Providing Lexicographers with Corpus Evidence for Fine-grained Syntactic Descriptions: Adjectives Taking Subject and Complement Clauses

Ulrich Heid and Hannah Kermes
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Azenbergstr.12, 70174 Stuttgart, Germany
{heid,kermes}@ims.uni-stuttgart.de

Abstract

This article deals with techniques for lexical acquisition which allow lexicographers to extract evidence for fine-grained syntactic descriptions of words from corpora. The extraction tools are applied to partially parsed text corpora, and aim to provide the lexicographer with easy to use syntactically pre-classified evidence. As an example we extracted German adjectives taking subject and complement clauses.

1 Introduction

Large monolingual dictionaries intend to cover the most important semantic and syntactic aspects of words. This includes quite prominently the description of syntactic subcategorization. For the more frequent syntactic constructions, data are easily accessible; idiomatic constructions often readily spring to the lexicographer's mind when the headword is being analyzed.

But there are facts that are less in sight and yet should be noted in a large dictionary. An example - which will be used in this article to illustrate our approach - is the construction of (German) adjectives with sentence complements (e.g. *that x happens is relevant for me*).

For such phenomena, which are less frequent in corpora but characteristic of the words under analysis, automatic sampling of evidence from large corpora seems appropriate, especially if the results are presented in a lexicographically useful way. The results of our acquisition programs are formatted in HTML and displayed in a web browser which allows the lexicographer to view the data sorted and presented according to different criteria (alphabetically, by frequency, by construction, etc.) with example sentences just one click away.

On the basis of a detailed linguistic analysis, we have extracted data for individual adjectives, automatically classified the data with respect to the syntactic constructions observed, and collected frequency data, for each construction type. The latter is important, partly because phenomena of lexical combinatorics (and thus preferences) seem to play a role, in addition to the syntactic construction, and partly because the relative (in-)frequency of certain constructions can be noted.

In the following, we first describe the phenomena, our tool and the criteria used for automatic classification, then we present sample results of the extraction and finally we compare some examples with the respective entries in *Duden Universalwörterbuch* (DUW, one volume, 4/2001) and in *Duden. Das große Wörterbuch der deutschen Sprache* (GWDS, 8/10 vols).

2 Phenomena

We looked at German adjectives subcategorizing finite subject and/or complement clauses. Although most of our data are from adjectives with *dass*-clauses, essentially the same procedures can be used to extract indirect interrogatives, and only minor changes are needed to cover infinitival clauses as well. Examples for constructions of adjectives with sentential subjects are displayed in (1) and (2).

(1) a. Daß er ihn sieht, ist klar.
That he can see him is evident.
b. *Ihn zu sehen, ist klar.
To see him is evident.

(2) a. *Daß er ihn sieht, ist schwer.
That he can see him is difficult.
b. Ihn zu sehen, ist schwer.
To see him is difficult.

The examples in (1) show that the adjective *klar* can have a sentential subject realized as a *dass*-clause (1a) but not as an infinitival clause (1b). The adjective *schwer*, on the other hand, can take a sentential subject realized as an infinitival clause (2b) but not as a *dass*-clause (2a).
The adjective *stolz* in (3) can take a prepositional phrase (PP) (3a), a *dass*-clause (3b) and an infinitival clause (3c) as complement. The sentential complements take the position of the prepositional object. In both (3b) and (3c) a pronominal adverb (*darauf*, "Korrelat") can optionally occur.

(3) a. Sie ist auf das Bild stolz.
She is proud of the picture.
b. Sie ist stolz [darauf], dass sie ausgewählt wurde.
She is proud [of it] that she was selected.
c. Sie ist stolz [darauf] ausgewählt worden zu sein.
She is proud to have been selected.

