Collocations of Complex Words: Implications for the Acquisition with a Stochastic Grammar

Heike Zinsmeister and Ulrich Heid

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Stuttgart, Germany
{zinsmeis,uli}@ims.uni-stuttgart.de

1 Introduction

This paper presents an investigation into the automatic extraction of noun+verb-collocations from German text corpora, by means of full parsing with a lexicalized statistical grammar. We argue that an approach based on full parsing has advantages over a partial analysis (see section 2.1).
The focus of this paper is not on the linguistic evaluation of the extracted collocations as such but on the collocational behaviour of compound nouns, as compared to the collocational preferences of their respective base nouns. In abstract, we expect the following results:

1. Compounds show the same collocational preferences as their bases[1] (inherited from these, cf. *Pause einlegen* 'take a rest' and *Atempause einlegen* 'take a rest', lit. 'take a breathing space');

2. Compounds have their own collocational preferences, not found with their bases (*Wahlkampf betreiben* 'pursue an election campaign' vs. $^{??}$*Kampf + betreiben* 'pursue a fight', *den Führerschein entziehen* 'take away so's driving license' vs. $^{??}$*Schein + entziehen* 'take away so's certificate'.

A special case of the second type are compounds with the same base that group together according to collocational behaviour, but do not share this behaviour with the base (*einen Diaabend/Vortragsabend/Ballettabend besuchen* (lit. 'attend a slide show/talk/ballet evening') vs. $^{??}$*einen Abend besuchen* 'attend an evening').
In linguistic terms, our hypothesis is that there is a correlation between transparent, productive compounding and inheritance of the collocational behaviour from the base noun (case 1, above). Conversely we expect compounds or compound groups (case 2) to have their own collocational preferences when they are "lexicalized", i.e. have a meaning not derivable compositionally from that of the base.
These facts have an impact on the acquisition technique based on a stochastic grammar. Our investigation was motivated by the fact that the lexicalized statistical grammar, which indexes all grammar rules by the lemma of the respective syntactic head, treated noun compounds as instances of the base noun and introduced the

[1] To avoid confusion of terminology, we use the term 'compound base' when we refer to the morphological head of a compound.

lemma of the base as the lemma of the compound. This was done to reduce the number of parameters, i.e. the number of unknown probability values, which have to be estimated in the grammar training.[2] Each occurrence of a compound was counted as an occurrence of the respective base. This is justified for compounds of the first type (see above), i.e. if the compounds inherit the collocational preferences of their bases. In cases of nin-inheritance, figures are blurred for bases and no data for compounds are available.

We investigated possibilities to improve the results by including compound lemmas of "lexicalized" compounds in the lexicon of the statistical grammar. We used a sample of base nouns and pertaining compounds and automatically extracted candidates of "lexicalized" compounds from the newly-trained grammar model. The result was evaluated manually.

In section 2, we will motivate the use of a stochastic grammar for the analysis of the corpora and briefly explain the parsing and data extraction methods used. In section 3, we will describe the experiment, and in section 4, we will finally discuss the results.

2 Collocation Extraction

2.1 Motivation for Full Parsing

We understand the term 'collocation' in a wide sense that refers to co-occurrence frequency and non-substitutability, but not necessarily to non-compositionality, although often the combinations cannot be translated word by word.

The identification of noun+verb-collocations like *seine Stimme abgeben* '*cast one's vote*' (lit. give away one's vote) is a difficult task for automatic extraction. In contrast to many other types of collocations, they do not necessarily occur adjacent to each other. They are not even restricted to a limited range of n adjacent items, which poses problems for classical n-gram approaches and for many flat, chunking-based approaches. We overcome this problem by making use of a full-fledged clausal analysis as preprocessing step to the collocation extraction.

The problem of non-adjacency holds especially for languages like German that allow for a relatively free word order of nominal arguments. There is no fixed order related to grammatical functions, instead word order and constituent order depend on various factors like information structure, animacy, definiteness, etc. The grammatical relation between a noun and a verb can therefore not be read off the linearization. Linguistic knowledge like case morphology and subcategorization information are needed to determine the relation. Not only nominal arguments do not occur in fixed positions, but also verbs take part in dislocations: German particle verbs split in Verb Second contexts. The finite verbal part occurs in second position whereas the particle remains clause finally at the right edge of the verb phrase. Consequently, it does not suffice to identify the finite verb. In addition, the rest of the clause has to be checked for a stranded particle to decide whether the finite verb is independent or just the verbal base of a split particle verb construction.

