A Collocation Database for German Verbs and Nouns

SABINE SCHULTE IM WALDE
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Azenbergstraße 12, 70174 Stuttgart, Germany
schulte@ims.uni-stuttgart.de

Abstract

The paper presents a database of collocations for German verbs and nouns. The collocations are induced from a statistical grammar model, whose parameters have been trained on 35 million words of German newspaper corpora. Concerning verbs, the database concentrates on subcategorisation properties and verb-noun collocations with regard to their specific subcategorisation relation (i.e. the representation of selectional preferences); concerning nouns, the database contains adjectival and genitive noun phrase modifiers, as well as their verbal subcategorisation. As a special case of noun-noun collocations, we present a list of 23,227 German proper name tuples. All collocation types are combined by a perl script which can be queried by the lexicographic user in order to filter relevant co-occurrence information on a specific lexical item. The database is ready to be used for lexicographic research and exploitation.

1 Introduction

The term *collocation* refers to the habitual co-occurrence of two lexical items within a specific grammatical relationship. The usage of collocations represents a crucial part of the meaning of words, cf. Harris (1968), and therefore constitutes an essential part of lexical dictionary entries. For example, within the lexical entry for the verb *essen* 'to eat', one would expect to find collocational nouns representing the transitive verb's direct object choice for food, such as *Brot* 'bread', *Fleisch* 'meat', *Eis* 'ice-cream', etc. The manual and computational work of lexicographers is supported by lexical resources such as collocational databases, which provide coherent combinations of lexical items.

In some approaches on collocation extraction, the definition of collocations is restricted to the non-compositional and idiosyncratic combination of lexical items. For example, Lin (1999) describes a method for a general automatic identification of non-compositional phrases, and Krenn and Evert (2001) extract German support verb constructions and figurative expressions. In contrast to the above approaches, our notion of collocations refers to their habitual usage.

This paper provides a lexical database of German verb and noun collocations. Concerning verbs, the database concentrates on subcategorisation properties and verb-noun collocations with regard to their subcategorisation relation (i.e. the representation of selectional preferences); concerning nouns, the database contains adjectival and genitive noun phrase modifiers, as well as their verbal subcategorisation. As a special case of noun-noun collocations, we present German proper name tuples.

The collocations are induced from a statistical grammar model, whose parameters have been trained on a German newspaper corpus: the collocation candidates refer to the empirical co-occurrence of two lexical items within a specific grammatical relationship; the collocation strength is based on the probabilistic co-occurrence counts and determined by the lexical association measure log-likelihood (Dunning, 1993). All collocation types are combined by a perl script which can be queried by the lexicographic user in order to filter relevant co-occurrence information on a specific lexical item. The database is ready to be used for lexicographic research and exploitation.

The work is closest to the word sketches for British English in (Kilgarriff and Tugwell, 2001b), the core of the lexicographic workstation WASP (Kilgarriff and Tugwell, 2001a). Related work by Lin (1998b; 1999) describes the automatic extraction of both habitual and non-compositional collocations for English and their usage in various NLP applications, such as the MUC tasks of named entity recognition and coreference resolution (Lin, 1998c), and semantic clustering (Lin, 1998a). Krenn and Evert (2001) and Evert and Krenn (2001) concentrate on the influence of lexical association measures on collocation induction, with reference to the extraction of support verb constructions and figurative expressions. Zinsmeister and Heid (2002) perform an extraction of German noun-verb collocations to compare the collocational preferences of compound nouns with those of the respective base nouns, Zinsmeister and Heid (2003) extract collocation triples of adjective-noun-verb combinations for lexicographic use, and Kermes and Heid (2003) use a chunker for the extraction of German verb-noun and adjective-verb collocations as well as tuples and triples of idiomatic expressions.

The paper is organised as follows. Section 2 describes the induction of collocations, followed by examples from the collocation database in Section 3. Section 4 refers to evaluation possibilities and realisations, and Section 5 describes related work on collocations.

