E. Lapshinova-Koltunski

DATA EXTRACTION FROM CORPORA AS A TOOL TO CLASSIFY PREDICATES BY THEIR SUBCATEGORISATION PROPERTIES

1. Introduction: aims and motivation

The number of NLP technologies for lexicon acquisition which are aimed at high quality results has been growing in the recent years. The lexical information retrieved with tools for the automatic extraction of data from text corpora can be stored in machine-readable lexicons and updated dynamically[1]. Many such tools aim at extracting words with their linguistic properties and tend to classify them syntactically or semantically.

In this work, we analyze the subcategorisation of different lexical units, by means of extracting them from text corpora along with their sentential complements. We elaborate an extraction architecture using available lexical and grammatical knowledge about the phenomena we extract. The lexical data are created to serve symbolic NLP, especially large symbolic grammars for deep processing (HPSG (cf. work in the LinGO project[2]) or LFG (cf. the PARGRAM project[3])).

Our aim is to classify the extracted data according to their subcategorisation properties. We also intend to compare properties of

[1] *Schulte im Walde S.* The induction of verb frames and verb classes from corpora // A. Lüdeling and M. Kytö (eds.). Corpus Linguistics. An International Handbook. Mouton de Gruyter, Berlin. 2006.

[2] *Copestake A., Lambeau F., Waldron B., Bond F., Lickinger D., Oepen S.* A lexicon module for a grammar development environment // Proceedings of the Linguistic Resources and Evaluation Conference 2004, Lisboa, Portugal, 2004. P. 1111–1114.

[3] *Butt M., Dyvik H., King T., Masuichi H., Rohrer C.* The Parallel Grammar Project // Proceedings of COLING–2002 Workshop on Grammar Engineering and Evaluation. P. 1–7.

different morphologically related predicates (verbs, nouns and multiwords) and analyze the phenomenon of «inheritance» of valency (e.g. in case of deverbal nouns, which share their subcategorisation properties with the underlying verbs).

2. Data and approaches

2.1. Data

In our work, we focus on four types of predicates: verbs, nouns, N+V multiwords and adjectives. For the description of valency phenomena related with these types of predicates, and for earlier extraction work, we refer to a number of studies on valency (including dictionaries), as well as to works on acquistion tools for predicates and their subcategorisation (cf. Table 1).

Table 1. Predicate types and related linguistic studies

predicates	**examples**	**existing dictionaries and studies**
verbs	*darüber/davon sprechen, dass... (to speak about that...)*	VDE[1], VALBU[2]
verbal multiwords	*zur Bedingung machen, dass... (to make it a condition that...)*	Krenn/Erbach 1994[3], Storrer 2006[4], Lapshinova/Heid 2007[5]

[1] *Herbst T., Heath D., Roe I.F. and Götz D.* A Valency Dictionary of English. A Corpus-Based Analysis of English Verbs, Nouns and Adjectives. Berlin/New York: Mouton de Gruyter, 2004.

[2] *Schumacher H., Kubczak J., Schmidt R., and Vera der Ruiter*. VALBU – Valenzwörterbuch deutscher Verben. Tübingen: Gunter Narr Verlag, 2004.

[3] *Krenn B., Erbach G.* Idioms and support verb constructions // J. Nerbonne, K. Netter, C. Pollard (Eds.). German in Head-Driven Phrase Structure Grammar. Stanford, CA: CSLI Publications, 1994. P. 297–340.

[4] *Storrer A.* Corpus-based investigations on German support verb constructions // C. Fellbaum (Ed.) Collocations and Idioms: Linguistic, lexicographic and computational aspects, London: Continuum Press. To appear, 2007.