The stochastic grammar we used in the experiments covers the described phenomena: it recognizes (split) particle verbs and identifies verbal arguments independently from linearization, thus achieving higher recall and higher precision than flat approaches.

2.2 Corpus Parsing

We used a statistical grammar that models linguistic knowledge and provides full sentence parses (Schulte im Walde 2001). A manually established context-free grammar with feature constraint annotations functions as backbone. It

[2] This is important since the major problem in training a lexicalized stochastic grammar is the large number of parameters which have to be estimated (cf. Schulte im Walde et al. 2001).

was trained on a newspaper corpus [3] by a statistical parser (LoPar, Schmid 2000): the parameters of the grammar are iteratively estimated. The first steps of training evaluate the grammar rules independently of lexical information. Frequent structures get high scores, less frequent structures low scores. In subsequent training steps each rule is multiplied by all potential lexical heads. These are the lemmas of the syntactic heads in terminal phrases which are then propagated to non-terminal structure. The probability mass of each rule is spread over the lexicalized rule variants. For the grammar rules this means that common structures might be 'unlearned' for specific lexical heads such that lexically determined structural preferences surface in the analysis. Lexicalization allows the grammar to learn lexical co-occurrences. These are head-head relations between mother nodes and their non-head daughter nodes, for example the relation between the verbal head of a clause and the nominal head of its subcategorized object.

2.3 Collocation Candidates

The collocation candidates are extracted from the trained grammar model which includes all the relevant frequency information. This allows for a relatively simple extraction algorithm, since the head-head co-occurrence data can be read off a file in text format and do not have to be collected from parse trees or parse forests. The crucial information is encoded as 'lexical co-occurrence' information: all mother categories are listed together with their non-head daughter categories annotated with the lexical heads of both. In addition, the estimated frequency of the particular lexical co-occurrence constellation is given. For example, the occurrence of *Hahn* ('tap') as accusative object of *abdrehen* ('turn off'): the example line below reads from right to left as follows. The verb *abdrehen* with the subcategorization frame *na* (expecting a nominative and an accusative argument) occurs in *VPA*s (active verb phrases) of the analyzed corpus together with an *NP.Acc* (accusative noun phrase) headed by *Hahn* with the estimated frequency of 7.62.

Lexical co-occurrence data: estimated frequency				
7.62	Hahn	NP.Acc	VPA-na	abdrehen

We extracted pairs of verbs and their internal arguments, i.e. their accusative objects in active clauses, by generalizing over various syntactic realizations: the frequency counts are collected for noun+verb-pairs independently of linearization, voice (i.e. active/passive alternation), and finiteness of the verb. This ensures exploiting the broad coverage of the grammar, thereby increasing recall and precision.

Adding up all occurrences of a given noun independently of the verbal head results in the estimated frequency of the noun in general. The same holds for a given verb, respectively. Counting all pair frequencies by generalizing over the lexical heads gives the total number of noun+verb-pairs that constitute the background on which the collocations are to be identified. From an abstract point of view, the lexical co-occurrence data of the model represents a compact corpus of noun+verb-pairs (syntactically homogeneous[4]: verb+direct object) which can easily be fed into a lexical association measure algorithm. In this sense the statistical grammar serves as an extraction tool in the process of identifying noun+verb-collocations.

We ordered the extracted noun+verb-pairs in two ways. First, with respect to their estimated frequency[5]. Second, with respect to their log-likelihood score to identify more reliably[6] such pairs that have a high relative association, i.e. potential collocations (cf. Dunning 1993).

[3] The lexicalized training was based on a newspaper corpus of approx. 25 million words.

[4] In the terms of (Evert and Krenn 2001).

[5] A related approach, a combination of estimated frequency and modeled probability in an EM-based classification model, has been used by (Prescher 2002) to extract collocation candidates.

