Semantic Modeling of Collocations for Lexicographic Purposes

Lothar Lemnitzer and Alexander Geyken

Berlin-Brandenburg Academy of Sciences and Humanities, Germany
{geyken,lemnitzer}@bbaw.de

The study which we will present in this paper aims at investigating and modeling lexical-semantic properties of collocations. Pairs of words that co-occur with statistic salience will be extracted automatically from a large German corpus with the help of the “Wortprofil”, a sketch-engine-like application. From these sets of co-occurring words, collocations in the narrow sense are selected manually. With these data, the following research questions will be addressed a) concerning the collocates: are we able to classify these into lexical-semantic classes and group them accordingly; b) concerning the bases: are we able to find significant numbers of shared collocates for lexical-semantically related bases and thus reach some form of generalization and regular patterns?

In our study we apply the Meaning-Text Theory of Mel’čuk, more precisely, the concept of Lexical Functions (LF). The idea to employ LFs for lexicographic work on collocations is not new. However, the combination of LF with semantic wordnets for the abstraction over individual bases (addressing question b above) is innovative as it has, to the best of our knowledge, not yet been used for modeling a larger subset of collocations in any language. In the study we report here, we have focused on a set of lexical items and their collocations in order to test the appropriateness of Lexical Functions and to model the phenomena and the intersection of collocates of related base words to generalize collocational patterns.

A practical goal of our work will gain a clearer view of how to use lexical-semantic features for the encoding of collocation in semasiological dictionaries such as the Digitales Wörterbuch der Deutschen Sprache (DWDS). A further goal and a contribution to lexicological theory is to better understand the
interdependence between regularity and arbitrariness of lexical choice. While the arbitrariness of lexical choice makes collocations hard to learn for the second-language learner, we assume that there are some (sub-)regularities, at least within groups of semantically related headwords. Applied properly to the task of language learning, such regularities should facilitate the acquisition of this part of the vocabulary.

Keywords: Collocation, Lexical Function, Corpus Linguistics, Lexical Semantics

1. Introduction

The study which we will present in this paper aims at investigating and modeling lexical-semantic properties of collocations. Following Hausmann (1984) and Mel’čuk (1995), we define collocation as a directed binary relation $R$ between a base $X$ and a collocate $Y$. For several headwords, pairs of co-occurring words will be extracted from a large German corpus with the help of a sketch-engine-like application (Wortprofil, cf. Didakowski & Geyken 2012, which is based on the ideas in Kilgarriff et al. 2004). From these sets of co-occurring word pairs, a subset of collocations has been selected manually for further consideration. On the basis of these data, the following research questions will be addressed: a) concerning the collocates: are we able to classify them consistently into lexical-semantic classes; b) concerning the bases: are we able to find significant numbers of shared collocates for lexical-semantically related bases and thus reach some form of generalization and regular patterns? To answer this question, we will draw on GermaNet (Kunze & Lemnitzer 2002), a lexical-semantic network of the German language built on the principles of Princeton WordNet (Fellbaum 1998), and work with semantically related lexical items.

In our study we apply the Meaning-Text Theory of Mel’čuk and colleagues, more precisely, the concept of Lexical Functions (LF, cf. Mel’čuk 1998). The idea to employ LF for lexicographic work on collocations is not new (cf. among others Mel’čuk, I. et al., 1984-1999, Polguère 2000). However, the combination of LF with a semantic network
for the abstraction over individual bases is innovative as it has, to the best of our knowledge, not yet been employed for modeling a larger subset of collocations in any language. In the study we report here, we have focused on some lexical items and their collocations in order to test whether Lexical Functions are an option to model the phenomena appropriately.