The examples in (1) to (3) show that different adjectives follow different selectional restrictions with respect to the possibility of taking sentential subjects or complements. The syntactic frames of adjectives seem related to the semantic and lexical properties of the adjective. Thus, the ability of adjectives to take sentential subjects or complements and what kind they can take, should be described in a dictionary.
Monolingual dictionaries tend to indicate these facts sporadically (see section 4.3), but not systematically. Even the specialized dictionary of adjective valency by Sommerfeldt/Schreiber [3/1983; 1/1974] includes only very few sentential complements (e.g. s.v. *würdig*). The authors classify complement clauses as variants of nominal or prepositional complements (p. 29); however, they do not mention the restrictions described above, nor any of the details described below.
There are basically two positions for sentential subjects: (i) the topic position at the beginning of the sentence (the so-called *Vorfeld*) (4a), and (ii) the extraposed position at the end of the sentence (the so-called *Nachfeld*) (4b+c).

(4) a. Daß die Bahn sich beteiligt, ist klar.
That the railway corporation participates, is clear.
b. Es ist klar, daß die Bahn sich beteiligt.
It is clear that the railway corporation participates.
c. Schließlich ist (es) klar, daß die Bahn sich beteiligt.
Finally (it) is clear, that the railway corporation participates.

If the sentential subject is extraposed the topic position is either filled by the expletive *es* (4b) or by another element (4c) with the expletive *es* optionally occurring between the verb and the adjective.
Additionally, the adjective can subcategorize datives (5a+b) or prepositional phrases (5c+d). Both dative and prepositional object can occur in the topic position with optional expletive *es* between the verb and the adjective (5a+c), as well as between the verb and the adjective, with the expletive *es* in topic position (5b+d).

(5) a. Mir ist (es) klar, daß ich mich erst informieren will.
b. Es ist mir klar, daß ich mich erst informieren will.
c. Für mich ist (es) klar, daß ich mich erst informieren will.
d. Es ist für mich klar, daß ich mich erst informieren will.
It is clear to me that I want to inform myself first.

Certain verbs subcategorize predicative adjectives; these verbs may embed the adjective and its complement (or subject) clause as in (6):

(6) a. Es ist klar, daß er kommt.
It is clear that he comes.
b. Es scheint/wird/.. klar, daß er kommt.
It seems/gets .. clear that he comes.
c. Er hält es für/nennt es/... klar, daß er kommt.
He takes is for/puts it to be/calls it/ .. clear that he comes.

The combination of verbs and adjectives, although free in principle, seems to be governed by preferences similar to collocational preferences. For example, *deutlich werden* is much more frequent than *deutlich sein.* A good dictionary should mention such preferences.
Some of the adjectives can appear in an elliptical construction, without a verbal predicate: sentences like those in (7) are rather frequent. However, the fact that only certain adjectives can enter this construction is idiomatic; at least there are clear preferences.

(7) a. Schon möglich, daß man sich mit solchen Sätzen schwertut.
It is well possible that one has difficulties with such sentences.
b. Verständlich, daß er sich in „Germany" wohlfühlt.
It is understandable that he feels at home in "Germany"
c. Wirklich schade, daß sie zumachen.
It is really a pity that you close.

The phenomena sketched above have recently been (at least partially) discussed by Sandberg 1998. He made a corpus analysis of the "Mannheimer Corpora" of the Institut fuer Deutsche Sprache. Sandberg [1998]discusses some adjectives in detail (including *klar, sicher, bekannt, deutlich* (clear, certain, known, evident)). We will discuss these adjectives in section 4. Sandberg, however, focuses on the presence or absence of the expletive *es*, therefore, he

does not cover all of the phenomena we touch upon. Studying the corpus data for *es*, Sandberg notes that a much broader collection of texts than the Mannheimer Corpus is needed to allow to make any claim based on frequency. Thus the figures extracted from our corpus (40 M words) are to be taken with caution as well.