[6] Cf. Evert et al. 2000.

The score expresses a degree of confidence with which we can reject the assumption that the co-occurrence of a noun+verb-pair is mere coincidence.

- estimated frequency:[7] $F_{co-occ}(n_i, v_j)$; for n_i functioning as the direct object of v_j.
- log-likelihood score:[8] $-2log\lambda = -2log\frac{likelihood\,L(independence\,assumption)}{likelihood\,L(dependence\,assumption)}$

3 Experiment

3.1 Data

The initial statistical grammar model generalized over nominal compounds by mapping compounds onto the lemma of their base noun. We extracted 200 nouns from the lexicon of the initial grammar that served as base nouns of the largest number of compound types (*c-types*). We count into the c-types only compounds made up of two common nouns, like $Zement_{NN}$*/werk* 'cement works'. In addition, we considered proper name-common noun like $Bonifatius_{NE}$*/werk*, and hyphenated compounds like *Chip-Werk, Mercedes-Werk*. Other types of compounds like $Früh_{ADJ}$*/werk* 'early works' or $Stell_{V}$*/werk* 'signal box' are ignored. If the first part of the compound was complex itself it was not analyzed further, e.g. *Atomkraftwerk* and *Braunkohlekraftwerk* count as two different c-types of the base *Werk*.

For the experiments on collocations of compounds versus bases, we manually determined 85 bases from the list of the 200 nouns: we preferred nominalizations of verbs (like *das Verbot* 'prohibition') and polysemous nouns (*die Leitung* 'wire','pipeline','management') but avoided real homographs (*der/die Leiter* 'executive','ladder') and too many concrete nouns. The 85 bases are related to a total number of 7,518 compound types.

3.2 Preprocessing

The first step of preprocessing includes a morphological analyzer (Schiller 1994). Inflectional variants are mapped onto their lemmas (e.g. *Häuser – Haus*, *abgab – abgeben*) and compounds are decomposed (e.g. *Atempause – Atem=Pause*). In German, compounds are written as a single word without blank (and normally also without a hyphen). It is therefore necessary to apply a morphological analysis to automatically relate compounds to their nominal bases. In addition, each token is assigned potential parts-of-speech and morpho-syntactic properties like case.

The lexicon of the parsing grammar is automatically created by morphologically analysing the corpus and mapping the output onto relevant grammar tags. Each entry includes the token and a list of triples consisting of terminal grammar tag, estimated frequency and lemma. In the initial grammar version the lemma of a compound is the lemma of its base. For example, *Bergwerks* is lemmatized as *Werk*.

In the experiment, we mapped the compounds on their compound lemma, for instance the token *Bergwerks* is mapped on the lemma *Bergwerk*. Supported by the morphological analyzer we related the 7,518 compound types to their tokens and replaced the initial (base) lemma tags by the corresponding compound lemma. The re-

[7] More precisely: $F_{co-occ}(h_d, C_d, C, h) += P_{outside}(e)P_{inside}(e')P_{inside}(c)P_{co-occ}(h_d|D_d, C, h)/P(T)$; where C is the mother category and h its lexical head; C_d is the category of the non-head daughter and h_d its head, respectively; c is some constituent of category C; e is the lexical co-occurrence event; and $P(T)$ is the overall probability of the parse tree (cf. Schmid 2000, p.14, his 'lexical choice' frequency F_{choice}).

[8] $-2log\lambda =$
$-2logL(c_{12}, c_1, p) + logL(c_2 - c_{12}, N - c_1, p)$
$-logL(c_{12}, c_1, p1) - logL(c_2 - c_{12}, N - c_1, p_2)$
whereby $L(k, n, x) = x^k(1-x)^{n-k}$ and $p = \frac{c_2}{N}$, $p_1 = \frac{c_{12}}{c_1}$, $p_2 = \frac{c_1 - c_{12}}{N - c_1}$ (cf. Manning and Schütze 1999, p. 172ff.). C_{12} is the co-occurrence frequency $f(n_i, v_j)$. C_1 is the total frequency of a given object $\sum_j f(n_i, v_j)$. C_2 is the total frequency of a given transitive verb $\sum_i f(n_i, v_j)$. N is the total frequencies of pairs $\sum_{ij} f(n_i, v_j)$.

vised lexicon was used forlexicalization and lexicalized training of the statistical grammar.

