Non-heads of Compounds as Valency Bearers: Extraction from Corpora, Classification and Implication for Dictionaries

Ekaterina Lapshinova-Koltunski
Universität Stuttgart

This paper describes an approach to the classification of nominal compounds based on their subcategorisation. German compound noun predicates, such as Grundproblem, Beweislast *and* Schlussfolgerung *subcategorizing for a subordinate clause are semi-automatically extracted from text corpora and classified according to which of their components, the head or the non-head, is the valency bearer. In over 40% of cases the subcategorisation of compounds is not determined by their heads. This kind of information should be included in subcategorisation lexicons as well as dictionaries for human users. We show that our semi-automatic approach can be applied in natural language processing, especially in lexicon and dictionary creation.*

1. Introduction

It is commonly accepted that the head of a compound is its valency bearer, i.e. that it determines the argument structure of the whole construct and is the semantic head of the whole compound (cf. Zwicky (1985) and Bauer (1988)). We review the subcategorisation of compounds and show that there are three types of nominal compound predicates based on the relationships between their constituents.

In this paper we deal with German nominal compounds subcategorizing a subordinate *dass*, *w-* or *ob* (that, wh- or if) clause, although our methods can be applied for other complements as well. Nominal compounds are very common in German, and they play a big role in the expression of nominal concepts. Thus, the lexical acquistion of such complex words is necessary for lexicography and natural language processing: there is a need to handle compound nouns automatically.

We describe a set of semi-automatic procedures focusing on the problem of automatic extraction and classification of lexical data according to their properties and their further use in building lexicons for natural language processing. The possible lexicographic implementation of the proposed pocedures is described in section 5, where we show the role of such procedures in the process of lexicon building.

2. Data and existing approaches

According to some linguists, compounding involves the interface between morphology and syntax (Spencer (1991)). Compounds have a constituent structure and most of them have a compositional reading (Johnston and Busa (1999)). One of the constituents acts as the head, and the other as the modifier. The head of a compound is characterized by the same properties as a phrase head: it determines the morphosyntactic categories and the subcategorisation of the whole compound (cf. Zwicky (1985) and Bauer (1985)). In this paper, we show that the commonly accepted assumption about the head and non-head constituents does not hold for all instances considered, and we present a set of procedures for creating or updating a subcategorisation lexicon.

2.1. *Data*

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:

(a) The subcategorisation of the whole construction is determined by its head:

(1) *das Grund**problem**, dass...* (“the basic problem, that...”) or
*das Forschungs**problem**, dass...* (“the research problem, that...”) vs.
*das **Problem**, dass...* (“the problem that...”).

(b) The subcategorisation of the whole construction is determined by its non-head:

(2) *das **Beweis**mittel, dass...* (“the means of evidence, that...”) or
*die **Beweis**last, dass...* (“the burden of proof that...”)
vs.
*der **Beweis**, dass...* (“the evidence that...”) and
das Mittel, dass... * (“the means that...”) or
die Last, dass... * (“the burden that...”).

(c) The compound has its own subcategorisation:

(3a) *die **Schlussfolgerung**, dass...* (“the conclusion that...”) vs.
*der **Schluss**, dass...* (“the conclusion that...”) or
*die **Folgerung**, that...* (“the argumentation (conclusion) that...”).

(3b) *der **Ehrgeiz**, dass...* (“the ambition that”) vs.
die Ehre, dass... * (“honour that...”) or
der Geiz, dass... * (“avarice that...”).

The subcategorisation of type (c) cases like those in (3) could be determined by both the head and the non-head constituents, e.g. in (3a), or by none of them and the compound can have its own properties (like in (3b)). Therefore, we speak about two subtypes of type (c) compounds: (c1) like in (3a), and (c2) like in (3b).

2.2. *Approaches to the linguistic description of compounds*

Nominal compounds have been the research topic of many linguists, e.g. Bergsten (1991), Ortner (1991), Levi (1978). They have received much attention in research towards natural language processing applications, for instance in Hippisley et al. (2005), Johnston and Busa (1999), Bouillon et al. (1992). Some authors concentrate on the semantic relations between a head noun and a modifier noun, others describe the principles of compositional treatment or the extraction of nominal compounds, but the problem of the subcategorisation relations between the constituents of a compound has to our knowledge not received any attention so far. None of the cited authors mentions that not only the head of a compound can determine its subcategorisation (cf. (b) and partially (c) types described in 2.1). This becomes apparent when the subcategorisation of compounds is compared with that of their elements. For instance, in (4) the non-head *Erfahrung* (“experience”) and not the head *Wert* (“value”) subcategorizes for the complement *wh*-sentence.

