A Model for a Multifunctional Dictionary of Collocations

Ulrich Heid
Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
D 70174 Stuttgart
heid@ims.uni-stuttgart.de

Rufus H. Gouws
University of Stellenbosch
Department of Afrikaans and Dutch
Private Bag X1
ZA Matieland 7602
rhg@maties.sun.ac.za

Abstract
We discuss user needs with respect to collocations, and we make proposals for the treatment of collocations in a multifunctional electronic dictionary. Most importantly, we suggest that collocations should be promoted to the status of secondary treatment units, which would allow the user to access collocational information and information about collocations depending on the actual needs. We furthermore list linguistic properties and relations of collocations which should be described in an electronic dictionary. Our model has been inplemented by Spohr (2005) in the framework of Description Logic.

1 Introduction

In many dictionaries, collocations are given as illustrative material, as a part of the indication of examples in the microstructure of an article. For the data collection and the formal modelling that should underlie a truly multifunctional electronic dictionary, we do not think that this treatment is adequate, and we thus propose a more detailed descriptive model. This report is based on work carried out in the framework of the project "Models for multifunctional electronic dictionaries", financed by the *Stellenbosch Institute for Advanced Study*, STIAS.

In section 1, we briefly position our work in the wide range of views on "collocations", and we give a few examples of what we mean by collocations. Section 2 summarizes the user's needs a multifunctional dictionary should satisfy; we give examples, then, of questions about collocations which should be answered by such a dictionary. In section 3, we summarize the types of facts about collocations which we propose for a multifunctional dictionary, subdividing them into properties and relations. Remarkably, data about collocations tend to be highly relational. In section 4, we point to new ways of making accessible the data we are interested in.

2 Basic assumptions

Theoretical and applied linguistics, lexicography and computational linguistics use the term "collocation" in many different ways, and with partly overlapping, partly divergent definitions. Sabine Bartsch's recent overview of these different strands (Bartsch 2004: 27-78) covers qualitative and quantitative definitional elements; her working definition is also usable for the purpose of the present study (Bartsch 2004: 76):

> *"Collocations are lexically and/or pragmatically constrained recurrent cooccurrences of at least two lexical items which are in a direct syntactic relation with each other."*

For the applied purposes of lexicography, we include in addition Hausmann's (2004) criterion of the presence, in the two-word co-occurrence, of an autosemantic base and a synsemantic collocate, i.e. of an element that receives its semantic interpretation only within the word combination (the collocate), whereas the other element (the base) can be interpreted without reference to the collocation. This view of collocations has been discussed and adopted among others in Evert (2005: 15-25), Heid (1998) and in the *Oxford Collocations Dictionary for Students of English*, OCDSE.

Other than the strictly frequency-based empiricist approach (Kjellmer 1994), we insist on the fact that there is a syntactic relationship in collocations. Table 1 gives a list of types for German, with examples and with a subdivision into base and collocate. Note that the types numbered 6 to 8 may be grouped under "verb+complement". Types for other Germanic and Romance languages are very similar.

No.	Type	Base	Collocate	Example
1	N + Adj	N	Adj	*tiefer Schlaf*
2	Adj + Adv	Adj	Adv	*tief rot*
3	V + Adv_{Adv}	V	Adv	*tief schlafen*
4	V + NP_{Advl}	V	NP_{Advl}	*Bauklötze staunen*
5	V + N_{Subj}	N_{Subj}	V	*Frage + sich stellen*
6	V + N_{Dat}	N_{Dat}	V	*Anforderung + genügen*
7	V + N_{Obj}	N_{Obj}	V	*Frage + aufwerfen*
8	V + PP_{Obj}	N_{inPP}	V	*zu + Darstellung + gelangen*
9	V + Adj_{Prd}	Adj_{Prd}	V	*verrückt spielen*
10	N + $\text{N}_{\text{Genitiv}}$	$\text{N}_{\text{Genitiv}}$	N	*Einreichung des Antrags*
11	N_{quant} + N	N	N_{quant}	*ein Schwarm Heringe*

Figure 1. Combination of categories and base- and collocate-features in German collocations

3 Multifunctional dictionaries

3.1 User needs and dictionaries

Lexicographers plan and compile dictionaries in terms of well-identified user needs and of dictionary functions needed to satisfy these needs. In a simplified way, one may distinguish between communication and knowledge-directed functions (cf. Bergenholtz/Tarp 2002): the former prevails when a dictionary assists its user to express himself or herself (production oriented) or to understand the language production of others (reception-oriented), whereas knowledge-directed functions are focused on learning new facts, about language or about the world. The presentation of the results of linguistic description, in dictionaries, is obviously affected by these needs and by the prevailing functions, and so is access to the data.