3.3 Extraction

We extracted pairs of direct object and verb together with the relevant estimated frequencies. The frequency values were used to calculate the log-likelihood ratio, $-2log\lambda$. The absolute value of the score tends to be higher if the noun frequency is higher. This means in general that a base+verb-pair achieves a much higher score than the average compound+verb-pair. The interesting cases are those in which the compound gets a higher log-likelihood value than its base with respect to a given verb, i.e. if it holds that $-2log\lambda(\text{comp}_{basej},\text{verb}_i) > -2log\lambda(\text{base}_j,\text{verb}_i)$ We set the threshold for the log-likelihood score of compound+verb-pairs at 10.83. This is a standard critical value of the χ^2-test. A score higher than 10.83 means that we can reject the independence assumption with 99.999 % confidence. The probability of error is thereby less than 0,001 (Manning and Schütze 1999, p.174). The picture gets blurred with nouns that have an overall low frequency in the preprocessed corpus. If such a noun occurs in a low-frequent pair, the log-likelihood score tends to be relatively high. In many cases it is higher than the threshold of 10.83, which means that there is a lot of noise in the extracted data and precision is not very good. 5,503 out of 18,051 compound+verb-pairs with an estimated frequency of lower or equal to 1.00 get a log-likelihood score of larger than 10.83. Among them there are even combinations proposed due to parsing errors, such as *Sonntagabend+rollen* 'roll sunday evening' which comes with a frequency of 0.88 and a log-likelihood score of 11.85. The pair would be wrongly included in the candidate set of noun+verb-collocates. To reduce the noise, we used the log-likelihood information only for compound+verb-pairs with an estimated frequency larger than 2.5.

To collect relevant compound+verb-pairs with frequencies lower than the threshold, we tested additional heuristics based on a comparison of the estimated frequencies of compounds and their bases. A zero frequency of a base+verb-pair was taken as an indicator for a lexicalized compound if (a) the compound co-occurs significantly often with the respective verb (this case is already covered in the log-likelihood analysis), or (b) if n compounds with a common base co-occur with the same verb which itself does not co-occur with the base, or (c) if a compound co-occurs with n verbs that are all not observed together with the respective base. In addition we checked whether or how often the compound may occur with the same verb as the base does. These heuristics are tested to find lexicalized candidates in spite of a sparse data situation.

4 Results

In the following, we first comment on the general results with respect to the expected types of different collocational behaviour. Then, we discuss the results of the experiments in more detail. The experiments are organized as follows. (i) We extracted compound+verb-pairs that came whith a frequency larger than 5.0 and the verb of which did not co-occur with the respective base. This turned out to be a quite reliable method for identifying lexicalized compounds. Due to the relatively high frequency threshold, this method cannot deal with sparse data. (ii) To reduce the sparse data problem, we lowered the frequency threshold for compound+verb-pairs to 2.5 and allowed co-occurrence of verb and the respective base but added restrictions on the log-likelihood scores. This method improved the recall. (iii) To get hold of lexicalized compounds that occurred only in low-frequency pairs, we tested a number of heuristics by grouping compound types according to a given verb or vice versa. This very rough method helped to handle sparse data.

4.1 Inherited Collocations

The nouns *Fest, Kampf*, and *Kraft* were analyzed in more detail as a sample of the 85 base nouns included in the experiment because many of their compounds occur frequently enough to provide interpretable collocational data.

The most prominent verbal collocates of *Fest* are *feiern* (estimated frequency of 88.24), *eröffnen* (20.45), *planen* (12.79), *veranstalten* (11.75), and *machen* (10.87). Considering all 94 compounds of *Fest* in the corpus and the collocations of these, 38.21% of all observed collocations (tokens) contain the verb *feiern*; *feiern* was observed in collocations of 49 of the 94 (52.13% of the) compound types. The next important collocates with compounds of *Fest* are *eröffnen, planen, veranstalten, machen, organisieren, besuchen*; these verbs account for another 24.87% of the analyzed occurrences.

