Syntactic subcategorization of noun+verb multiwords: description, classification and extraction from text corpora

Ekaterina Lapshinova-Koltunski [1], Ulrich Heid [2]
Universität Stuttgart, IMS

Abstract

It is commonly accepted that verbal idioms are predicates, i.e. have their own subcategorization (**?**). For support verb constructions, it is assumed that they "inherit" subcategorization from their nominal component (**?**). In this paper, we reexamine the subcategorization of preposition+noun+verb-multiwords, and in particular the dividing line between inheritance and idiosyncrasy. We present semi-automatic precision-oriented tools to extract relevant data from text corpora, and we inspect the first, preliminary results. From the data analyzed so far, it seems that the (semantic) dividing line between SVCs and idioms and the subcategorization-related one between "inheritance" and autonomous predicate status are not isomorphic, as some (non-idiomatic) SVCs can have idiosyncratic subcategorization properties.

Keywords : multiword expressions, subcategorization, extraction, classification.

1. Introduction

In this paper we deal with German multiword expressions (MWEs) consisting of a nominal and a verbal part (verb+PP groups), and we focus on the following two questions: (i) among the linguistic properties of such combinations, is there a need to identify and record their subcategorization behaviour as "multiword predicates", and if so, for which subtypes? (ii) If needed, how can data about this property be reliably extracted from text corpora, and are there automatic methods to do so?

Examples of the MWEs we are interested in are mainly support verb constructions and verbal idioms, such as *in Aussicht stellen* ("to announce", cf. (1)), *zu Protokoll geben* ("to put on record", cf. (2)), *ins Auge fallen* ("to catch sb's eye", cf. (3)), all of which are typically followed by (subcategorized) sentential complements. We focus on sentential complementation, although our methods and conclusions transfer to other subcategorized complements as well.

(1) *Darauf gingen Rektoren und Präsidenten umso bereitwilliger ein, als ihnen in Aussicht gestellt wurde, daß die Hochschulleitungen gestärkt werden sollten* (FAZ) ['rectors and university presidents accepted this all the more happily, because it was announced to them that university management would be reinforced']

(2) *[...] da er [...] zu Protokoll gab, daß heute noch bei [...] beschäftigte Leute dafür verantwortlich gewesen seien* (FAZ) ['because he put on record that people still employed by [...] had been responsible for this']

[1] IMS, Universität Stuttgart, katerina@ims.uni-stuttgart.de
[2] IMS, Universität Stuttgart, heid@ims.uni-stuttgart.de

(3) *[...] daß Ihnen nicht ins Auge gefallen ist, daß durch [...] ein unangemessener Aufwand [...] zugemutet wird* (debate in the German *Bundestag*) ['that it didn't catch your eyes, that [...] imposes an inappropriate effort on [...]']

Our perspective is that of symbolic NLP, especially large symbolic grammars for deep processing, such as HPSG (cf. work in the LinGO project (**?**)) or LFG (cf. the PARGRAM project (**?**)). For these, detailed linguistic knowledge about MWEs is necessary.

In the remainder of this paper, we discuss German MWE data of the kind of (1) to (3) (cf. section 2.1) and the descriptive state of the art (section 2.2), before we present methods for the extraction from corpora of sentences which allow us to derive information about the syntactic subcategorization of the MWEs (section 3). In section 4, we review the extraction results and the methods with a view to the abovementioned questions. We conclude in section 5

2. Data and existing approaches

Phraseological research distinguishes between collocations and idioms, sometimes with an intermediate category of partial idioms ((**?**): 38) or with a subclassification of collocations into transparent vs. opaque ones ((**?**): 8). Collocations are assumed to include support verb constructions (SVCs, cf. (**?**), (**?**), etc.). As the borderline between these categories is not a clearcut one (cf. the detailed discussion of the state of the art, wrt. collocations and their "neighbours" in (**?**): 27 – 78), we will in the following use the term *noun+verb-MWE* to refer to the targeted class of items in a general way, and the terms *SVC*, *collocation* and *idiom* to refer to the common classificatory intuition.

2.1. Data

With respect to the subcategorization of sentence complements with German prepositional noun+verb-MWEs, especially with collocations and idioms, the following cases can be observed. Our classification relies on the relationship between the subcategorization of the MWE and that of its noun component: (a) and partly (b) share it, whereas (c) and (d) don't. All examples contain support verbs (cf. the list proposed by (**?**)).