[5] *Lapshinova E., Heid U.* Syntactic subcategorization of noun+verb multiwords: description, classification and extraction from text corpora // Proceedings of the 26th International Conference on Lexis and Grammar. Bonifacio, Corsica, 2007.

predicates	examples	existing dictionaries and studies
nouns	*die Erklärung, warum... (the explanation why...)*	VDE, Sommerfeldt/Schreiber 1983[1],Sommerfeldt/Schreiber 1996[2]
adjectives	*dafür zuständig, dass... (responsible for that...)*	VDE, Sommerfeldt/Schreiber 1983[3],Sommerfeldt/Schreiber 1996[4]

2.2. Predicates in NLP

Subcategorisation information is an important research topic in modern NLP. This kind of information should be included into lexicons for NLP, as many syntactic theories use it for sentence building. An NLP application, which builds on these theories, needs more detailed subcategorisation entries, containing the description of the argument structure of all words under analysis, and their morphological relations. Most existing tools for predicate acquisition concentrate on valency patterns for English verbs and only a few analyze other predicate types and other languages. An overview of works describing tools for different languages is outlined in Table 2.

As mentioned above, we examine four types of predicates for German, and we elaborate semi-automatic procedures for their extraction and classification, which can be applied for creating subcategorisation lexicons for German or for enhancing existing ones (e.g. IMSLex[5]).

---

[1] *Sommerfeldt K., Schreiber H.* Wörterbuch zur Valenz und Distribution deutscher Substantive. Leipzig: VEB Bibliographisches Institut, 1983b.

[2] *Sommerfeldt K., Schreiber H.* Wörterbuch der Valenz etymologisch verwandter Wörter: Verben, Adjective, Substantive. Tübingen: Niemeyer, 1996.

[3] *Sommerfeldt K., Schreiber* H. Wörterbuch zur Valenz und Distribution deutscher Substantive…1983b.

[4] *Sommerfeldt K., Schreiber H.* Wörterbuch der Valenz etymologisch verwandter Wörter… 1996.

[5] *Fitschen A.* Ein computerlinguistisches Lexikon als komplexes System. Ph.D. thesis IMS, University of Stuttgart // Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS). Vol. 10. № 3. 2004.

Table 2. Languages and NLP-tools for predicates acquisition

langs.	**authors/publications**
EN	Uschioda 1993[1], Manning 1993[2], Briscoe/Carroll 1997[3], Carroll/Fang 2004[4], etc.

[1] *Ushioda A., Evans D.A., Gibson T., Waibel A.* The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora // Proceedings of the Workshop on the Acquisition of Lexical Knowledge from Text. Columbus, OH, 1993. P. 95–106.

[2] *Manning C.D.* Automatic Acquisition of a Large Subcategorization Dictionary from Corpora // Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. Columbus, OH, 1993. P. 235–242.

[3] *Briscoe T., Carroll J.* Automatic Extraction of Subcategorization from Corpora // Proceedings of the 5th ACL Conference on Applied Natural Language Processing, Washington, DC, 1997. P. 356–363.

[4] *Carroll G., Fang A.* The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser // Proceedings of the 1st International Joint Conference on Natural Language Processing. Sanya City, China, 2004. P. 107–114.

langs.	authors/publications
other languages	
DE	Schulte im Walde 2002[1], Eckle/Kohler 1999[2], Wauschkuhn 1999[3], Spranger 2004[4], etc.
FR	Chesley/Salmon-Alt 2006[5]
NL	Spranger/Heid 2003[6]
CZ	Sarkar/Zeman 2000[7]
P	de Lima 2002[8]

[1] *Schulte im Walde S.* A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG // Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain, 2002. P. 1351–1357.

[2] *Eckle-Kohler J.* Linguistic Knowledge for Automatic Lexicon Acquisition from German Text Corpora. Berlin: Logos Verlag, 1999.

[3] *Wauschkuhn O.* Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora. PhD thesis, Institut für Informatik, Universität Stuttgart, 1999.

[4] *Spranger K.* Beyond Subcategorization Acquisition – Multi-Parameter Extraction from German Text Corpora // G. Williams, S. Vessier (Eds.). Proceedings of the 11th Euralex International Congress. 2004. Vol. 1. P. 171–176.

[5] *Chesley P., Salmon-Alt S.* Automatic Extraction of Subcategorization Frames for French//Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 2006.

[6] *Spranger K., Heid U.* A Dutch Chunker as a Basis for the Extraction of Linguistic Knowledge // T. Gaustad (ed.). Computational Linguistics in the Netherlands 2002. Selected Papers from the 13th CLIN Meeting.

[7] *Sarkar A., Zeman D.* Automatic Extraction of Subcategorization Frames for Czech // Proceedings of the 18th International Conference on Computational Linguistics. Saarbrücken, Germany, 2000. P. 691–697.

