzen für Verben und Mehrwortausdrücke
- 1.2 für Nominalprädikate:
- 1.2.1 Suche nach Nominalprädikaten im VF
- 1.2.2 Fortsetzung mit Schritt 3

2 Spezifizierte Queries: Identifikation von Prädikaten
- 2.1 Suche nach Verben
- 2.2 Suche nach Mehrwortausdrücken

3 Spezifizierte Queries und Skripte: Klassifikation von Pädikaten
- 3.1 Klassifikation der Verbalprädikate: V1, V2, V3
- 3.2 Klassifikation der Nominalprädikate:
- 3.2.1 nach ihren Subkategorisierungseigenschaften: N1, N2, N3
- 3.2.2 nach ihrer morphologischen Struktur: einfach vs. zusammengesetzt (Komposita)

4 Vergleich der Subkategorisierungseigenschaften morphologisch verwandter Prädikate
- 4.1 Identifikation und Klassifikation der *ung*-Nominalisierungen: (Nung1), (Nung2), (Nung3)
- 4.2 Identifikation, Extraktion und Klassifikation der Basisverben: (Vbase1), (Vbase2), (Vbase3)
- 4.3 Klassifikation der Relationen zwischen Nominalisierungen und ihren Basisverben: R1, R2, R3

5 Zusätzliche Verfahren
- 5.1 Klassifikation der Nominalkomposita: C1, C2, C3-1, C3-2
- 5.2 Klassifikation der Mehrwortausdrücke: M1, M2, M3, M4

Abbildung 1: Schrittverlauf der Extraktions- und Klassifikationsverfahren

Ergebnisse: Extraktion and Klassifikation der Prädikate

Extraktion und Klassifikation der Nominalprädikate Die Extraktionsergebnisse weisen auf, dass unsere Tools eine wesentliche Anzahl von Nomen finden können, die Nebensätze subkategorisieren (über 15.000 Typen[8] und über 59.000 Tokens). Diese Nomina können nach ihren Subkategorisierungseigenschaften in drei Klassen unterteilt werden. Die erlangten Ergebnisse ermöglichen den Vergleich der Anteile von deklarativen und interrogativen Nebensätzen (die mit einem Nomen im VF vorkommen), der zeigt, dass Nominalprädikate Präferenzen für *dass*-Sätze haben, wie z.B. in Tabelle 5.

Nebensatz	***dass***	***w-/ob***	**GESAMT**
Typen	10232	5455	15687
in %	65,23	34,77	100,00
Tokens	40028	19219	59247
in %	67,56	32,44	100,00

Tabelle 5: Anteil der *dass*- und *w-/ob*-Sätze mit Nominalprädikaten im VF

Die Auswertung der Klassifikationsergebnisse für Nomina zeigt auch, dass Nominalprädikate meistens ein *dass*-Satzkomplement subkategorisieren. Die N1- und N3-Nomina (die einen *dass*-Satz erlauben) kommen in unseren Daten am häufigsten vor. Die N2-Nomina, die keine *dass*-Sätze erlauben, sind eher selten, vgl. Tabelle 6.

Klasse	**N1**	**N2**	**N3**	**GESAMT**
Typen	4116	3858	7713	15687
in %	26,24	24,59	49,17	100,00
Tokens	36228	5128	17891	59247
in %	61,15	8,65	30,20	100,00

Tabelle 6: Anteil der N1-, N2- und N3-Nominalprädikaten

Die Extraktion und Klassifikation der Nominalkomposita liefern Ergebnisse, die unsere Annahme bestätigen, dass der Kopf eines Kompositums nicht immer der Valenzträger ist (vgl. Komposita von Typen C2 und C3). Wir extrahieren Nominalkomposita, die automatisch in drei Klassen eingeteilt werden können. Die Klassifikation basiert auf Relationen zwischen Valenzeigenschaften der Komposita selbst und ihrer Konstituente. Obwohl die meisten Komposita zur Klasse C1 gehören, machen die C2- und C3-Komposita über 30% der analysierten Fälle aus, was eine unerwartet bedeutende Menge darstellt, vgl. Tabelle 7[9].

Extraktion und Klassifikation der Mehrwortprädikate Die Ergebnisse zeigen, dass einige MWAs eigene Subkategorisierungseigenschaften haben können, die nicht von ihren Nominalkonstituenten übernommen werden. In Bezug auf ihre Valenzeigenschaften verhalten sich diese Mehrwortprädikate wie Idiome, obwohl ihre Semantik

[8]Unter Kontexttypen verstehen wir 'Query Matches' oder Abfragentreffer, die unsere Tools liefern.

[9]Die Extraktionsergebnisse für Komposita werden als Typen in ihrer Standardbedeutung angegeben, d.h. Lemmatypen.

Klasse	**C1**	**C2**	**C3-1**	**C3-2**	**GESAMT**
Typen	423	53	131	21	628
in %	67,36	8,44	20,86	3,34	100,00

Tabelle 7: Anteil der C1-, C2- und C3-Komposita im VF

nicht immer idiomatisch ist: das syntaktische Verhalten stimmt nicht mit der semantischen Unterscheidung überein, die aus der Phraseologie kommt. Der Vergleich der Nomina, die sowohl innerhalb als auch außerhalb der Funktionsverbgefüge vorkommen, erlaubt die MWAs nach ihren Präferenzen für *dass*- und *w-/ob*-Nebensätze und mit Bezug auf die Vererbungshypothese zu klassifizieren. Tabelle 8 illustriert die Ergebnisse unserer Klassifikation.

Klasse	**M1+M2**	**M3+M4**	**GESAMT**
Typen	1701	1452	3151
in %	53,98	46,02	
Tokens	11687	7787	19474
in %	60,01	39,99	100,00

Tabelle 8: Anteil der M1+M2- und M3+M4-Mehrwortausdrücke

Extraktion und Klassifikation der Vererbungsrelationen Die Analyse der Extraktions- und Klassifikationsergebnisse für die Relationen zwischen den morphologisch verwandten Prädikaten (Verben und ihren Nominalisierungen) zeigen, dass ihre Subkategorisierungseigenschaften nicht immer übereinstimmen (Relationen vom Typ R2 und R3). Einerseits bestätigen unsere Ergebnisse (vgl. Tabelle 9) die Annahme, dass deverbale Nomina ihre Valenzeigenschaften von ihren Basisverben übernehmen (Relationen vom Typ R1), anderseits weisen die Ergebnisse auf, dass der Vererbungsprozesss in manchen Fällen begrenzt ist, z.B. dass die Nominalisierung nur *dass*-Satzkomplemente übernimmt.

Klassen	**R1**	**R2**	**R3**	**TOTAL**
Typen	72	75	13	160
in %	45,00	46,87	8,13	100,00

Tabelle 9: R1-, R2- und R3-Relationen für die häufigsten Verb-Nominalisierung-Paare

Diese Phänomene finden Erklärung in semantischen Eigenschaften sowohl der Prädikate als auch der subkategorisierten Nebensätze, vgl. (Bäuerle/Zimmermann 1991), (Karttunen 1977), (Schwabe 2004), (Fischer 2005) and (Oppenrieder 2006). Die Selektionsrestriktionen einiger Nominalisierungen, z.B. *Erfahrung* oder *Vorstellung*, unterscheiden sich von den Selektionsrestriktionen ihrer Basisverben, bspw. *erfahren* oder *vorstellen*. Dies beeinflusst die Wahl der Satzkomplemente. Weiterhin wird der Prozess der Subkategorisierungsvererbung von kontextuellen Parametern (z.B. Einbettung unter Modalverben oder Vorkommen in negativen Kontexten) beeinflusst.

Evaluierung der Tools Die Evaluierung der Einzelschritte unserer Verfahren zeigen, dass die Extraktions- und Klassifikationstools gute Ergebnisse liefern können, vgl. Tabelle 10. Die Genauigkeit (Precision) und die Trefferquote (Recall) der Klassifikationsschritte hängt von der Genauigkeit der Extraktionsergebnisse ab. Um die höhere Genauigkeit der Ergebnisse zu erreichen, durchlaufen wir eine Reihe von Filterungsschritten, z.B. Anwendung lexikalischer und syntaktischer Restriktionen, usw. Für die höhere Trefferquote sind zusätzliche Extraktionen nötig, z.B. aus geparsten Corpora oder weiteren Kontexten. Dies kann die Trefferquote erhöhen, während die Präzision der Extraktion stark sinken kann. Dies ist in unserer Arbeit unerwünscht, da wir einen Ansatz entwerfen, der auf Precision abzielt.