2 Collocation Induction

The collocations are induced from a statistical grammar model, which is based on the framework of head-lexicalised probabilistic context-free grammars (Schulte im Walde *et al.*, 2001). The core of the grammar model is a context-free grammar for German, which incorporates the lexical heads of each rule into the grammar. The statistical parser `LoPar` (Schmid, 2000) performs unsupervised training on the lexicalised grammar, using 35 million words of a large German newspaper corpus from the 1990s. The trained grammar provides frequencies for the lexicalised rules and lexical choice parameters (relations between lexical heads with reference to a grammar rule).

The trained statistical grammar model serves as source for the induction of collocations: The model provides frequencies f for any two lexical items l_1 and l_2 co-occurring within a grammar-specific relationship r: $f(l_1, r, l_2)$. For any pair of lexical items within a specific relationship $\langle l_1, r, l_2 \rangle$, the collocation strength of the pair with respect to their relation is calculated by the lexical association measure *log-likelihood*. Dunning (1993) introduced the likelihood ratio as a useful tool for measuring similarity in text analysis, especially with respect to the behaviour of rare events. Among others, Evert and Krenn (2001) confirm the reliability of the log-likelihood measure in collocation induction, next to lexical associations based on raw frequencies and the *t-score*, and emphasise its usage for low frequency data.

A mathematical re-formulation of Dunning's log-likelihood ratio for the lexical association of the lexical items l_1 and l_2 in the relationship r (cf. `www.collocations.de`) is given in Equation (1):

$$log - likelihood(l_1, r, l_2) = 2 * \sum_{ij} O_{ij} \ * \ log \ \frac{O_{ij}}{E_{ij}} \qquad (1)$$

O_{ij} and E_{ij} refer to entries in the contingency table for the lexical items l_1 and l_2, cf. Table 1: O_{ij} represent the empirical frequencies in the statistical grammar model, as observed for the lexical items l_1 and l_2 within the relationship r. The expected frequencies E_{ij} are calculated as the product of the respective O_{ij} marginals, normalised by the total frequency N of all relationship r tuples, with $N = O_{11} + O_{12} + O_{21} + O_{22}$.

	l_2	$\neg l_2$
l_1	$O_{11} = f(l_1, r, l_2)$ $E_{11} = \frac{(O_{11}+O_{12})*(O_{11}+O_{21})}{N}$	$O_{12} = f(l_1, r, \neg l_2)$ $E_{12} = \frac{(O_{11}+O_{12})*(O_{12}+O_{22})}{N}$
$\neg l_1$	$O_{21} = f(\neg l_1, r, l_2)$ $E_{21} = \frac{(O_{21}+O_{22})*(O_{11}+O_{21})}{N}$	$O_{22} = f(\neg l_1, r, \neg l_2)$ $E_{22} = \frac{(O_{21}+O_{22})*(O_{12}+O_{22})}{N}$

Table 1: Contingency table for co-occurrence counts

3 Collocation Database

The collocation database contains various collocation types for German verbs and nouns, where the types of collocations refer to different relationships with respect to the verbs and nouns. Concerning verbs, the database concentrates on subcategorisation properties and verb-noun collocations with regard to their specific subcategorisation relation (i.e. the representation of selectional preferences); concerning nouns, the database contains adjectival and genitive noun phrase modifiers, as well as their verbal subcategorisation. As a special case of noun-noun collocations, we present German proper name tuples. Following, we present example entries of the collocation database. Each entry is accompanied by the respective log-likelihood value (LLH).

Subcategorisation properties of verbs represent an essential part of our linguistic knowledge, since the verb is central to the meaning and the structure of a sentence. We therefore place emphasis on subcategorisation-specific aspects of collocations: the lexical association between verbs and nouns with regard to their specific subcategorisation relation.