[8] *de Lima E.* The Automatic Acquisition of Lexical Information from Portuguese Text Corpora with a Probabilistic Context-Free Grammar. PhD thesis. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2002.

3. Methods and tools for extracting predicates – the role of appropriate context

In this study, we focus on two main questions: how can data about subcategorisation properties be extracted from text corpora and how can lexical items be classified according to their subcategorisation properties?

3.1. Input

Our input is a corpus of newspaper texts of German[1] which are sentence-tokenized, pos-tagged and lemmatized (TreeTagger/lemmatizer and the STTS tagset[2]. To assess the need for chunking, we used YAC, a recursive chunker for German[3]. Extraction queries in the form of regular expressions rely on the Stuttgart CorpusWorkBench (CWB[4]).

---

[1] Texts from Germany, Austria and Switzerland, a total of ca. 950M words: Austrian ('AT', ca. 500M) and Swiss ('CH', ca. 180M), which are part of the German reference corpus DeReKo. The corpora from Germany include extracts (1992-2000) from *die tageszeitung* ('taz', 111M), *Frankfurter Rundschau* ('FR', 40M), *Frankfurter Allgemeine Zeitung* ('FAZ', 71M), *Stuttgarter Zeitung* ('StZ', 36M), *DIE ZEIT* ('ZEIT', 86M) as well as literary texts from the 'Gutenberg' Archive ('DE Lit.', 138M).

[2] *Schmid H.* Probabilistic Part-of-Speech Tagging Using Decision Trees // International Conference on New Methods in Language Processing. Manchester, UK, 1994. P. 44–49; *Schmid H.* Improvements in Part-of-Speech Tagging with an Application to German // 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, 1999. P. 13–26.

[3] *Kermes H.* Off-line (and On-line) Text Analysis for Computational Lexicography. Ph.D. thesis IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), 2003. Vol. 9. № 3.

[4] *Evert E.* The CQP Query Language Tutorial. IMS, Stuttgart, 2005 // **URL http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/html/**

3.2. Context

Depending on the types of predicates we analyze, different sentence structures are used in the extraction procedures: verb-final (-last) sentences (VL), passive sentences (P) and sentences with a noun in Vorfeld (VF). We deliberately work with those sentence structures which present a relatively regular order of NPs and PPs.

3.2.1. Verb-final sentences

VL sentences (mainly subclauses) make up ca. 20–25% of all corpus text. In this context in German, we have a regular sequence of elements: the subcategorized subclause usually follows the verb and nominal or adjectival predicates or predicate elements tend to be immediately left-adjacent to the verbal one. The subclause following the verb is typically subcategorized by the verb or, if relevant, by nominal and adjectival elements preceeding the verb, or by a noun+verb multiword. (cf. (1)).

(1) *Weil Clinton und seine Anwälte erst* ***in Erfahrung bringen wollen****, was Lewinsky zu sagen hat. (As Clinton and his lawyers* ***want to bring into experience*** *(to find out), what Lewinsky wants to say.)*

3.2.2. Passive constructions

Passive constructions can also be used for extracting all kinds of predicates. Like in VL, we have a regular sequence of elements here, as the subclause following the verb is typically subcategorized by the verb, a noun (as in (2)), an adjective following the verb or a verbal multiword.

(2) *Wenn also* ***die Frage gestellt wird****, wo Erziehung stattfindet.*
(If one ***poses the question****, where education takes place).*

3.2.3. Vorfeld

VF is mainly used for extracting nominal predicates. German grammars[1] state that if a noun in VF is followed by a sentential complement, this complement can only be subcategorized by the noun, as seen in (3).

(3)*Aber all die **Erklärungsversuche**, warum der Teufel sich an X heranmacht, sind auf der Glatze gedrehte Locken.*

*(But all the **explanation-attempts** (attempts to explain) why the devil chats up X are as futile as giving a bald man a comb).*

3.3. Extraction procedures and identification of types

The extraction steps in predicate identification and classification proceed from the general to the specific[2].

