CDB – A Database of Lexical Collocations

Brigitte Krenn *

*Brigitte Krenn
Austrian Research Institute for Artificial Intelligence (ÖFAI)
Freyung 6/3/1a, Vienna, Austria
brigitte@ai.univie.ac.at

Abstract

CDB is a relational database designed for the particular needs of representing lexical collocations. The relational model is defined such that competence-based descriptions of collocations (the competence base) and actually occurring collocation examples extracted from text corpora (the example base) complete each other. In the paper, the relational model is described and examples for the representation of German PP-verb collocations are given. A number of example queries are presented, and additional facilities which are built on top of the database are discussed.

1. Introduction

Language usage is full of partially rigid word combinations, which we will call lexical collocations here. A proper treatment of this kind of linguistic entities is particularly important for domain-specific applications, but also for computational linguistics applications in general, such as machine translation, lexical selection in generation and parse pruning.

The term collocation here is used for word combinations that are lexically determined and constitute particular syntactic dependencies such as verb-object, verb-subject, adjective-noun relations, etc. A specific feature of a broad range of collocations is that the syntactic constructions involved obey only partially the generative rules of grammar. Thus appropriate representations need to provide means to account for so called grammatical irregularities in collocations. In current approaches to collocation dictionaries or databases, a hand-crafted local grammar is specified for each collocation representing morphological and syntactic properties of the components, position and type of external modifications, permissible syntactic transformations, etc. See for instance (Segond and Tapanainen, 1995), (Breidt et al., 1996), (Keil, 1997), (Tschichold and Hacken, 1998).

A serious drawback of this kind of approach is that explicit descriptions of collocations do not meet the tendency of collocations to vary with respect to domain and speaker. As a consequence, these purely competence-based approaches either over- or undergenerate. An attempt to overcome these shortcomings is presented in (Dufour, 1998). Dictionary entries of collocations are represented by linguistic descriptions containing features which are associated with numeric weights reflecting the intuitions of the human annotator. Unfortunately these weights are hard to interpret.

2. Goal of the paper

In the paper, an alternative approach to the description of collocations is presented. A relational database and its implementation is described where the representation of a collocation consists of

- a competence base: an abstract, partially over-generating competence-grammatical description, and
- an example base: a collection of real-world occurrences of the particular collocation.

The relational model is determined by the task of linking collocation instances (types) and linguistic descriptions of collocation-class-specific information with collocation examples (token) derived from various corpora. In addition, book-keeping information such as corpus name and sentence number relative to the corpus is stored, in order to allow the origin of a particular example to be traced back, so that larger contexts can be accessed. The database is extendible in depth with respect to linguistic descriptions and in width with respect to collocation instances and corpus data stored. A relational approach is required as thus flexible views on the data can easily be provided. This is particularly important, because the database, on the one hand, is conceived as a research tool which supports the development of collocation theories and, on the other hand, it is intended to function as a collocation lexicon for parsers and generators.

For a distinction of collocation instances and collocation examples see section 3. The relational model is described in section 4. and the database management system in section 5. Example queries are presented in section 6. The functionality on top of CDB is discussed in section 7.

3. The Data

3.1. Selection of Collocation Instances

The current database contains more than 1 000 German PP-verb collocations (types). The collocations have been selected from a set of preposition-noun-verb (PNV) triples which have been extracted from an 8 million word
3.2. Selection of Collocation Examples

The collocation instances are applied for identifying collocation examples from corpora. In particular, carrier sentences of PP-verb collocations are extracted from arbitrary corpora. The current database contains examples from a subcorpus (8 million words) of the Frankfurter Rundschau Corpus and from a German newsgroup corpus (10 million words). The newsgroup corpus is part of a corpus collected in the FLAG-project at the German Research Institute for Artificial Intelligence (DFKI), Saarbrücken (http://www.dfki.de/pas/f2w.cgi?ltp/flag-e).