Another example of the same type is the noun *Kampf*. The collocates shared by many of its compounds are *führen* (13.82% of the total occurrences), *eröffnen* (3.65%), *verlieren, gewinnen, fortsetzen, beenden, liefern, entbrennen, ausfechten, austragen, entscheiden* (together 10.58 % of the occurrences). Interestingly, the second most frequent combination with *Kampf* is *jmdm den Kampf ansagen ('to challenge sb')*. This collocation is lexicalized and a combination with *ansagen* seems impossible with any compound of *Kampf*.

The above examples illustrate cases where the collocational behaviour of compounds is partly inherited from that of the bases. The nouns analyzed are mainly monosemous. A polysemous case is the noun *Kraft*. Its compounds fall into two groups:

(a) `'power, strength, force'`: *Triebkraft, Symbolkraft, Ausdruckskraft, Durchsetzungskraft, ...*

(b) `'employee, personnel'`: *Nachwuchskraft, Honorarkraft, Führungskraft, Halbtageskraft, ...*

Along with the two distinct semantic groups, collocations also group together. With group (a), prominent verbs are *haben, stärken, bündeln, verlieren, verleihen, beweisen, entfalten*, whereas group (b) has *einsetzen, einstellen, freisetzen, suchen, anstellen*. Very few – unspecific and likely not collocationally relevant – verbs show up with compounds of both groups: *brauchen, geben, entwickeln*.

4.2 Lexicalized Cases without Inheritance

From the estimated frequency figures for collocations, separately for bases and for compounds, it is easy to extract those cases where a given compound has a highly frequent collocation with a verb and where this verb does not collocate with the respective base at all. This case is the inverse of *den Kampf ansagen*, the non-inherited idiom observed above. A few prominent examples are listed in the following:

- *Autobahn, Fahrbahn + sperren*, but not **Bahn + sperren*
- *Bußgeld verhängen*, but not **Geld + verhängen*
- *Hilfestellung + leisten*, but not **Stellung + leisten*

The examples all contain lexicalized compounds which are morphologically transparent, but not (cf. *Hilfestellung*) or only partially (cf. *Bußgeld*) semantically transparent. Among the bases concerned are mainly very general ones (e.g. *Art, Werk, Wert, Punkt*) which give rise to semantically opaque compounds (like *Handwerk, Kunstwerk, Feuerwerk, Standpunkt, Sportart* etc.).

4.3 Frequency-Based Extraction

To evaluate the first extraction results, we picked 40 automatically derived candidates for lexicalized compounds and evaluated the noun+verb-pairs in comparing the collocational preferences of these compounds with the collocation preferences of their bases. The candidates were determined by a threshold of the frequency f, $f > 5$, for the lexical co-occurrence event. The compound co-occurs with a verb that does

not co-occur with the base. 29 of the candidates occured mainly in idiomatic collocations. 9 candidates showed a mixed behaviour: they had an overlap with their base but are rather idiomatic. Only 2 out of the 40 candidates mainly inherited the collocation preferences of their base. We take this result as a confirmation of the hypothesis that the analysis of collocational behaviour can be used for identifying candidates of lexicalized compounds. The following table shows the result of the (manual) evaluation.

Forming idiomatic collocations	
Alarmanlage	Anhaltspunkt
Autobahn	Besatzungsmitglied
Bußgeld	Eigentumsverhältnis(-se)
Fahrbahn	Feindbild
Feuerwerk	Gangart
Größenordnung	Handwerk
Hilfestellung	Höhepunkt
Meinungsbildung	Notdienst
Spielabend	Sportangebot
Sportart	Standpunkt
Stellenwert	Streckenführung
Streitwert	Umweltschutz
Urstand	Verkehrsführung
Verwarnungsgeld	Waffenstillstand
Zeitpunkt	

Mixed but rather idiomatic	
Arbeitskampf	Arbeitskraft
Autofahrer	Grenzwert
Kopfgeld	Motorradfahrer
Ozonwert	Sozialhilfe
Wahlkampf	

