Collocation Mining: Exploiting Corpora for Collocation Identification and Representation

Brigitte Krenn
ÖFAI, Vienna, Austria
brigitte@ai.univie.ac.at *

Abstract

The work presented provides computational linguistics methods and tools for collocation identification from arbitrary text, and methods and tools for representing collocations in a relational database integrating competence (collocation-type-specific linguistic analysis) and performance information (corpus sentences). The work differs from existing approaches to collocation identification in systematically utilizing collocation type-specific linguistic information. With respect to collocation representation, the work is the first to systematically and in a large scale combining competence-based descriptions of collocations with actual occurrences in text.

1 Introduction

Lexical collocations are an important phenomenon in natural language which is still far from being understood. In the work presented, approaches to the treatment of to two major aspects of lexical collocations are presented: (1) the importance of collocation-type-specific linguistic information in statistics-based collocation identification, and (2) the necessity of corpus examples for the analysis and description of collocations. In the former case, an approach to collocation identification is presented, where statistical techniques and knowledge on distinctive linguistic properties of collocations as well as textual properties of the extraction corpus are taken into account. In the latter case, a representation model and a database for collocations are described, where linguistic descriptions of collocations and data on real-world occurrences are combined.

By collocations we mean word combinations which are lexically determined and which constitute particular syntactic structures such as Funktionsverbgefüge FVG, figurative expressions, idioms, phraseological units and the like. The term collocate is used for the lexically determined elements of a collocation. Collocation phrase refers to the phrase constituted by one or more collocates.

Methods and strategies for linguistically motivated, corpus-based collocation identification are presented in section 2. A relational model and database for storing collocations are described in section 3.

2 Collocation Identification

In the following, methods and procedures are presented for selecting sets of collocation candidates from arbitrary corpora, and for distinguishing collocational from noncollocational word combinations. Finally, identification results for German based on a newspaper and a newsgroup corpus are discussed.

2.1 Initial Candidate Selection

Corpora used for collocation identification are automatically annotated with parts-of-speech and rudimentary syntactic structure. For this task, a statistical part-of-speech tagger and partial parser ([8]) are employed. A set of collocation candidates is then extracted applying collocation-specific syntactic constraints. For identification of German PP-verb collocations such constraints are: preposition and noun must be constituents of a single PP, PP and verb must co-occur in a sentence. The thus specified set of collocation candidates consists of pairs comprising a particular preposition-noun-verb (PNV) triple and the occurrence frequency of the triple in the extraction corpus.

Because of the use of collocation-class-specific syntactic information, our initial set of collocation candidates contains less noise than a candidate set which is chosen on the basis of PoS-information only or even worse just according to a predefined numeric span where collocation partners are selected by means of distance ex-

*This work has been sponsored by the *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)*, Grant No. P12920. Financial support for ÖFAI is provided by the Austrian Federal Ministry of Education, Science and Culture.

pressed by the number of words inbetween the collocates. The latter two approaches are still state-of-the-art even though the advantage of employing more detailed linguistic information for collocation identification is nowadays widely agreed on. For an overview of models for corpus-based collocation identification see [7].

2.2 Identification Models

In the work presented, linguistically motivated considerations also play an important role in the further process of collocation identification. Two identification models will be discussed.

(1) Phrase Entropy: Based on the observation that phrases which are part of a collocation are usually grammatically more rigid than phrases which aren't, a model has been developed which distinguishes collocational PNV-triples from noncollocational ones by means of phrase entropy.

(2) Lexical Keys: This is a completely lexicon-based approach. Selected verbs are employed as lexical keys to identify a particular class of PP-verb collocations.

Both models and model combinations are compared to a selection of commonly used statistical methods for collocation identification such as mutual information [2], Dice-coefficient [9], log-likelihood [4] or relative entropy [3].

2.2.1 Phrase Entropy

Entropy H is a suitable means for modeling the (in)variation of PP-instances related to a particular PN-combination.[1]

$$H = -\sum_{i=1}^{n} p(X = x_i) \log p(X = x_i)$$

Entropy measures the informativity of a probability distribution $p(X)$: the larger the entropy the more information (uncertainty) is contained in the distribution. Distributions with distinct peaks are less informative than flat distributions. Applied to collocation phrases, this means: phrases constituted by collocational word combinations will have a lower entropy value than phrases constituted by noncollocational word combinations. Given a PN-tuple and its related PP-instances[2], the instances with identical surface strings are grouped together. Each PN-combination is associated with k classes of m instances. The occurrence of classes with overproportionally large m leads to a decrease of the entropy value, and thus is an indicator for collocativity.