(a) Noun+verb-MWE which subcategorize for a sentential complement which is also subcategorized by the nominal component of the MWE ("inheritance" from the noun):

(4) *zur Bedingung machen, daß* ("make it a condition")
vs. *die Bedingung, daß ... ist akzeptabel* ("the condition that... is acceptable")

(b) Noun+verb-MWEs which subcategorize for a type of sentential complement which may, under certain contextual conditions, be also present with the nominal MWE component:

(5) *in Erfahrung bringen, ob ...* ("to find out if")
vs. *er hat (die) Erfahrung, daß/*ob/w-* ("he has (the) experience that/*whether/wh-")
vs. *haben Sie (eine) Erfahrung, *daß/ob/w- ?* ("do you have (any) experience *that/whether/wh- ?")

(c) Noun+verb-MWEs which subcategorize for a sentential complement, even though neither their nominal nor their verbal component do so [1], and which are semantically transparent, i.e. do not qualify for the status of idioms:

(6) *zum Ausdruck bringen, daß ...* ("to express")
vs. **der Ausdruck, daß ...*

[1] 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.

(d) Noun+verb-MWEs which subcategorize for a sentential complement, even though none of their components do so, and which are commonly seen as idioms, either because they contain "cranberry" lexemes (7) or because they are non-compositional (8):

(7) *in Abrede stellen, daß ...* ("to deny that") vs. **die Abrede* [2]

(8) *ins Auge fallen, daß ...* ("to catch sb's eye")

Many of the MWEs in types (a) to (c) contain deverbal nouns (e.g. *Erfahrung*, *Ausdruck*); many of these share their subcategorization properties with the underlying verbs (e.g. *erfahren* ("to find out") + *daß/ob/w-*). The extent to which verb, nominalization and SVC are parallel in this respect needs yet to be explored in detail. For the purpose of further discussion, we group the cases (a) and (b) together (cases with "inheritance"), as well as (c) and (d) ("non-inheritance" cases).

2.2. Approaches to the linguistic description of noun+verb-MWEs

Whether or not support verb constructions should be considered as predicates with their own subcategorization properties has been discussed by (**?**) who suggested that the predicative noun provided its subcategorization properties to the SVC. This approach is applicable to the (a) cases above, or to SVCs such as *make + proposal*, *come to + decision*, *cherish + hope*, etc. [3]. For the (b) cases to be captured by the same rule, reasoning about the truth values of the sentence complement within and outside the SVC context would be needed (see (**?**)). Recently, (**?**) noted exceptions to the 'inheritance' rule (cf. type (c)), such as *jmdm einen Rat erteilen* ('to give advice to sbdy'), however without deeper analysis. In Storrer's example neither *Rat* nor *erteilen* take a dative (exception: other SVCs with *erteilen*), but the SVC requires one; the verb *raten*, however, does take a dative. For idiomatic expressions (cf. (d)), (**?**):60ff and (**?**):40ff suggest a treatment as valency bearers in their own right.

Lexicography often provides less formal subcategorization patterns: (**?**): 357 use dependency structure to (manually) describe the syntactic form of noun+verb-MWEs, which in principle would allow for a subcategorization description. (**?**) give informal indications of complements (e.g. *"jemandem {Informationen|eine Antwort|...} erteilen"* (p. 455)), but do not address the issue neither. Only Krenn's database of SVCs (cf. (**?**): 190s) contains a field for the subcategorization frame of SVCs, illustrated by a three-place frame (NP_{nom}, (NP_{dat}), NP_{acc}) for *zur Verfügung stellen* ('put at someone's disposal', btw. another example of type (c)); (**?**):303 and (**?**): 983s mention the need for subcategorization frames of SVCs, but neither they nor Krenn extract data about this property automatically.

3. Methods and tools for the extraction of noun+verb-MWEs in context

3.1. Input

Our work starts from sentence-tokenized German newspaper corpora which are pos-tagged and lemmatized [4]. To assess the need for chunking (see section 4.3), we used YAC, a recursive chunker for German (cf. (**?**)). Regular expressions for data extraction rely on the Stuttgart CorpusWorkBench (CWB [5]). We use texts from Germany, Austria and Switzerland, a total of ca. 950M words [6].