Inheritance of collocation	
Mißtrauensantrag	Pressekonferenz

4.4 Log-Likelihood Scores

To evaluate the rest of the extraction experiments, we pre-determined lexical compounds from the compound list without taking the noun+verb-collocations into consideration. The test was whether a compound can be replaced by the base without a relevant change in meaning, for example *Verteidigerstellung* ('position of defender') can be replaced by *Stellung* which expresses the same meaning in a less specific way, i.e. it is not a lexicalized compound, whereas *Problemstellung* 'way of looking at a problem' is not a specific type of *Stellung'*, i.e. it is a lexicalized compound. Table 4.4 lists the manually determined lexicalized compounds. The compounds which were extracted by the log-likelihood-based method are given in bold letters. We set the following thresholds:

- $-2log\lambda$(compound,verb) > 10.83
- F_{co-occ}(compound,verb) > 2.5
- $-2log\lambda$(compound,verb$_i$) $> -2log\lambda$(base,verb$_i$)

See also the more detailed results in table 2.

Base	#	lexicalized c-types
Abend	65	*Elternabend,* ***Feierabend,*** ***Lebensabend***
Art	35	***Eigenart,*** ***Gangart,*** *Mundart,* ***Sportart,*** ***Spielart,*** *Tonart*
Fest	94	none
Kampf	64	***Wahlkampf***
Stellung	46	*Fragestellung,* ***Hilfestellung,*** *Problemstellung, Schlüsselstellung, Themenstellung*
Werk	192	***Bauwerk,*** *Bergwerk, Fachwerk,* ***Feuerwerk,*** ***Handwerk,*** ***Kraftwerk,*** *Laufwerk, Mundwerk,* ***Netzwerk,*** ***Triebwerk,*** ***Schuhwerk***

Compounds of *Stellung* tend to be transparent as long as the first part of them are common nouns. Due to ambiguities in the morphological analysis the test items include words like *Feststellung, Klarstellung* or *Zufriedenstellung.* They are in fact not compounds but nominalizations of complex verbs like *feststellen, klarstellen*, and *zufriedenstellen.* We expect them to be opaque in meaning, and also to show individual collocation preferences. They are indeed extracted as lexicalized items by our tests.
Compounds with the base *Art* deviate from the right-hand head rule in that the semantic base of the compound is its first part, for instance *Baumart* 'treetype' is not a kind of *Art* 'kind'

but a kind of *Baum* 'tree'[9]. We expect therefore that there is no significant match in the collocational behaviour of compounds and base. This is in fact born out. Among the 25 pairs with $-2log\lambda$ (= L(c,v)) > 10.83 there are only 4 wich have a positive count for a related compound. None of the 9 most prominent pairs of *Art*+verb co-occur with a compound of *Art*.
Compounds involving proper names or hyphenated compounds were evaluated independently. The latter often feature a proper name or an abbreviaton as first compound part. Both types tend to be transparent and inherit their collocational preferences from their bases. Table 1 shows some of the extraction results.

4.5 Grouping Heuristics

To improve the recall on lexicalized compounds, we tested additional simple grouping heuristics. We extracted sets of compounds with 5 or more members that co-occur with a verb that has zero-frequency with the corresponding base. This method lead to 77 additional candidates, among them e.g. *Mundwerk* and *Gottesdienst*. Another heuristic was extracting compounds that co-occur with a set of 4 or more verbs that have all zero-frequency with the corresponding base. We found 11 additional candidates this way, including *Badenwerk* (which is a proper name) and *Dauerwelle*.

4.6 Using the Results in Collocation Acquisition

The manually analyzed data support our initial assumption, that semantically transparent compounds tend to inherit the collocational preferences of their bases whereas lexicalized compounds show their own preferences.The initial stochastic grammar model was based on the assumption that the collocational behaviour of compound nouns is inherited from that of base nouns. The technical advantage of this approach was that it helped to reduce the number of parameters to be estimated. The exemplary data analyzed in section 4.1 confirmed that the initial approach was empirically correct for certain types of compounds. For lexicalized compounds, this approach did not lead to useful data. Therefore, lexicalized compounds have to be extracted from the data (e.g. by the methods discussed in this paper) and defined in the lexicon for further applications. Another problematic type involves polysemous base nouns. Further processing (e.g. clustering) of the detailed data on compounds (as we show them in this paper) should lead to a first rough outline of "senses" of the bases.