Hausmann (2004, 1989) has convincingly argued for major differences in the access to collocations depending on productive vs. receptive use: in a production dictionary collocations have to be placed under (and made accessible via) their bases, whereas access via the whole expression or via each of its elements should be made possible in a reception-oriented dictionary.

3.2 Multifunctionality: paper vs. electronic dictionaries

Obviously, users would be ideally served by dictionaries satisfying their exact needs; thus different types of dictionaries would be needed for different functions. This may lead to as many as eight bilingual dictionaries for a given language pair. Often, publishers however try to cover several kinds of needs with one single poly- or multifunctional dictionary, for example a monolingual dictionary intended for reception and (to some extent) for production. However, this practical view of multifunctionality has rightly been criticized as sometimes not really been adapted to the actual user's needs of the typical user and usage situation. In printed dictionaries, it is difficult to ensure a rapid and unimpeded access to microstructural entries included to assist the user in one of a number of functions for which the dictionary makes provision. To achieve this, a micro-architecure with clearly identifiable search-zones that accommodate the relevant entries is a prerequisite.

In the electronic medium, other than in paper dictionaries, a truly multifunctional dictionary (system) seems however to be possible. This is not only due to less severe space constraints, but especially also to a wide range of access possibilities and a different approach to typological features. We can imagine a large enough collection of linguistic data, along with the definition of "filters" (indeed seen like views onto a database) that would allow us to extract from the data collection those elements which are needed to satisfy a given kind of user with very specific needs and reference skills in a very specific user and usage situation. Along with this, we see a need for different kinds of layouts supporting the different ways of access to the data (access structures, in the terms of Hausmann/Wiegand 1989) mentioned above. A multifunctional electronic dictionary would then be based on Gouws' (2005) notion of "Mutterwörterbuch". Such a dictionary includes a combination of different dictionaries, with clear instructions to ensure access to the desired data.

3.3 Multifunctionality: Examples of user needs with respect to collocations

In the following, we list a few possible needs of users with respect to collocations. The list is by no means (intended to be) complete, it is rather meant to show the range and divergence of such needs. A general list of usage types (covering the first two of our examples) can be found in Bahns 1996: 38s.

In a typical reception situation, especially when reading a foreign language, a user may come across a collocation, and he/she may not be able to clearly understand its meaning. This is not infrequent with what Tutin/Grossmann 2003 call "collocations opaques", such as FR *peur bleue* ("terrible fear"). As *peur* ("fear") is known to the user, he/she may look the collocation up s.v. *bleu*, being interested in a paraphrase of meaning, a synomym, or a translation. Their interest does however not go beyond these types of information.

A completely different situation is found in text production. Writing about the tax return he/she still has to "do", the user wants to know whether the right verb is *do, make, file, submit* or *deliver*? Under *tax return*, the dictionary should provide the appropriate verb, its syntactic construction, etc.

In another writing situation, the need for a collocation may even be evident form the start. To produce a German equivalent of the French sentence in (1), knowledge of the equivalence in (2) is useless, because of syntactic incompatibility of the verb *erinnern* and the intended syntactic construction in the translation (3), a passive, where *Geschwindigkeitsbegrenzungen* is a subject:

1. *ces limitations de vitesse sont signalées à l'avance et rappelées aux conducteurs par une signalisation latérale*
2. *rappeler qc à qn. ↔ jemanden an etwas erinnern*
3. *Diese Geschwindigkeitsbegrenzungen werden im Voraus bekanntgegeben und ...*

Under *erinnern*, the dictionary should provide (potentially collocational) synonyms which have a syntactic valency pattern such that the thing someone is reminded of can be a passive subject. This is true of *jemandem etwas in Erinnerung rufen*, a collocation which is passivizable (... *und sie werden den Fahrern durch Streckensignale in Erinnerung gerufen*). In this case, the dictionary user is interested in the syntax of the collocation, its morphosyntactic form (no article), its register and style, etc. Ideally, a bilingual dictionary would offer such information, accessible from *rappeler*.