The verb-noun collocations are regarded as a particular strength of the collocational database, since the relationship between verb and noun refers to a fine-grained combination of subcategorisation frame types and the respective frame roles. The German grammar contains 38 subcategorisation frame types. Possible arguments in the frames are nominative (n), dative (d) and accusative (a) noun phrases, reflexive pronouns (r), prepositional phrases (p), expletive *es* (x), non-finite clauses (i), finite clauses (s-2 for verb second clauses, s-dass for *dass*-clauses, s-ob for *ob*-clauses, s-w for indirect *wh*-questions), and copula constructions (k). In the case of prepositional phrase arguments in the frame, the prepositions in addition refer to case and preposition, such as ‘mit$_{Dat}$’, ‘für$_{Akk}$’. The verb-noun collocations are defined with respect to any nominal argument slot within the frame types:

- Considering each role of a specific verb-frame combination, the collocations represent nominal selectional preferences of the verb. Table 2 illustrates examples for the object nouns of the verbs *kaufen* ‘to buy’ and *reden* with the preposition *über*$_{Akk}$ ‘to talk about’. The verb-noun collocations in the former table contain things one can buy, as expected. The latter table illustrates that the range of things to talk about is diverse, with specific attention towards politics and arts.

- Considering the collocations with respect to a specific noun, they represent properties of the noun. Table 3 illustrates an example for the noun *Buch* ‘book’, accompanied by the verbs which most prominently subcategorise the noun as direct object. The verbs refer to different properties of a book, e.g. to its content which is written and read, to the publication process, and to the item which is borrowed and given back.

In addition to the verb-noun collocations, nouns as the content holder of utterances are described by their collocational choices. The collocations describe the nouns in question by typical adjective and genitive modifiers. Table 4 demonstrates an example of adjectival modifiers for the noun *Nacht* ‘night’. As for the subcategorisation by verbs, the range of adjectives refers to different properties of the noun, such as the time aspect with respect to the last or the coming night, the appearance of the night being dark, hot or cold, quiet or disturbed, and the manner of spending the night, e.g. drinking or sleepless. An example of typical genitive modifiers is given in Table 5 for the noun *Zeichen* ‘symbol’. In this case, most modifiers refer to different kinds of states, e.g. time states, abstract mind states such as hope and confidence, but also to a specific kind of symbol, e.g. *D* as abbreviation for *Deutschland* ‘Germany’. As a special case of noun-noun collocations, we induce a list of 23,227 German proper name tuples; the 20 most prominent combinations in the newspaper corpora are given in Table 6.

Noun		LLH
Grundstück	‘site’	191.945
Haus	‘house’	167.313
Aktie	‘share’	143.489
Zeug	‘stuff’	120.558
Wohnung	‘appartment’	63.638
Karte	‘map’	62.493
Produkt	‘product’	61.167
Gelände	‘site’	55.414
Fleisch	‘meat’	54.858
Katze	‘cat’	52.053
Gemüse	‘vegetables’	51.447
Auto	‘car’	51.024
Buch	‘book’	48.355
Panzer	‘tank’	47.814
Ware	‘goods’	41.086
Sache	‘thing’	39.127
Immobilie	‘real estate’	38.464
Gut	‘manor’	38.021
Milch	‘milk’	36.630
Schuh	‘shoe’	35.729

Noun		LLH
Geld	‘money’	97.070
Inhalt	‘content’	57.786
Problem	‘problem’	54.780
Politik	‘politics’	51.668
Thema	‘topic’	38.516
Ding	‘thing’	38.189
Koalition	‘coalition’	35.140
Freiheit	‘freedom’	33.847
Kunst	‘art’	27.027
Perspektive	‘perspective’	22.387
Umfang	‘extent’	20.269
Möglichkeit	‘possibility’	19.327
Konsequenz	‘consequence’	19.246
Film	‘movie’	18.734
Sekte	‘sect’	18.032
Sex	‘sex’	17.083
Islam	‘Islam’	16.018
Besetzung	‘occupation’	15.418
Detail	‘detail’	14.819
Zölle	‘customs’	14.706

Table 2: Verb-noun collocations for objects of *kaufen* ‘to buy’ and *reden über*$_{Akk}$ ‘to talk about’