The corpora used for example selection are part-of-speech tagged and annotated with rudimentary NP and PP structure (phrase chunks). The particular part-of-speech tagger and phrase chunker employed are described in (Brants, 1996) and (Skut and Brants, 1998), respectively. Both, tagger and chunker are trained on the Negra Corpus (cf. http://www.coli.uni-sb.de/sfb378/negra-corpus/). An example for the information available with an automatically preprocessed corpus sentence is shown by the partially bracketed sentence

Der Verband stellt ihm lediglich die Infrastruktur zur Verfügung.

(The union makes to him only the infrastructure available.)

Two NPs and a PP have been identified by the chunker, the other elements of the sentence are left unattached. Each word is annotated with a unique part-of-speech.\(^1\)

\[
\begin{align*}
(P_P & \text{ zur APPRAR T} \\
 & \text{Verfügung NN } )_{PP} \\
(P_N & \text{ die ART} \\
 & \text{Infrastruktur NN } )_{NP} \\
(N_P & \text{ der ART} \\
 & \text{Verband NN } )_{NP} \\
(NP & \text{ Der ART} \\
 & \text{stellt VVFIN} \\
 & \text{ihm PPER} \\
 & \text{lediglich ADV} \\
\end{align*}
\]

3.3. Corpus-Based Update of Collocation Instances

While collocation instances are employed for selecting collocation examples from corpora, the examples in turn are used to detect further potential for generalizations of collocation instances. For example, the collocation instances zur Verfügung stellen (to the availability put, 'make available') and zu Verfügung stellen (to availability put, 'make available') can be generalized to zur? Verfügung stellen with zur? being a regular pattern representing the disjunction of the words zu and zur. Generalizations of collocation instances already stored in the database involve changes of the identification numbers of the collocation instances involved as well as changes of all entries in the database which are related to the generalized instances. In particular, the relations COLLOCATION-INSTANCE, CI-ANALYSIS and COLLOCATION-REALIZATION are affected.

4. The Relational Model

The relational model of CBD is defined by four base relations which are linked via keys. The relations COLLOCATION-INSTANCE and CI-ANALYSIS constitute the competence base, the relations COLLOCATION-REALIZATION and CR-STRUCTURE constitute the example base. The relations and their attributes are described in sections 4.1 to 4.4. An extra attribute for comments is defined for each relation.

4.1. COLLOCATION-INSTANCE CI

Summing up, collocation instances in our terms are generalized representations of the major lexical elements, the collocates, of a collocation. Since the current database only contains PP-verb collocations, collocation instances are represented by generalized PNV-triples. Prepositions may be generalized over the plain variant and a fusion between preposition and determiner such as zu (to) and zur (to the) in zur? Verfügung stellen. Nouns are represented as full forms, verbs as infinitives. This kind of encoding reflects the linguistic observations that prepositions in PNV-collocations, in some cases, may vary between plain preposition and fusion of preposition and determiner, and that nouns usually do not vary in their inflection, whereas verbs typically do.

<table>
<thead>
<tr>
<th>ci-id</th>
<th>ci-string</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>zur? Verfügung stellen</td>
</tr>
<tr>
<td>2014</td>
<td>zur? Verfügung stehen</td>
</tr>
<tr>
<td>2015</td>
<td>zur? Verfügung haben</td>
</tr>
<tr>
<td>1745</td>
<td>in Betrieb gehen</td>
</tr>
<tr>
<td>1746</td>
<td>in Betrieb nehmen</td>
</tr>
<tr>
<td>1751</td>
<td>außer Betrieb setzen</td>
</tr>
<tr>
<td>1752</td>
<td>außer Betrieb gehen</td>
</tr>
<tr>
<td>2802</td>
<td>unter Lupe nehmen</td>
</tr>
<tr>
<td>2823</td>
<td>am Herzen liegen</td>
</tr>
</tbody>
</table>