Maximum likelihood estimates are used, i.e.,

$$PPentropy_{PN_j} =$$

$$-\sum_{i=1}^{k} \frac{f(PPinstance_{i_{PN_j}})}{f(PN_j)} \log \frac{f(PPinstance_{i_{PN_j}})}{f(PN_j)}$$

with $f(PPinstance_{iPN_j})$ the number of occurrences of a particular $PPinstance\ i$ constituted by a certain PN-tuple j, and $f(PN_j)$ the number of PN-tuples j in the extraction corpus.

2.2.2 Lexical Keys (kwic-model)

The approach is well suited for FVG, because typical support verbs are easily specified. We use the following list of verbs: *bleiben* (stay), *bringen* (bring), *erfahren* (experience), *finden* (find), *geben* (give), *gehen* (go), *gelangen* (get), *geraten* (get), *halten* (keep), *kommen* (come), *nehmen* (take), *setzen* (set), *stehen* (stand), *stellen* (put), *treten* (step), *ziehen* (draw), cf. [1]. See also [5], where a method of automatic identification of support-verbs is presented.

2.3 The Data

2.3.1 Candidate Sets

The corpora employed in the current study are an 8 million word portion of the Frankfurter Rundschau Corpus[3], and a 10 million word subset of the FLAG-corpus, a newsgroup corpus which has been collected in the FLAG-project at DFKI.[4] Different candidate samples have been drawn from the corpora and used in the experiments:

P.N.VVPP-Trigrams: The set comprises all sequences where a preposition P, a noun N and a past participle of a main verb VVPP are adjacent in the extraction corpus. The set contains a high proportion of FVG.

P.N.V(Full Form)-Triples: The set consists of PNV-combinations where V is a main verb, and PN and V co-occur within a sentence. The set is much larger than the previous one, and not biased with respect to FVG.

[1] For a mathematical discussion of entropy see [3] or any other standard book on information theory.

[2] i.e. all PPs in the extraction corpus which are constituted by the particular P and N irrespective of their occurrence within collocations or noncollocations;

[3] The Frankfurter Rundschau Corpus is part of the ECI Multilingual CD-ROM I.

[4] `http://www.dfki.de/pas/f2w.cgi?ltp/flag-e`

P.N.V(Base Form)-Triples: All main verbs are reduced to their base forms. Thus occurrence frequencies of morphologically flexible combinations increase, resulting in a larger sample and a grown proportion of high frequency co-occurrences.

2.3.2 Test Sets

In order to test the models for collocation identification, triples which occur less than three times are deleted from the candidate sets. The cut-off threshold has been chosen because: (i) applying statistical models to low frequency data leads to poor results; (ii) a certain number of low frequency data is kept as it makes the task harder for the models. Applying the threshold reduces the candidate sets, for instance, for full form data extracted from the newspaper and the newsgroup corpus by 97 % and 94 %, respectively. The remaining PNV-triples[5] have been manually classified into collocations and noncollocations, and the collocations have been grouped into Funktionsverbgefüge and figurative expressions employing the criteria shown in figure 1. The thus selected set of "true" collocations is used for evaluating the output of the identification models.

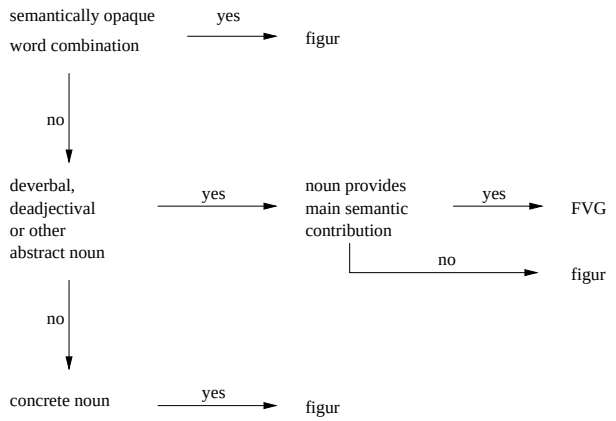

Figure 1: Criteria for manual distinction of figurative expressions and FVG

For testing, the collocation candidates are grouped into three samples A, B and C which differ with respect to the range of occurrence frequencies, i.e., set A contains all tuples which occur 10 times and more, the tuples in set B occur 5 times and more, and set C comprises the tuples that occur at least three times. Thus the difficulty for the statistical models increases from set A to C. For a detailed description of the experiments see [6].