Verb		LLH
schreiben	‘to write’	1,172.622
lesen	‘to read’	573.643
veröffentlichen	‘to publish’	274.126
führen	‘to keep account of’	107.207
herausbringen	‘to publish’	88.072
verfassen	‘to write’	77.820
publizieren	‘to publish’	52.625
vorstellen	‘to present’	50.766
kaufen	‘to buy’	48.720
zuklappen	‘to close’	46.816
herausgeben	‘to publish’	35.326
füllen	‘to fill’	33.704
mitbringen	‘to bring’	31.214
verfilmen	‘to film’	28.364
ausleihen	‘to borrow’	27.513
zurückgeben	‘to give back’	27.487
wälzen	‘to read (intensively)’	22.865
übersetzen	‘to translate’	18.813
zurückschicken	‘to send back’	17.991
rezensieren	‘to review’	17.825

Table 3: Noun-verb collocations for verbs subcategorising *Buch* ‘book’ as direct object

Adjective		LLH
schlaflos	‘sleepless’	664.577
ganz	‘whole’	322.272
lang	‘long’	194.687
durchzecht	‘to spend the night drinking’	115.659
lau	‘tepid’	115.366
dunkel	‘dark’	98.603
still	‘quiet’	96.963
heilig	‘holy’	88.313
ruhig	‘quiet’	76.759
durchwachen	‘to stay awake all night’	72.451
letzt	‘last’	69.724
durchzechen	‘to spend the night drinking’	68.247
heiSS	‘hot’	66.826
darauffolgen	‘following’	59.217
rauschen	‘great’ (idiomatic)	57.369
unruhig	‘disturbed’	55.044
vorletzt	‘last but one’	44.006
neu	‘new’	43.896
vergehen	‘last’	40.477
kalt	‘cold’	38.652

Table 4: Adjectival modifiers to noun *Nacht* ‘night’

Noun$_{Gen}$		LLH
Zeit	‘time’	166.272
Trauer	‘mourning’	111.050
Solidarität	‘solidarity’	110.368
Schwäche	‘weakness’	107.896
Hoffnung	‘hope’	101.726
Dank	‘thanks’	54.810
Protest	‘protest’	53.644
Verfall	‘decline’	39.737
Stern	‘star’	37.870
Ermutigung	‘encouragement’	37.621
Wille	‘will’	35.720
Jubiläum	‘anniversary’	33.582
Bereitschaft	‘willingness’	27.524
Versöhnung	‘conciliation’	27.289
Zuversicht	‘confidence’	27.029
D	‘D(eutschland)’	26.329
Resignation	‘resignation’	24.010
Unzufriedenheit	‘unhappiness’	24.010
Wachstum	‘increase’	22.740
Freundschaft	‘friendship’	22.676
Wende	‘change’	22.318
Ernsthaftigkeit	‘seriousness’	21.163
Migration	‘migration’	19.631
Würde	‘dignity’	19.038

Table 5: Genitive modifiers to noun *Zeichen* ‘symbol’

Proper Name	LLH
New York	8,955.388
Helmut Kohl	6,586.359
Saddam Hussein	5,611.021
George Bush	3,976.309
Bill Clinton	3,961.956
Bad Homburg	3,568.071
Theo Waigel	3,145.698
Boris Jelzin	2,860.349
Oskar Lafontaine	2,825.231
Steffi Graf	2,778.741

Proper Name	LLH
Willy Brandt	2,694.888
Bad Vilbel	2,444.396
Rose Hausen	2,315.475
Gregor Gysi	2,256.899
Erich Honecker	2,243.533
Nelson Mandela	2,175.772
Rita Süssmuth	2,151.375
Tel Aviv	2,093.286
Björn Engholm	1,908.901
Joschka Fischer	1,887.982

Table 6: (German) Proper name tuples

4 Evaluation

The evaluation of automatically produced semantic information is a difficult task. Introspection (especially by the lexicographer producing the lexical information) is unreliable, since it cannot prove the value of the data in an objective way. An evaluation grounded on the usage of the data, cf. Kilgarriff and Tugwell (2001b), is a proof of the usefulness of the data, but cannot judge the data in an objective (numerical) way either. In few cases, existing manual resources such as dictionaries and thesauri are available. In most other cases, the only objective way to judge about the semantic usefulness of the data is to integrate the information into NLP applications and hope for an improvement. For example, in some languages the framework of SENSEVAL provides an opportunity to utilise and evaluate semantic information for improving word sense disambiguation.