(4) *Über **Erfahrung**swerte, ob dies in der Schweiz möglich ist,*
verfügt man nicht.
(“Of the **experience** value, if it is possible in Switzerland, they don’t dispose (They don’t have the experience, if it is possible in Switzerland).”)
die Erfahrung, ob... (“the experience if...”) vs.
**der Wert, ob...* (“the value if...”)

We examine all three types of compounds mentioned in 2.1, and try to elaborate semi-automatic procedures for the extraction and classification of compound predicates, which can be applied for creating subcategorisation lexicons for German or for enhancing existing ones[1] In figure 1, we give an example of IMSLex entries for nouns subcategorizing for sentential complements.

1 Such as IMSLex (Lezius et al. (2000), CISLex (Guenthner and Maier (1996) and Guenthner (1996)) and HaGenLex Hartrumpf et al. (2003)).

Lexicon entry	Examples from corpora
Erfahrung(s-comp(C_daß))	*Erfahrung, dass*
Erfolg(s-comp(C_daß))	*Erfolg, dass*
Ergebnis(s-comp(C_daß))	*Ergebnis, dass*
Ergebnis(s-comp(C_wh/ob))	*Ergebnis, wie*

Figure 1: Examples of IMSLex entries for nouns with *dass*, *wh-* and *ob* subclauses

3. Methods and tools for the extraction of nominal compounds in context

3.1. *Input*

Our input is a corpus of newspaper texts of German[2] which are sentence-tokenized, pos-tagged and lemmatized.[3] Extraction queries in the form of regular expressions rely on the Stuttgart CorpusWorkBench (CWB[4]).

3.2. *Extraction procedures and identification of types*

Predicate identification

The first extraction procedures were performed on nominal predicates in "Vorfeld" (left of the finite verb) cf. example (4) and the query in figure 2. If a noun in "Vorfeld" is followed by a sentential complement (in this research, a *dass, w-* or *ob* clause), this complement can only be subcategorized by the noun.

	Query building blocks	Comments	Extracted sentence
1.	<s>	sentence beginning	
2.	[pos! ="NN|V.FIN"]{0,3}	prenominal material	
3.	[pos="APPR|APPRART"]?	optl. preposition or prep/art	*Über*
4.	(<np> ... </np>)	noun phrase	*Erfahrungswerte*
5.	","	comma	,
6.	(pos="PW.*"&	relative pronoun, but not	
	word! ="wobei|womit")|	"wobei" and "womit" or	
7.	(word="ob")|	conjunction "ob" or	*ob*
8.	(word="dass")	conjunction "daß"	
9.	[pos! ="$.|V.FIN"]*	subclause: non-verbal part	*dies in der Schweiz möglich*
10.	[pos="V.FIN"]	finite verb of subclause	*ist*
11.	","	comma	,
12.	[pos="V.FIN"]	finite main verb	*verfügt*
13.	within s;	rest of main clause	*man nicht.*

Figure 2: Query for extraction of nouns in Vorfeld position subcategorizing a *wh-*, *ob-* or *daß*-clause

2 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 and have been made available to us by the Institut für deutsche Sprache, Mannheim, in a cooperative project. 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).

3 We use Schmid's TreeTagger/lemmatizer and the STTS tagset (Schmid (1994) and (1999)).

4 Cf. Evert (2005).

The query in figure 2 starts with a noun phrase (or a prepositional phrase), (lines 1 to 4), which is followed by a subordinate clause (lines 5 to 11), marked by commas, and the main clause, which starts with a finite (main) verb (line 12) (cf. Lapshinova and Heid (2007)).

With the help of the morphological tool SMOR (Schmid et al. (2004)), we sort the list of nominal predicates extracted by the query in figure 2 into two groups: simplex predicates (*Problem, Beweis, Schluss*) and complex predicates (*Grundproblem, Beweismittel, Schlussfolgerung*). Both lists are used for the further subclassification of compounds.

Subclassification

To classify the extracted candidates according to (a) to (c) (cf. section 2.1), we use a list of "known" simplex predicates[5] to build a query for extracting compounds of types (a) and (b). For this purpose, we modify the query in figure 2 by lexically specifying the noun phrase in line 4 (cf. figure 3). To extract type (a) compounds, we use the nouns from the list as heads (i.e. assume that they are preceded by another element, cf.line 4a.) and for type (b) we use them as non-head predicates (line 4b).