[5] 10 432 for the newspaper corpus, 22 814 for the newsgroup corpus

2.4 Results

The following results have been achieved based on a number of experiments where PP-entropy and the kwic-model as well as their combination are compared with the other statistical models (henceforth association models). χ^2 tests have been applied to judge the significance of the differences between the models.

For sets A and B of full form data, the entropy model leads to significantly better accuracy results in identifying FVG and figurative expressions than the best association model. The difference between the models is insignificant for set C. Recall of FVG from set C of full form data, however, is 3 times higher employing the entropy model compared to the best association model. The results for the base form data are much more heterogeneous.

When the entropy model is combined with the kwic-model, identification of FVG from base form data becomes significantly more accurate compared to simply applying the entropy model. The major advantage of the combined model over the simple entropy model is that the kwic-strategy leads to a strong reduction of the candidate data, resulting in higher identification accuracy.

In general: The kwic-model increases recall of FVG and figurative expressions, but it does not improve identification accuracy of figurative expressions. The kwic-model is an appropriate means for identifying FVG form data with high FVG-density. This is particularly the case for P.N.VVPP-trigrams. Moreover the lexical keys employed better suit FVG than figurative expressions.

The previous results have been based on the newspaper corpus. The tendencies found are also supported by the results achieved employing the newsgroup corpus.

3 Collocation Representation

In this part of the paper, a relational model and database is presented where collocations are represented by means of competence-based linguistic descriptions and sets of example occurrences extracted from different corpora. The major characteristics of the representation model are:
(1) Feature-based description of collocations: Each collocation (type) is associated with a set of attribute-value pairs, representing on the one hand general features such as morphosyntactic and syntactic proper-

ties of the words constituting a collocation, and on the other hand features which are typical for a particular class of collocations.
(2) Collections of real-world occurrences: Sentences containing realizations of a collocation (tokens) are automatically extracted from corpora and described with respect to general and collocation-specific features.

3.1 The Relational Model

Four base relations are defined which are linked via keys. See figure 2.

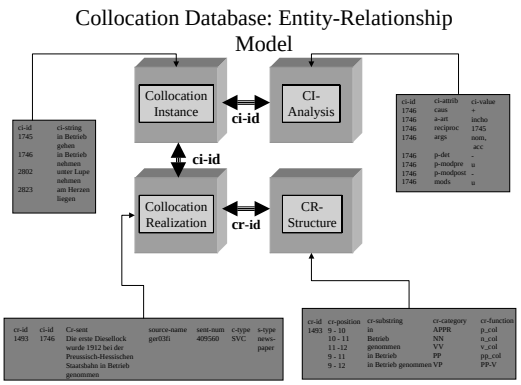

Figure 2: Base relations (cubes) and links (labeled arrows) constituting the database model; there is one data file (rectangle) for each relation

3.1.1 COLLOCATION-INSTANCE CI

Collocation instances are generalized representations of the collocates of a collocation. 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, stehen*}. Nouns are represented as full forms, verbs as infinitives. This kind of encoding reflects the linguistic observations that prepositions in some PNV-collocations 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. The instances are stored in the attribute **ci-string.** Each instance is associated with an identification number (attribute **ci-id**). See table 1 for illustration.[6]

[6]The examples translate as follows: zur? Verfügung stellen ('make available'), zur? Verfügung stehen ('be available'), in Betrieb gehen ('go into operation'), in Betrieb nehmen ('put into operation'), unter Lupe nehmen ('have a close eye on someone or something'), am Herzen liegen ('have at heart')

ci-id	ci-string
2012	zur? Verfügung stellen
2014	zur? Verfügung stehen
1745	in Betrieb gehen
1746	in Betrieb nehmen
2802	unter Lupe nehmen
2823	am Herzen liegen

Table 1: The relation COLLOCATION-INSTANCE and its attributes

3.1.2 CI-ANALYSIS

The relation is designed for representing collocation-class-specific linguistic descriptions. 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 collocation class. Thus we account for the fact that collocation-class-specific properties vary from class to class, and we also account for the expectation that linguistic descriptions of individual collocation classes will change with increasing theoretic insight into the phenomenon. In table 2, CI-ANALYSIS for the FVG *zur? Verfügung stellen* is presented. The collocation is causative, and has inchoative[7] Aktionsart. The FVG 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 FVG is underspecified (u). Postnominal modification in the PP-collocate is blocked (-), i.e., no postnominal modification is allowed. The reciprocal collocation is *zur? Verfügung stehen.*

ci-id	ci-attrib	ci-value
2012	caus	+
2012	a-art	incho
2012	reciproc	2014
2012	args	NPnom (NPdat) NPacc
2012	roles	agent recipient theme
2012	p-det	u
2012	p-modpre	u
2012	p-modpost	-
2012	pass	-
2012	mods	u