Prädikate	**Nomina**	**MWAs**	**Verben**
Precision in %	99,00	81,06	96,10

Tabelle 10: Precision-Ergebnisse für die Prädikatenextraktion

Folgerungen

Die Untersuchung extrahierter und klassifizierter Prädikate zeigt, dass Nomina (einschließlich Komposita und Nominalisierungen) und Mehrwortausdrücke eigene Subkategorisierungseigenschaften haben, die in der Erstellung der Lexika berücksichtigt werden sollten. Die Subkategorisierungseigenschaften morphologisch verwandter Wörter, i.e. Verben und ihrer Nominalisierungen, stimmen nicht in allen Fällen überein. Die Übereinstimmungen werden hier als Subkategorisierungsvererbung bezeichnet. Die Unterschiede können entweder als Reduktion der Subkategorisierungseigenschaften (z.B. wenn das Nomen Teil der verbalen Valenzeigenschaften verliert) oder als Erweiterung der Subkategorisierungsvererbung klassifiziert werden. Das unerwartete Subkategorisierungsverhalten deverbaler Nomina kann entweder mit ihren semantischen Eigenschaften (ihren Selektionsrestriktionen) oder ihren kontextuellen Parametern erklärt werden. Die Analyse der Nominalkomposita und Mehrwortausdrücke (zum grössten Teil Funktionsverbgefüge) zeigt, dass sie eigene Subkategorisierungseigenschaften aufweisen können, die unabhängig von den Eigenschaften ihrer Konstituenten sind, vgl. MWAs vom Typ M3 und M4, z.B. *zum Ausdruck kommen* und *in Abrede stellen*, oder Komposita vom Typ C3-2, z.B. *Wortspiel, Sehnsucht*. Außerdem kann in den Komposita vom Typ C2, z.B. *Erklärungsversuch* und C3-1 (*Schlussfolgerung*) nicht nur der Kopf, sondern auch der Nichtkopf als Valenzträger auftreten[10]. Dieses Verhalten widerspricht der Annahme, dass der Kopf eines Kompositums seine Subkategorisierungseigenschaften bestimmt.

Diese Phänomene sollten (semi)-automatisch behandelt werden können. Semiautomatische Extraktion und Klassifikation der oben genannten Fälle ist mit einer precision-orientierten Methode möglich, die im Rahmen dieser Arbeit entwickelt wurde. Unsere Methode basiert auf der Akquisition der Subkategorisierungseigenschaf-

[10]Beispiele der extrahierten und klassifizierten Prädikate werden im Anhang A aufgelisted. Weitere Beispiele werden auf der Ressourcen-Seiten des IMS (Universität Stuttgart) zur Verfgung gestellt, nachdem die vorliegende Arbeit begutachtet ist.

ten von Prädikaten aus annotierten Corpora. Annotationen schließen u.a. Tokenisierung, Wortart-Tagging, Lemmatisierung, morphologische Analyse und teilweise Chunking ein. Die Extraktions- und Klassifikationsarchitektur besteht aus aufeinander folgenden Schritten, die allgemeine und spezifische Queries umfassen. Die allgemeinen Queries basieren auf Wortstellungsmodellen der deutschen Sprache, während die spezifischen Queries syntaktische und lexikalische Restriktionen enthalten, die unsere Suche auf die gezielten Prädikattypen (Verben, Nomina, MWAs) einschränken. Darüber hinaus, verwenden wir morphologische Tools, um Komposita, *ung*-Nominalisierungen oder Basisverben zu identifizieren. Die Klassifikation der Prädikate und ihrer Relationen in die Unterklassen besteht aus Corpus-Queries und automatischen Verfahren zum Abgleich der Subkategorisierungseigenschaften.

Die beschriebene Architektur kann in verschiedenen Bereichen ihre Anwendung finden, z.B. in Wörterbuch- oder Lexikonbildung, Fremdsprachenunterricht, Übersetzungen oder NLP-Systeme, bspw. formalen Grammatiken (HPSG oder LFG) oder IE-Ansätzen.

Bibliography

[Admoni 1970] Admoni, V. Der deutsche Sprachbau. München: Beck, 1970.

[Agel 2000] Agel, V. (2000). Valenztheorie. Tübingen: Narr (Narr Studienbücher).

[Agel 2003] Agel, V., Eichinger, L.M., Eroms, H.-W., Hellwig, P., Heringer, H.J., Lobin, H. (2003). Valenz und Dependenz. *Ein internationales Handbuch der zeitgenösischen Forschung/An international Handbook of Contemporary Research.* 1. Halbband/Volume 1. Berlin/New York: Walter de Gruyter.

[Allerton 1982] Allerton D.J. (1982). *Valency and the English Verb*. London: Academic Press.

[Baker *et al.* 2003] Baker, C.F., C.J. Fillmore, B. Cronin (2003). The Structure of the FrameNet Database. In: *International Journal of Lexicography* 16, pp. 281-296.

[Baker *et al.* 1998] Baker, C.F., C. J. Fillmore, and J. B. Lowe (1998). The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Canada.

[Baker 2000] Baker, C.F. and J.Ruppenhofer (2002). FrameNet's Frames vs. Levin's Verb Classes. In J. Larson and M. Paster (eds.). *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, pp. 27-38.

[Balabanov 2007] Balabanov, A. (2007). Extraktion von Verbsubkategorisierungsrahmen aus Verbletztsätzen des 'TiGer-Korpus'. Studienarbeit. IMS, Universität Stuttgart.

[Baroni/Kilgariff 2006] Baroni, M., A. Kilgariff (2006). Large Linguistically-Processed Web Corpora for Multiple Languages. *Proceedings of the EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 3-7, 2006.

[Bartsch 2004] Bartsch, S. (2004). Structural and functional properties of collocations in English, A corpus study of lexical and pragmatic constraints on lexical co-occurence. Tübingen: Narr.

[Baschewa 2004] Baschewa, E. (2004). Objekte und Objektsätze im Deutschen und im Bulgarischen. *Eine kontrastive Untersuchung unter besondere Berücksichtigung der Verben der Handlungssteuerung*. Frankfurt am Main: Peter Lang GmbH.

[Bauer 1978] Bauer, L. (1978). The grammar of nominal compounding. Odense: Odense University Press.

[Bauer 1983] Bauer, L. (1983). English Word-formation. Cambridge: Cambridge University Press.

[Bauer 1988] Bauer, L. (1988). Introducing Linguistic Morphology. Edinburgh: University Press.

[Bäuerle/Zimmermann 1991] Bäuerle, R., T.E.Zimmermann (1991). Fragesätze. In von Stechow, A., D. Wunderlich (eds.). Ein internationales Handbuch der zeitgenössischen Forschung/An International Handbook of Contemporary Research. Berlin/New York: Walter de Gruyter, pp 333-348.

[Bausewein 1990] Bausewein, K. (1990). Akkusativobjekt, Akkusativobjektsätze und Objektsprädikate im Deutschen. Untersuchungen zu ihrer Syntax und Semantik. Tübingen: Niemeyer.

[Bergenholtz/Tarp 1994] Bergenholtz, H., S. Tarp (1994). *Mehrworttermini und Kollokationen in FachwörterbÜchern*. In Schaeder, B., H. Bergenholtz (eds). Fachlexikographie. Fachwissen und seine Repräsentation in Wörterbüchern. Tübingen: Narr.

[Bielicky/Smrz 2008] Belicky, V., O.Smrz (2008). Building the Valency Lexcion of Arabic verbs. In *Proceedings of LREC-2008*. Marrakech, Marroc.

[Bisetto/Scalise 2005] Bisetto, A. and S. Scalise (2005). The classification of compounds. Lingue e linguaggio IV.2, pp. 319-332.

[Bealer 1998] Bealer, G. (1998). Propositions. *Mind*, 107, pp. 1-32.

[Bergsten 1991] Bergsten N. (1991). *A Study on Compound Substantives in English*. Almquist and Wiksell, Uppsala.

[Bianco 1996] Bianco, M.T. (1996). Valenzlexicon Deutsch-Italienisch - Dizionario della valenza verbale (Deutsch im Kontrast). Groos.

[Bick 2003] Bick, E. (2003). A CG & PSG Hybrid Approach to Automatic Corpus Annotation. In *Proceedings of Shallow Processing of Large Corpora (SProLaC 2003)*, pp. 1-12.

[Bloomfield 1933] Bloomfield, L. (1933), *Language*, New York: Holt.

[Blumenthal/Rovere 1998] . Blumenthal, P., R. Giovanni (1998). PONS. Wörterbuch der italienischen Verben. Konstruktionen, Bedeutungen, Übersetzungen. Klett.