3.4 Collocations: towards treatment unit status

The last usage scenario above is interesting in different respects. First, because of the non-standard access path (via a synonym relation from a "single word item"). Second, and equally important, because of the fact that a collocation here is in the centre of the search. In most dictionaries, collocations come as illustrative material, in the microstructure of an entry devoted to one of its two components. Here, however, it should be "promoted" to the same headword status as "normal" lemmata: what is needed, is information about the collocation, not only its mention. The collocation becomes a fully-fledged lexicographic treatment unit.

This seems to show that an electronic dictionary should allow the user to decide how much information about a given linguistic object (e.g. a collocation) he/she wants to see, and a collocation that may be presented as an illustrative example for one use, may well be promoted to a lemma-like status in another usage situation. Procedures of non-lemmatic addressing can elevate the collocation to much more than merely an illustrative entry. Thus collocations should be ***second level treatment units***, which can be accessed without going through a macrostructural treatment.

4 Data about collocations: A wider view

With a view to computational modelling, we distinguish two kinds of data about collocations, namely (1) their properties at different levels of linguistic description and (2) the relationships they have with other lexical items.

4.1 Properties of collocations

Some of the linguistic properties of collocations have often been discussed in the literature; we list the main properties in the following.

1. *Category combination and distribution of base and collocate:*
   Both are needed to allow for efficient access to collocation; see the table in figure 1 above for examples.
2. *Grammatical category of the collocation as a whole:*
   Some collocations can be interpreted as having as a whole a grammatical category, which makes them paradigmatically exchangeable with other expressions, e.g. synonyms, be they single word units or collocations or idiomatic expressions. For example, verb+adverb-collocations obviously are exchangeable with simple verbs (*tief* + *schlafen* ↔ *schlafen, dormir à poings fermés*).
3. *Preferences with respect to morphosyntax and distribution:*
   To be able to correctly insert collocations into a sentence context, users must know about morphosyntactic preferences of these collocations, e.g. with respect to number, determination (definite, indefinite, null article), modification of the noun group in noun+verb-collocations, number of the noun in noun+adjective-collocations, etc. As most such preferences are more tendencies than categorical values, ideally a percentage, with respect to a given corpus (or to different corpora) should be given. In English, *high hopes* prefers the plural; in German, the collocation *Veto+einlegen* ("to veto") prefers a possessive determiner: *ein Veto einlegen*. More examples of such cases are given in Evert et al. 2004, Heid/Ritz 2005, Ritz 2005 etc. Distributional examples (preference to appear in participle form, in a relative clause etc.) have been discussed by Siepmann 2005: 433s, Siepmann 2003: 244s.
4. *Syntactic subcategorization:*
   Many noun+verb-collocations, especially support verb constructions, are characterized by the fact that the noun is the predicate (and the verb only serves to insert the predicative noun into the sentence). However, this "inheritance" is not present in all collocations, and thus, a description in terms of syntactic (and ideally semantic) subcategorization is need-

ed. We favour a three-layered representation which distinguishes grammatical categories (NP, AP, VP), their grammatical functions (e.g. subject, direct object, etc.) and their semantic roles. For the latter, we find a frame semantic classification useful, but other ways of denoting and distinguishing the semantic roles (be it I, II, III, as in the ECDS) are sufficient for the purpose.

This allows the user to get a clear picture of the relationships between syntactic complements and semantic roles in the syntactic environment of collocations. If a similar "valency" description is used for synonymous single word items, the relationships between single word and multiword items can be read off the data. Figure 2 gives a simplified overview of a few examples.