4.7 Iterative Approach

We applied an iterative approach for including lexicalized compounds into the lexicon. In general, compounds were lemmatized with the lemma of their base. Only a part of the compounds was lemmatized with the respective compound lemma (in our experiment 7,518 compound types, belonging to 85 base types). After lexicalization and training of the new model, collocation mismatches between the compounds and theor corresponding bases were analyzed. The compound lemmas of lexicalized compounds were then included in the lexicon. Non-lexicalized compounds, on the other hand, were again mapped on the lemma of their bases. The procedure will be repeated with further similar samples. This procedure will keep the number of parameters to be estimated in subsequent grammar trainings manageable and will iteratively lead to an improvement of the extraction results.

5 Summary and Outlook

For the extraction of noun+verb-collocations from German corpora, we use a stochastic grammar; it produces mostly syntactically homogeneous material, with higher precision and higher

[9] There is a whole class of nouns that behave the same.

Verb	base	L(b,v)	F(b,v)	comp	L(c,v)	F(c,v)
bezeichnen	Antrag	0.21	6.66	CDU-	16.09	3.00
bauen	Bahn	17.11	6.01	U-	19.56	3.71
benutzen	Bahn	14.56	4.12	U-	18.87	2.99
betreten	Bahn	0.00	0.00	U-	24.42	3.00
fahren	Bahn	41.52	8.85	S-	58.19	7.07
fahren	Bahn	41.52	8.85	U-	42.69	5.80
beschäftigen	Kraft	0.02	0.99	ABM-	30.04	3.67
engagieren	Kraft	0.00	0.00	ABM-	36.41	3.00
ablehnen	Vertrag	21.25	16.40	Maastricht-	30.41	3.98
erwarten	Wert	0.01	5.00	Ozon-	27.68	3.00

Table 1: Sample results for hyphenated compounds

recall than an approach not based on grammatical knowledge of equal detail. To keep the parameter estimation problem in the training step manageable, we ran an experiment on the collocational preferences of compound nouns, as compared to those of their bases. The results suggest that collocate selection is mostly shared between bases and those compounds which are productively built, and which are thus not only morphologically, but also semantically transparent. Lexicalized compounds, however, tend to have their own collocations which do not (or only very partially) overlap with those of their bases. This has implications for the improvement of the stochastic grammar as a collocation extraction tool: it makes sense to construct a lexicon of the most frequent compounds with deviant collocational behaviour and to use the inheritance hypothesis for all other compounds. An iterative methodology is appropriate for this lexicon construction work, which provides at the same time lists of lexicalized compounds which may be of use also for other NLP tasks.

References

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

Evert, Stefan, Ulrich Heid, and Wolfgang Lezius. 2000. Methoden zum Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In Werner Zühlke and Ernst G. Schukat-Talamazzini (eds.), *KONVENS-2000 Sprachkommunikation*, pp. 215–220. VDE-Verlag.

Evert, Stefan, and Brigitte Krenn. 2001. Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th ACL Meeting*, Toulouse, France.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge (MA): MIT Press.

Prescher, Detlef. 2002. *EM-basierte maschinelle Lernverfahren für natürliche Sprachen*. PhD thesis, University of Stuttgart. AIMS Vol.8, No.2.

Schiller, Anne. 1994. DMOR - User's Guide. Technical report, IMS, University of Stuttgart.

Schmid, Helmut. 2000. Lopar: Design and Implementation. Arbeitspapiere des Sonderforschungsbereichs 340 149, IMS, University of Stuttgart.

Schulte im Walde, Sabine. 2001. The Statistical Grammar Framework. Ms., University of Stuttgart.

Schulte im Walde, Sabine, Helmut Schmid, Mats Rooth, Stefan Riezler, and Detlef Prescher. 2001. Statistical Grammar Models and Lexicon Acquisition. In Christian Rohrer, Antje Rossdeutscher, and Hans Kamp (eds.), *Linguistic Form and its Computation*, pp. 387–440. Stanford, CA: CSLI Publications.