The paper proceeds as follows. In section 2 we will briefly mention related approaches. In section 3, we will outline the context of and the resources for our work, in terms of the corpora and language technology tools we have been using. In section 4 the theoretical framework of our work will be introduced. This framework draws mainly on the works of Igor Mel’čuk and his colleagues, i.e. Meaning-Text Theory in general and Lexical Functions (LF) in particular. Subsection 1 of Section 5 is devoted to example headwords from German: a noun and a verb. With these examples we will show the merits, but also the shortcomings of the theoretical approach taken. In subsection 2 we will present a method of comparing the collocational profiles of semantically related bases and the potential of it for the purpose of lexicographic description and lexical modeling. The approach that we present in section 5.2 constitutes the innovative approach of the work. With such a comparative analysis that has been made possible with an “intersect” function of the Wortprofil we are able to build a bridge to other lexicological theory with a stronger focus on the regularity of lexical combinations, such as the theory of the “Generative Lexicon”. We will close our paper with a view on our future plans.

2. Related Work

The use of Lexical Functions to model collocations on a larger scale and as part of lexicological and lexicographical description is not new. Actually for more 30 years Mel’čuk and his colleagues have been describing a large part of the French vocabulary in the four volumes of the *Dictionnaire explicatif et combinatoire du français contemporain* (cf. Mel’čuk, I. et. al., 1984-1999). This work has been taken up for French by Alain Polguère and his *Dico* project (Polguère 2000, 2009) and for other Romance languages, for example by the Spanish *REDES* dictionary (REDES 2004 and Bosque 2011) and an online dictionary of Spanish collocations (cf. Vincze et al. 2011). For
German, there is one larger study which draws on Lexical Functions for the description of collocational relations of a larger part of the vocabulary, i.e. the so-called “Gefühlsworterschatz” (vocabulary of emotions, cf. Mel’čuk & Wanner 1996).

On a more theoretical, lexicological level Wanner et al. 1996 raise the question of how to model semantic features of collocates in order to achieve a grouping of collocates which is theoretically better grounded and more explicit than the “intuitive” groupings that can be found in traditional collocations dictionaries. In the course of their practical work (in the field of the Spanish vocabulary of emotions, cf. Moreno et al. 2013) they employ some simplifications to the set of original Mel’čukian Lexical Functions. Concerning the other part of the collocation, the lexical items which form their bases, Bosque raises the question of how to take semantic relatedness between such items into account. His search for semantically motivated groups of collocational bases is driven by both theoretically and practically concerns. A set of shared collocations can be considered as an additional indicator for the semantic relatedness of two lexical units. Practically, collocates that some related groups of lexical units share could be encoded more economically in a dictionary, for example by drawing on lexical inheritance. However, the author had, at the time of writing, no access to a (Spanish) lexical-semantic network.

Using corpora to extract lexical co-occurrences, i.e. a superset of collocations, has also been common practice since the advent of corpora which are large enough for such a kind of aggregative statistics. The most famous example of that kind of work is the sketch-engine approach by Adam Kilgarriff and colleagues (cf. Kilgarriff and Tugwell 2002 and Kilgarriff et al. 2004) who have taken up work by e.g. Kenneth W. Church and colleagues (cf. Church & Hanks 1991). Concerning the computational lexicographical efforts to model word co-occurrence for German on a large-scale level, two approaches deserve to be mentioned. First, at the Institut für Deutsche Sprache a large database of co-occurrence profiles is located (CCDB, corpora.ids-mannheim.de/ccdb, cf. Belica 2011). Second, the Wortschatz-Projekt at the Universität Leipzig extracts pairs of co-occurring words from a large German corpus collection (wortschatz.uni-leipzig.de). Both projects have in common that they largely employ
automatic processing of corpora with statistical tools. Thus, they capture co-occurrences and not collocations in a narrower sense. CCDB additionally provides groups of collocates that are shared by two base words. However, these groups are not labeled and making sense of these groups is left to the user of these data. Both projects therefore provide helpful input for a lexicographical treatment of collocations, but this treatment, i.e. selection of collocates and their description, is not the scope of either of them.