[Böhmová et al. 2001] Böhmová, A., J. Hajic, E. Hajicová and B. Hladká (2001). The Prague Dependency Treebank: Three-Level Annotation Scenario. In Anne Abeillé (ed.), Treebanks: Building and Using Syntactically Annotated Corpora. Kluwer Academic Publishers.

[Bond/Shirai 1997] Bond,F. and S.Shirai (1997). Practical and Efficient Organization of a Large Valency Dictionary. In *Proceedings of the 4th Natural Language Processing Pacific*, Phuket, Thailand.

[Bouillon *et al.* 1992] Bouillon, P., K. Bösefeldt, G. Russell (1992). Compound Nouns in a Unification-Based MT System. In *Proceedings of the Third Conference on Applied Natural Language Processing*. Trento, Italy, pp. 209-215.

[Bouma/Villada 2002] Bouma, G., B. Villada (2002). Corpus-based acquisition of collocational prepositional phrases. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University.

[Bourigault *et al.* 2005] Bourigault, M.-P.J., C.Fabre, C.Frerot, S.Ozdowska (2005). Syntex, analyseur syntaxique de corpus. In *Actes des 1èmes journées sour le Traitement Automatique des Lingues Naturelles*, Dourdan.

[Braasch *et al.* 2002] Braasch, A., A. Buhr, C. Navarretta, S. Nimb, S. Olsen, B.S. Pedersen, N.Sørensen (2002). *SprogTeknologisk Ordbog - Lingvistike Specifikationer*. Technical Report, version 5, Center for Sprogteknologi, Denmark.

[Breidt 1993] E. Breidt (1993) Extraction of v-n collocations from text corpora: a feasibility study for German, in: *Proceedings of the Workshop on very large corpora: Academic and industrial perspectives*. Columus, OH: ACL.

[Brent 1991] Brent, M. R. (1991). Automatic Semantic Classification of Verbs from their Syntactic Contexts: an Implemented Classifier for Stativity. In: *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, Germany, pp. 222-226.

[Brent 1993] Brent, M. R. (1993), From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. In *Computational Linguistics* 19(2), pp. 243-262.

[Bresnan 1972] Bresnan, J. (1972). The theory of complementation in English syntax. Cambridge, MA: MIT dissertation.

[Bresnan 1982a] Bresnan, J. (1982) *The Mental Representation of Grammatical Relation*, Cambridge, MA: The MIT Press.

[Bresnan 1982b] Bresnan, J. (1982) Control and Complementation. In: *The Mental Representation of Grammatical Relation*. Cambridge, MA: The MIT Press, pp. 282-390.

[Bresnan 2001] Bresnan, J. (2001) Lexical-functional syntax. Oxford: Blackwell.

[Brinkmann 1971] Brinkmann, H. (1971) Die deutsche Sprache: Gestalt und Leistung. Düsseldorf: Schwann, 1971. - XXXI.

[Briscoe/Carroll 1997] Briscoe, T., J. Carroll (1997). Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, Washington DC, pp. 356-363.

[Briscoe/Carroll 2002] Briscoe, E. and J. Carroll. (2002). Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Gran Canaria. pp. 1499-1504.

[Burger 1998] Burger, H. (1998). Phraseologie. Eine Einführung am Beispiel des Deutschen. Berlin: Erich Schmidt Verlag.

[Burnard 2007] Burnard, L. (2007), ed. Reference Guide for the British National Corpus (XML Edition) Published for the British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services.

[Busse/Dubost 1983] Busse, W., J.-P. Dubost (1983). Französisches Verblexikon. Die Konstruktion der Verben im Französischen. Klett, Auflage 2.

[Butt et al. 2002] M. Butt, H. Dyvik, T. King, H. Masuichi, C. Rohrer: "The Parallel Grammar Project", in: *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pp. 1-7.

[Camacho/Santana 2004] Camacho R.G., L. Santana (2004). Argument Structure of Deverbal Nouns in Brazilian Portuguese. Journal of Language and Linguistics. Vol.3 No. 2, 2004.

[Carroll/Rooth 1998] Carroll, G., M. Rooth (1998). Valence Induction with a Head-Lexicalized PCFG. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*. Granada, Spain.

[Carroll/Fang 2004] Carroll, J., A. Fang (2004). The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*. Sanya City, China, pp. 107-114.

[Chesley/Salmon-Alt 2006] Chesley, P, and S. Salmon-Alt (2006), Automatic Extraction of Subcategorization Frames for French. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

[Chomsky 1965] Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.

[Chomsky 1970] Chomsky, N. (1970). Remarks on Nominalization. In Jacobs, R.A. and P.S. Rosenbaum (eds.), *Readings in English Transformational Grammar*, Ginn and Co., Waltham, Mass.

[Chomsky 1973] Chomsky, N. (1973). Conditions on transformations. A festschrift for Morris Halle, ed. by Steven Anderson and Paul Kiparsky. New York: Holt, Rinehart and Winston.

[Church/Hanks 1989] Church, K.W., P. Hanks (1989). Word Association Norms, Mutual Information and Lexicography. 27th ACL, Vancouver, pp. 76-83.

[Copestake et al. 2004] Copestake, A., F. Lambeau, B. Waldron, F. Bond, D. lickinger, S. Oepen:"A lexicon module for a grammar development environment", in: *Proceedings of the Linguistic Resources and Evaluation Conference 2004*, Lisboa, Portugal, 2004, pp. 1111 - 1114.

[Crouch *et al* 2006] Crouch, D., M. Dalrymple, T. King, J. Maxwell, and P. Newman (2006). XLE documentation. URL http://www2.parc.com/isl/groups/nltt/xle/doc/xle toc.html

[Dalrymple 2001] Dalrymple, M. (2001). Lexical Functional Grammar. Volume 34 of *Syntax and Semantics*. Academic Press, New York.

[Daniels 1963] Daniels, K. (1963). *Substantivierungstendenzen in der deutschen Gegenwartssprache. Nominaler Ausbau des verbalen Denkkreises*. Düsseldorf.

[O'Donovan et al. 2005] O'Donovan R., A.Cahill, A.Way, M.Burke, J.Genabith (2005). Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. In: *Proceedings of ACL-2005*.

[Eckle-Kohler 1998] Eckle-Kohler, J. (1998). Methods for quality assurance in semi-automatic lexicon acquisition from corpora. In *Proceedings of EURALEX'98*, Liége, Belgium.

[Eckle-Kohler 1999] Eckle-Kohler, J. (1999). Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora/Linguistic Knowledge for Automatic Lexicon Acquisition from German Text Corpora. Berlin: Logos Verlag.

[Egli 1974] Egli, U. (1974). Ansätze zur Integration der Semantik in die Grammatik. Kronberg: Scriptor.

[Ehrich 1991] Ehrich; V. (1991). Nominalisierungen. In von Stechow, A., D. Wunderlich (eds). Semantik. *Ein internationales Handbuch der zeitgenösischen Forschung/An International Handbook of Contemporary Research*. Berlin/New York: Walter de Gruyter.

[Eisenberg 1994] Eisenberg, P. (1994). *Grundriss der deutschen Grammatik*. 2. überarbeitete Aufl. Stuttgart/Weimar.

[ELDIT] ELDIT http://dev.eurac.edu:8081/MakeEldit1/Eldit.html

[Emons 1974] Emons R. (1974). Valenzen englischer Prädikatsverben. Tübingen: Narr.

[Emons 1978] Emons, R. (1978). Valenzgrammatik für das Englische: eine Einführung (Anglistische Arbeitshefte 16). Tübingen: Max Niemeyer Verlag.

[Engel/Schumacher 1976] Engel, U., H. Schumacher (1976). Kleines Valenzlexikon deutscher Verben. Tübingen: Narr. 306 S.

[Engel 1977] Engel, U. (1977). Grammatik in Lehrbüchern für Deutsch als Muttersprache. In: Engel, U., S. Grosse (eds.). Grammatik und Deutschunterricht. Jahrbuch 1977 des Instituts für deutsche Sprache. Düsseldorf: Pädagogischer Verlag Schwann. *Sprache der Gegenwart 44*, pp. 102-135.

[Engel 1988] Engel, U. (1988). *Deutsche Grammatik*. Heidelberg: Groos.

[Engel 1991] Engel, U. (1991). *Deutsche Grammatik*. 2. Aufl. Heidelberg: Groos.

[Engel 1994] Engel, U. (1994). *Syntax der deutschen Gegenwartssprache. Grundlagen der Germanistik 22*. 3. Aufl. Berlin.

[Engel 1996] Engel, U. (1996). Tesnière mißverstanden. In: Greciano, G., Schumacher, H. (eds). *Lucien Tesnière - syntaxe structurale et op'erations mentales. Akten des deutsch-französischen Kolloquiums anläßlich der 100. Wiederkehr seines Geburtstages, Straßbourg 1993* (= Linguistische Arbeiten 348). Tübingen, 53 - 61.