Predicate identification. We apply general queries to extract verbs, nouns, adjectives or multiwords along with their sentential complementation. General queries are underspecified (lines 6-8) and only contain constraints for the searched predicate type. An example of a general query for extracting verbs in VL is shown in table 3. We search for sentences which start with a conjunction, a relative or an interrogative pronoun (line 2) followed by some optional words (excluding finite verbs) and a finite verb (line 4) at the end of the main clause, followed by comma and the subcategorized *dass-, ob* or *w-*clause.

---

[1] *Zifonun G., Hoffmann L., Strecker B.* Grammatik der deutschen Sprache. Band 2. Berlin/New York: de Gruyter, 1997; *Helbig G., Buscha J.* Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht.Berlin, Langenscheidt, 2005.

[2] *Lapshinova E.* Extracting Predicates Subcategorizing for Wh-Clauses: an Architecture for a Semi-automatic System // Proceedings of the 12th ESSLLI Student Session. Dublin, Ireland, 2007.

Classification into subtypes. To subclassify predicates according to their subcategorisation properties, we apply specific queries which contain further constraints for subtypes including lexical specification. For instance, we lexically modify line 3 in table 3 to exclude nominal predicates known to our system in front of the verbal complex if we search for verbal predicates subcategorizing for a subclause. Nominal predicates in this position could in fact subcategorize for the subcaluse themselves, and these sentences could be ambiguous with respect to the valency bearer. To prevent this, we add the constraint [lemma! =RE($imslex)] which prevents our system from extracting predicative nouns in front of the verb.

Table 3. Query for predicates in a VL sentence

	Query building blocks	**comments**	**extracted sentence**
1.	<... >	sentence context	*Sie zeigten mir, da saßen Leute,*
2.	[pos="KOU.*|PREL.*|PW.*"]	conj., relat. or inter. pronoun	*die*
3.	[pos! ="V.*FIN"&word! =",|-"]*	optional, no finite verbs or punctuation	
4.	<vc>...</vc>	verbal complex	*wußten*
5.	" "	comma	*,*
6.	[(pos="PW.*")&	relative pronoun or	*worauf*
7.	(word="ob")|	conjunction "ob" or	
8.	(word="dass")	conjunction "daß"	
9.	[pos! ="V.*FIN"]*	suclause: non-verbal	*es*
10.	[pos="V.FIN*"]	finte verb od subclause	*ankommt*
11.	[pos="$."]	sentence end	*.*
12.	within s;	within a sentence	

4. Results and interpretation

Our first extraction procedures delivered interesting results about the subtypes of predicates based on their subcategorisation. In the following section we show some of our results on the subclassification of verbal, nominal and multiword predicates. The analysis of adjectives and their subclassification is a task for future work.

4.1. Verbs and their complements

Our experiments showed that not all verbs can take all three types of sentential complementations analyzed here. We assume that the choice for a sentential complement is determined by the verb semantics. The verb gives the information about how the complements are realized, e.g. as a noun phrase, *dass*, *ob* or *w*-clause[1]. Our extraction tests show that, for instance, verbs subcategorizing alternatively for both an *ob* and a *w*-clause, have preferences for one of the two sentential complements. Thus, 72% of extracted German verbs of discussing, found with all three types of complement clauses, have a subcategorized *ob*-clause and only 15% of them select a *w*-clause as a complement. Some linguists point out that the type of complement clause can also depend on the context of the main clause: its modality or polarity[2]. For example, verbal predicates taking a *w*- or an *ob*-clause, in interrogative contexts, show preferences for *dass*-clauses in declaratives. However there are few studies on contextual preferences of predicates.

We assume that verbs subcategorizing for *dass*, *ob* and *w*-clauses can be classified into the following groups:

(V1) verbs, which are able to subcategorize for both *dass* and *ob/w-* sublcauses: *erklären, dass/ob/w- (explain that/if/wh-) or beweisen, dass/w- (prove that/wh-)*;

(V2) verbs, which are able to subcategorize for only *dass* or only *ob/w-*: *fragen, *dass/ob/w- (ask *that/if/wh-) or zusichern, dass/*ob/ *w- (that/*if/*wh-)*;

---

[1] *Bausewein K.* Akkusativobjekt, Akkusativobjektsätze und Objektsprädikate im Deutschen. Untersuchungen zu ihrer Syntax und Semantik. Tübingen: Niemeyer, 1990.