In our own work, we largely follow the Sketch Engine approach of co-occurrence pattern extraction, with some modifications and adaptations for German (see below, section 3). The necessary adaptations are mainly due to the particularities of German syntax, in particular the variability in word order (for details, cf. Didakowski & Geyken 2013, section 2). As such, the approach of the Wortprofil is close to the “collostructural” approach (cf. Stefanowitsch & Gries 2003). However, while the linking of complex lexical patterns with grammatical pattern is the main concern of the collostructural approach, the grammatical analysis and classification of the Wortprofil does not make similar claims as to the theoretical groundedness.

As a lexicographical project we invest time and effort in the selection proper collocations from the sets of co-occurrence pairs that the Wortprofil provides, and in the syntactic and semantic classification and grouping of these items.

With our approach to classify German collocations by LFs, we largely follow Wanner and colleagues. However, we do not want to restrict ourselves to a particular subject field nor to a particular part of speech. It is essential for our further lexicographical enterprise to draw on theoretical framework(s) with which we are able to cover the full German vocabulary.

With the view on the lexical-semantic relatedness of collocational bases and the detection of shared collocation, we elaborate the ideas of Bosque. We will use a full-fledged lexical-semantic resource (GermaNet, Kunze & Lemnitzer 2002) and a tool which enables us to view the difference as well as the intersection of the word profiles of two (or more) lexical items (see below, section 3).

3. The Context of our Lexicographical Work
The project in which the reported research originates is a long term academy project at the *Berlin-Brandenburgische Akademie der Wissenschaften* (Berlin-Brandenburg Academy of Sciences, BBAW in short). The ultimate goal of the researchers and developers working in the project is to compile a large synchronic dictionary as a component of a digital lexical information system on the internet. This lexicographic work draws on existing lexical resources of the BBAW – so-called legacy dictionaries which have subsequently been digitized and integrated into the information system.

The project is divided into three phases of 6 years and will be running until 2024 (cf. Klein & Geyken 2010 and Geyken 2013). During the first phase, three dictionaries (*Deutsches Wörterbuch* (DWB); *Etymologisches Wörterbuch des Deutschen*, ed. by Wolfgang Pfeifer, and the *Wörterbuch der deutschen Gegenwartssprache*, WDG) have been integrated into the
DWDS lexical system (www.dwds.de, see fig. 1). The latter dictionary, WDG, is the current focus of our lexicographic work.

Related to the lexical resources are large German reference corpora: the DWDS-Kernkorpus of the 20th/21st century, a balanced reference corpus of 110 million tokens (cf. Geyken 2007). In addition, an opportunistic corpus of texts from 10 influential national newspapers has been compiled. It currently consists of 3.5 billion tokens. The corpora are continuously being updated and extended.

The central resource of the DWDS, called DWDS-Wörterbuch and based on the WDG, is currently being updated. The main focus is on the provision and lexical description of lexical items that have come into use after 1977, the date when the last volume of the WDG was published. For this purpose, a list of around 40,000 lemmas has been compiled, based on the aforementioned corpora and on Google books word lists (cf. Geyken & Lemnitzer 2012).

Full articles are currently being written for these headwords. One major concern of this work is to provide rich sets of collocations. It is our firm conviction that collocations are essential for the use of the dictionary by both language learners and translators. That is why they are one major part of the sense descriptions, along with the definition and corpus citations.

Fig. 3 shows the lexicographical work place and authoring environment of the DWDS project, in particular the collocational groups for the headword Bau (‘building’).

For the corpus based extraction of pairs of co-occurring words, a so called word profile generator, similar to the sketch engine, has been developed by the DWDS project team (cf. Geyken et al. 2009, Didakowski & Geyken 2013). This application provides lists of co-occurring words for various syntactic relations (e.g. “nominal head and adjective attribute”, “verbal head and direct object”; “coordination”, for a full list of relations cf. Didakowski & Geyken 2013) and links these word-pairs with corpus citations. The extraction and ordering of the word-pairs are based on statistical computations over a morpho-syntactically and syntactically annotated German sub-corpus containing about 1.8 billion tokens (fig. 2).