[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] We do not put a determiner in the citation forms, as it depends on the context; if there is a *that*-clause subcategorized by the noun, there is a preference fo a definite determiner.

[4] We use (**?**)'s TreeTagger/lemmatizer and the STTS tagset, cf. the URL:
`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`.

[5] Cf. (**?**).

[6] Austrian (referred to by 'AT' in our tables in section 4, ca. 300M) and Swiss ('CH' ca. 180M) texts 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*

3.2. Linguistic assumptions and extraction architecture

Identification. For most of the work we report here, we concentrate on occurrences of noun+verb MWEs in verb final sentences (i.e. German subclauses, ca. 20-25% of all corpus text), because in these constructions the nominal element of the MWE tends to be immediately left-adjacent to the verbal one, (cf. (9) and (**?**)), and the subclause following the verb is typically subcategorized by the MWE as a whole or by its noun.

	Query building blocks	comments	extracted sentence
1.	[pos="KOU.*|PREL.*|PW.*"]	conjunction, relative or interrogative pronoun	weil
2.	[pos!="V.*FIN"&word!=",|-"]*	optional words, no finite verbs, no punctuation	Clinton und seine Anwälte erst
3.	[pos="APPR|APPRART"]	preposition or prep. + article	in
4.	[]{0,1}	optional word	
5.	[pos="NN"]	support(ed) noun	Erfahrung
6.	[pos!="NN|PPER"&lemma!="da.*"]{0,3}	up to 3 words, no noun, personal pron. nor pron. adverb	
7.	<vc>	verbal complex:	
8.	[lemma=RE($verblist)]	with verb from SV list	bringen
9.	[pos="VM.*|VA.*|V.FIN"]{0,3}	auxiliaries	wollen
10.	</vc>	end of verbal complex	
11.	","	comma	,
12.	[	start of complement clause	
13.	(pos="PW.*"&word!="wobei|womit")	interrog. pronoun, no sent. advs. "wobei", "womit"	was
14.	| (word="ob")	or conjunction "ob"	
15.	| (word="dass")]	or conjunction "daß"	
16	within s;	within a sentence	Lewinsky zu sagen hat, um ...

Figure 1. Query for extraction of multiword expressions in verb-final position subcategorizing a wh-, ob- or daß-clause

The extraction patterns (cf. figure 1) search for verb-final sentences containing a MWE (cf. lines 3 to 10). We extract complex sentences consisting at least of a main clause and one subordinate *daß*-, *wh*- or *ob*-clause. The multiword predicate whose subcategorization we are exploring is followed by a comma (line 11) and a conjunction *daß* or *ob* or a *wh*-word which introduces a subordinate clause (cf. alternatives in lines 13 to 15). The right hand column of figure 1 and 2 contains a matching sentence, as an example.

Subclassification. To automatically classify the candidates according to (a) to (d) (cf. section 2.1), we compare the results of the query in figure 1 with data obtained from another context, namely NPs in the Vorfeld, i.e. left of the finite verb, cf. example (10) and the query in figure 2.

(9) *[...], wie sie etwa dadurch zum Ausdruck käme, daß ...* ("how it roughly was expressed by the fact that... (lit.: how it ... came to the expression ...")

(10) *Die im [...] vorgesehene Bedingung, daß die Einschnitte nicht erfolgen, [...] scheint...* ("the condition foreseen in [...], that the cuts do not happen, seems [...]")

If a noun in Vorfeld position is followed by a sentential complement, this complement can only be subcategorized by the noun. Consequently, cases which fulfil both context tests, (9) and (10),

tageszeitung ('taz', 111M), *Frankfurter Rundschau* ('FR', 40M), *Frankfurter Allgemeine Zeitung* ('FAZ', 71M), *Stuttgarter Zeitung* ('StZ', 36M), as well as literary texts from the 'Gutenberg' Archive ('DE Lit.', 138M).

belong to type (a) of our classification; cases where *daß* and *ob/w-* are switched between the two models belong to type (b). Cases with different subcategorization in models (9) and (10) belong to types (c) and (d); a complete automatic distinction between these two types is not intended here (cf. e.g. (**?**) for work on the automatic separation of idioms and (rather compositional) collocations); to single out some of the (d)-cases, we can obviously identify "cranberry" lexical items that appear only in the idioms.