Concerning this work, the collocational data is evaluated in parts. The subcategorisation frame descriptions underlying any verb-noun collocations are formally evaluated by comparing the automatically generated verb frames of over 3,000 verbs against manual definitions in the German dictionary *Duden – Das Stilwörterbuch* (Dudenredaktion, 2001). The F-score is 65.30% with and 72.05% without prepositional phrase information: the automatically generated data is both easy to produce in large quantities and reliable enough to serve as proxy for human judgement (Schulte im Walde, 2002). However, the evaluation does only refer to the structural verb frame types; so far, no semantic information has been compared to dictionary entries.

The proper names are evaluated against their appearance in the training corpus: 200 proper names are randomly chosen from the list of 23,227 German proper name tuples. The proper names are looked up in the training corpus: in case they are correctly induced from the corpus data, they are judged correct, otherwise they are false positives. The overall precision of the proper name database is 65.33%.

For the main part of the semantic collocation data we do not provide an evaluation yet, and SENSEVAL does not include German and therefore drops out of the evaluation possibilities. But the data are ready to be used in lexicographic research and exploitation, in order to prove them useful by utilisation.

5 Related Work

This work was inspired by and is therefore closest to the *word sketches* for British English as described in (Kilgarriff and Tugwell, 2001b). Kilgarriff and Tugwell define a collocation database on basis of 26 grammatical relations between two lexical items, as found in the British National Corpus. The strength of their collocations is estimated by a salience measure combining mutual information and the logarithm of the co-occurrence count. In addition to presenting the collocations and a measure of strength, the co-occurrences are linked to corpus positions, to facilitate the recovery of the related word pair. The word sketches have been used for years and proven valuable by lexicographers in a dictionary project. Compared to (Kilgarriff and Tugwell, 2001b), the German collocation database is less extensive with respect to the number of different relationships, and the linking to corpus positions is not implemented. In contrast, the German grammar specialises in the subcategorisation behaviour of the verbs, which results in a fine-grained lexical collocation resource of verb frames and selectional preferences.

Lin (1998b; 1999) uses a dependency parser to extract collocations from corpora. In (Lin, 1998b), he concentrates on the extraction of habitual collocations, in (Lin, 1999) on the extraction of non-compositional collocations. In both cases, the same methodology is applied: the strength of the collocations is determined by mutual information. Lin (1998b) evaluates the collocation tuple extraction by comparing all extracted collocations to those in a treebank for a different corpus, but he does not evaluate the semantic content of the collocations. Lin (1999) compares the non-compositional collocations to an English Idioms Dictionary, which results in precision and recall values of approx. 15%. He justifies the low evaluation results by showing that also manual dictionaries evaluated against each other show remarkably low PR-results. In (Lin, 1998a), he compares thesaurus entries based on the similarity of word collocations with entries in the manually constructed thesauri WordNet and Roget and shows a significantly closer similarity to WordNet than Roget. (Lin, 1998c) successfully applies the collocation information to concrete NLP tasks, the named entity recognition and coreference resolution in MUC-7.

Evert and Krenn (Krenn and Evert, 2001; Evert and Krenn, 2001) study the extraction of collocations from corpora from a specific point of view. They extract collocation candidates for adjective pairs, support verb constructions and figurative expressions and compare the application of different measures of lexical association in order to filter non-compositional collocations. For the evaluation, they provide an extensive set of the collocation types, manually annotated with the collocation judgement.

Zinsmeister and Heid (2002) perform an extraction of noun-verb collocations by full parsing, whose results represent the basis for comparing the collocational preferences of compound nouns with those of the respective base nouns. The insights are used to improve the lexicon of the statistical parser. Zinsmeister and Heid (2003) present an approach for German collocations with collocation triples: Five different formation types of adjectives, nouns and verbs are extracted from the most probable parses of German newspaper sentences, using the same statistical grammar model as underlying this work. The collocation candidates are determined automatically and then manually filtered for lexicographic use. Kermes and Heid (2003) utilise a recursive chunker to annotate German corpus data with complex phrase structures. The chunks specify lemma information, morpho-syntactic features and coarse semantic properties. Manually defined search routines extract verb-noun and adjective-verb collocations as well as tuples and triples of idiomatic expressions.

