A Representation Scheme and Database for German Support-Verb Constructions

Brigitte Krenn

In this paper, we present a description model for support-verb constructions (SVCs), and a relational database for storing both competence-driven and corpus-based descriptions of SVCs. We use a syntactically annotated corpus for extraction of actual occurrences of SVCs.

In diesem Papier wird ein Beschreibungsmodell für Funktionsverbgefüge (FVG) vorgestellt. Kompetenz- und Performanzdaten zu FVGen sind in einer relationalen Datenbank gespeichert. Für die Beschaffung von Daten des Sprachgebrauchs wird ein syntaktisch annotiertes Korpus benutzt.

KONVENS-98

1 Introduction

Support-verb constructions (SVCs) are an essential feature of language usage. In order to develop grammar theories and natural language applications that properly account for SVCs, a better understanding of these phenomena is indispensable. As a precondition for theoretical linguistics research and as input for processing components linguistically marked-up SVC data are required.

In this paper, a description model for SVCs is presented, see section 3, and a relational database for storing both competence-driven and corpus-based data is described, see section 5. Section 2 gives a brief introduction to SVCs.

2 Support-Verb Constructions

2.1 Characteristics

SVCs are particular **verb-object collocations** that typically function as paraphrases for main verbs or adjective-verb combinations. They consist of lexically determined elements, at least a noun and a verb.

The so called **predicative noun** is usually deverbal or deadjectival. It comes with its own argument structure, and contributes the major part of the meaning to the SVC. In many cases, the noun is accompanied by a preposition, see the examples given in this paper. The noun normally combines with more than one verb. We distinguish SVC instances and SVC types. A type consists of an NP or PP and a set of verbs. Each individual noun-verb combination constitutes an instance, see example (1). Henceforth we use SVC for both instance and type whenever the interpretation is clear from the context.

The main function of the verbs is to express **Aktionsart** and **causativity**. Causativity increases the argument structure by one, the causer of a process or state. Aktionsart expresses a particular phase of a process or state. We distinguish 4 Aktionsarten: inchoative (begin of process or state), terminative (end of process or state), continuative (continuation of process or state) and neutral.[1] The Aktionsart of an SVC, however, is not exclusively determined by the verb lemma, see example (1). Here, a change of the preposition is required to express terminativity, *außer Betrieb* {*setzen* (termin, caus), *gehen* (termin)} ('to take out of service', 'to go out of service').

(1) *in Betrieb* { *gehen* (incho), *nehmem/setzen* (incho, caus), *sein* (neut), *bleiben* (contin), *lassen* (contin, caus) }
('to go into operation', 'to put into operation'/'to start up', 'to be running', 'to keep on running', 'to keep (something) running')

2.2 Grammatical Stability versus Grammatical Flexibility

A particular characteristic of collocations is that they obey on the one hand the generative rules of grammar, and on the other hand they are restricted with respect to morphological and syntactic variation, and modification. Grammar rules that apply to verb-object constructions also apply to SVCs, i.e. predicative noun and whole SVC can be modified, some SVCs can be passivized, SVCs can occur as adjectival prenominal modifiers, etc. The grammatical potential, however, is restricted at individual SVC instances. In addition, SVCs are productive. New instances and types emerge, vary in individual usage and between language domains. A representation model thus must allow for description of unchanging properties of an SVC, but also provide means to represent less predictive aspects.

[1] For a thorough discussion of Aktionsart and causativity in SVCs see [12].

3 A Representation Model for Support-Verb Constructions

In the following, we present a representation scheme that is well suited to account for the double-sided nature of SVCs as grammatically flexible verb-object combinations and as lexically fixed constructions. SVCs are described with respect to morphological variation in the predicative phrase, structural variation of the SVC, and modification of the predicative noun and the SVC. Apart from this description scheme which provides the competence grammatical background, an extensive collection of real occurrences in texts is given which is useful for investigating and modelling the actual realizations on the basis of the grammatical potential.

3.1 Lexical Representation of the Collocation

In order to account for morphological (in)variation of the collocates, the verb, the predicative noun and the preposition if existing are represented either by their lemma or, in the case of complete morphological invariability, by the appropriate full form. Additional information is specified like number and case of the predicative noun, parts-of-speech of the collocates, Aktionsart (a-art) and causativity (caus) of the SVC, see figure 1, *zur Verfügung stellen* ('to make available').[2]

The preposition occurs in two realizations *zu* and *zur*, i.e. without determiner or with determiner included *zur* (APPRART) = *zu* (APPR) + *der* (ART). The form of the noun is fixed, the verb is represented by its lemma indicated by small capitals. The case of the noun is fixed to dative due to the preposition, its number is fixed to singular. The SVC is causative, its Aktionsart is inchoative.