Predicate/Collocation	Valency		
propose	SUBJ	OBJ2	to-XCOMP
	NP	NP	INF
	SENDER	ADDRESSEE	TOPIC
make + proposal	SUBJ	*to*-POBJ	to-XCOMP
	NP	PP(to)	INF
	SENDER	ADDRESSEE	TOPIC
get + proposal	SUBJ	*from*-POBJ	to-XCOMP
	NP	PP(from)	INF
	ADDR.	SENDER	TOPIC

Figure 2. Related single and multiword predicates

5. *Semantic annotation*:
   The semantic classification of collocations has been much discussed in the literature (cf. in particular work in the paradigm of Meaning ↔ Text-Theory, such as the ECDs (Mel'čuk et al. 1984)). For production purposes, such a description is vital, even though typical contextualized examples often may already give good hints. An approach like Mel'čuk's Lexical Functions is certainly suited for this purpose. In this paper, we deliberately concentrate on other topics which have been less analyzed.
6. *Pragmatic marks*:
   Not only the collocation as a whole, but also certain aspects of its use (e.g. specific morphosyntactic forms) may be marked with respect to style, register, region or time. We thus foresee a marking on the collocation as a whole (two unmarked words could be combined into a marked collocation, as is the case with Afrikaans *voordrag + lewer* (give+talk), which is formal, even though both elements are unmarked), on its elements, and on its morphosyntactic usage properties.

All of the above can be modelled as simple attributes of the collocations, with an attribute name and a value (possibly a quantified one, in the case of preferences).

4.2 Relations and links between collocations and single word lexical items

If we accept that collocations are treated on the same foot as "single word items", we also need to foresee a set of relations between collocations and single words or multiwords. In fact, the synonymy discussed above between *erinnern* and *in Erinnerung rufen* is an instance of this phenomenon.

More generally, we expect the following relations to be modelled in a multifunctional dictionary:

1. *Lexical semantic relations involving collocations:*
   These relations include synonymy, antonymy, possibly other (e.g. taxonomic) relations. Together with the subcategorization description discussed above, this relational description provides access to the paraphrasing potential of collocations. We assume that there are lexical semantic relations which involve collocations. Obviously, this view presupposes that collocations have the same status as single word lexemes; then, the (quasi)-synonymy of *[to] propose* and *[to] make* a proposal can be postulated. We prefer the classification as quasi-synonymy, as there isn't a full identity between the two expressions.
2. *Morphological relations:*
   Both, collocates and bases may entertain morphological (i.e. word formation) relations with other words. Being able to relate the collocations accordingly adds to the text production and paraphrasing use of the dictionary. The following are a few examples:
   • Word formation relations between collocates:
   Antrag einreichen (submit + proposal), *Einreichung des Antrags*,
   Antragseinreichung, Antragseinreicher, etc.
   • Word formation relations between bases:
   cause einlegen – Rauchpause/Denkpause/Atempause/Mittagspause/... einlegen.
   Note that less frequent morphologically related collocations may not require a full entry in the dictionary; a classified link to the most frequent collocation (e.g. from *Mittagspause einlegen* to *Pause einlegen*) may be sufficient. Relations of the first type may also lead to additional quasi-synonymy relations, as e.g. *Antrag einreichen* and *Einreichung des Antrags* are very closely related.
3. *Combinations of collocations:*
   The analysis of large amounts of text tends to throw up longer sequences of words of unexpectedly high frequency. Some of these are collocational chains (Hausmann 2004) or combinations of collocations which share a base ("collocational clusters", Spohr 2005). If these are particularly frequent, they may usefully serve to illustrate common uses of their component collocations (and they need to be linked with those), thereby providing additional "context". Examples:
   • *scharfe Kritik üben* (collocational cluster: "criticize massively", *scharfe Kritik* + *Kritik üben*; shared base: *Kritik*, cf. Heid 1994:231, and with acquisition techniques, Zinsmeister/Heid 2003);
   • *eifersüchtig über seine Rechte wachen* (*eifersüchtig wachen über* + *über seine Rechte wachen*; *wachen über*: collocate of *Recht* and base of *eifersüchtig*; collocational chain).

4. *Links between collocations and their components:*
   Obviously, collocations should be linked with the entries of their components; ideally, such links should lead to readings of base and collocate, rather than to lemmas; or, if this is impossible, they should link to a specific collocational reading of the collocate.