From these sets of co-occurrences, which can be considered as collocation candidates, the lexicographers manually select those word pairs which
qualify as collocations in a narrower sense. One criterion is the recurrence and conventionality of use of these pairs (“Usualität”, cf. Steyer 2003 for an in-depth discussion).

Fig 2: Word profile for the base word *Gespräch* (‘conversation’). In particular, the 100 most typical co-occurring words for the relation “nominal-base adjectival modifier” are shown. The font size of the co-occurring word indicates the salience o the word pair. As statistics, the logDice is chosen (cf Rychly et al 2008), other statistics can be selected from a menu. When clicking on one of the co-occurring words, the user gets a list of concordance lines from the corpora in which the word pair appears.

One important feature of the word profile application is the capability of comparing the sets of co-occurring words for two bases and to use the set operations of difference (diff) and intersection (intersect). Diff returns the words that typically co-occur with base A and not or significantly less typically with base B. More important to our work, the intersect function returns the words that are in the lists of both base words. The ranking of these common collocates is based on the respective rank of this word in
either list. Intersection can be done iteratively on an arbitrary long list on base words. We can thus get those collocates which are in the range of \( n \) (semantically related) base words. In section 5.2 below we present examples. The set functions of the word profile are still in an experimental stage at the time of submission of this paper, but should be available by the time of publication through our website www.dwds.de.

Both the selection of collocations from the set of co-occurring words and the grouping of these collocations by syntactic and semantic relations should be informed by lexico-semantic theory. Such a theoretical framework should define the concept of a collocation in a way that it can be used to guide both the selection and the grouping of collocates on semantic principles. We will look at Meaning Text Theory in general and Lexical Functions in particular as such a theoretical framework in the following section.

4. Collocations as Lexical Functions

As has been said above (section 1), we consider a collocation to be a binary, directed relation. The elements of this relation are a base (lexical unit) and a collocate. Some relation, to be further specified, holds between the two.

A more formal definition of collocation given by Mel’čuk (1995, p. 179):

A collocation \( \textbf{AB} \) of \( \textit{L} \) is a semantic phraseme of \( \textit{L} \) such that its signified ‘\( \textit{X}’ \) is constructed out of the signified of the one of its two constituent lexemes — say, of \( \textbf{A} […] \) such that the lexeme \( \textbf{B} \) expresses ‘\( \textit{C}’ \) contingent on \( \textbf{A} \).

A and B are lexical items, \( \textit{L} \) denotes a language (e.g. German), \( \textit{X’} \) and \( \textit{C’} \) are the meanings (of the collocate as a whole or of one of its components).

Such a formal definition can be employed to distinguish the set of collocations (for a particular headword) from the set of freely co-occurring word pairs on one side and from fixed idioms (phrasemes) on the other side.

The decision, however, on whether a particular word pair is a collocation and neither a free phrase nor a phraseme is often hard to draw in practice.
Whether a word-pair is classified as a collocation depends on whether a) the collocate provides a identifiable meaning to the phrase (in contrast to a proper idiom, where all parts lose their independent meaning) and b) this meaning is contingent of the co-occurrence with the base (in contrast to a free combination of the two words).

The (generalized) contribution which some collocates provide to some collocations is conceptualized, in the framework of Meaning Text Theory, as a Lexical Functions. In Mel’čuk (1995, p. 183), Lexical Function is defined as follows:

\[
\text{a Lexical Function } f \text{ is a function that associates with a given lexical unit } L, \text{ which is the argument, or keyword, of } f, \text{ a set } \{L_i\} \text{ of (more or less) synonymous lexical items — the value of } f — \text{ that are selected contingent on } L \text{ to express a specific meaning corresponding to } f. \text{ Thus } f(L) = \{L_i\}. \text{ To put it differently, an LF, particularly a simple standard LF, is a very general and abstract meaning that can be expressed in a large variety of ways depending on the lexical unit to which this meaning applies.}
\]

Mel’čuk provides (e.g. in Mel’čuk 1995, p. 186ff) a list of around 60 standard Lexical Functions. On the highest level, these functions are divided into paradigmatic functions (e.g. synonymy and antonymy) and syntagmatic functions. Only the latter apply to collocations and are therefore relevant for our investigation.