[2] See *Bäuerle R., Zimmermann T.E.* Fragesäetze. Semantik. Ein internationales Handbuch zeitgenössischen Forschung. Berlin/New York: de Gruyter, 1991. P. 333–348.

(V3) verbs, which are able to subcategorize for both *dass* and *ob/w-* under different contextual conditions (e.g. with changed modality or polarity): *denken, dass (think that)* in a declarative sentence vs. *denken, ob (think if)* in an interrogative sentence.

The analysis of the properties (semantic or contextual) that influence the choice for complementation is a task for our future work.

4.2. Nominal predicates and the «head» problem of compounds

The extraction results showed that both simplex and compound nominal predicates in VF can subcategorize for a *dass-*, *w-* and *ob-*clause. With the help of the morphological tool SMOR[1], we sort them into two groups: simplex predicates, like *Problem* («problem»), *Beweis* («proof, evidence»), *Schluss* («conclusion»), and complex predicates, like *Grundproblem* («base problem»), *Beweismittel* («means of evidence»), *Schlussfolgerung* («conclusion»)[2].

In Table 4 we outline the frequency of compound predicates, compared to the frequency of simplex predicates extracted in VF from the corpora (220 M words).

Table 4. Simplex and compound predicates with a *dass-*, *ob-* or a *wh-*clause.

Occurrencis in corpora	**tokens**		**types**	
	simplex	**compound**	**simplex**	**compound**
'FR'+'FAZ'+'taz'	84,30%	15,70%	88,00%	12,00%

As shown in table 4, the phenomena we analyze are not frequent (about 12–15,7% of all nominal predicates in VF are compounds).

---

[1] *Schmid H., Fitschen A., Heid U.* SMOR: A German computational morphology covering derivation, composition, and inflection // Proceedings of LREC–2004. Lisbon, Portugal.

[2] More details in *Lapshinova-Koltunski E., Heid U.* Head or Non-head? Semi-automatic procedures for extracting and classifying subcategorisation properties of compounds // Proceedings of LREC–2008. Marrakech, Morocco, 2008.

With respect to the relations between the subcategorisation of a compound and that of its head constituent, the following types of compounds can be observed:

(C1) The subcategorisation is determined by the head: *das Forschungs**problem**, dass... (the research problem, that...)* vs. *das **Problem**, dass... (the problem that...).*

(C2) The subcategorisation is determined by the non-head: *die **Beweis**last, dass... (the burden of proof that...)* vs. *der **Beweis**, dass... (the evidence that...)* and *die Last, dass...* (the burden that...).*

(C3–1) The subcategorisation is determined by both the head and the non-head: *die **Schlussfolgerung**, dass... (the conclusion that...)* vs. *der **Schluss**, dass... (the conclusion that...)* or *die **Folgerung**, that... (the conclusion that...).*

(C3–2) The compound has its own subacetgorisation properties: *der **Ehrgeiz**, dass... (the ambition that)* vs. *die Ehre, dass...* (honour that...)* or *der Geiz, dass...* (avarice that...).*

Cases (C2) and (C3) do not match the commonly accepted assumption that the head of a compound is its valency bearer. Such cases should receive a specific treatment in NLP.

In Table 5, we summarize the proportion of type (C1) to (C3) occurrences in the newspaper corpora mentioned above. The figures show that type (C1) cases are the most frequent in text corpora. Compounds of type (C2) and (C3), however, make up over 40% of all compound cases in VF, which is a considerable amount.

Table 5. Occurrence of (C1) to (C3) types in VF

Types	**C1**	**C2**	**C3**
occurences in corpora	56,6%	11,8%	31,6%

4.3. «Inheritance» relationships of multiwords

The data analyzed so far allow us to classify prepositional noun+verb MWEs with respect to the relationship between the subcategorisation of MWEs and that of their noun component: (M1) and partly (M2) «inherit» it, whereas (M3) and (M4) don't[1]:

(M1) «Inheritance»: MWE and its nominal component can subcategorize for a sentential complement: *zur Bedingung machen, daß* («make it a condition») vs. *die Bedingung, daß* ("the condition that")

(M2) «Inheritance»+«switching» of true values: MWE and its nominal component (under certain contextual condtions) can subcategorize for a type of sentential complement:

MWE: *in Erfahrung bringen, ob ...* («to find out if»)

affirm.: *er hat (die) Erfahrung, daß/*ob/w-* («he has (the) experience that/*if/wh-»)

interr.: *haben Sie (eine) Erfahrung, *daß/ob/w- ?* («do you have (any) experience *that/if/wh- ?»)