5 Accessing data in an implementation of the model

We have encoded a small collocation list (ca. 1000 German collocations) in the description logic formalism OWL-DL (Spohr 2005, Spohr/Heid 2006), following the descriptive model discussed in section 3. OWL-DL was chosen because of its usefulness for consistency checking, its inferencing capacity, and its well-understood formal properties (decidability, monotonicity).

We distinguish between a model of descriptions (meta model of data categories and allowed types of linguistic descriptions) and a lexical model which contains lexical entities. Properties and relations can be directly expressed in OWL-DL, and the formal features of the relations (transitivity, symmetry, existence of an inverse) can be defined in a simple way; thus we distinguish between full synonymy (symmetric and transitive) and quasi-synonymy (symmetric, but not transitive). Class hierarchies of properties and of relations can be used in underspecified queries, e.g. when lexical objects (of any type) are searched for that are in some sort of morphological relation (be it derivational or compounding) with a given collocation element. As the DL model of the lexicon is, formally speaking, a graph, queries to the lexicon can be formulated as searches for (partial) graphs. This allows us to formulate queries of any degree of specifity, and combining any partial information about a (set of) collocation(s).

The specific problem explained above, in section 2.2, with French-German translation, can be seen either as a query in a German monolingual dictionary ("give me a synonym of any kind, single word or multiword, of *erinnern*, where the participant denoting the object remembered can be a syntactically realized as passive subject"), or, if two monolingual dictionaries are combined to form a bilingual one (cf. Spohr/Heid 2006), as a query for a German equivalent of *rappeler* with the above mentioned constraints.

Both queries do not (necessarily) start from an element of the collocation, but rather from a potential synonym (*erinnern*) or its equivalent (*rappeler*). Besides the synonymy relation, they involve syntactic constraints (syntax/semantic mapping). A user of a printed collocations dictionary would have to read through the whole article of, say, *Erinnerung*, to find an appropriate collocation. The design we propose for our electronic dictionary allows us to search the respective data directly.

6 Conclusions and future work

We have outlined requirements and data types for a multifunctional collocation dictionary. Our model goes beyond existing collocational dictionaries, both printed (e.g. OCDSE 2002, because of the relational component) and electronic (DAFLES does not link collocational and morphological information, except via the headword; ELDIT has no way to promote collocations to the level of treatment units, and DICE has no morphological and few lexical semantic relations). In addition, searching the possibilities are more limited in the electronic dictionaries mentioned. They provide access by the base, by Lexical Functions (DICE, DAFLES) and by category combinations (DICE). Our proposal is closest to Mel'čuk's ECDs and to their implementation in the DiCo/LAF model (cf. e.g. Steinlin 2004), but it does not, for the moment, keep track of a semantic classification of collocations. Unfortunately, a direct comparison is difficult, as the DiCo data model is not fully published; it emphasizes however as well word-formation relationships around the keywords it contains; it is not clear to us how far DiCo/LAF keeps track of word-formation relations within collocations.

Our proposal is in line with the (rather general) guidelines of EAGLES and Xmellt (Calzolari et al. 2002), which suggest a similar modeling of syntactic subcategorization. The standards are obviously not as detailed with respect to collocational description as our proposal, which is in intended to serve a wide variety of usage situations and user types.

As the tests with a small set of collocations have proven successful, we intend in the future to broaden the collocational fragment covered by using results from semi-automatic data extraction from large corpora (cf. Ritz 2005). In addition, a tighter integration with lexical data for single word items is planned, for example with data on syntactic subcategorization. The data model will be completed with proposals for user-friendly query. In addition, procedures for feeding data from the dictionary into NLP applications will be prepared.

References

A. Dictionaries

Kjellmer, G. (1994), *A dictionary of English collocations*. Based on the Brown corpus, (Oxford: OUP), 1994, (3 vols)

Mel'čuk, I. A., Arbatchewsky-Jumarie, N., Elnitsky, L., Iordanskaja, L., Lessard, A. (1984), *Dictionnaire explicatif et combinatoire du français contemporain. Recherches Lexico-Sémantiques.* (I) Montréal 1984; vols. II, III et IV since.

OCDSE (2002), Crowther, J. et al., *Oxford Collocations Dictionary for students of English*, (Oxford: Oxford University Press) 2002.