A few examples might suffice to illustrate the concept (examples are German and English). Lexical Functions are presented as follows: The name of the function appears in capital letters. Following the name is the argument of the function, i.e. the word to which the function is applied (in other terminology, the base of the collocation). After an equals-sign and enclosed in single quotes the values of the function are given, i.e. the collocates.

\[
\text{Adjectival → Intensifier (=M_{AGN})}
\]

\[
\text{MAGN(smoker)} = \text{‘heavy’}
\]
MAGN(Raucher) = ‘stark’
MAGN(regret) = ‘big, bitter, deep, profound’
MAGN(Bedauern) = ‘aufrichtig, tief, lebhaft, schmerzlich, echt’
Verbal → sth. Begins (INCEP)

INCEP(fire) = ‘break out, start’
INCEP(Feuer) = ‘ausbrechen’

These so-called standard Lexical Functions fulfill two requirements
a) they are applicable to a large number of headwords and b) they are
universal, i.e. applicable to headwords in any language (yielding language
specific sets of collocates). Thus, Lexical Functions should be an ideal tool
for classifying sets of collocates for many headwords, which is exactly what
we want to do.

Using this small and finite set of simple LFs, one can construct complex
Lexical Functions such that the collocates and the base together form
a meaning which is not further decomposable (Mel’čuk 1995, p. 193).
Examples:

Adjectival → as it should not be (=ANTIVER)

ANTIVER (shame) = ‘false’
ANTIVER (Scham) = ‘falsch’

Verbal → cause to begin (CausIncep)

CausIncep (fire) = ‘set on, start’
CausIncep (Feuer) = ‘entfachen’

Mel’čuk also introduces configurations of Lexical Functions (Mel’čuk
1995, p. 194). Since these are not relevant for the following, we will not go
into further detail.

Another group of Lexical Functions which have been introduced lately
into the theory are non-standard Lexical Functions. They differ from the
standard Lexical Functions in that they are a) applicable to a small and
finite set of arguments (i.e. collocation bases) and b) (supposedly) not universal. Mel’čuk (2012, p. 45) defines them as follows:

A non-standard LF $f_{\text{non-stand}}$ describes a non-standard collocation where the semantic relation between the base and the collocate is not institutionalized in the language; to specify this relation, $f_{\text{non-stand}}$ must be described in the same metalanguage as that used for lexicographic definitions[...]

He gives the following examples ($\sim$ is used as a placeholder for the base word):

WITH NO DAIRY PRODUCTS ADDED(coffee) = ‘black [\~]\’
OHNE MILCH(Kaffee) = ‘schwarz [\~]\’
USED TOO MUCH(example) = ‘hackneyed [\~]\’
ZU OFT GEBRAUCHT(Beispiel) = ‘überstrapaziert [\~]\’

Note that these LFs have a “name” for which the descriptive language of the dictionary is used while the standard LFs have an abstract Latin name.

Alain Polguère devotes a longer text to the theoretical status and the practical use of non-standard Lexical Functions in the context of the Dico project (Polguère 2007). Anne-Laure Jousse tries to “re-formalize” some of the non-standard LFs which have been used in Dico in order to make them applicable for language technology tasks such as machine translation and text generation (cf. Jousse 2007). We will come back to non-standard Lexical Functions in our examples (section 5.1)

5. The Lexicographical Tasks

5.1 Classifying and Grouping the Collocates

As has been stated earlier, the task at hand is to provide sets of collocations for a large number of headwords in a dictionary of contemporary German. Currently, the only way to group the select collocations is by the syntactic relation between the base word and the collocates (e.g. nominal head and
adjectival modifier). This leads to long lists of (alphabetically ordered) collocates (see fig. 3). This leads, in the best case, to “intuitive” groupings in the sense of Wanner.