predicative phrase		case	number	verb	a-art	causativity
zu(r)	Verfügung	dat	sg	STELLEN	incho	+
APPR(ART)	NN			VV		

Figure 1: lexical description of an SVC

3.2 Syntactic Structure and Modification of the Predicative Phrase

SVCs structurally divide into two parts, the verb and the NP or PP comprising the predicative noun. According to competence grammar, standard

[2] The part-of-speech tags are taken from [18]. APPR stands for preposition, ART for article, NN for noun, VV for verb.

prenominal modifiers in German NPs or PPs are NP_{gen} and ADJP, postnominal modifiers are NP_{gen}, PP and ADVP. This general modification pattern is clearly restricted in SVCs.

Looking at the modification of randomly selected SVCs from a 7 million word corpus, we get strong evidence that predicative nouns typically occur without modification. 148 of 170 PPs that are constituted by the preposition *in* (in/into) and the noun *Kraft* (force) consist of the two words only. For *außer* (out of) and *Kraft* we have 21 to 22 total, for *ins Gespräch* ('into a conversation') we have 56 to 60, *in Betrieb* ('into operation') 116 to 137, *zur Verfügung* ('at the disposal of') 622 to 665, *in Führung* ('in the lead') 101 to 124, *in Brand* ('on fire') 61 to 63, *in Vergessenheit* ('into oblivion') 20 to 25, *in Anspruch* ('claims') 199 to 207, *in Frage* ('into question') 166 to 257.

In the case of *ins Gespräch* the corpus data also reveal that the noun is either modified by a PP_{mit} (3 instances) ('into a conversation with') or the SVC is modified by *miteinander* (1 instance) ('into a conversation with each other'). These data correspond to human introspection but are too rare to be statistically meaningful. In general, data on modification of predicative phrases tend to be rare; for instance in a sample of 665 PPs containing *zur Verfügung*, there were only two instances of prenominal modification, namely *zur eigenen Verfügung* ('at one's own disposal').

Another typicality of predicative phrases is restricted determination, see for instance *zur/zu Verfügung*. The occurrence of an article like *der* or *ein* would be odd, **zu der/einer Verfügung* ('at the/*a disposal of'). A possessive pronoun, however, is perfect, *zu seiner Verfügung* ('at his disposal').

In the competence part of the representation, see figure 2, modification and determination is explicitly blocked like with *in Frage*, spelled out with respect to dominant variants, see postnominal modification (mod_{post}) at *ins Gespräch*, or left underspecified in more flexible cases, like prenominal modification (mod_{pre}) and determination at *ins Gespräch*. Especially for flexible cases, corpus data are well suited to provide information on the distribution of actual occurrences.

det	mod_{pre}	mod_{post}	predicative phrase
-	-	-	in Frage
		PP_{mit}	ins Gespräch

Figure 2: determination and modification in the predicative phrase

3.3 Syntactic Structure and Modification of the SVC

SVCs can also be modified as a whole. A number of modifiers can be found adjacent to the predicative phrase like {*automatisch, endlich, gerade, sofort, sogar, später*} *in Kraft* ({automatically, finally, just, immediately, even, later} into force). Variation of SVC modifiers is less predictable than variation of predicative phrase modifiers. For a corpus-based approach, large amounts of data need to be processed for identification of a representative modifier selection for an SVC as occurrence frequencies are low.

The predicative noun, if it is deverbal or deadjectival, has its own argument structure. From a purely syntactic perspective, the verb can be seen as head of the SVC. The predicative phrase requires one argument slot, verb and noun arguments merge. See for instance *in Betrieb nehmen* ('to put into operation'). Here the arguments of *nehmen*(NP_{nom}, NP_{acc}, PP_{in}) and *betreiben*(NP_{nom}, NP_{acc}) (to run) which is the verb underlying *Betrieb* merge, and the predicative noun enters the PP.

In Betrieb nehmen expresses causativity. For a non-causative reading the causer, realized as NP_{nom}, needs to be eliminated. This can be achieved either by passivization (sentence (2)) or by exchanging the verb (sentence (3)). Here no causer argument is available.

(2) *die Anlage wurde in Betrieb genommen*
('the system was put into operation')

(3) *die Anlage ist in Betrieb gegangen*
('the system went into operation')

For a representation of the argument structure of SVCs see figure 3, where the obligatory arguments are listed. The predicative phrase is excluded as its appearance can be retrieved from the word level description, recall figure 1. Modifiers are only listed if they are typical for a particular SVC.

argument structure	SVC
NP_{nom} NP_{acc}	in Betrieb nehmen
NP_{nom}	in Betrieb sein

Figure 3: argument structures of SVCs

4 Extraction Corpus

The examples given in this paper are derived from the Frankfurter Rundschau Corpus, a German newspaper corpus which is part of the ECI Multilingual CD-ROM 1. Parts of the corpus have been automatically annotated with parts-of-speech (incl. punctuation) and phrasal chunks like NP, PP, ADJP. The tools used are described in [1] and [16].