Table 1: The relation COLLOCATION-INSTANCE and its attributes

Collocation instances are stored in the attribute ci-string. Each instance is associated with an identification...
number (attribute ci-id). See table 1 for an illustration of the relation COLLOCATION-INSTANCE.²

4.2. CI-ANALYSIS

The relation is designed for representing collocation-class-specific linguistic descriptions. It is part of the competence base, because the representations are determined by linguistic theory. Thus it comes closest to standard representations of collocations. In the current version of CDB, the relation is specified for Funktionsteilbeugungen (FV) representing those characteristics for which a broad consensus exists in the literature. For a summary of current research on FV see (Krenn, 2000), p. 74ff.

In order to account for the fact that collocation-class-specific properties vary from class to class, and that linguistic descriptions of individual collocation classes are expected to change with increasing theoretic insight into the phenomenon, three basic attributes are specified for CI-ANALYSIS, namely ci-id, ci-attrib, and ci-value, with the values of ci-attrib and ci-value being pairwise defined for each data-record.

For the description of FV, eight values for the attribute ci-attrib have been specified. Caus stands for causativity, i.e., the existence of the thematic role of a causer or cause in the argument structure of the FV. A-art stands for Aktionsart of the collocation. Four values are distinguished: inchoative – representing the beginning stage of a process or state, continuative – representing the durational aspect of a process or state, terminative – representing the final stage, and neutral. The value of reciproc is the identification number of the causative or noncausative partner collocation, e.g. zur Verfügung stehen (‘make available’, causative) and zur Verfügung stehen (‘be available’, non-causative) are reciprocal. The value of args is the list of syntactic arguments required by an FV. P-det, p-modpre and p-modpost specify properties of the predicative phrase, in our examples the PP-collocate, with p-det referring to determination, p-modpre and p-modpost referring to pre- and postnominal modification, respectively. In mods, modifiers applying to the whole FV are specified. Possible values for the previous four attributes are: <realizations> which is a variable for a list of permissible or typical realizations; blocked (-) which states that the attribute has a null value; unspecified (u) which states that the range of the permissible values is in accordance with the general rules of grammar. As this assumption is too general in most cases, the particular realizations are read from the corpus examples.

The description of the FV zur? Verfügung stellen in CI-ANALYSIS reads as follows, see table 2. The collocation is causative, and has inchoative Aktionsart. The FV has three syntactic arguments: an NPnom, an NPacc, an NPdat. The surface realization of the NPdat is optional as far as competence grammar is concerned. Information on determination, prenominal modification of the PP-collocate and modification of the whole FV is underspecified. Postnominal modification in the PP-collocate is blocked, i.e., no postnominal modification is allowed. The reciprocal collocation is zur? Verfügung stehen.

4.3. COLLOCATION-REALIZATION CR

The relation is defined for storing sentences identified from corpora which contain occurrences of a particular collocation instance. For each example sentence, the following information is represented: the surface realization cr-sent including the part-of-speech and chunk tags; a unique identification number cr-id; the number of the collocation instance ci-id of which the sentence is an example; the kind of source source-type the example has been retrieved from, such as newspaper corpus, corpus of computer manuals, newsgroup corpus, etc.; the name of the corpus source-name; and the number of the sentence sent-num relative to the other sentences in the corpus from which the collocation example has been selected. In addition, an attribute c-type is specified representing the collocation type. The current database entries fall into two groups of collocation types: Funktionsteilbeugungen and figurative expressions (figur). For a definition of FV see (Bußmann, 1990). Figurative expressions are PNV-combinations that require figurative interpretation. An example entry in the relation COLLOCATION-REALIZATION is given in table 3. For the convenience of the reader, the example sentence is presented without part-of-speech and chunk tags. In order to fit the table to the column, the value of cr-sent is represented by a placeholder (*).³