[Engel 2004] Engel, U. (2004). *Deutsche Grammatik*. Neubearbeitung. München.

[Ehrich 1991] Ehrich, V. (1991). Nominalisierungen. In von Stechow, A., D. Wunderlich (eds.). Ein internationales Handbuch der zeitgenössischen Forschung/An International Handbook of Contemporary Research. Berlin/New York: Walter de Gruyter. pp 333-348.

[Ehrich/Rapp 2000] Ehrich, V., I. Rapp (2000). Sortale Bedeutung und Argumentstruktur: ung Nominalisierungen im Deutschen. *Zeitschrift für Sprachwissenschaft 19*, pp. 245-303.

[Erben 1972] Erben J. Deutsche Grammatik: ein Abriss/Johannes Erben. München: Hueber.

[Estival et al. 1995] Estival, D. (1995). The Construction of Test Material TSNLP Report (WP 3.1). URL: http://cl-www.dfki.uni-sb.de/tsnlp/publications.html#wp3.1

[Evert 2005] Evert E. (2005). The CQP Query Language Tutorial. IMS, Stuttgart. URL http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/html/

[Eynde/Blanche-Benveniste 1978] Eynde, K. Van den and C.Blanche-Benveniste (1978). Syntaxe et mécanismes descriptifs: présentation de l'approche pronominale Cahiers de Lexicologie 32, pp. 3-27.

[Eynde 2001] Eynde, K. Van den and M. Piet (2001). La syntaxe du verbe, l'approche pronominale et le lexique de valence PROTON, Preprint nr.174, Departement of Linguistics, K.U.Leuven, pp. 36 (published in modified form in French Language Studies 13, 2003). URL http://bach.arts.kuleuven.be/pmertens/papers/proton.pdf

[Eynde/Mertens 2006] Le dictionnaire de valence DICOVALENCE: manuel d'utilisation. Université de Leuven, version 1.2 URL http://bach.arts.kuleuven.be/dicovalence/manuel_061117.pdf

[Fazly/Stevenson 2006] A. Fazly, S. Stevenson: "Automatically constructing a lexicon of verb phrase idiomatic combinations", in: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL-2006,* (Trento/New Brunswick: ACL) 2006, pp. 337 – 344.

[Fellbaum 1998] Fellbaum, C. (ed.) (1998). WordNet: An Electronic Lexical Database. MA: The MIT Press.

[Fellbaum et al 2006] Fellbaum, C., A. Geyken, A. Herold, F. Koerner and G. Neumann (2006). Corpus-based studies of German idioms and light verbs. *International Journal of Lexicography* 19:4, pp. 349-360.

[Fillmore 1968] Fillmore, C. (1968). The case for case. In E. Bach and R. T. Herms (eds). *Universals in Linguistic Theory.* New York: Holt, Rinehart and Winston, pp. 1-88.

[Fillmore 1977] Fillmore, C.J. (1977). Scenes-and-Frames Semantics. In Antonio Zampolli (ed.). Linguistic Structures Processing, volume 59 of *Fundamental Studies in Computer Science*. Amsterdam: North Holland Publishing.

[Fillmore 1982] Fillmore, C.J. (1982). Frame Semantics. Linguistics in the Morning Calm, pp 111-137.

[Fillmore 2003] Fillmore, C.J. (2003). Valency and Semantic Roles: the Concept of Deep Structure Case. In Agel, Eichinger, L.M., Eroms, H.-W., Hellwig, P., Heringer, H.J., Lobin, H. (eds). Valenz und Dependenz. *Ein internationales Handbuch der zeitgenösischen Forschung/An international Handbook of Contemporary Research.* 1. Halbband/Volume 1. Berlin/New York: Walter de Gruyter. pp. 357-475.

[Filllmore 2007] Fillmore C.J. (2007). Valency Issues in FrameNet. In: T. Herbst and K. Götz-Votteler (eds.), Valency. Theoretical, Descriptive and Cognitive Issues. Berlin - New York: Mouton de Gruyter. pp. 129-156.

[Fillmore 2008] Fillmore C.J. (2008). A Valency Dictionary of English. Review Article. *International Journal of Lexicography Advance Access*. October 2008.

[Fischer 1999] Fischer, K. (1999). Verb Valency - Attempt at Conceptual Clarification. In: *The Web Journal of Modern Language Linguistics* 4-5. Published by the School of Modern Languages, University of Newcastle upon Tyne.

[Fischer 2005] Fischer, M. (2005). Ein Zweifelfall: *zweifeln* im Deutschen. *Linguistische Berichte*. Hamburg: Helmut Buske Verlag.

[Fitschen 2004] Fitschen, A. (2004). Ein computerlinguistisches Lexikon als komplexes System. Universitat Stuttgart: IMS, 10(3). AIMS.

[Forst 2003] Forst M. (2003). Treebank Conversion - Establishing a testsuite for a broad-coverage LFG from the the TIGER treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora (LINC '03)*, Budapest, Hungary.

[Fritzinger *et al.* 2010] Fritzinger, F., F.Richter, M.Weller (2010). Pattern-based Extraction of Negative Polarity Items from Dependency-parsed Text. Submitted for *LREC-2010*.

[Gahl 1998] Gahl, S. (1998). Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus. In *Proceedings of the 17th international conference on Computational linguistics* - Volume 1, Montreal, Quebec, Canada.

[Georgala 2003] Georgala, E. (2003). A Statistical Grammar Model for Modern Greek: The Context-free Grammar. In *Proceedings of the 24th Annual Meeting of the Linguistics Department of the Aristotle University of Thessaloniki*. Thessaloniki, Greece.

[Gove 1977] Gove, P.B. (ed.) (1977), Webster's seventh new collegiate dictionary. Springfield, MA: G.& C. Merriam.

[Götz-Votteler 2007] Götz-Votteler, K. (2007). Describing semantic valency. In Herbst, T., K. Götz-Votteler (eds.). Valency. Theoretical, Descriptive and Cognitive Issues. (Trends in Linguistics. Studies and Monographs). Walter de Gruyter.

[Grebe 1973] Grebe P. Duden "Grammatik der deutschen Gegenwartssprache". Mannheim: Bibliograph. Inst., 1973.

[Grimshaw 1979] Grimshaw, J. (1979). Complement selection and the lexicon. Linguistic Inquiry, 10, pp. 279-326.

[Grimshaw 1990] Grimshaw, J. (1990). *Argument Structure*. Cambridge: The MIT Press.

[Grishman *et al.* 1994] Grishman,R., C.Macleod, A.Meyers (1994). COMLEX Syntax: Building a Computational Lexicon. In *Proceedings of Coling 1994: The 15th International Conference on Computational Linguistics*, pp.268-272.

[Groenendijk/Stokhof 1984] Groenendijk, J., M. Stokhof (1984). Studies on the semantics of questions and the pragmatics of answers. PhD thesis, Department of Philosophy, University of Amsterdam.

[Grossmann/Tutin 2003] Grossmann, F., A. Tutin (2003). Les collocations – analyse et traitement. Amsterdam: De Werelt. Travaux et Recherches en Linguistique Appliqueé, E1.

[Gurevich et al. 2007] Gurevich, O., R. Crouch, T.H. King, V. de Paiva (2007). Deverbal Nouns in Knowledge Respresentation.In *Journal of Logic and Computation Advance Access*. December 20.

[Hamblin 1973] Hamblin, CL. (1973). Questions in Montague English. *Foundations of Language* 10, pp. 41-53.

[Hanks *et al.* 2006] Hanks, P., A. Urbschat and E. Gehweiler (2006). German Light Verb Constructions in Corpora and Dictionaries. *International Journal of Lexicography*, Vol. 19 Nr. 4. Advance access publication, 4 November 2006, Oxford University Press.

[Happ 1976] Happ, H. (1978). Théorie de la valence et enseignement du français. *Le Français Moderne* 46, pp. 97-134.

[Harris 1968] Harris, Z. (1968). Mathematical Structures of Language. Wiley (Interscience), New York.

[Hartrumpf *et al.* 2003] Hartrumpf, S., H. Helbig, R. Osswald (2003). "The Semantically Based Computer Lexicon HaGenLex - Structure and Technological Environment". Traitement automatique des langues, 44(2), pp. 81-105.

[Heid 1998] Heid, U. (1998). Building a Dictionary of German Support Verb Constructions. In: *Proceedings of the 1st International Conference on Linguistic Resources and Evaluation*, Granada. pp. 69-73.

