

Collocations of Complex Nouns: Evidence for Lexicalisation

Heike Zinsmeister, Ulrich Heid

Institute for Natural Language Processing

University of Stuttgart

Azenbergstr. 12

D-70174 Stuttgart

Germany

{zinsmeis,heid}@ims.uni-stuttgart.de

Abstract

This paper combines a corpus-based study of noun+verb collocations with an attempt to distinguish compositional, regularly formed compounds from lexicalised ones. We claim that morphologically regular, compositional compounds share most of their collocational preferences with their compound heads, whereas lexicalised compounds have their own collocational preferences, distinct or only marginally overlapping with those of their heads. We test this claim on corpus data for noun+verb collocations of 85 German simple nouns and their compounds (1,058 types). The extraction relies on stochastic parsing, since this procedure provides syntactically homogeneous data and frequency counts, which are input to the log-likelihood association measure, to compute significance scores for the combinations. The experiments seem to confirm our hypothesis: a collocational analysis of this kind can serve to identify lexicalised compounds. Being able to identify lexicalised compounds is relevant for lexicography in different respects: in the definition of the nomenclature of a dictionary, because lexicalised compounds should be included with priority, and in work on the micro structure, because collocations of a compound which are shared with the compound head may be treated extensively under the head, with only a reference or a pointer under the compound.

1 Introduction

This paper combines a corpus-based study of noun+verb collocations with an attempt to distinguish compositional compounds from lexicalised ones. The study focuses on collocations of noun+verb pairs where the noun is the syntactic head of the verbs direct object. We compare the collocational preferences of compound nouns with those of their heads. We take the term collocation to denote binary word combinations where one element cannot be easily substituted; many such combinations co-occur significantly often in text; even though many of these combinations have no word by word translations, we do not include semantic non-compositionality as a defining criterion. This definition is roughly equivalent to those of (Hausmann, 1989; 2004) and (Benson, Benson and Ilson, 1986).

The work described here is prompted by the observation that certain morphologically complex nouns, especially compounds, share most or all of their collocates with the respective compound heads (cf. *Atempause einlegen* ‘take a rest’, lit. ‘take a breathing space’ and *Pause einlegen* ‘take a rest’).

These facts may be more regularly distributed than it may seem at first sight. It could be that collocational preferences are governed by similar non-lexeme-specific preferences of a semantic nature, as productivity in the combinatory potential. (Lüdeling, 2002) discusses

collocational and idiomatic examples, where next to *Zähne putzen* ('brush one's teeth') also *seine Beißerchen putzen* ('brush one's little pearls') is found in texts, as well as collocations with compounds, such as *seine Schneidezähne putzen* ('brush ones incisors'). This phenomenon seems to be related with the productive modifiability of idiomatic expressions, as found in *da stehen dem Schwarzkittel die Borsten zu Berge* (lit. 'that makes the boars bristles stand on end'), where *Borsten* replaces *Haare* (and the wild boar the person).

For the domain of collocations of complex nouns, it would be lexicographically relevant to classify such productive uses, and to find a correlation between the collocational behaviour of compounds and their semantic compositionality. For this correlation, we have a working hypothesis; we suggest that there is a correlation between collocational preferences and the lexicalisation of a compound: (i) non-lexicalised, productively built (fully compositional) compounds would then share a large number of collocations with their compound heads; (ii) whereas lexicalised compounds (which often but not necessarily are non-compositional) would have their own collocations, distinct or only marginally overlapping with those of the compound head.

The collocational behaviour of compounds would then be usable as an indicator of their lexicalisation. A similar approach was pursued by (Lin, 1999) to automatically identify non-compositional expressions. "[T]he metaphorical usage of a non-compositional expression causes it to have a different distributional characteristics than expressions that are similar to its literal meaning".¹

If the above hypotheses turn out to be confirmed by the empirical data, this has several implications for lexicography: (i) dictionaries may mark non-lexicalised noun compounds and possibly indicate that their collocates are likely to be shared with those of the compound head; alternatively, only non-shared collocations could be marked; (ii) when it comes to space saving in a printed dictionary, non-lexicalised compounds may be left out from the nomenclature more easily than lexicalised ones (or their micro-structural indications may be reduced); collocational preference data might be useful, then, for lexicographers to decide about the inclusion into the dictionary.