Table 2: The relation CI-ANALYSIS specified for the FV zur? Verfügung stellen

<table>
<thead>
<tr>
<th>cr-id</th>
<th>ci-attrib</th>
<th>ci-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>caus</td>
<td>+</td>
</tr>
<tr>
<td>2012</td>
<td>a-art</td>
<td>incho</td>
</tr>
<tr>
<td>2012</td>
<td>reciproc</td>
<td>2014</td>
</tr>
<tr>
<td>2012</td>
<td>args</td>
<td>NPnom (NPdat) NPacc</td>
</tr>
<tr>
<td>2012</td>
<td>p-det</td>
<td>u</td>
</tr>
<tr>
<td>2012</td>
<td>p-modpre</td>
<td>u</td>
</tr>
<tr>
<td>2012</td>
<td>p-modpost</td>
<td>-</td>
</tr>
<tr>
<td>2012</td>
<td>mods</td>
<td>u</td>
</tr>
</tbody>
</table>

Table 3: The relation COLLOCATION-REALIZATION and its attributes; * here is a placeholder for the sentence 3800 Quadratmeter Fläche auf drei Etagen stehen in dem Neubau nun dort zur Verfügung, wo einst Kühe in Ställen untergebracht waren.

³A translation of the sentence is 3800 square meters are now available on three floors in the new building where once cows were kept.
In **CR-STRUCTURE**, individual realizations of PNV-collocation instances are represented by 5 kinds of substrings which are:

1. the prepositional collocate,
2. the nominal collocate,
3. the verbal collocate,
4. the substring beginning with the prepositional collocate and ending with the nominal collocate,
5. the string stretching from the PP-collocate (4.) to the verbal collocate.

In all cases, the words, the part-of-speech and the chunk tags are stored.

The following attributes are specified for **CR-STRUCTURE**: **cr-id**, the identification number of the particular collocation example; **cr-substring**, one of the five substrings described above; **cr-position-begin** (cpb), the begin position of the particular substring; **cr-position-end** (cpe), the end position of the particular substring; **cr-category**, a label representing the syntactic category of the particular substring; **cr-function**, a label representing the function of the particular substring, such as: \( v \_\text{col} \) for the verbal collocate; \( p \_\text{col} \), \( n \_\text{col} \) and \( pp \_\text{col} \) for the prepositional, the nominal and the PP-collocate, respectively; \( V \_Pp \) and \( PP \_V \) are the possible function labels for string 5.

Table 4 shows the entries in **CR-STRUCTURE** for example sentence 508 (table 3). Again, part-of-speech and chunk labels are omitted.

<table>
<thead>
<tr>
<th>cr-id</th>
<th>cpb</th>
<th>cpe</th>
<th>cr-substring</th>
<th>cr-cat</th>
<th>cr-func</th>
</tr>
</thead>
<tbody>
<tr>
<td>508</td>
<td>7</td>
<td>7</td>
<td>stehen</td>
<td>VV</td>
<td>v_col</td>
</tr>
<tr>
<td>508</td>
<td>13</td>
<td>13</td>
<td>zur</td>
<td>APPR-ART</td>
<td>p_col</td>
</tr>
<tr>
<td>508</td>
<td>14</td>
<td>14</td>
<td>Verfügung</td>
<td>NN</td>
<td>n_col</td>
</tr>
<tr>
<td>508</td>
<td>13</td>
<td>14</td>
<td>zur Verfügung</td>
<td>PP</td>
<td>pp_col</td>
</tr>
<tr>
<td>508</td>
<td>7</td>
<td>14</td>
<td>stehen in dem Neubau nur dort zur Verfügung</td>
<td>VP</td>
<td>V-PP</td>
</tr>
</tbody>
</table>

Table 4: The relation **CR-STRUCTURE** and its attributes

The entries in **COLLOCATION-REALIZATION** and **CR-STRUCTURE** are automatically generated from the example sentences. The entries in **CI-ANALYSIS**, on the other hand, are largely hand-crafted representing a competence-based linguistic description of a certain collocation class. Similarly, the entries in **COLLOCATION-INSTANCE** are initially constructed independently from the collocation realizations. In the long term, prevalent regularities in the corpus examples related to a particular collocation instance are employed to modify the entries in **CI-ANALYSIS** and **COLLOCATION-INSTANCE**.