	Query building blocks	Comments	Sample extracted words
4a.	[lemma="...|.+beispiel|.+frage|..."]	nouns from "known" lists	Parade**beispiel**, Journalisten**frage**
4b.	[lemma="...|Beweis.+|Erklärung.+|..."]	nouns from "known" lists	**Beweis**mittel, **Erfahrung**swert

Figure 3: Example of queries for types (a) and (b)

In some cases, both the head and the non-head constituents can subcategorize for a subclause. For example, in (3a) both the non-head *Schluss* ("conclusion") and the head *Folgerung* ("argumentation (conclusion)") subcategorize for a sentencial complement (cf. 2).

We can assume that the nominal compound *Schlussfolgerung* belongs to type (c1), and the subcategorized subclause is determined by both its constituents. We suppose that compounds which occur both in the type (a) and type (b) lists belong to type (c1). The compounds that occur neither in the type (a) nor type (b) lists are mostly type (c2) cases (e.g. *Ehrgeiz*), whose elements do not take sentential complements at all (cf. figure 4).

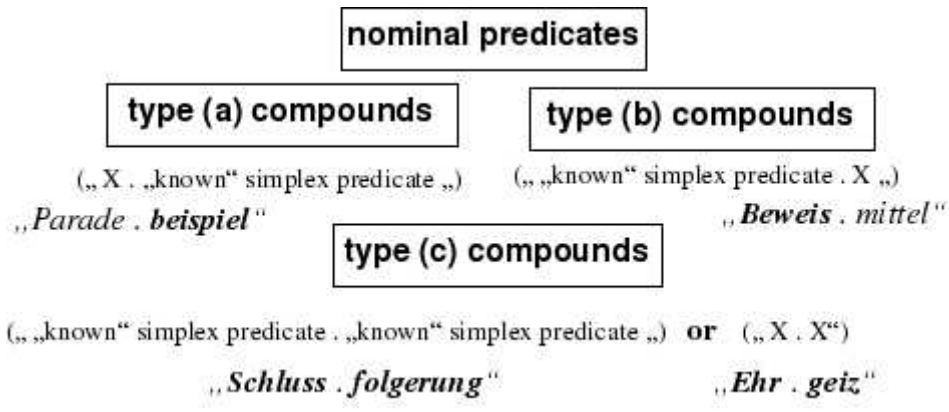

Figure 4: Subclassification of nominal compounds

4. Results and interpretation

Many of the extracted compounds contain deverbal nouns as consituents, e.g. *Beweis* ("proof"), *Auswahl* ("choice"). *Erklärung* ("expaination")). Many of them share their

5 Here we use both the list of extracted simplex predicates and an NLP Lexicon for German.

subcategorization properties with their base verbs, e.g. *beweisen, dass/w-/ob* ("to prove that/wh-/if") or *erklären, dass* ("to explain that").

Interestingly, most (a) types have deverbal nouns as heads and thus as their valency bearers, whereas most (b) types contain deverbal non-heads and most (c) types are mixed, a deverbal noun can be both the head and the non-head component (cf. table 1). In the cases listed in table 1, the subcategorized subclause is determined by the deverbal nominal constituents *Beweis, Schluss* and/or *Folgerung*. The verbs underlying the nominalisations are listed in the right column of the table.

Type	Compound	Deverbal noun	Underlying verb
a	Wahrheits**beweis**, ob...	Beweis, ob...	beweisen, ob...
	("trouth proof if...")	("proof if...")	("to prove if...")
b	**Beweis**last, dass...	Beweis, dass...	beweisen, dass...
	("burden of proof that...")	("proof that...")	("to prove that...")
	Schlussfolgerung, dass...	Schluss, dass...	schliessen, dass...
c	("conclusion that...")	Folgerung, dass...	folgern, dass...
		("conclusion/deduction that")	("to conclude that...")

Table 1: Compounds containing deverbal nominal constituents

We assume that there are correspondences between the subcategorisation of the deverbal valency-bearer of nominal compounds and that of the verbs which underlie the deverbal constituents (cf. the work of Schierholz (2001) on the selection restrictions of verbs and their nominalisations in a prepositional phrase). The analysis of the relationships between the subcategorisation of verbs and their nominalisations is a task for our future work.

Our extraction experiments show that some constituent parts of compounds occur more frequently. In table 2, we give some (b) type examples whose non-head deverbal components are frequent in corpora.

Non-head	Head	Frequency
	Führung	43%
Beweis-	Last	24%
	etc.	33%
	Ansatz	16%
Denk-	Anstoß	13%
	etc.	71%

Table 2: Sample German compounds of type (b) frequent non-heads

The occurrence figures were obtained by extraction in "Vorfeld", as well as additional tests in "Mittelfeld". Table 3 shows that some non-deverbal head constituents of a compound of type (b) can also be frequent.