2 Preparing the Extraction: Creating Homogeneous Material

The identification of noun+verb-collocations is a difficult task for automatic extraction from German corpora. In contrast to many other types of collocations, they do not necessarily occur in adjacent word sequences. They are not even restricted to a window of n adjacent items. The problem of non-adjacency holds especially for languages like German that allow for a relatively free word order of nominal arguments. In German, furthermore, finite particle verbs may occur in discontinuous form, such as in *Eine Pause_{ACC} legte_V er_{NOM} heute nicht ein_{PARTICLE}* 'He did not take a break today'.

We overcome this problem by making use of a clausal analysis as a preprocessing step to the collocation extraction. We used a statistical grammar (Schulte im Walde, 2003) that covers the phenomena described above: it recognises (split) particle verbs and identifies verbal arguments independently from linearisation. A manually established context-free grammar is trained by a statistical parser (Schmid, 2000) on a newspaper corpus of 35 million words. During training the grammar rules are enriched with estimated frequency counts and

with information about their lexical heads. The trained grammar co-occurrence frequencies of lexical heads that are related by grammatical structure (such as object+verb). From an abstract point of view, this lexical co-occurrence data represents a syntactically homogeneous data set of direct object+verb pairs (in the terms of (Evert and Krenn, 2001)).

3 Experiment

Our experiments are organised as follows. For 85 heads that are related to a total number of 7,518 compound types, we (i) extracted compound+verb-pairs that came with a token frequency higher than 5.0 (1,058 types) and the verb of which did not co-occur with the respective compound head. This turned out to be a quite reliable method for identifying lexicalised 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 (2,024 types) and allowed co-occurrence of the verb and the respective head but added restrictions on the log-likelihood scores (Dunning, 1993) of the word pairs.² This method improved the recall.

4 Results and Discussion

We manually inspected data for most of the 85 nouns (and their compounds) in our test sample. In this section, we first describe a few clear cases: in Section 4.1.1, we show nouns which share most of their collocates with their head nouns, whereas in Section 4.1.2, we give examples of compounds which have collocations distinct from those of the respective compound heads. To provide a first evaluation, we picked 40 candidates by frequency and analysed the respective data manually (Section 4.2.1). Finally, we used log-likelihood figures to compare collocation preferences of compounds and of their heads; the results are described in Section 4.2.2 and illustrated in Table 1.

4.1 Qualitative Evaluation

4.1.2 Shared Collocations. As a sample, the nouns *Fest* and *Kraft* are analysed in more detail, because many of their compounds occur frequently enough to provide interpretable collocational data. We mainly analysed the collocational preferences of noun+noun compounds, but the results seem to carry over to verb+noun compounds as well (cf. *Führungskraft_N* vs. *Schreibkraft_N*).

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* (types), such as *Abschiedsfest*, *Brezelfest*, *Fußballfest* found in the corpus, 38.21 % of all observed collocations (tokens) of these compounds 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 analysed occurrences. .

The nouns analysed above are mainly monosemous. A polysemous case is the noun *Kraft*. Its compounds fall into two groups: (a) power, strength, force: *Triebkraft*, *Sym-*

bolkraft, Ausdruckskraft, etc. and (b) employee, personnel: *Nachwuchskraft, Honorarkraft, Führungskraft*, etc. 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*.

4.1.2 Lexicalised Cases: no Sharing. From the estimated frequency figures for collocations, separately for heads 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 head at all. This case is the inverse of *den Kampf ansagen*. A few prominent examples are: *Autobahn, Fahrbahn + sperren*, but not **Bahn + sperren*; *Bußgeld verhängen*, but not **Geld + verhängen*; *Hilfestellung + leisten*, but not **Stellung + leisten*. These examples all contain lexicalised compounds which are not (cf. *Hilfestellung*) or only partially (cf. *Bußgeld*) semantically compositional. Among the heads 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.).