The illustration of related work on collocations shows that our approach of German lexical collocations is not the first one, but differently to previous approaches our database contains more variable collocation types and pays specific attention towards the variety of verb subcategorisation aspects. The database is in general more restricted than the English pendants, but more detailed with respect to a fine-grained lexical resource of verb frames and selectional preferences. Most approaches on collocation extraction suffer from the difficulty of evaluating the collocation information.

6 Summary

This paper presented a database of collocations for German verbs and nouns. Specific attention is paid towards the variety of verbal subcategorisation aspects, ranging from selectional preferences of verbs with respect to a particular subcategorisation environment, to nominal properties as given by their diverse modifiers. As a special case of noun-noun collocations, we presented a list of 23,227 German proper name tuples with 65.33% precision.

All collocation types are combined by a perl script which can be queried by the lexicographic user in order to filter relevant co-occurrence information on a specific lexical item. The database is ready to be used for lexicographic research and exploitation. So far, an evaluation is provided for the underlying structural verb-frame definitions and the proper name database.

References

Dudenredaktion, editor. *DUDEN – Das Stilwörterbuch*. Number 2 in ‘Duden in zwölf Bänden’. Dudenverlag, Mannheim, 8th edition, 2001.

Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

Stevan Evert and Brigitte Krenn. Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Metting of the Association for Computational Linguistics*, Toulouse, France, 2001.

Zellig Harris. Distributional Structure. In Jerold J. Katz, editor, *The Philosophy of Linguistics*, Oxford Readings in Philosophy, pages 26–47. Oxford University Press, 1968.

Hannah Kermes and Ulrich Heid. Using Cunked Corpora for the Acquisition of Collocations and Idiomatic Expressions. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research*, Budapest, Hungary, 2003. This volume.

Adam Kilgarriff and David Tugwell. WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation. In *Proceedings of the MT Summit VII*, Santiago de Compostela, Spain, 2001a.

Adam Kilgarriff and David Tugwell. WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Proceedings of the ACL Workshop on Collocations*, Toulouse, France, 2001b.

Brigitte Krenn and Stefan Evert. Can we do better than Frequency? A Case Study on Extracting PP-Verb Collocations. In *Proceedings of the ACL Workshop on Collocations*, Toulouse, France, 2001.

Dekang Lin. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th International Conference on Computational Linguistics*, Montreal, Canada, 1998a.

Dekang Lin. Extracting Collocations from Text Corpora. In *Proceedings of the First Workshop on Computational Terminology*, Montreal, Canada, 1998b.

Dekang Lin. Using Collocation Statistics in Information Extraction. In *Proceedings of the 7th Message Understanding Conference*, 1998c.

Dekang Lin. Automatic Identification of Non-compositional Phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, MD, 1999.

Helmut Schmid. Lopar: Design and Implementation. Arbeitspapiere des Sonderforschungsbereichs 340 *Linguistic Theory and the Foundations of Computational Linguistics* 149, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2000.

Sabine Schulte im Walde. Evaluating Verb Subcategorisation Frames learned by a German Statistical Grammar against Manual Definitions in the *Duden* Dictionary. In *Proceedings of the 10th EURALEX International Congress*, pages 187–197, Copenhagen, Denmark, 2002.

Sabine Schulte im Walde, Helmut Schmid, Mats Rooth, Stefan Riezler, and Detlef Prescher. Statistical Grammar Models and Lexicon Acquisition. In Christian Rohrer, Antje Rossdeutscher, and Hans Kamp, editors, *Linguistic Form and its Computation*. CSLI Publications, Stanford, CA, 2001.

Heike Zinsmeister and Ulrich Heid. Collocations of Complex Words: Implications for the Acquisition with a Stochastic Grammar. In *International Workshop on 'Computational Approaches to Collocations'*, Vienna, Austria, 2002.

Heike Zinsmeister and Ulrich Heid. Significant Triples: Adjective+Noun+Verb Combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research*, Budapest, Hungary, 2003. This volume.