(M3) «Non-inheritance»: MWEs (semantically trasparent) can subcategorize for subclauses, and neither their nominal nor their verbal component do so[2]: *zum Ausdruck bringen, daß ...* («to express») vs. **der Ausdruck, daß ...*

(M4) «Non-inheritance»: non-compositional idioms or MWEs with «cranberry» lexemes: *in Abrede stellen, daß ...* («to deny that») vs. **die Abrede*[3] *, ins Auge fallen, daß ... («to catch sb's eye»).*

---

[1] Cf. *Lapshinova E., Heid U.* Syntactic subcategorization of noun+verb multiwords… 2007.

[2] We also group in this class MWEs whose noun has a sentence complement, but in a massively different subcategorization frame: *Beweis* («proof») takes a *für*-PP or a sentential complement with a(n optional) correlate (*dafür*), whereas *unter Beweis stellen* («to provide evidence for»), which also has a sentence complement, can never take the correlate nor a *für*-PP.

[3] The only non-MWE reading of *Abrede* is that of 'oral agreement', which is found in 22% of the occurrences of the lemma in our corpus, but always without a sentential complement.

In Table 6, we summarize some absolute frequency figures from one of our extraction exercises, based on the newspapers 'FR', 'FAZ' and 'taz'.

Judging from the small sample in table 6, we can assume that our tools throw up useful results. MWEs of types (M1) or (M2) can show up significantly both within an MWE (+MWE) and outside (-MWE). The identification of the «switching» of truth values for complement clauses characteristic of (M2) can be observed with *in Erfahrung bringen*: it accepts *daß*-clauses outside an MWE, but within the MWE, it shows up consistently with *wh-* and *ob*-clauses. *Abrede* and *Vergessenheit* seem to take sentential complements only within MWEs.

Table 6. Sample German noun vs. MWEs and their subclauses, by word order models

	MWE, noun	*dass*		*wh-*		*ob*	
		-MWE	+MWE	-MWE	+MWE	-MWE	+MWE
		VF	VL	VF	VL	VF	VL
(M1)	*in **Aussicht** stellen*	55	60	0	0	0	0
+	*zur **Bedingung** machen*	37	59	0	0	0	1
(M2)	*in **Erfahrung** bringen*	40	39	2	17	0	13
(M3)	*in **Rechnung** stellen*	5	53	2	1	0	0
+	*zum **Ausdruck** kommen*	0	24	0	0	0	0
(M4)	*in **Abrede** stellen*	0	25	0	0	0	0
	*in **Vergessenheit** geraten*	0	34	0	0	0	0

5. Conclusion and future work

The extraction results obtained within the present study show that there are limits to the correspondences or «inheritance» of subcategorisation (e.g. compounds or multiwords). There is a need for tools to identify such cases by means of data extraction from corpora, as subcategorisation information is important for NLP applications. For this purpose, a precision-oriented semi-automatic extraction is possible which can operate on a tokenized, tagged and lemmatized text.

As future work, we would like to extend the kinds of extracted complements beyond subclauses and also analyze the relationships between the subcategorisation of verbs and their nominalisations within a compound or a MWE or outside (cf. Table 7).

Table 7. Nominal predicates or predicate elements vs. base verbs

	nominalisation	**base verb**
simplex	*Beweis, dass... (proof (evidence) that...)*	
compound	*Beweismittel, dass...(means of evidence that...)*	*beweisen, dass...*
MWE	*unter Beweis stellen, dass...(to give proof that...)*	*(to prove that..)*

6. Acknowledgements

This research is part of the author's PhD work supported by the DFG-funded Research Graduate Program GK-609 «Linguistic Representations and their Interpretation». Many thanks to Dr. Ulrich Heid for his helpful comments and suggestions for improvement.