4.2 Quantitative Evaluation

4.2.1 A Frequency-Based Evaluation. To evaluate our procedures against the hypothesis that lexicalised compounds do not share (many of their) collocates with their compound heads, we inspected the collocations found with 40 candidates classified as lexicalised compounds by our tools. 29 of the candidates occurred mainly in non-shared, idiosyncratic collocations such as *Alarmanlage* and *Anhaltspunkt*. Nine candidates such as *Arbeitskampf* and *Arbeitskraft* showed a mixed behaviour: they have an overlap with the collocations of their head but are rather idiomatic. Only two out of the 40 candidates mainly share the collocation preferences of their head. We take this result as a confirmation of the hypothesis that the analysis of collocational behaviour can be used for identifying candidates of lexicalised compounds.

4.2.2 Log-Likelihood Scores To evaluate the rest of the extraction experiments, we manually determined lexical compounds from the compound list without taking the noun+verb-collocations into consideration and compared them with the collocational results. The test used in the manual selection exercise was whether a compound can be replaced by its head without a change in meaning that goes beyond hyperonymy: e.g. *Verteidigerstellung* can be replaced by its hypernym *Stellung*. We conclude from this behaviour that *Verteidigerstellung* is not a lexicalised compound, whereas *Problemstellung* is indeed lexicalised, as it is not a specific type of *Stellung*. More generally, compounds of *Stellung* tend to be non-lexicalised as long as the first part of them are common nouns.^{3,4}

Table 1 shows sample data from our experiments. Each line of the table starts with a verb(al collocate) and has, in its 2nd column, a noun, in its 5th column a compound of that noun. Columns 3 and 4 contain figures for the log-likelihood ratio of the collocation between the simple noun and the verb, and for its absolute frequency, respectively. Similarly, columns 6 and 7 contain log-likelihood and frequency figures for the collocation of the verb and the compound. We have set some compounds in bold face, namely those which we consider as

lexicalised. The data for *Hilfestellung* are, e.g. very clear: as mentioned above, *Stellung* + *leisten* does not occur in our corpus (thus “0.00” in columns 3 and 4), whereas *Hilfestellung leisten* occurs with an estimated frequency of 32.55 (column 7) and has a high log-likelihood value (column 6: 306.30). *Marktstellung* (market position), instead, is considered as non-lexicalised, with respect to *Stellung*: the collocation with *ausbauen* has considerable log-likelihood values with each (18.92 with *Stellung* and 39.30, with *Marktstellung*).

verb v	head h	ll(h,v)	f(h,v)	compound c	ll(c,v)	f(c,v)
haben	Abend	0.43	15.65	Feierabend	23.54	6.66
verbringen	Abend	168.73	22.07	Lebensabend	527.35	39.84
organisieren	Fest	20.54	6.00	Sommerfest	32.18	4.64
einläuten	Kampf	0.00	0.00	Wahlkampf	27.76	3.50
treffen	Stellung	0.00	0.00	Feststellung	20.07	3.93
leisten	Stellung	0.00	0.00	Hilfestellung	306.30	32.55
ausbauen	Stellung	18.92	6.53	Marktstellung	39.30	4.00
verlieren	Stellung	17.68	12.75	Monopolstellung	18.15	3.00
verbessern	Stellung	1.52	2.01	Rechtsstellung	34.07	3.00
einnehmen	Stellung	52.26	12.20	Spitzenstellung	96.02	8.52
abbrennen	Werk	0.00	0.00	Feuerwerk	166.53	10.57
lernen	Werk	0.00	0.00	Handwerk	135.75	18.40
beherrschen	Werk	0.00	0.00	Handwerk	80.84	10.94
verstehen	Werk	0.08	1.66	Handwerk	96.54	16.99
aufbauen	Werk	4.77	3.96	Netzwerk	77.23	8.99

Table 1: Sample extraction results

5 Summary and Outlook

For the extraction of noun+verb-collocations from a German corpus, we use a stochastic grammar; it produces syntactically homogeneous material that is subsequently sorted by means of an association measure. We ran an experiment on the collocational preferences of compound nouns, as compared to those of their compound heads. The results suggest that collocate selection is mostly shared between heads and those compounds which are built according to productive morphological rules, and which are not only morphologically, but also semantically compositional. Lexicalised compounds, however, tend to have their own collocations which do not (or only partially) overlap with those of their heads.