In the course of our investigations, we have looked at around 30 base words, distributed over the major word classes, all with rich sets of collocations. We will exemplify our major points with two examples: The noun Bau – in the sense of building as an artifact, ignoring the other senses of this lexeme – and the verb adressieren (to address). We will show which (non)-standard Lexical Functions can help to order the lists of collocates.

**Fig. 3:** Screenshot of the DWDS dictionary writing system with parts of the article *Bau* (in the sense of ‘building’ as an artifact)

**Bau – Standard LFs**

VER(Bau)= ‘klassischer (classical), prächtiger (magnificent), repräsentativer (presentable), solider (solid), stolzer (lordly), traditioneller (traditional) [~]’

(VER = as it should be)

POS₂(Bau)= ‘ansehnlicher (good-looking), imposanter (monumental), schmucker (pretty), schöner (beautiful), stattlicher (lordly), trutziger (fortified) [~]’

(POS² = positive valuation)

ANTIVER(Bau)= ‘beschädigter (damaged), leerstehender (empty),
verwitterter (weather-beaten) [~]’

(ANTIVER = as it should not be)

ANTIPOS$_2$(Bau) = ‘dürftiger (paltry), hässlicher (ugly), kalter (cold),
klotziger (bulky), schmuckloser (inornate) [~]’

(ANTIPOS$_2$ = negative valuation)

PLUSVER(Bau) = ‘gewaltiger (huge), gigantischer (gigantic), großer
(large), massiver (massive), riesiger (gigantic) [~]’

(PLUSVER = more that it should be)

LIQU(Bau) = ‘[~] brennt ab (to burn down), brennt nieder (to burn
down), stürzt ein (to collapse), verfällt (fall into disrepair)’

(LIQU = cease to exist)

CAUSLIQU(Bau) = ‘[~] abreißen (to tear down), abtragen (to tear
down), entkernen (to remove the core)’

(CAUSLIQU = cause that sth. ceases to exist)

CAUSREAL(Bau) = ‘[~] errichten (build), hochziehen (bring up the
walls)’

(CAUSREAL = cause that sth. becomes real)

FINREAL(Bau) = ‘[~] fertigstellen (to finish), vollenden (to accomplish)’

(FINREAL = finalize the process of becoming real)

INCEPREAL(Bau) = ‘[~] beginnen (to start), entwerfen (to project),
planen (to plan)’

(INCEPREAL = begin the process of becoming real)

PERMREAL(Bau) = ‘[~] finanzieren (to finance), genehmigen (to
approve)’

(PERMREAL = allow sth. to become real)

• Non-standard LF

FORM(Bau) = ‘eingschossiger (one-story), einschiffiger (single-nave),
einstöckiger (one-story), geschlossener (compact), kubischer (cubic),
langgestreckter (stretched), lichtdurchfluteter (suffused with light),
luftiger (aerial), quaderförmiger (ashlar-formed), schlanker (lean),
würzelförmiger (cubic), zweigeschossiger (two-story) [~]’

FORM(Bau) = ‘[~] fällt [klein, groß, massiv...] aus (turn out ), überragt
[andere Bauten] (to tower above)’

FUNKTION(Bau) = ‘multifunktionaler (multi-functional), öffentlicher
Semantic Modeling of Collocations for Lexicographic Purposes

(public), sakraler (ecclesiastical)

FUNKTION(Bau) = ‘[~] beherbergt [jmdn., etw.] (to house), lockt (to attract), zieht [Besucher] an (to attract)

STIL(Bau) = ‘antiker (antique), barocker (baroque), gotischer (gothic), gründerzeitlicher (relating to a 19th century building boom), historischer (ancient, vintage), klassizistischer (classical), moderner (modern), postmoderner (post-modern) [~]’

(Stil = style)

URHEBER(Bau) = ‘städtischer (municipal), kommunaler (communal) [~]’

(URHEBER = author, originator)

MATERIAL(Bau) = ‘hölzerner (of wood) [~]’

STATUS(Bau) = ‘denkmalgeschützter (landmarked) [~]’

Note that, in contrast to the notation of Mel’čuk and Polguère, we consider the syntactic relation as the first level of ordering. A syntactic relation can realize several Lexical Functions. Therefore we reproduce, in this list, a lexical function for each kind of syntactic relation.