3 Acquisition tools for fine-grained lexical and syntactic description

3.1 Corpus-linguistic tools

Our corpora are tokenized and part-of-speech tagged with Helmut Schmid's TreeTagger (cf. [Schmid 1994a] and [Schmid 1994b])[1]. Lemma and agreement information is annotated using the IMSLex morphology [Lezius et al. 2000][2].

The corpora are then partially parsed by the Three Level Incremental Partial Parser (TLIPP), a fully automatic tool based on a symbolic regular expression grammar (cf. [Kermes & Evert 2001]). The rules of the grammar are written as queries in the CQP corpus query language, as it is used in the IMS Corpus Workbench (CWB) [Christ 1994][3].

TLIPP is designed to provide a basis for the extraction of linguistic information, e.g. for lexicographic use, yet it delivers no full parse. The idea is to build up relatively flat annotations of certain (maximal) syntactic constituents incrementally: adverbial phrases (AdvP), adjectival phrases (AP), noun phrases (NP), prepositional phrases (PP) and verbal complexes (VC)[4]. In addition, certain lexical and structural features of chunks and phrases (e.g., head lemmas, agreement information and lexical properties) are annotated. Annotating syntactic constituents rather than chunks is necessary, as complex phrasal structures involving (recursive) embedding in pre-head position are rather common in German. Chunking in the sense of annotating non-recursive kernels of phrases cannot cover these structures sufficiently, especially, if the annotation is meant to help the extraction of linguistic information.

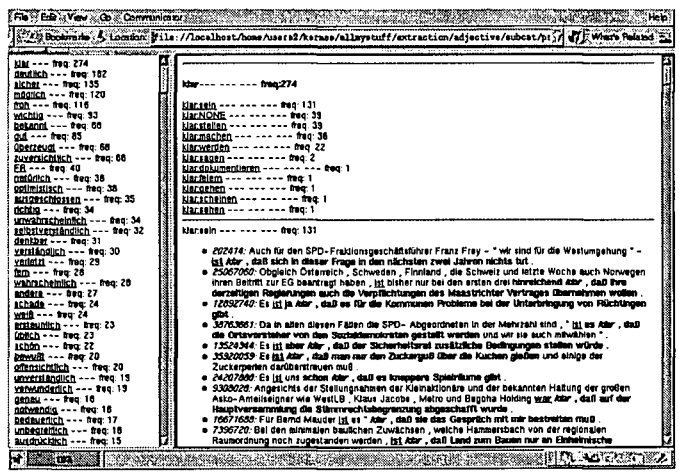

Figure 1: Sample screen displaying the adjective *klar* (clear).

3.2 Tools for extraction and presentation

When the corpus has been annotated with syntactic constituents. Queries performed on this corpus can use the structural mark-up introduced by TLIPP and the feature attributes of the annotations. We have designed an extraction module that can additionally apply multiple filters to exclude or include results that meet certain linguistic criteria.
The extraction tool is further able to sort the results according to different lemmas or according to different elements of the query, without having to change the query itself. The results are displayed in HTML, for example sorted by the lemma chosen. Example sentences are optionally displayed and linked to the corresponding sorting item. See
Figure 1, for a sample screen, where the left column contains a list of adjective candidates, the top right window the combinations of the adjective *klar* with predicative verbs and the bottom right window examples for *klar* and *sein.* The material was extracted from a corpus[5] containing two years of the newspaper *Frankfurter Rundschau* (40 Million words).

3.3 The case of adjectives: the criteria and queries for the automatic classification of corpus samples

In section 2, above, we introduced the phenomena we wish to cover in the case study on adjectives. For convenience, we summarize the criteria we applied in the extraction tool in Table 1 below, along with pointers to the example sentences from section 2.