[Heid/Gouws 2006] Heid, U., R. Gouws (2006). A model for a multifunctional electronic dictionary of collocations. In *Proceedings of the XIIth Euralex International Congress*, Torino, pp. 979-988.

[Heid 2005] Heid, U. (2005). Corpusbasierte Gewinnung von Daten zur Interaktion von Lexik und Grammatik: Kollokation - Distribution - Valenz. In: F. Lenz and S. J. Schierholz (eds.). Corpuslinguistik in Lexik und Grammatik. Tübingen, Stauffenburg.

[Heid 2006] Heid, Ulrich (2006) *Valenzwörterbücher im Netz.* in Petra C. Steiner; Hans C. Boas and Stefan J. Schierholz, editors, Kontrastive Studien und Valenz, Festschrift für Hans Ulrich Boas, Frankfurt: Peter Lang, pp. 69-89.

[Heid 2007] Heid, U. (2007). Valency data for Natural Language Processing: What can the *Valency Dictionary of English* provide? In: T. Herbst and K. Götz-Votteler (eds.), Valency. Theoretical, Descriptive and Cognitive Issues. Berlin - New York: Mouton de Gruyter. pp. 365-382.T. Herbst and K. Götz-Votteler (eds.), Valency. Theoretical, Descriptive and Cognitive Issues. Berlin - New York: Mouton de Gruyter, pp. 365-382.

[Heid/Weller 2008] Heid. U. and M.Weller (2008). Tools for collocation extraction: preferences for active vs. passive. In *Proceedings of LREC-2008*. Marrakech, Morocco.

[Heid *et al.* 2008] Heid, U., F. Fritzinger, S. Hauptmann, J. Weidenkaff and M. Weller (2008). Providing Corpus Data for a Dictionary for German Juridical Phraseology. In A. Storrer, A. Geyken, A. Siebert and K.-M. Würzner (eds.). Text Resources and Lexical Knowledge - Selected Papers from the 9th Conference on Natural Language Processing, KONVENS 2008. Berlin, New York: Mouton de Gruyter.

[Helbig 1971] Helbig, G. (1971). Beitrge zur Valenztheorie. The Hague [u.a.]: Mouton.

[Helbig 1984] Helbig, G. (1984). Probleme der Beschreibung von Funktionsverbgefügen im Deutschen. In Helbig, G. *Studien zur deutschen Syntax*. Bd. 2, Leipzig.

[Helbig 1992] Helbig, G. (1992). *Probleme der Valenz- und Kasustheorie*. Number 51 in Konzepte der Sprach- und Literaturwissenschaft. Tübingen: Max Niemeyer Verlag.

[Helbig/Buscha 2005] Helbig, G., J.Buscha (2005). Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht.Berlin, Langenscheidt.

[Helbig/Schenkel 1969] Helbig, G. Schenkel, W. (1969). Wörterbuch zur Valenz und Distribution deutscher Verben. Leipzig.

[Helbig/Schenkel 1973] Helbig, G. Schenkel, W. (1973). Wörterbuch zur Valenz und Distribution deutscher Verben. Leipzig. 458 S.

[Helbig/Schenkel 1991] Helbig, G. Schenkel, W. (1991). Wörterbuch zur Valenz und Distribution deutscher Verben. Leipzig.

[Helbig 2005] Helbig, H. (2005). Meaning representation with multilayered extended semantic networks. In *Proceedings of the 18th Florida Artificial Intelligence Conference (FLAIRS-05)*, pp. 32-37.

[Herbst 1983] Herbst, T.(1983). Untersuchungen zur Valenz englischer Adjektive und ihrer Nominalisierungen. Tübingen: Narr.

[Herbst 1999] Herbst, T. (1999). English Valency Structures - A first sketch. EESE 2/99.

[Herbst 2004] Herbst, T. (2004). Valency theory and the *Valency Dictionary of English*. A few remarks on the linguistic and lexicographic principles. In: Herbst, T., D. Heath, I.F. Roe and D. Götz (2004). A Valency Dictionary of English. A Corpus-Based Analysis of English Verbs, Nouns and Adjectives. Berlin/New York: Mouton de Gruyter.

[Herbst et al. 2004] Herbst, T., D. Heath, I.F. Roe and D.Götz (2004). A Valency Dictionary of English. A Corpus-Based Analysis of English Verbs, Nouns and Adjectives. Berlin/New York: Mouton de Gruyter.

[Herbst 2007] Herbst, T. (2007). Vaelncy complements or valency patterns. In Herbst, T., K. Götz-Votteler (eds.). Valency. Theoretical, Descriptive and Cognitive Issues.

[Herbst/Götz-Votteler 2007] Herbst, T., K. Götz-Votteler (2007). Valency. Theoretical, Descriptive and Cognitive Issues. Berlin/New York: Walter de Gruyter.

[Heringer 1968] Heringer, H.J. (1968). Die Opposition von 'kommen' und 'bringen' als Funktionsverben. Untersuchungen zur grammatischen Wertigkeit und Aktionsart. *Sprache der Gegenwart 3*, Schwann, Düsseldorf.

[Heringer 1970] Heringer, H.J. (1970). Theorie der deutschen Syntax, München: Hueber.

[Heringer 1996] Heringer H.-J. (1996). Deutsche Syntax dependentiell. Tübingen: Stauffenburg-Verl.

[Hintikka 1967] Hintikka J. *The Semantics of Questions and the Questions of Semantics*. Case Studies of the Interrelations of Logic, Syntax and Semantics. Amsterdam (=Acta Philosophica Fennica 28, 4).

[Hippisley *et al.* 2005] Hippisley, A, D. Cheng und K.Ahmad (2005). The head-modifier principle and multilingual term extraction. In *Natural Language Engineering*, Vol. 11 (2), pp. 129-157.

[Höhle 1986] Höhle T.N. (1986). Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder. In: A.Schöne, (Ed.). Kontroversen alte und neue. *Akten des VII. Internationalen Germanistenkongresses Göttingen*, 1985, Max Niemeyer Verlag: Tübingen (Bd. 3), pp. 329 - 340.

[Hull/Gomez 2000] Hull, R.D., F.Gomez. Semantic interpretation of deverbal nominalisations. Natural Language Engineering 6 (2). Cambridge University Press. pp. 139-161.

[Ienco et al. 2008] Ienco, D., S.Villata, C.Bosco (2008). Automatic extraction of subcategorization frames for Italian. In *Proceedinga of LREC-2008*. Marrakech, Marocco.

[Jacobs/Rosenbaum 1970] Jacobs, R.A., P.S. Rosenbaum (1970). Readings in English Transformational Grammar. Ginn and Company, Waltham, Massachusetts.

[Jacobs 2003] Jacobs, J. (2003). Die Problematik der Valenzebenen. In Agel, Eichinger, L.M., Eroms, H.-W., Hellwig, P., Heringer, H.J., Lobin, H. (eds). Valenz und Dependenz. *Ein internationales Handbuch der zeitgenösischen Forschung/An international Handbook of Contemporary Research.* 1. Halbband/Volume 1. Berlin/New York: Walter de Gruyter. pp. 378-399.

[Johnson *et al.* 2002] Johnson, C. R., C. J. Fillmore, M. R. L. Petruck, C. F. Baker, M. Ellsworth, J. Ruppenhofer, and E. J. Wood (2002). FrameNet: Theory and Practice. ICSI Berkeley, 2002. URL http://www.icsi.berkeley.edu/ framenet/book/book.html

[Johnston/Busa 1999] Johnston, M., F. Busa (1999). Qualia structure and the compositional interpretation of compounds. In E. Viegas (ed.), Breadth and depth of semantic lexicons. Dordrecht: Kluwer.

[Kamber 2006] Kamber, A. (2006). Funktionsverbgefüge - empirisch. Eine korpusbasierte Untersuchung in fremdsprachendidaktischer Perspektive. Ph.D. Thesis.

[Kaplan/Bresnan 1982] Kaplan, R., J. Bresnan (1982). Lexical functional grammar: A formal system for grammatical representation. In Joan Bresnan (ed.) The Mental Representation of Grammatical Relations. MIT Press, Cambridge, MA, pp. 173-281.

[Karttunen 1977] Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*. 1, pp. 3-44.

[Keil 1997] Keil, M. (1997) Wort für Wort. Repräsentation und Verarbeitung verbaler Phraseolexeme (Phraseo-Lex). Tübingen: Niemeyer.

[Kermes 2003] Kermes H. (2003). Off-line (and On-line) Text Analysis for Computational Lexicography. Ph.D. thesis IMS, University of Stuttgart. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)*, volume 9, number 3.

[Kinyon/Prolo 2002] Kinyon, A., C.A. Prolo (2002). Identifying Verb Arguments and their Syntactic Function in the Penn Treebank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, pp. 1982-1987.