The lexical unit Bau (in the sense of building) has an exceptionally rich set of collocations. Ordering them semantically makes them much more usable for typical users of our resource, e.g. language learners and translators looking for appropriate lexical items and collocations to denote an aspect of it or a state of affairs which involves a building. The word Bau is interesting in a further respect: As an artifact it does not only have some attributes and is involved in various processes (see the standard LFs above) but it is also to be described in terms of its form, function, authorship etc. (see the non-standard LFs above). Those of the readers who are familiar with the “Generative Lexicon“ approach authored by James Pustejovsky (cf. Pustejovsky 1991) will realize that our choice on non-standard Lexical Functions resembles some, if not most, of the “qualia“ which have been introduces in this theory to describe aspects of the meaning of artifacts.

adressieren

In the following, we will present, as a further example, one verb. We want to show that the toolbox of (non)-standard Lexical Functions suffices to
describe lexical units of non-nominal parts of speech.

The verb adressieren (‘address’)

This verb has, in our dictionary, four senses (S1-S4)

1. mit einer Adresse versehen (Brief etc.) (‘to direct (a letter etc.) to a specific person’)
2. eine Botschaft, Nachricht an jmd., jds. Adresse richten (‘to communicate sth. to a specific person or audience’)
3. ein Problem, Thema ansprechen (‘to take on as a topic for discussion or inquiry’)
4. eine Zelle auf einem Speichermedium oder einen Rechner im Netz ansprechen (‘to specify a location in memory or a server in a network’)

Each of these senses triggers some collocations. Some of these lexical units co-occur with the base in more than one meaning, which is an additional indication that these senses are semantically related.

In some of the word-pairs with adressieren as one part, the word adressieren is the head of the collocation, in some others, the verb is a collocate. For completeness, we will list LFs in both directions.

VER(adressieren)= ‘richtig (correctly) [~]’ (S1)
(VER = as it should be)
BON(adressieren)= ‘persönlich (personally) [~]’ (S1, S2)
(BON = well)
MAGN(adressieren)= ‘direkt (directly) [~]’ (S1 – S4)
(MAGN = intensifier)

CAUS1REAL(Botschaft)= ‘[~] adressieren ’ (S2)
(CAUSREAL1 = cause sth. to become a reality)
CAUSREAL(Nachricht)= ‘[~] adressieren ’ (S1-S2)
CAUS1MANIF(Kritik)= ‘[~] adressieren ’ (S2)
(CAUS1MANFIF = cause sth. to become manifest)
CAUS1MANIF (Problem)= ‘[~] adressieren ’ (S3)
CAUS1MANIF (Thema)= ‘[~]adressieren ’ (S4)
OPER3(Anschrift)= ‘adressieren [an ~]’ (S1)
(OPER=sth. happens; the syntactic pattern is: sb. addresses sth. to a particular place/person)
OPER1(Paket)= ‘[~] adressieren ’ (S1)
OPER1(Brief)= ‘[~] adressieren ’ (S1)

Of course, one could change the perspective of direction of the latter collocations. In such a manner, we would have *adressieren* as the argument and *Botschaft* (‘message’), *Nachricht* (‘notice’), *Kritik* (‘criticism’) as arguments of a lexical function. The lexical function in this case would be a non-standard LF. We could give it a natural language label such as DIRECT. The same holds for the collocates *Brief, Paket* (‘letter, package’). However there is a subtle distinction. With *Botschaft*, we denote a non-material object, while with *Brief* we address the medium on which such a non-material message is written. Again, we find such a difference (let us call this the difference between the medium and the message) in the qualia structure of the resp. nouns in the framework of the “Generative Lexicon”-theory. This is a further argument to look into that theory and to look for links between both theories: Meaning-Text-Theory and the Generative Lexicon.