DAFLES, Dictionnaire d'Apprentissage du Français Langue Étrangère ou Seconde: http://www.kuleuven.ac.be/ilt/dafles.htm

DICE, Diccionario de Colocaciones del Español: http://www.dicesp.com/

DiCo, Dictionnaire en Linge de Combinatoire du Français: www.olst.umontreal.ca/dicouebe/

ELDIT, Elektronisches Lernerwörterbuch Deutsch-Italienisch: www.eurac.edu/Eldit

B. Other Literature

Bahns, J. (1996), *Kollokationen als lexikographisches Problem. Eine Analyse allgemeiner und spezieller Lernerwörterbücher des Englischen*, (Tübingen: Niemeyer), 1996, *Lexicographica Series Maior 74.*

Bartsch, S. (2004), *Structural and functional properties of collocations in English*, A corpus study of lexical and pragmatic constraints on lexical co-occurence, Tübingen, Narr.

Bergenholtz, H., Tarp, S. (2002), 'Die moderne lexikographische Funktionslehre. Diskussionsbeitrag zu neuen und alten Paradigmen, die Wörterbücher als Gebrauchsgegenstände verstehen', *Lexicographica* 22 (2002), pp. 145-155.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., McLeod, C., Zampolli, A. (2002), 'Towards best practice for multiword expressions in computational lexicons' in *Proceedings of the Linguistic Ressources and Evaluation Conference 2002*, Las Palmas de Gran Canaria, Spain 2002, pp. 1934-1940.

Evert, S., Heid, U., Spranger, K. (2004), 'Identifying Morphosyntactic Preferences in Collocations', in *Proceedings of the Linguistic Resources and Evaluation Conference 2004*, Lisboa, Portugal, 2004, pp. 907-911.

Gouws, R. H. (2005), 'Die zweisprachige Lexikographie Afrikaans-Deutsch – eine metalexikographische Herausforderung', in Perkov et al. (eds.), *Proceedings: Colloquium on bilingual lexicography with German*, Reihe Germanistische Linguistik: Hildesheim, Georg Olms.

Grossmann, F., Tutin, A. (2003) (Eds.), *Les collocations – analyse et traitement*, Amsterdam, De Werelt [= Travaux et Recherches en Linguistique Appliqueé, E1]

Hausmann, F. J. (2004), 'Was sind eigentlich Kollokationen?', in Steyer, K. (Hg.) *Wortverbindungen – mehr oder weniger fest*, Institut für Deutsche Sprache Jahrbuch 2003, 2004, pp. 309-334.

Hausmann, F. J., Wiegand, H. E, (1989), 'Component Parts and Structures of Monolingual Dictionaries: A Survey', in Hausmann et al. 1989-1991 (eds.): *Wörterbücher. Dictionaries. Dictionnaires. An International Encyclopedia of Lexicography*, Berlin, Walter de Gruyter, pp. 328-360.

Ritz, J. (2005), 'Entwicklung eines Systems zur Extraktion von Kollokationen mittels morphosyntaktischer Features', (Stuttgart: IMS), 63 pp., [= Diploma thesis].

Siepmann, D. (2005), 'Collocation, colligation and encoding dictionaries. Part I: Lexicological Aspects', *International Journal of Lexicography*, Vol. 18, No. 4, (Oxford: Oxford University Press) 2005.

Spohr, D. (2005), 'A Description Logic Approach to Modelling Collocations', (Stuttgart: IMS), ms., 90 pp., [= Diploma thesis] 2005

Spohr, D., Heid, U. (2006), 'Modeling Monolingual and Bilingual Collocation Dictionaries in Description Logics'. To appear in *Proceedings of Workshop on Multiword Lexemes and Multilinguality*, EACL, April 2006, Trento, Italy

Steinlin, J., Polguère, A., Kahane, S., El Ghali, A. (2004), 'De l'article lexicographique à la modélisation objet du dictionnaire et des liens lexicaux', in *Proceedings of the Euralex International Congress 2004*, Lorient, France, 2004, pp. 177-186.

Zinsmeister, H., Heid, U. (2003), 'Significant Triples: Adjective+Noun+Verb Combinations', in *Proceedings of Complex 2003*, Budapest, 2003.