This study is part of work on dictionary updating tools, which compare e.g. the headword list of an existing dictionary with the most frequent lemmas from corpora (see Evert et al. 2004). Many corpus words proposed for inclusion into the dictionary are compounds, and often frequency is not the only criterion to really include them. If an automatic tool such as the one sketched here could provide hints as to the lexicalisation status of the compounds, another criterion would be operationalised: most lexicographers would want lexicalised compounds to be captured in their dictionary, whereas not all non-lexicalised ones are seen as indispensable.

Endnotes

1. Recent work by e.g. (Pearce, 2001) and (Baldwin et al., 2003) has the same objective. All of them employ more sophisticated mathematical models than ours. Our focus, however, is on the lexicographic aspects of the resulting data.
2. The log-likelihood score (Dunning, 1993) is not fully adequate for our data since it calculates discrete frequency counts whereas our data consists of continuous estimated frequencies. As the error is expected to be small we used the standard implementation here (see www.collocations.de).
3. Due to ambiguities in the morphological analysis the test items include words like *Feststellung* or *Klarstellung*. They are in fact not compounds but nominalisations 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 lexicalised items by our tools.
4. Compounds with the head *Art* (kind) deviate from the right-hand head rule in that the semantic head of the compound is its first part, for instance *Baumart* is not a kind of *Art* but a kind of *Baum*. There is a whole class of nouns that behave the same. We expect therefore that there is no significant match in the collocational behaviour of these compounds and their head. This is in fact born out. Among 25 pairs there are only four which have a positive count for a related compound. None of the nine most prominent collocations of *Art+verb* co-occur with any compound of *Art*.

References

- Baldwin T., Bannard, C., Tanaka, T. and Widdows, D.** 2003. An Empirical Model of Multiword Expression Decomposability in F. Bond et al. (eds.), *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.
- Benson, M, Benson, E. and Ilson, R.** 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Amsterdam and Philadelphia: John Benjamins.
- Evert, S. and Krenn, B.** 2001. Methods for the Qualitative Evaluation of Lexical Association Measures in *Proceedings of the 39th ACL Meeting*, Toulouse.
- Evert, S. Heid, U., Säuberlich, B., Debus-Gregor, E. and Scholze-Stubenrecht, W.** 2004 Supporting Corpus-based Dictionary Updating in *Proceedings of Euralex 2004* (this volume).
- Hausmann, F. J.** 1989. Kollokationen in deutschen Wörterbüchern. Ein Beitrag zur Theorie des lexikographischen Beispiels in *Lexikographie und Grammatik*.
- Hausmann, F. J.** 2003. Was sind eigentlich Kollokationen? in *Akten der Jahrestagung 2003*, Mannheim: IDS.
- Lin, D.** 1999. Automatic identification of non-compositional phrases in *Proceedings of the 37th ACL Meeting*.
- Lüdeling, A.** 2002. Special Topic Session: *The productivity of collocations*, Colloc 2002, Vienna.
- Pearce, D.** 2001. Synonymy in Collocation Extraction in *WordNet and Other Lexical Resources: Applications, Extensions and Customizations* (NAACL 2001 Workshop). Pittsburgh: Carnegie Mellon University.
- Schmid, H.** 2001. Lopar: Design and Implementation. *Arbeitspapiere des Sonderforschungsbereichs 340*, 149. Stuttgart: University of Stuttgart.
- Schulte im Walde, S.** 2003. Experiments on the Automatic Induction of German Semantic Verb Classes PhD thesis, Stuttgart: University of Stuttgart, published as *AIMS* 9(2).