## 5. The Core Machinery

The database management system tsdb(1) (Oepen et al., 1998) is used as core machinery. Tsdb(1) has been developed in the TSNLP-project at DFKI, Saarbrücken.\(^4\)

The database has been chosen because of its small and flexible kernel. The complete database consists of:

- a binary file comprising the engine and a library of interface functions;
- the relations file storing the names of the base relations and the headings, i.e., the names of the permissible attributes and the types of their values;
- a data file for each base relation comprising the body (the data) of the relation.

The relations file and the data files are plain ASCII. The user is free to define the data format. Thus new relations and databases can be easily set up, which is an important feature for experimenting with new strategies for the description of collocations. Headings and bodies can be easily changed or extended by manipulation of the relations file and by string operations on the data files.

Retrieval by string manipulation (regular expression matching) is supported. The database can be easily connected to arbitrary applications. Another important criterion for choosing tsdb(1) was that the core engine is non-commercial and runs on different platforms. Because of the organization of the data in plain ASCII files, portation to other databases is also easy. The use of a commercial database is suggested when the number of collocation-instances and related examples becomes very large. For the future, portation of CDB to the DiET database management system is envisaged.\(^5\)

## 6. Example Queries

Other than the base relations presented in section 4., query results are derived relations. Queries to base relations enable establishing new relations on the data. Thus they provide new views on descriptions of collocations, and are a means for flexible examination of the data. In the following a selection of example queries is presented.

- **retrieve ci-string**.
  
  Returns the list of collocation instances (types) stored in the database.

- **retrieve ci-string where ci-string ~ "stellen"**.
  
  Returns the list of collocation instances containing the verb *stellen*.

- **retrieve ci-string where ci-string ~ "zur .* stellen"**.
  
  Returns the list of collocation instances where the first word is *zur* and the last word is *stellen*.

- **retrieve cr-string where ci-string "zur? .* stellen"**.

---


\(^5\)http://www.dfki.de//pas/f2w.cgi?ltc/diet-e


Returns a list of pairs containing the collocation instances (ci-string) and the example sentences (cr-sent) for which the value of ci-string matches strings which start with the word "zu" or "zur" and end with the word "stellen".

retrieve cr-sent source-name where ci-string = "zur Verfügung stellen".

Returns a list of pairs containing example sentences (cr-sent) for the collocation instance "zur Verfügung stellen" (ci-string = "zur Verfügung stellen") and the name of the corpus (source-name) within which the particular example has been found.

retrieve cr-id cr-position-begin cr-position-end cr-substring where ci-id = <number>.

Returns for each corpus example which is related to the collocation instance with identification number <number> a list of quadruples containing: the identification number of the example (cr-id), the begin and end position (cr-position-begin, cr-position-end) of the substring in the example sentence, and the example string itself (cr-substring). Five quadruples are returned for each example sentence providing information on the prepositional, the nominal and the verbal collocate, as well as the string beginning with the preposition and ending with the noun, and the string which spans from the PP-collocate to the verbal collocate.

retrieve cr-id cr-substring where cr-function = "PP_col" & ci-id = <number>.

Returns a list of pairs containing the identity number (cr-id) of the corpus example and the particular substring (cr-substring) for which the following conditions hold: the substring must be the PP-collocate (PP-col) and the sentence from which the substring is taken must be an example for the collocation instance with identification number ci-id = <number>.

retrieve cr-id cr-substring where cr-function = "V-PP" & ci-id = <number>.

The present query is similar to the previous except that now the substring (cr-substring) spanning from the verb- to the PP-collocate (cr-function = "V-PP") is retrieved. Only those examples are retrieved where the verbal collocate precedes the PP-collocate. Retrieval of substrings where the PP precedes the verb requires the following condition: cr-function = "PP-V".