No.	Criterion	Examples
1	two-place vs. three place	(5)
2	sentence = subject vs. sentence = complement	(1), (2) vs. (3)
3	[+/- *es*] or [+/- *dar-*]	(3), (4)
4	Predicative verb	(6)
5	Use without verbs	(7)

Table 1: Criteria for the classification of German adjective complements

We did not limit the extraction results to certain adjectives or verbs as we wanted to find out: (i) whether predicative adjectives really only occur with the well-known list of so-called predicative verbs, including most prominently sein (be), werden (become), bleiben (remain), nennen (call), (ii) whether additional verbs can occur with predicative adjectives, (iii) whether and what kind of prefences predicative adjectives have with respect to verb selection, (iv) whether similar constructions with other verbs have an idiomatic or collocational character, (v) which predicative adjectives can occur without a main verb and in what constructions, and (vi) whether all predicative adjectives can occur with a dass-clause in extraposed and in topicalized position, and whether they prefer one or the other construction?
We applied the following queries:

1. Predicative adjectives with a dass-clause in extraposed position. The query searches for adjectival phrases (AP), which are not part of an NP, followed by a dass-clause. The corresponding verb is the next verb left of the AP within the sentence boundary. Excluded are results with a reflexive pronoun and a pronominal adverb between the AP and the verb. These results form a subclass to be stored in a separate list.

2. Predicative adjective with a dass-clause in extraposed position with no main verb. The query searches for APs in sentence initial position followed by a dass-clause.
3. Predicative adjectives with a dass-clause in topicalized position. The query searches for a dass-clause, followed by a verbal complex (VC), any number of adjuncts and arguments, and an AP in sentence final position.

4. Results and Interpretation

4.1 General lexicographic aspects

The extraction tools provide results of good quality. In order to exclude noise (constructions such as the consecutive construction *so+ADJ, daß* (so ADJ that), mistagged adverbials, etc.) we refined the queries that were informally listed in section 3.3.
HTML presentations as shown in
Figure 1allow the lexicographer to examine the results in an easy and comfortable way. Currently the results are sorted by constructions (e.g. all examples of topicalized *daß*-clauses, sorted by adjective, then by verb). A summary is currently being developed. It will allow the lexicographer to get a quick overview of all constructions of a given adjective.
The corpora used so far (newspaper texts from *Frankfurter Rundschau* (2 years, 40 million words) and *Stuttgarter Zeitung* (2 years, 36 million words) provide usable evidence for the most frequent adjectives (see below for details). Larger corpora (several hundreds of millions of words) will soon allow to document the syntactic behaviour of adjectives more thoroughly, giving information about less frequent adjectives as well.

4.2 Sample results: Adjectives

In the following, we will illustrate the extraction results with a few selected answers to the questions asked above, in section 3.3.
We first comment on the availability of topicalized (that x ... is ADJ) vs. extraposed ([it] is ADJ that x) constructions. The adjective *klar* (clear) occurs with the verbs *sein* and *werden* in both constructions. The extraposition of the sentential argument is preferred. Other adjectives prefer this construction type to an even greater extent, especially the adjectives *bekannt*, *sicher*, *möglich* and *wichtig* (see
Table 2 for frequency data). Note that the topicalized construction seems always possible: it is found equally with low frequency adjectives, if, however, sporadically. A lexicographer should thus give an extraposed construction as an example, at least in entries of less frequent adjectives; with high frequency adjectives, like *klar*, it may be helpful to also give a topicalized example, to remind dictionary users of this possibility. In an electronic product, it may be useful to make figures like those in
Table 2 accessible to the user as well.

Adjective	predicative verb	freq. top[6]	freq. ex[7]
klar (clear)	sein	131	20
	werden	22	3
bekannt (known)	sein	49	1
	werden	30	0
sicher (certain)		113	2
möglich (possible)		56	3
wichtig (important)		83	3

Table 2: Frequency data for sentential arguments in extraposed and topicalized position (in 40 million words)

Another descriptive question raised in section 3.3 concerns the selectional restrictions with respect to the predicative verbs and possible preferences.