[Kiparsky/Kiparsky 1970] Kiparsky, P., C. Kiparsky (1970). *Fact*. In M. Bierwisch and K.E. Heidolph (eds), Progress in Linguistics, pp. 143-73, The Hague: Mouton.

[Kipper et al. 2000] Kipper, K., H.T.Dang, and Palmer, M. (2000). Class-based construction of a verb lexicon. In *Proceedings of Seventeenth National Conference on Artificial Intelligence AAAI 2000*, Austin, TX.

[Kipper-Schuler 2005] Karin Kipper-Schuler (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. PhD. Thesis. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.

[Klotz 2007] Klotz, M. (2007). Valency Rules? The case of verbs with propositional complements. In Herbst, T., K. Götz-Votteler (eds). Valency. Theoretical, Descriptive and Cognitive Issues. Berlin/New York: Walter de Gruyter.

[Korhonen 2002] Korhonen, A. (2002). Subcategorization Acquisition. PhD thesis published as Techical Report UCAM-CL-TR-530. Computer Laboratory, University of Cambridge.

[Krenn/Erbach 1994] Krenn B. and G. Erbach (1994). Idioms and support verb constructions. In: J. Nerbonne, K. Netter, C. Pollard (Eds.): German in *Head-Driven Phrase Structure Grammar*, (Stanford, CA: CSLI Publications), [= CSLI Lecture Notes], pp. 297-340.

[Krenn 2000] Krenn, B. (2000). The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations. Saarbrücken: DFKI, Universität des Saarlandes.

[Krifka 1991] Krifka, M. (1991). Massennomina. In von Stechow, A., D. Wunderlich (eds.). Ein internationales Handbuch der zeitgenössischen Forschung/An International Handbook of Contemporary Research. Berlin/New York: Walter de Gruyter. pp. 399-418.

[Kupiec 1992] Kupiec, J. (1992). Robust Part-of-Speech Tagging Using a Hidden Markov Model. Computer Speech and Language 6, pp. 225-242.

[Lapata 2002] Lapata, M. 2002. The disambiguation of nominalisations. Computational Linguistics, 28(3), pp. 357-388.

[Lapshinova 2007] Lapshinova, E. (2007). Extracting Predicates Subcategorizing for Wh-Clauses: an Architecture for a Semi-automatic System. To appear in *Proceedings of the 12th ESSLLI Student Session*. Dublin, Ireland, August 6-17.

[Lapshinova/Heid 2007] Lapshinova, E., U.Heid (2007). Syntactic subcategorization of noun+verb multiwords: description, classification and extraction from text corpora. In: *Proceedings of the 26th International Conference on Lexis and Grammar*. Bonifacio, Corsica, October 2-6.

[Lapshinova/Heid 2008] Lapshinova, E., U.Heid (2008). Head or Non-head? Semi-automatic procedures for extracting and classifying subcategorisation properties of compounds. In *Proceedings of LREC-2008*. Marrakech, Morroco, Mai 28-30.

[Lapshinova 2008] Lapshinova, E. (2008). NonÂheads of compounds as valency bearers: extraction from corpora, classification and implication for dictionaries. In *Proceedings of EURALEX-2008*. Barcelona, Spain.

[Lapshinova 2009] Lapshinova-Koltunski E. (2009). Classification of "Inheritance" Relations: a Semi-Automatic Approach. In Gelbukh, A., S. Torres and I. Lopez (eds.). *Journal of Research on Computing Science*, Special Issue on *Advances in Computer Science and Engineering*.

[Lees 1963] Lees, R. (1963). The Grammar of English Nominalizations. Mouton & Co., The Hague.

[Lenci et al. 2008] Lenci, A., B.McGillivray, S.Montemagni, V.Pirrelli. Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora. In *Proceedinga of LREC-2008*. Marrakech, Marocco.

[Lester 2008] Lester, M. (2008). McGraw-Hill's Essential ESL Grammar: A Handbook for Intermediate and Advanced ESL Students. McGraw-Hill Professional.

[Levi 1978] Levi J.N. (1978). The syntax and semantics of complex nominals. New York: Academic Press.

[Levin 1993] Levin, B. (1993). English Verb Classes and Alternations: A preliminary Investigation. University of Chicago Press. Chicago II.

[Lewis 1970] Lewis, D.K. (1970). General Semantics. In: *Synthese* 22, 18-67. - Reprinted in: D. Davidson and G.Harman (eds.) *Semantics of Natural Language*. Dordrecht: Reidel (1972), pp. 169-218.

[Lezius et al. 2000] Lezius, W., S. Dipper and A. Fitschen (2000). IMSLex - Representing Morphological and Syntactical Information in a Relational Database. In U. Heid, S. Evert, E. Lehmann and C. Rohrer. (Hrsgg.), *Proceedings of EURALEX*, Stuttgart, Germany, pp. 133-139.

[Li/Abe 1998] Li, H., Abe, N. (1998). Generalising case frames using a thesaurus and MDL principle. *Computational Linguistics*, 24, pp. 217-244.

[de Lima 2002] de Lima, E. (2002), *The Automatic Acquisition of Lexical Information from Portuguese Text Corpora with a Probabilistic Context-Free Grammar*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

[McCawley 1970] McCawley, J. (1970). Where Do Noun Phrases Come From? In Jacobs, R.A. and P.S. Rosenbaum (eds.), *Readings in English Transformational Grammar*, Ginn and Co., Waltham, Mass.

[McCawley 1982] McCawley, J. (1982). Thirty Million Theories of Grammar. Univ. of Chicago Press, Chicago.

[Mackenzie 1997] Mackenzie, J.L. (1997). Nouns are avalent - and nominalizations too. In Karen van Durme (ed.). The valency of nouns. Odense: Odense University Press, pp. 89-118.

[Macleod *et al.* 1997] Macleod,C., A.Meyers, R.Grishman, L.Barrett, and R.Reeves (1997). Designing a Dictionary of Derived Nominals. In *Proceedings of Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria.

[Macleod *et al.* 1998a] Macleod, C., R. Grishman, A. Meyers (1998). COMLEX Syntax. Computers and the Humanities. 31(6), pp. 459-481.

[Macleod *et al.* 1998b] Macleod, C., R. Grishman, A. Meyers, L. Barrett, R. Reeves (1998). NOMLEX: A Lexicon of Nominalizations. In *Proceedings of EURALEX-98*, Liege, Belgium.

[Macleod 2002] Macleod, C. (2002). Lexical Annotation for Multi-word Entries Containing Nominalizations. In: *Proceedings of LREC-2002*. Las Palmas, Spain.

[Manning 1993] Manning, C. D. (1993), Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH, pp. 235-242.

[Manning/Schütze 1999] Manning, , H. Schütze (1999). Foundations of Statistical Natural Language Processing. Cambridge: The MIT Press, Massachusetts.

[Maragoudakis et al 2001] Maragoudakis, M., K.L.Kermanisdis, G.Kokkinakis (2001). Learning subcategorization frames from corpora: A case study for modern Greek. Thechnical report, Wire Communication Laboratory.

[Marchand 1969] Marchand, H. (1969). The Categories and Types in Present-Day English Word-Formation. München: Oscar Beck.

[Matthews 1981] Matthews, P.H. (1981). *Syntax*. Cambridge: Cambridge University Press.

[Matthews 2007] Matthews, P.H. (2007). The scope of valency in grammar. In Herbst, T., K. Götz-Votteler (eds.). Valency. Theoretical, Descriptive and Cognitive Issues.

[Maxwell 1995] Maxwell, K. (1995). Automatic Translation of English Compounds: Problems and Perspectives. In Alberto, P. and P. Bennett (eds.). Lexical Issues in Machine Translation, CEC). Luxembourg: Commission of the European Community.

[Meinschaefer 2004] Meinschaefer, J. (2004). The syntax and argument structure of deverbal nouns from the point of view of a theory of argument linking. International Conference on Deverbal Nouns. Lille, September 2004.

[Melloni 2007] Melloni, C. (2007). Polysemy in Word Formations: the Case of devrbal Nominals. Ph.D.Thesis. University of Verona.

[Mel'čuk 1988] Mel'čuk, I. (1988). Dependency syntax: theory and practice. Albany, NY: State Univ. of New York Press.

[Mel'čuk 1996] Mel'čuk, I. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In L. Wanner (ed.): Lexical Functions in Lexicography and Natural Language Processing, Amsterdam/Philadelphia: Benjamins, pp. 37-102.