The query in figure 2 has a noun phrase (or a PP) at the beginning of the sentence (lines 1 to 4), followed by a subordinate clause (lines 5 to 11) and the finite (main) verb.

	Query building blocks	comments	extracted sentence
1.	<s>	sentence beginning	
2.	[pos!="NN|V.FIN"]{0,3}	prenominal material	Allein
3.	[pos="APPR|APPRART"]?	optl. preposition or prep/art	
4.	(<np> ... </np>)	noun phrase	die Ankündigung
5.	","	comma	,
6.	(pos="PW.*"& word!="wobei|womit")|	relative pronoun, but not "wobei" and "womit" or	
7.	(word="ob")|	conjunction "ob" or	
8.	(word="dass")	conjunction "daß"	dass
9.	[pos!="$.|V.FIN"]*	subclause: non-verbal part	er
10.	[pos="V.FIN"]	finite verb of subclause	komme
11.	","	comma	,
12.	[pos="V.FIN"]	finite main verb	hatte
13.	within s;	rest of main clause	den Börsenkurs vergangene Woche in die Höhe getrieben.

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

4. Results and interpretation

4.1. Sample results

In table 1, we summarize some absolute frequency figures from one of our extraction exercises, based on the newspapers *FR*, *FAZ* and *taz* (see note 6, above, for the abbreviations), a total of ca. 220 M words.

MWE type	MWE, noun	dass				wh-				ob			
		-SV		+SV		-SV		+SV		-SV		+SV	
		VF	VL	VL	MF	VF	VL	VL	MF	VF	VL	VL	MF
a+b	in **Aussicht** stellen	55	28	60	47	0	0	0	0	0	0	0	0
	zur **Bedingung** machen	37	137	59	25	0	0	0	0	0	0	1	0
	in **Erfahrung** bringen	40	46	39	10	2	1	17	0	0	0	13	1
c+d	in **Rechnung** stellen	5	3	53	7	2	1	1	0	0	0	0	0
	zum **Ausdruck** kommen	0	8	24	63	0	0	0	0	0	0	0	1
	in **Abrede** stellen	0	0	25	17	0	0	0	0	0	0	0	0
	in **Vergessenheit** geraten	0	0	34	12	0	0	0	2	0	0	0	0

Table 1. Sample German noun vs. SVC pairs and their complement clauses, by word order models

Overall, the phenomena we analyze are quite rare; thus we get low frequencies, especially for cases (b), (c) and (d). In the table, we analyze *in Aussicht stellen* ('announce'), *zur Bedingung machen* ('use as a condition'), *in Erfahrung bringen* ('find out'), as well as *in Rechnung stellen* ('account for'), *zum Ausdruck kommen* ('be expressed'), *in Abrede stellen* ('deny') and *in Vergessenheit geraten* ('fall into oblivion').

For the first three MWEs, we have both a sufficient number of true MWE cases (columns '+SV') and enough non-MWE uses of the nouns (i.e. Vorfeld occurrences, columns '-SV'). With the

last four MWEs, we have few or no Vorfeld occurrences, which suggests that they are limited to SVC or idiom uses, i.e. belong to types (c) or (d). In fact, for "cranberry" nouns in idioms, we trivially expect no occurrence outside the MWE (cf. *Abrede, Vergesseneheit*) [7].

To operate a rough classification into the types (a)/(b) vs. (c)/(d), as introduced in section 2.1 above, we determined the numbers of occurrences of each subclause type (*daß, ob, w-*), for both support verb uses (+SV) and non-support-verb uses (-SV), and for the German word order models. For the latter, we give separate figures: VF in table 1 stands for 'Vorfeld', and points to the non-support-verb test discussed in figure 2, above. Obviously, non-support-verb uses of the nouns may also appear in other word order models, especially the verb-last model (column VL under -SV). Similarly, even though our extraction tools mainly rely on verb-last contexts, we provide the number of Mittelfeld occurrences (MF) of SVCs in a separate column.