adjective	verb	frequency
klar (clear)	sein	151
	werden	25
deutlich (clear)	werden	63
	sein	2
sicher (certain)	sein	115
	gelten als	11
möglich (possible)	sein	59
	halten für	29
bekannt (known)	sein	50
	werden	30

Table 3: Frequency data for selectional preferences with respect to the predicative verbs (in 40 million words)

The figures in Table 3 show that most adjectives (*klar, sicher, möglich, bekannt*) occur most likely with the predicative verb sein. The degree of preference differs, however, *klar* and *sicher* showing the clearest preference. *Möglich* occurs quite frequently with *halten für* as well, including idioms such as *x sollte y nicht für möglich halten (e.g. man sollte nicht für möglich halten, daß...*(you wouldn't believe that ...)). The verb+adjective combination with *bekannt* is at the borderline to form an autonomous verb. *Deutlich* shows a clear preference for *werden* over *sein*. For a large dictionary such collocational preferences are highly relevant, as they prove to be stable across the two corpora analysed and seem to be confirmed by larger corpora.

Finally, we summarize in

Table 4 the most prominent cases of adjectives encountered significantly often without a verb, in an elliptical (or: absolute) construction. We indicate the frequency in *Frankfurter Rundschau* (FR) and Stuttgarter *Zeitung* (STZ), the adjective lemma and possible modifying adverbs.

Frequency in FR	frequency in STZ	adjective lemma	adverbs
39	42	klar (clear)	-
25	17	gut (good)	wie, nur/bloß
23	11	möglich (possible)	gut, schon, durchaus, nicht
14	7	verständlich (understandable)	fast (1*)
12	14	schade (pity)	zu, wirklich, nur
8	5	schön (nice)	wie
7	1	erstaunlich (astonishing)	umso erstaunlicher
0	5	dumm (bad)	zu
2	1	bedauerlich (regrettable)	wirklich

Table 4: Adjectives frequently used without embedding predicative verbs

It can be noted that the adjectives occur either alone or modified by an adverb. Selectional restrictions are responsible for (i) what kind of adjective can occur in such a construction, (ii) which of these adjectives can be modified by an adverb, and (iii) what kind of adverbs can function as modifiers for which adjectives. An example of such a construction is *(Schon) möglich, daß er das nicht will ([it is] (well) possible that he doesn't like this).* The adverb *möglich* was found 34 times in such a construction. It occurred either alone or modified by the adverbs *schon, gut, durchaus* and *nicht.* These combinations are clearly collocational in nature, and a dictionary should list them.

4.3 A comparison with monolingual dictionaries

Here we can only sketch the results of an informal comparison with *Duden. Das Große Wörterbuch der deutschen* Sprache (GWDS), and (the related) *Duden Universalwörterbuch* (DUW). This comparison was done manually, for the most frequent adjectives, *klar, bekannt, deutlich, möglich, sicher, wichtig.*

We observe the following:

- neither of the dictionaries has any topicalized example for any of the six adjectives above;
- DUW mentions the *daß*-clause construction for all adjectives except *bekannt*; GWDS mentions it in the entries of all six adjectives;
- three-place constructions (possible with all six adjectives) are explicitly given in DUW for *klar*, in GWDS for *klar, sicher* and *wichtig*;
- the dictionaries sporadically indicate other predicative verbs than *sein*: especially GWDS has combinations like *bekannt werden/vorkommen, deutlich machen*, etc.
- the elliptical construction of *möglich* is present in both dictionaries (*gut möglich, leicht/sehr wohl möglich, daß* in GWDS and DUW), but, e.g., the entry of *klar,* which also is quite frequently used in an elliptical construction, does not include this information.

5 Conclusion

The first results of the extraction work carried out on adjectives with *daß*-clauses seem to indicate that it is worth while exploiting the potential of a layered partial parsing of large corpora and subsequent specific corpus query for lexicography. A quite broad documentation becomes available to the lexicographer concerning specific lexical and syntactic issues. It does not overload the lexicographer to provide additional material, as the extraction, sorting, pre-classifying and presentation are done automatically (typically off-line). Since it is possible to view the data by lemma as well as across a given phenomenon, the relative importance of a single phenomenon for a class of items can be checked easily by the lexicographer.