[Meyers *et al.* 1994] Meyers, A., C. Macleod, and R. Grishman (1994). Standardization of the Complement Adjunct Distinction. In: *Proceedings of the 7th EURALEX International Congress*. Goteborg, Sweden.

[Merlo/Stevenson 2001] Merlo, P., S. Stevenson (2001). Automatic Verb Classification Based on Statistical Distributions of Argument Structure. In: *Computational Linguistics* 27(3), pp. 373-408.

[Messiant et al. 2008] Messiant, C., A.Korhonen, T.Poibeau (2008). LexSchem: A Large Subcategorization Lexicon for French Verbs. In *Proceedinga of LREC-2008*. Marrakech, Marrokko.

[Miller et al. 1990] Miller, G.A., R.Beckwith, C.Fellbaum, D.Gross, K.Miller (1990). WordNet: An online lexical database. *International Journal of Lexicography*.

[Moltmann 2003] Moltmann, F. (2003). Propositional Attitudes without Propositions. *Synthese* 135, pp. 77-118.

[Moffett 2002] Moffett, Marc (2002). Are 'that' - clauses really singular terms? Unpublished MS.

[Montague 1974] Montague, R. (1974). The Proper Treatment of Quantification in Ordinary English. In R. Thomason (ed.). Formal Philosophy. Selected Papers of Richard Montague. Yale University Press. New Haven, Conn.

[NOMLEX] NOMLEX http://nlp.cs.nyu.edu/nomlex/index.html

[Nunes 1993] Nunes, M. (1993). Argument linking in English derived nominals. In R. V. Valin, (ed.). *Advances in Role and Reference Grammar*. John Benjamins, Amsterdam, pp. 375-432.

[O'Donovan et al. 2005] O'Donovan, R., M. Burke, A. Cahill, J. van Genabith, A. Way (2005). Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. In *Computational Linguistics* 31(3), pp 329-365.

[Olsen 2002] Olsen, S. (2002). Some aspects of the syntactic encoding of nouns in a computational lexicon - the STO project. In *Proceedings of the Tenth EURALEX International Congress,* Copenhagen, Denmark.

[Oppenrieder 2006] Oppenrieder, W. (2006). Subjekt- und Objektsätze. In Àgel, V. L. M. Eichinger, H-W. Eroms, P. Hellwig, H. J. Heringer, H. Lobin (eds.). Denedency and valency/Dependenz and Valenz. An International Handbook of Contemporary Research/Ein internationales Handbuch der zeitgenössischen Forschung. Berlin/New York: Walter de Gruyter, pp. 900-913.

[Ortner 1991] Ortner L. (1991). Substantivkomposita: Komposita und kompositionsähnliche Strukturen 1. In *Deutsche Wortbildung*. Band 4.

[Osswald 2004] Osswald, R. (2004). Die Verwendung von GermaNet zur Pflege und Erweiterung des Computerlexikons HaGenLex. In C. Kunze, L. Lemnitzer, and A. Wagner (eds). LDV Forum - Anwendungen des deutschen Wortnetzes in Theorie und Praxis, 19(1/2), 2004, pp. 43-51.

[Pado et al. 2008] Pado, S., M. Pennacchiotti and C.Sporleder (2008). Semantic role assignment for event nominalisations by leveraging verbal data. In: *Proceedings of COLING-2008*.

[Palmer 2000] Palmer, M. (2000). Consistent criteria for sense distinctions. Special Issue of *Computers and the Humanities*, SENSEVAL98: Evaluating Word Sense.

[Palmer et al. 2005] Palmer, M., D.Gildea, and P.Kingsbury (2005). The Proposition Bank: An Annotated Resource of Semantic Roles. *Computational Linguistics* 31(1), pp. 71-106.

[Persson 1975] Persson, I. (1975). Das system der kausativen Funktionsverbgefüge. Eine semnatisch-syntaktische Analyse einiger verwandter Konstruktionen. Lund : LiberLaromedel Gleerup.

[Pinker 1989] Pinker, S. (1989). Learnability and Cognition: The Acquisition of Argument Structure. Cambridge: The MIT Press.

[Poibeau/Messiant 2008] Poibeau T., C. Messiant (2008). Do we still Need Gold Standards for Evaluation? In: *Proceedings of LREC-2008*. Marrakech, Marokko.

[Pollard/Sag 1994] Pollard, C., I. Sag (1994). Head-Driven Phrase Structure Grammar. University Of Chicago Press, 1 edition. Chicago.

[Pustejovky 1995] Pustejovsky, J. (1995). The Generative Lexicon. MIT Press, Cambridge, Massachusetts.

[Rappaport 1983] Rappaport, M. (1983). On the nature of derived nominals. In Rappaport, M., B. Levin, and A. Zaenen (eds). *Papers in Lexical-Functional Grammar*, Bloomington: Indiana University Linguistics Club, pp. 113-142.

[Richter/Sailer 2002] Richter, F. and M. Sailer (2002). Cranberry Words in Formal Grammar. In *Empirical issues in formal sytax and semantics*. Presses, pp. 155-171.

[Ritz/Heid 2006] Ritz. J. and U. Heid (2006). Extraction tools for collocations and their morphosyntactic specificities. In *Proceedings of the Linguistic Resources and Evaluation Conference, LREC-2006*, Genova, Italia [CD-ROM].

[Ruppenhofer et al. 2006] Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck, C. R. Johnson, J. Scheffczyk (2006). FrameNet II: Extended Theory and Practice.

[Sag *et al.* 2001] Sag, I., T.Baldwin, F.Bond , A.Copestake , D.Flickinger (2001). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*. Mexico City, Mexico.

[Sarkar/Zeman 2000] Sarkar, A. and D. Zeman (2000). Automatic Extraction of Subcategorization Frames for Czech. In:*Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Germany, pp. 691-697.

[Schierholz 2001] Schierholz S.J. (2001). Präpositionalattribute: syntaktische und semantische Analysen. Tübingen. Niemeyer.

[Schierholz 2005] Schierholz S.J. (2005). Valenzwörterbücher für Substantive. In Mogensen, J.E., H. Gottlieb, A. Zettersten (eds.). *Symposium on Lexicography XI in Copenhagen* (= Lexicographica. Series Maior 115). Tübingen, pp. 475-487.

[Schiffer 1996] Schiffer, S. (1996). Language-created language-independent entities. *Philosophical Topics*, 24, pp. 149-167.

[Schlobinski 1992] Schlobinski, P. (1992). Funktionale Grammatik und Sprachbeschreibung. Eine Untersuchung zum gesprochenen Deutsch sowie zum Chinesischen. Westdeutscher Verlag (Opladen).

[Schmid 1994] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*. Manchester, UK, pp. 44-49.

[Schmid 2000] Schmid, H. (2000). *LoPar, design and Implementation*. IMS, Universität Stuttgart. URL http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/LoPar.html

[Schmid 1999] Schmid, H. (1999). Improvements in Part-of-Speech Tagging with an Application to German. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann and D. Yarowsky (eds), *Natural Language Processing Using Very Large Corpora*. volume 11 of *Text, Speech and Language Processing*. Kluwer Academic Publishers, Dordrecht, pp. 13-26.

[Schmid *et al* 2004] Schmid, H., A. Fitschen and U. Heid (2004). SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of LREC-2004*. Lisbon, Portugal.

[Schuler 2005] Schuler, K. K. (2005). Verbnet: a Broad-Coverage, Comprehensive Verb Lexicon. *Doctoral Thesis*. University of Pennsylvania.

[Schulte im Walde 2000] Schulte im Walde, S. (2000). Clustering Verbs Semantically According to their Alternation Behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, August 2000.

[Schulte im Walde 2002] Schulte im Walde, S. (2002). A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, pp. 1351-1357.

[Schulte im Walde 2006] Schulte im Walde, S. (2006). Experiments on the Automatic Induction of German Semantic Verb Classes. In *Computational Linguistics* 32(2), pp. 159-194.

[Schulte im Walde 2009] Schulte im Walde, S. (2009). The induction of verb frames and verb classes from corpora. In A. Lüdeling and M. Kytö (eds). *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

[Schumacher 1986] Schumacher, H. (1986) Verben in Feldern: Valenzworterbuch zur Syntax und Semantik deutscher Verben. de Gruyter.

[Schumacher 2004] Schumacher, H., J.Kubczak, R.Schmidt and Vera der Ruiter (2004). VALBU - Valenzwörterbuch deutscher Verben. Tübingen: Gunter Narr Verlag.

[Schwabe 2004] Schwabe, K. (2004). On the Semantics of German Declarative and Interrogative Root and Complement Clauses. In: P. Denis, E. McCready, A. Palmer, and B. Reese (eds.). *Proceedings of the 2004 Texas Linguistics Society Conference: Issues at the Semantics-Pragmatics Interface*. Somerville: Cascadilla Proceedings Project, pp. 79-91.