Availability of phrasal information allows to identify typical modification patterns within the predicative phrase and with respect to the complete SVC. Additional evidence for verb collocates can be found in NP/PP chunks containing an SVC as adjectival modifier like *die in Kraft gesetzten Verbote* ('the prohibitions brought into force').[3]

Summing up, the corpus is used for extraction of example sentences containing prespecified SVCs, but it is also used to identify SVC candidates by means of statistics, cf. [10].

5 A Database for the Representation of SVCs

5.1 Characteristics of the Database

We use a relational database that has been developed for storing and retrieving linguistic data. The database is portable and highly flexible in its interfaces. The implementation in ANSI C supports Unix platforms, Macintosh and Intel-based personal computers, cf. [14]. We have chosen the database for the following reasons:

Adequacy As the database has been designed with a focus on the development of data for natural language research and applications, the database kernel is small and flexible, retrieval by string manipulation (regular expression matching) is supported, the interface allows connection to arbitrary applications.

Flexibility The database consists of a binary file comprising the engine and a library of interface functions, a relations file where the data objects (relations) with appropriate arguments and types are stored, and one file for each relation comprising one record per line with fields of variable width. Both relations and data files are plain ASCII. The user is free to define the

[3] These phrases are rare. Their detection depends on the syntactic information available. In our corpus, they can be found by looking for predicative phrases in APPR* NN ADJA NN patterns. ADJA stands for attributival adjective.

data format. Relations and attributes can be easily changed or extended by manipulation of the relations file and string operations on the data files.

Exchangeability The database is non-commercial and runs on different platforms. If required, the data can be easily exported to other databases.

5.2 The Data Scheme

Three major types of information are represented in the data scheme, namely book-keeping information, linguistic description and examples. In terms of database relations we have: collocation-instance, ci-structure, collocation-realization and cr-structure which are described below.

Book-keeping information Corpus name and sentence number are stored in order to trace back the origin of a particular example. In addition, comments can be added to each SVC description and example.

Linguistic description For each SVC, a number of attributes are specified for storing the linguistic descriptions presented in section 3.

Examples For each SVC example sentences derived from corpora are stored. Each word is annotated with parts-of-speech and its position and function within syntactic structure whereby structure is restricted to phrasal chunks. This is because annotation relies on automatic processing as large amounts of data are required for example-based representation of SVCs. Shallow approaches to processing like part-of-speech tagging and phrase chunking offer for the time being the best trade-off between time efficiency and robustness of processing, and reliability of the linguistic information. The predicative phrase and the verb receive additional annotation so that they can be accessed without the need to retrieve the whole sentence.

collocation-instance (ci) This relation defines attributes for the representation of the morphological realization of the collocates, for Aktionsart and causativity.

ci-structure This relation consists of attributes for the representation of parts-of-speech of the collocates, the syntactic structure of the predicative phrase, modification information on the predicative phrase and the whole SVC, and the argument structure of the SVC. Summing up, this relation represents the competence-based syntactic analysis of an SVC.

collocation-realization (cr) Here the corpus-derived example sentences and the according book-keeping information are stored.

cr-structure This relation links the example sentence with a description of its predicative phrase and verb collocate.

6 Related Work

The work presented in this paper relates to research on corpus-based identification of collocations, cf. [17], [3], [6], [7], [8], [15], [4]. In contrast to these approaches where collocation extraction is based on plain word forms or part-of-speech annotated forms, the work presented here makes use of a syntactically annotated corpus. The major benefits of such an approach are:
(1) The accuracy of frequency counts increases, i.e. more syntactically plausible collocation candidates are found. In [17] it has already been shown that syntactic analysis of phrases containing candidate collocations helps to discard false candidates.
(2) Grammatical restrictions related to collocations can be automatically derived from the corpus, especially with respect to modification of the predicative phrase and particular types of modification of the SVC.

A completely different approach to collocation identification is taken in the DEFI project, here general-purpose dictionaries are automatically exploited with respect to multi-word units (MWU). The lexeme database is then used for online identification of MWUs in corpora, cf. [5]. While DEFI aims at providing tools for automatic translation of MWUs, our database aims at providing linguistically rich and usage adequate monolingual data as a basis for theoretical linguistics investigations into the nature of collocations and for various nlp applications.