In the near future, we expect to improve the tools further, as far as adjectives are concerned. In the medium term, we hope to be able to compile a library of extraction procedures of this kind, and make it available as options to provide data in a tool suite for automatic excerption (cf. [Heid et al. 2000]).

This work is similar to that of Kilgarriff (cf. [Kilgarriff & Tugwell 2001]), in the WASPS project. It focuses, however, more on specific syntactic phenomena (for which there is not yet enough documentation).WASPS is aimed more at providing summaries of the most important phenomena for any particular item. The procedures and techniques can in principle be used to collect material for a broader set of phenomena as well. Examples of those under study include the use of nouns in singular and plural, noun+noun collocations and detailed data about support verb constructions.

6 References

[Abney 1991] Abney, S. (1991). Parsing by chunks, in: R. Berwick, S. Abney, and C. Tenny (eds.) *Principle-Based Parsing.* Kluwer Academic Publishers.

[Abney 1996] Abney, S. (1996). Partial parsing via finite-state cascades, in: *Proceedings of the ESSLLI '96 Robust Parsing Workshop.*

[Christ 1994] Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system, in: *Papers in Computational Lexicography COMPLEX '94.* Budapest, Hungary.

[Heid et al. 2000] Heid, U., Evert, S., Docherty, V., Worsch, W. and Wermke, M. (2000). A data collection for semi-automatic corpus-based updating of dictionaries, in: *Proceedings of the 9th EURALEX International Congress.* Stuttgart, Germany.

[Kermes & Evert 2001] Kermes, H., and Evert, S. (2001). YAC – A Recursive Chunker for Unrestricted German Text. Ms., submitted to LREC 2002.

[Kilgarriff & Tugwell 2001] Kilgarriff, A., and Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. in: *Proceedings workshop "COLLOCATION: Computational Extraction, Analysis and Exploitation"*, pp.32-38. 39th ACL & 10th EACL, Toulouse, July 2001.

[Lezius et al. 2000] Lezius, W., Dipper, S., and Fitschen, A. (2000). IMSLex – representing morphological and syntactical information in a relational database, in: *Proceedings of the 9th EURALEX International Congress.* Stuttgart, Germany.

[Sandberg 1998] Sandberg, Bengt (1998). *Zum* es *bei transitiven Verben vor satzförmigem Akkusativobjekt.* Narr, Tübingen [=Tübinger Beiträge zur Linguistik, 443].

[Schmid 1994a] Schmid, H. (1994a). Part-of-speech tagging with neural networks, in: *Proceedings of the 15th International Conference on Computational Linguistics (Coling '94)*, pp. 172-176. Kyoto, Japan.

[Schmid 1994b] Schmid, H. (1994b). Probabilistic part-of-speech tagging using decision trees, in: *International Conference on New Methods in Language Processing,* pp. 44-49. Manchester, UK.

[Sommerfeldt & Schreiber 1983] Sommerfeldt, Karl-Ernst and Schreiber, Herbert (1983). *Wörterbuch der Valenz und Distribution deutscher Adjektive*. Max Niemeyer, Tübingen (1/1974, Leipzig: VEB Verlag Enzyklopaedie).

7 Endnotes

1 For more information see:
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
2 For more information see: http://www.ims.uni-stuttgart.de/projekte/IMSLex.
3 For more information see:
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/index.html.
4 Steve Abney conducted a similar approach for English using a cascaded finite-state parser [Abney 1991; Abney 1996].
5 The text was made available via the European Corpus Initiative, ECI; it covers all issues of Frankfurter Rundschau of 1992 and 1993.
6 Frequency of topicalized sentential arguments
7 Frequency of extraposed sentential arguments.