[Schwabe/Fittler 2009a] Schwabe K., R. Fittler (2009). Semantic Characterizations of German Question-Embedding Predicates. In: P. Bosch, D. Gabelaia, and J. Lang (eds.): *TbiLLC 2007, LNAI 5422*, Springer-Verlag, Berlin/Heidelberg, pp. 229-241.

[Schwabe/Fittler 2009b] Schwabe K., R. Fittler (2009). Syntactic force of consistency conditions for German matrix predicates. In: *Proceedings of the 10th Symposium on Logic and Language*. Budapest, Hungary, 26-29 August, 2009, pp. 157-167.

[Selva et al. 2002] Selva, Th., S. Verlinde, J. Binon. (2002). Le DAFLES, un nouveau dictionnaire pour apprenants du français. Actes du dixième congrès EURALEX'2002 (European Association for Lexicography). Copenhagen.

[Seretan 2008] Seretan, V. (2008). Collocation Extraction Based on Syntactic Parsing. Ph.D. thesis, University of Geneva. URL: http://www.latl.unige.ch/personal/vseretan/publ/PhDThesisVioletaSeretan.pdf

[Siegel 1998] Siegel, E.V. (1998). Linguistic Indicators for Language Understanding: Using Machine Learning Methods to combine Corpus-based Indicators for Aspectual Classification of Clauses. PhD thesis, Department of Computer Science, Columbia University.

[Sommerfeldt/Schreiber 1983] Sommerfeldt, K. and H. Schreiber (1983b). Wörterbuch zur Valenz und Distribution deutscher Substantive. Leipzig: VEB Bibliographisches Institut.

[Sommerfeldt/Schreiber 1996] Sommerfeldt, K. and H. Schreiber (1996). Wörterbuch der Valenz etymologisch verwandter Wörter: Verben, Adjective, Substantive. Tübingen Niemeyer.

[Somers 1987] Somers, H.L. (1987). Valency and Case in Computational Linguistics. Edinbugh: Edinburgh University Press.

[Spencer 1991] Spencer A. (1991). Morphological Theory. Cambridge, Blackwell.

[Spranger 2004] Spranger, K. (2004) *Beyond Subcategorization Acquisition - Multi-Parameter Extraction from German Text Corpora.* in Geoffrey Williams and Sandra Vessier, editors, Proceedings of the 11th Euralex International Congress volume 1, pp. 171-176.

[Spranger/Heid 2003] Spranger, K. and U. Heid (2003). A Dutch Chunker as a Basis for the Extraction of Linguistic Knowledge. In: Tanja Gaustad (ed.) *Computational Linguistics in the Netherlands 2002*. Selected Papers from the 13th CLIN Meeting.

[Stechow 1991] von Stechow, A., D. Wunderlich (1991). Semantik. *Ein internationales Handbuch der zeitgenösischen Forschung/An International Handbook of Contemporary Research*. Berlin/New York: Walter de Gruyter.

[Storrer/Schwall 1993] Storrer, A., Schwall, U. (1993). Description and Acquisition of Multiword Lexemes. In *Proceedings of EAMT Workshop*. pp. 35-50.

[Storrer 2007] Storrer A. (2007). Corpus-based Investigations on German Support Verb Constructions. In Fellbaum, C. (ed.). Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London: Continuum Press.

[Surdeanu et al. 2003] Surdeanu, M., S. Harabagiu, J. Williams, P. Aarseth (2003). Using Predicate-Argument Structures for Information Extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*. Richardson, Texas.

[Tarvainen 1981] Tarvainen, K. (1981). *Einführung in die Dependenzgrammatik.* (= Reihe Germanistische Linguistik 35). Tübingen.

[Tesnière 1959] Tesnière, L. (1959). *Éléments de syntaxe structurale. Deuxiéme édition.* Paris, 2. Aufl.

[Tesnière 1980] Tesnière, L. (1980). *Grundzüge des strukturalen Syntax.* Herausgegeben und übersetzt von Ulrich Engel. Stuttgart.

[Teubert 1979] Teubert, W. (1979). Valenz des Substantivs: attributive Ergänzungen. und Angaben. Düsseldorf.

[Teubert 2003] Teubert, W. (2003). Die Valenz nichtverbaler Wortarten: Substantiv. In Agel, V., Eichinger, L.M., Eroms, H.-W., Hellwig, P., Heringer, H.J., Lobin, H. (eds). Valenz und Dependenz. *Ein internationales Handbuch der zeitgenösischen Forschung/An international Handbook of Contemporary Research.* 1. Halbband/Volume 1. Berlin/New York: Walter de Gruyter, pp. 820-835.

[Tichý 1978] Tichý, P. (1978). Questions, Answers, and Logic. *American Philosophical Quarterly* 15, 4, pp. 275-284.

[Trawiński *et al.* 2008] Trawiński, B. M. Sailer, J.-Ph. Soehn, L. Lemnitzer and F. Richter (2008). Cranberry Expressions in English and in German. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 35-38. European Language Resources Association (ELRA): Marrakech, Morocco.

[Ushioda et al 1993] Ushioda, A., D.A. Evans, T. Gibson, A. Waibel (1993). The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora. In *Proceedings of the Workshop on the Acquisition of Lexical Knowledge from Text*, Columbus, OH, pp 95-106.

[VALLEX] VALLEX 1.0. URL http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0

[Vendler 1967] Vendler, Z. (1967). *Linguistics in Philosophy*. Cornell University Press, Ithaca.

[Vetulani *et al.* 2006] Vetulani, Z., T. Obrebski, G. Vetulani (2006). Syntactic Lexicon of Polish Predicative Nouns. In Calzolati (ed.) *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy. pp. 1734-1737.

[Vetulani *et al.* 2007] Vetulani, Z., T. Obrebski, G. Vetulani (2007). Towards a Lexicon-Grammar of Polish: Extraction of Verbo-Nominal Collocations from Corpora. *FLAIRS Conference*, pp. 267-268.

[Vetulani *et al.* 2008] Vetulani, G., Z. Vetulani, T. Obrebski (2007). Verb-Noun Collocation SyntLex Dictionary - Corpus-Based Approach. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Marocco.

[Villada Moirón 2005] Villada Moirón, M.B. (2005). Data-driven identification of fixed expressions and their modifiability. Ph.D.Thesis.

[Wauschkuhn 1999] Wauschkuhn, O. (1999). Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora. PhD thesis, Institut für Informatik, Universität Stuttgart.

[Wechsler 2008] Wechsler, S. (2008). A diachronic account of English deverbal nominals. In Chang,C.B., H.J. Haynie (eds.). *Proceedings of the 26th West Coast Conference on Formal Linguistics*, Cascadilla Proceedings Project, Somerville, MA, pp. 498-506.

[Welke 1988] Welke, K. (1988). Einführung in die Valenz- und Kasustheorie. Leipzig: Bibliogr. Inst.

[Wierzbicka 1972] Wirzbicka, A. (1972). Semantic Primitives. Frankfurt: Athenäum.

[Žabokrtský 2005] Žabokrtský, Z. (2005). Valency Lexicon of Czech Verbs. Doctoral Thesis. Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague.

[Winhart 2002] Winhart, H. (2002). Funktionsverbgefüge im Deutschen. Zur Verbindung von Verben und Nominalisierungen. Ph.D. Thesis.

[Wojtak 1992] Wojtak, B. (1992). *Verbale Phraseolexeme in System und Text*. Tübingen: Niemeyer. (Reihe Germanistische Linguistik: 125).

[Wotjak/Heine] Wojtak, B., A. Heine (2007). Syntaktische Aspekte der Phraseologie I: Valenztheoretische Ansätze. In: Burger, H., D. Dobrovol'skij, P. KÃ¼hn, N.R. Norrick (eds). An International Handbook of Contemporary Research/Ein internationales Handbuch der zeitgenössischen Forschung. Berlin, New York: Walter de Gruyter, pp. 41-53.

[Zaenen/Engdahl 1994] Zaenen, A. and E. Engdahl (1994). Descriptive and Theoretical Syntax in the Lexcion. In: Beryl T. S. Atkins, A. Zampolli (Eds): *Computational Approaches to the Lexicon.*, Oxford University Press, 1994, pp. 181-212.

[Zifonun *et al.* 1997] Zifonun,G., L.Hoffmann, B.Strecker (1997). Grammatik der deutschen Sprache. Band 2. Berlin/New York: de Gruyter.

[Zwicky 1985] Zwicky A.M. (1985). "Heads". In: *Journal of Liguistics*. volume 21, pp. 1-29.