Non-head	Head	Frequency
Vorgangs-		41%
Sicht-	Weise	22%
etc.		37%

Table 3: Sample German compounds of type (b) frequent heads

A partial evaluation of the described automatic procedures (provided on corpora of 268M words) shows that we can achieve high precision and recall in identifying predicates subcategorizing for a subclause. We achieved an average precision of 95,5% in extracting nominal predicates with sentential complements in "Vorfeld". The successful automatic identification of nominal compounds is proven by a precision of 94,2%, a recall of 95,7% and an f-measure of 94,9% achieved in the analysis of the morphological sorting procedures. Type (b) compounds can be extracted with a recall of 84,5% (for more details, see Lapshinova-Koltunski and Heid (2008)).

5. Compounds in lexicon building

The system for extraction and classification proposed in this paper is relevant for building or updating subcategorisation lexicons and subcategorisation indications of dictionaries for human users. The treatment of compounds with the described set of procedures limits the need for listing all compounds in a dictionary. The automatically extracted compound nominals can be classified and added to entries according to the "detected" valency-bearer noun (cf. figure 5):

Type (a) compounds are listed in alphabetical order, their subcategorisation indications only contain a reference to the respective item of their head.

Type (b) compounds are listed in alphabetical order, their subcategorisation indications are spelled out and additionally contain a reference to the non-head.

Type (c) compounds are listed in alphabetical order and contain their own subcategorisation item.

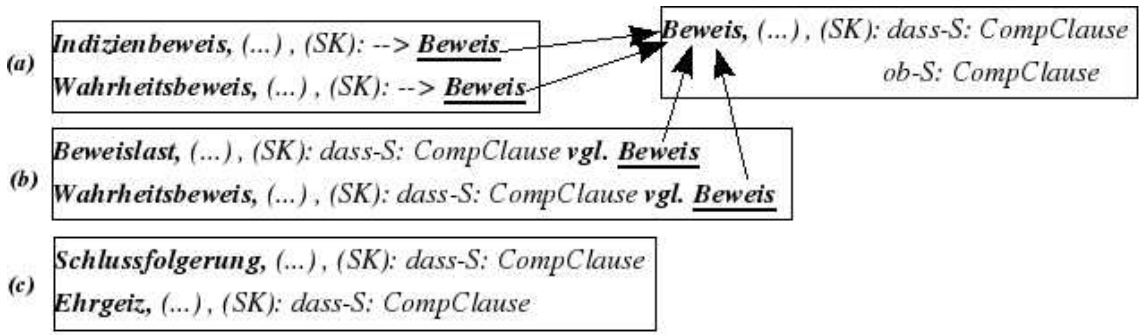

Figure 5: Examples of lexicon entries for (a) to (c) compounds

Most dictionaries do not contain this kind of information. In the DUDEN big dictionary of German, for instance, we find entries for nominal compounds (e.g. for *Beweisführung, Beweislast, Beweismittel* or *Denkansatz, Denkanstoß, Denkbild*), which contain, however, no information about the subcategorisation (cf. figure 6).

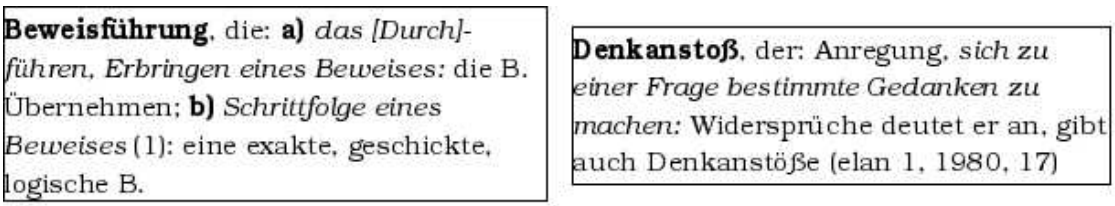

Figure 6: Examples of DUDEN dictionary entries for compounds

Teaching and multilingual processing

A subcategorisation dictionary containing special notes for types (b) and (c) can have an application in language teaching as well as in multilingual natural language processing. If this kind of information is available, a user can differentiate these cases from the "inheritance" cases of type (a), as known from the general rule. This differentiation is important not only in the process of language learning but also for translation of compounds without a loss of information.

6. Conclusion and future work

In this paper, we have shown a set of procedures for a semi-automatic treatment of compounds, which can extract and classify German complex nominals into three groups (types (a) to (c)). The system can operate on a tokenized, tagged, lemmitized (and partially parsed) text.