Specifications of queries which involve attributes of CI-ANALYSIS are left to the interested reader.

7. Additional Functionality

7.1. Consistency Checks

Values of attributes in tsdb(1) can be of tree basic types: string, integer or key. For each attribute in a base relation one of the three value types is defined, and consistency check are made when loading tsdb(1). But more specific appropriateness checks on attribute-value pairs are outside the scope of the core machinery, and need to be handled by external programs. Such a program has been implemented for checking the values of the attributes ci-attrib and ci-value in CI-ANALYSIS. The program operates directly on the file storing the data sets of CI-ANALYSIS.

7.2. Further Processing of the Query Results

Another task for external programs is evaluation and further analysis of selected query results. The additional functionality is provided by programs which apply to files to which the query results are written by tsdb(1).

At present, the following functionality is available:

- Generalizations over collocation instances

  - Construction of new values for ci-string: results from queries to CR-STRUCTURE are employed to specify generalizations over the values of ci-string;

  - Generalizations over verbs, i.e., creating a list of nominal and/or prepositional collocates for a particular verb: the lists of potential partner collocates are built on query results to ci-string;

  - Generalizations over nouns, i.e., constructing lists of verbs, prepositions, and verb-preposition combinations co-occurring with a certain nominal collocate: the lists of potential partner collocates are built on query results to ci-string;

- Pretty-printing of corpus examples, two options are available

  1. Part-of-speech and chunk tags are stripped off, and the plain word string is returned.

  2. The corpus data are returned as formatted labeled bracketings.

The programs apply to results from queries to COLLOCATION-REALIZATION and CR-STRUCTURE.

- Occurrence statistics

  - Average distance between preposition and noun in the PP-collocates of a particular collocation;

  - Average distance between PP- and verb-collocate of a particular collocation.

In both cases, the programs operate on results of queries to CR-STRUCTURE. Two strategies are pursued: 1. the distance between collocates is measured by the number of words in-between; 2. statistics are made over the syntactic structure of the PP-collocate and the lexical realizations of occurring determiners and modifiers; the distance between PP- and verb-collocate is measured in terms of intervening phrases.

8. Conclusion

Collocations in the current database are represented mainly on a syntax-oriented basis. The reason for this is that a principled approach is still out of sight, even though the co-occurrence of syntactic generativity and collocation-specific rigidity in collocations is apparent. A step towards an understanding of this kind of interrelation has
been made in the work presented by specifying a representation scheme and implementing a database which account for generative and static aspects of collocations in an integrative way, combining competence-based syntactic description and real-world data in a large scale. This has become feasible, because of the availability of efficient tools for shallow syntactic processing, and because of the existence of respective training corpora, as well as the on-line availability of huge amounts of text.

Syntactic description of collocations is an important first step towards a better understanding of the phenomenon. The representation of semantic information is another crucial even more genuine step towards a theory of collocations. Especially under the assumption that the particularities in syntactic structure of collocations are no more than a reflex of underlying semantics- and pragmatics-driven processes. In this respect, the availability of semantic databases like WordNet⁶ and their application to semantic tagging becomes useful for further enhancement of the representations of collocations.

Moreover, description at pragmatic level is necessary, in order to account for the commonness of a word combination, and to investigate the pragmatic function of a collocation and the stylistic implications of its usage. The current database already contains some information of this kind, such as information on the origin of a particular collocation realization (cf. the attribute cr-source), and the encoding of Aktionsart and causativity at FVG. With respect to the former, more data and an enlargement of the pool of example corpora is necessary. With respect to the latter, strategies for automating the assignment of Aktionsart and causativity need to be defined, and methods developed which enable automatic identification of utterances where FVG are used, and cases where verbal counterparts are employed.

All in all, the ground for a more appropriate treatment of collocations is settled, new directions of research take shape, but many of the details are still topic of further research.

9. References


---