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

[7]beginning of process or state

3.1.3 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** (crs) 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; the name of the corpus **source-name**;[8] 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: Funktionsverbgefüge and figurative expressions. 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 (∗).[9]

cr-id	ci-id	crs	s-name	sent-num	c-type	s-type
508	2014	∗	ger03fi	409585	FVG	news-paper

Table 3: The relation COLLOCATION-REALIZATION and its attributes; ∗ 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.*

3.1.4 CR-STRUCTURE

In CR-STRUCTURE, the 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.

Words, part-of-speech and chunk tags are also stored.

The following attributes are specified: **cr-id**, the identification number of the particular collocation example; **cr-substring** (cr-s), 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** (cr-cat), a label representing the syntactic category of the particular substring; **cr-function** (cr-func), a label representing the function of the particular substring, such as: v_col for the verbal collocate; p_col, n_col and pp_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.

cr-id	cpb	cpe	cr-s	cr-cat	cr-func
508	7	7	S1	VV	v_col
508	13	13	S2	APPRART	p_col
508	14	14	S3	NN	n_col
508	13	14	S4	PP	pp_col
508	7	14	S5	VP	V-PP

Table 4: The relation CR-STRUCTURE and its attributes; S1 = stehen, S2 = zur, S3 = Verfügung, S4 = zur Verfügung, S5 = stehen in dem Neubau nur dort zur Verfügung; S1 to S5 have been introduced in the paper to fit the table to the column

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. 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.

3.2 Example Queries

Other than the base relations presented in section 3.1, 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. See the following example queries.

```
retrieve ci-string where ci-string ~
"stellen".
```

Returns the list of collocation instances containing the verb *stellen*.

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

[8] source is abbreviated by s in table 3

[9] A translation of the sentence is *3800 square meters are now available on three floors in the new building where once cows were kept.*

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

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".

Flexible access to all kinds of information represented is possible, and a variety of generalizations over the data can be made. Both tasks are indispensable preconditions for closer investigations of collocation phenomena.

4 Conclusion

It could be confirmed that PP-entropy is a good alternative to the association measures for identifying FVG and figurative expressions from high and medium frequency full form data, but also for identifying FVG from high frequency base form data, and for identifying figurative expressions from medium frequency base form data. The particular strength of the kwic-based approach lies in its ability to improve the identification accuracy for FVG when combined with a statistics-based approach. All in all, the results provide evidence that recall and precision of the individual models for collocation identification strongly depend on the type of collocations to be identified, and the frequency distribution of the word combinations in the sample used for collocation extraction. In the newsgroup corpus, for instance, lexical recurrence is higher than in the newspaper corpus, which can be considered as a reflex of the different styles of the two corpora.

The database currently contains descriptions for approximately 1 000 collocations (467 FVG and 560 figurative expressions). Each collocation is described by a number of corpus examples (sentences containing a word-combination that has been automatically identified as collocational with respect to a particular corpus) and by a linguistic description which contains information on syntactic structure as well as a collocation-type-specific analysis. This way, linguistic analysis and actually occurring data complement each other. As the representations are stored in a relational database, different views on the data can be generated, and thus, together with the extraction component, a tool is available that allows for systematic studying collocations and their usage or function in text. The collocation database can also be viewed as a dictionary-cum-corpus and is thus a valuable resource in syntactic processing, generation and machine translation. Moreover, the example base can be used for training statistical models of collocations.

References

[1] 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.

[2] K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76 – 83, Vancouver, Canada, 1989.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 1991.

[4] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61 – 74, 1993.

[5] Gregory Grefenstette and Simone Teufel. A corpus-based method for automatic identification of support verbs for nominalisations. In *Proceedings of the EACL*, Dublin, Ireland, 1995.

[6] Brigitte Krenn. *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations.* Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 7. German Research Center for Artificial Intelligence and Saarland University, Saarbrücken, Germany, 2000.

[7] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA., 1999.

[8] 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.

[9] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):3 – 38, 1996.