The work also relates to investigations into the description and representation of collocations as described in [11], [2], or [9]. While these papers present approaches to competence-based description and representation of collocations and idioms, the work presented here is the first to combine competence-based and corpus-based descriptions of collocations which is indispensable for a better understanding of collocation phenomena. With respect to this aspect, the work is comparable to the project DiET, diagnostic and evaluation tools for natural language processing applications, that aims at using competence and performance data (corpus data) for evaluation of nlp systems, cf. [13].

7 Conclusion

The representation scheme and database presented in this paper provide a basis for detailed investigations into the nature of collocations. The approach originally designed for SVCs is expected to be easily extendible to other kinds of collocations as the major characteristics of collocations – grammatical flexibility versus lexical fixation – are already modeled with respect to SVCs. Due to the flexibility of the database interfaces, extension of the representation scheme for instance with a semantic representation for SVCs or extension with data from other languages shall be unproblematic. The current database comprises competence-based descriptions of 65 SVCs and related examples retrieved from the Frankfurter Rundschau Corpus.

References

[1] Thorsten Brants. Tnt – a statistical part-of-speech tagger. Technical report, Universität des Saarlandes, Computational Linguistics, 1996.

[2] E. Breidt, F. Segond, and G. Valetto. Formal description of multi-word lexemes with the finite-state formalism IDAREX. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 1036 – 1040, Copenhagen, Denmark, 1996.

[3] Elisabeth Breidt. Extraction of n-v-collocations from text corpora: A feasibility study for german. In *Proceedings of the 1st ACL-Workshop on Very Large Corpora*, 1993.

[4] Vincent J. Docherty, Ulrich Heid, and Judith Eckle-Kohler. Computational linguistic support for the revision of a commercial dictionary – corpus-based updates of entries and of collocation information. In *Beiträge zur 6. Fachtagung der Sektion Computerlinguistik DGfS-CL*, Heidelberg, Deutschland, Oktober 1997.

[5] Nicolas Dufour. A database for computerized multi-word unit recognition. In *Proceedings of ISP-3*, Stuttgart, Germany, 1998.

[6] Katerina T. Frantzi and Sophia Ananiadou. Extracting nested collocations. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 41 – 46, Copenhagen, Denmark, 1996.

[7] Masahiko Haruno, Satoru Ikehara, and Takefumi Yamazaki. Learning bilingual collocations by word-level sorting. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 525 – 530, Copenhagen, Denmark, 1996.

[8] Saturo Ikehara, Satoshi Shirai, and Hajime Uchino. A statistical method for extracting uniterrupted and interrupted collocations from very large corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 41 – 46, Copenhagen, Denmark, 1996.

[9] Martina Keil. *Wort für Wort - Repräsentation und Verarbeitung verbaler Phraseologismen.* [Sprache + Information.] Niemeyer, Tübingen, 1997.

[10] Brigitte Krenn. Acquisition of phraseological units from linguistically interpreted corpora. A case study on German PP-verb collocations. In *Proceedings of the 3rd International Symposium on Phraseology*, Stuttgart, Germany, March 1998.

[11] Brigitte Krenn and Martin Volk. DiTo-Datenbank. Datendokumentation zu Funktionsverbgefügen und Relativsätzen. Technical report, Dokument D-93-24, Deutsches Forschungszentrum für Künstliche Intelligenz, 1993.

[12] Nadia Mesli. *Les Locutions Verbales dans les Ecrits de M. Luther.* PhD thesis, Université d'Aix-Marseille I, 1989.

[13] Klaus Netter et al. DiET – diagnostic and evaluation tools for natural language processing applications. In *Proceedings of LERC-98*, Granada, Spain, 1998.

[14] Stephan Oepen, Klaus Netter, and Judith Klein. TSNLP — Test Suites for Natural Language Processing. In John Nerbonne, editor, *Linguistic Databases*, CSLI Lecture Notes 77. Center for the Study of Language and Information, 1998.

[15] Sayori Shimohata, Toshiyuki Sugio, and Nunji Nagata. Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, 1997.

[16] Wojciech Skut and Thorsten Brants. Chunk Tagger. Stochastic Recognition of Noun Phrases. In *ESSLI Workshop on Automated Acquisition of Syntax and Parsing*, Saarbrücken, Germany, August 1998.

[17] Frank Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143 – 177, 1993.

[18] Christine Thielen and Anne Schiller. Ein kleines und erweitertes Tagset fürs Deutsche. In *Tagungsberichte des Arbeitstreffens Lexikon + Text 17./18. Februar 1994, SchloßHohentübingen. Lexicographica Series Maior*, Tübingen, 1995. Niemeyer.