The described architecture can find an application for the creation and enhancement of subcategorisation lexicons for German. The semi-automatic procedures allow us to treat compounds compositionally which saves time and effort for listing the most frequent compounds in a lexicon.

Our future work will also include an extension of the kinds of extracted complements beyond subclauses, as well as the comparison of the subcategorisation behaviour of compounds, their deverbal constituents and the underlying verbs (*Beweismittel - Beweis - beweisen*).

References

Bauer, L. (1988). *Introducing Linguistic Morphology*. Edinburgh: University Press.

Bergsten, N. (1991). *A Study on Compound Substantives in English*. Uppsala: Almquist and Wiksell.

Bouillon, P.; Bösefeldt, K.; Russell, G. (1992). "Compound Nouns in a Unification-Based MT System". *Proceedings of the Third Conference on Applied Natural Language Processing*. Trento, Italy. 209-215.

Duden: das große Wörterbuch der deutschen Sprache: in zehn Bänden [3]. Mannheim: Dudenverlag, 1999.

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

Feldweg, H.; Hinrichs, E. (1996). "Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen". *Lexicographica. Series Maior,* 73. Tübingen: Niemeyer.

Guenthner, F. (1996). "Electronic Lexica and Corpora Research at CIS". *International Journal of Corpus Linguistics* 1 (2). 287-302.

Guenthner, F.; Maier, P. (1996). "Das CISLex-Wörterbuchsystem". In Feldweg, H.; Hinrichs, E. (eds.). *Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*. Tübingen: Niemeyer. 69 - 82.

Hartrumpf, S.; Helbig, H.; Osswald, R. (2003). "The Semantically Based Computer Lexicon HaGenLex - Structure and Technological Environment". *Traitement automatique des langues* 44 (2). 81-105.

Hippisley, A.; Cheng, D.; Ahmad, K. (2005). "The head-modifier principle and multilingual term extraction". *Natural Language Engineering* 11 (2). 129-157.

Johnston, M.; Busa, F. (1999). "Qualia structure and the compositional interpretation of compounds". In Viegas, E. (ed.). *Breadth and depth of semantic lexicons*. Dordrecht: Kluwer. 167-87.

Kermes, H. (2003). "Off-line (and On-line) Text Analysis for Computational Lexicography". *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)* 9 (3). Stuttgart: IMS.

Lapshinova, E.; Heid, U. (2007). "Syntactic subcategorization of noun+verb multiwords: description, classification and extraction from text corpora". *Proceedings of the 26th International Conference on Lexis and Grammar* [online]. Bonifacio, October 2007. *http://infolingu.univ-mlv.fr/english/Colloques/Bonifacio/proceedings.html*

Lapshinova-Koltunski, E.; Heid, U. (2008). "Head or Non-head? Semi-automatic procedures for extracting and classifying subcategorisation properties of compounds". *Proceedings of LREC-2008* (to appear). Marrakech.

Levi, J. N. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.

Lezius, W.; Dipper, S.; Fitschen, A. (2000). "IMSLex - Representing Morphological and Syntactical Information in a Relational Database". In Heid, U.; Evert, S.; Lehmann, E.; Rohrer, C. (Hrsgg.). *Proceedings of EURALEX*, Stuttgart. 133-139.

Ortner, L. (1991). "Substantivkomposita: Komposita und kompositionsähnliche Strukturen 1". *Deutsche Wortbildung*. Band 4. Düsseldorf : Schwann.

Schierholz, S. J. (2001). *Präpositionalattribute: syntaktische und semantische Analysen*. Tübingen: Niemeyer.

Schmid, H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees"[online]. In *International Conference on New Methods in Language Processing*. Manchester. 44-49. *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/*

Schmid, H. (1999). "Improvements in part-of-speech tagging with an application to German". In Armstrong, S.; Church, K.; Isabelle, P.; Manzi, S.; Tzoukermann, E.; Yarowsky, D. (eds). *Natural Language Processing Using Very Large Corpora*. Vol. 11 of *Text, Speech and Language Processing*. Dordrecht: Kluwer Academic Publishers. 13-26.

Schmid, H.; Fitschen, A.; Heid, U. (2004). "SMOR: A German computational morphology covering derivation, composition, and inflection". *Proceedings of LREC-2004*. Lisbon, Portugal.

Spencer, A. (1991). *Morphological Theory*. Cambridge: Blackwell.

Zwicky, A. M. (1985). "Heads". *Journal of Liguistics* 21. 1-29.