Verb	base	L(b,v)	F(b,v)	comp	L(c,v)	F(c,v)
haben	Abend	0.43	15.65	Feier	23.54	6.66
verbringen	Abend	168.73	22.07	Lebens	527.35	39.84
haben	Art	8.33	60.47	Eigen	12.40	5.07
einschlagen	Art	0.00	0.00	Gang	143.74	11.94
ankündigen	Art	0.16	1.00	Gang	65.07	9.13
einlegen	Art	0.00	0.00	Gang	54.18	6.58
trainieren	Art	0.00	0.00	Kampfsport	50.74	2.86
haben	Art	8.33	60.47	Spiel	10.84	2.85
kennenlernen	Art	0.95	1.00	Sport	78.07	7.00
ausüben	Art	1.32	1.09	Sport	27.27	2.91
organisieren	Fest	20.54	6.00	Sommer	32.18	4.64
organisieren	Fest	20.54	6.00	Straßen	32.71	3.90
einläuten	Kampf	0.00	0.00	Wahl	27.76	3.50
dominieren	Kampf	0.00	0.00	Wahl	17.56	2.91
treffen	Stellung	0.00	0.00	Fest	20.07	3.93
leisten	Stellung	0.00	0.00	Hilfe	306.30	32.55
erhoffen	Stellung	0.00	0.00	Hilfe	20.47	2.78
bieten	Stellung	0.00	0.00	Hilfe	25.36	5.41
verlangen	Stellung	0.00	0.00	Klar	27.73	4.48
fordern	Stellung	0.00	0.00	Klar	14.63	3.80
zerstören	Stellung	0.00	0.00	Luftabwehr	39.55	3.00
ausbauen	Stellung	18.92	6.53	Markt	39.30	4.00
verlieren	Stellung	17.68	12.75	Monopol	18.15	3.00
verbessern	Stellung	1.52	2.01	Rechts	34.07	3.00
einnehmen	Stellung	52.26	12.20	Spitzen	96.02	8.52
vornehmen	Stellung	0.00	0.00	Weichen	47.68	4.99
erwarten	Stellung	0.00	0.00	Weichen	14.86	2.96
abschalten	Werk	1.88	0.98	Atomkraft	291.55	23.13
bauen	Werk	64.22	22.22	Atomkraft	74.23	12.37
instandsetzen	Werk	0.00	0.00	Bau	41.84	2.98
zerstören	Werk	0.05	1.00	Bau	17.06	2.99
abbrennen	Werk	0.00	0.00	Feuer	166.53	10.57
entfachen	Werk	0.00	0.00	Feuer	64.64	4.87
veranstalten	Werk	0.00	0.00	Feuer	23.42	3.00
legen	Werk	0.25	1.07	Hand	502.89	57.32
lernen	Werk	0.00	0.00	Hand	135.75	18.40
erlernen	Werk	0.00	0.00	Hand	109.22	10.00
beherrschen	Werk	0.00	0.00	Hand	80.84	10.94
verstehen	Werk	0.08	1.66	Hand	96.54	16.99
erschweren	Werk	0.00	0.00	Hand	17.31	3.00
betreiben	Werk	1.63	3.02	Kernkraft	19.68	3.00
liefern	Werk	2.17	3.69	Kernkraft	16.12	2.68
abschalten	Werk	1.88	0.98	Kraft	74.44	7.00
stilllegen	Werk	22.25	5.00	Kraft	58.82	6.00
betreiben	Werk	1.63	3.02	Kraft	36.89	6.00
besetzen	Werk	12.88	6.77	Kraft	14.98	2.99
schaffen	Werk	0.02	4.12	Kunst	29.08	7.48
zerstören	Werk	0.05	1.00	Kunst	13.47	2.99
zerstören	Werk	0.05	1.00	Lebens	31.16	4.00
aufbauen	Werk	4.77	3.96	Netz	77.23	8.99
tragen	Werk	4.63	0.51	Schuh	27.31	3.91
verlieren	Werk	2.20	0.94	Trieb	23.22	3.96
billigen	Werk	0.49	1.00	Vertrags	31.27	3.21

Table 2: Sample extraction results