4.2. Interpreting the figures

Judging from the small sample in table 1, we can assume that our tools throw up useful results. By comparing the -SV and +SV columns, we can determine MWEs of types (a) or (b) as those which show up significantly under both conditions. A separation into (a) vs. (b), i.e. an identification of the "switching" of truth values for complement clauses characteristic of (b) can be observed with *in Erfahrung bringen*: the SVC seems to accept *daß*-clauses the same way as the noun does outside SVCs, but the SVC reading in addition shows up consistently with indirect questions (both *wh-* and *ob*), which is not the case with the non-SVC use of the noun.

Moreover, the table shows which kinds of subclauses the MWEs prefer: for example with *in Aussicht stellen, in Rechnung stellen, in Abrede stellen, in Vergessenheit geraten*, *ob*-clauses have not been found. *Abrede* and *Vergessenheit* seem to take sentential complements only within MWEs.

More "idiomatic" support verb constructions (our type (c)) are *zum Ausdruck bringen/kommen* ("to express"/"to be expressed"), *zur Sprache bringen/kommen* ("to mention"/"to be mentioned"), or *zu Protokoll geben* ("to put on record").

A full quantitative evaluation of the tools would require the preparation of a gold standard corpus. We have not yet been able to do this. A partial evaluation of the non-support-verb cases (in Vorfeld position), on the *FR* corpus (40 M words), provided a precision of 67,2 %, a recall of 56,2 %, and thus an F-measure of 61,2 %.

4.3. Linguistic information required for the extraction work

We start from tokenized, tagged and lemmatized text, and we make use of the grammatical properties of verb-final sentences (cf. figure 1) and of the Vorfeld (see figure 2), built into the queries.

The modelling of verb-final sentences depends to some extent on detailed models of NPs and PPs, and thus profits from additional preprocessing by means of chunking or partial parsing. For the Vorfeld cases, no partial parsing is needed: we compared figures of extraction patterns with and without NP and PP boundaries annotated (absolute frequency indicated under '+chunks' and '–chunks' in table 2). The chunked corpora provide slightly less data, but most of the cases found without the use of chunking ('diff' in table 2) proved to be true positives ('TP', absolute figures and percentages).

[7] For *Abrede*, see note 2 above; with *Vergessenheit*, we have, besides the targeted idiom, also *der Vergessenheit anheimfallen* ('fall into oblivion, idiomatic'), *der Vergessenheit entreißen* ('avoid that sth falls into oblivion', idiomatic), and *nach langer Vergessenheit* ('after a long period of oblivion'), a total of 11 % of the data.

Source	type	+chunks	–chunks	diff.	TP
taz	Vorfeld+wh-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+wh-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 2. Extraction results in Vorfeld position with and without chunking

5. Conclusions and future work

Our work is still in an early stage, but our first experiments show that certain support verb constructions have their own subcategorization properties which are not 'inherited' from their nominal elements. With respect to subcategorization, such SVCs thus behave like idioms, even though their semantics is not fully idiomatic: the syntactic behaviour is not fully parallel to the (semantic) distinctions known from phraseology.

This situation calls for tools to identify such cases by means of data extraction from corpora. We propose precision-oriented semi-automatic extraction which can operate on tokenized, tagged and lemmatized text; partial parsing is not essential for this task. We focus on the precision of the automatic extraction tools, relying on the availability of enough data to compensate for the fact that we deal with a low frequency phenomenon, and for the lower recall caused by the precision oriented approach.

The comparison of the observed frequency data of both, SVC occurrences and uses of the noun outside SVCs, allows us to broadly classify the MWE candidates, in terms of their preferences for *daß-*, *ob-*, *w-*clauses, and with respect to the "inheritance" hypothesis; we think that the observed classes (a)/(b) vs. (c)/(d), i.e. [with inheritance] vs. [without inheritance], are stable even despite the theoretically problematic status of null occurrences.

Future work will include an extension of the kinds of extracted complements beyond subclauses, and the use of more data, to achieve substantial coverage. We will also address the comparison with the subcategorization behaviour of the verbs underlying the targeted nominalizations (*erfahren – Erfahrung – in Erfahrung bringen*).

Thanks

This work is part of Ekaterina Lapshinova-Koltunski's PhD work, which is supported by the DFG-funded Research Training Group GK-609 "Linguistic Representations and Their Interpretation"; Ulrich Heid's contribution is in addition part of the project *Collocations en Contexte* which is partly funded by the Agence Universitaire pour la Francophonie, Réseau LTT.