In this section we outlined plans to employ Lexical Functions in order to categorize and group collocates of specific headwords.

In the following section we will look at the bases of collocations. The idea is to “factor out” collocates that are shared by semantically related base words.

5.2. Comparing the Collocational Sets of Semantically Related Bases

A crucial aspect of our work is to generalize collocational relations over sets of bases in such a way that shared collocates of semantically related bases are described only once. If we model the lexical-semantic relations between the bases (see section 1 above) we would be able to encode some collocations at the highest node of the hierarchy and to inherit these collocations down the nodes in a hierarchy (i.e. the more specific terms, the hyponyms).

In section 3 we have described the set functions of the Wortprofil tool which compare the word profiles of two words – and, iteratively, of larger sets of base words. While the “diff” function returns collocates that are
Lothar Lemnitzer and Alexander Geyken

typical for one base word (and not the other) respectively, the “intersect” function returns those collocates which are typical for both words.

With this tools and with the German wordnet we have powerful means that, for the first time to the best of our knowledge, allow to capture collocates for semantically related words. We will demonstrate this in the following by some examples: first, we will intersect the collocates of Bau (‘building’) with all hyponyms which are registered in GermaNet, for example Schloss (‘castle’) or Altbau (‘old building’); second we will look at two synonyms, i.e. Vorstellungsgespräch and Bewerbungsgespräch (both: ‘job interview’), third we take two adjectives which partially overlap in their meanings. i.e. kariert (‘plaid’) and kleinkariert (‘small-minded’ and also ‘plaid’).

**Bau and its Hyponyms**

For around 250 (co-)hyponyms for Bau that we get from GermaNet, the Wortprofil yielded lists of collocates (details can be gleaned from the set of raw data that we provide together with this paper). By intersecting these lists, we obtained the following lists of “top collocates” (the ~ is again used as a placeholder for the base word):

- altes ~ [old] (130), ~ stehen [stand] (121), kleines ~ [small] (97), es gibt ~ [there is] (91), neues ~ [new] (78), großes ~ [big] (75), ehemaliges ~ [former] (71), in ~ gehen [enter into] (66), ~ bauen [build] (63), ~ errichten [construct] (59)

From these examples we learned that a) there are collocates which connect a significant number of the bases; b) these collocates are fairly common ones and refer to regular relations which are triggered by semantic features that are shared by the bases, e.g. form and function and c) that these regular relations can be modeled adequately with Lexical Functions.

Further down the list, there are collocates which connect interesting subsets of the buildings, e.g. essen (‘to eat’) which is a shared collocate of 11 bases, all referring to buildings with a gastronomic function, e.g. Bistro and Wirtshaus (‘inn’). Investigating such subregularities will help to obtain
more coherent semantic groups of base words.

*Bewerbungsgespräch and Vorstellungsgespräch*

In the following and in fig. 4 we present two synonyms: *Bewerbungsgespräch* and *Vorstellungsgespräch* (‘job interview’). The shared collocates (in the whiter area in the middle of the window in fig. 4) signify typical actions in relation to the interview at different stages, e.g. *einladen* (‘to invite’), *vorbereiten* (‘to prepare’), as well as to objects which are used in or in combination with such an event, e.g. *Assessment-Center, Einstellungstest* (‘aptitude test’). With pairs of (near) synonyms, it might also be interesting to look at the differences in collocational profiles. Again, Lexical Function will suffice to model most of these shared collocates.

*Fig. 4.* Contrastive word profile for the base lexemes *Bewerbungsgespräch* and *Vorstellungsgespräch* (both: ‘job interview’)*
**kariert vs. kleinkariert**

Both adjectives share a meaning, i.e. plaid, a characteristic of cloth and pieces of clothing. The shared collocates therefore refer to such objects, e.g. *Jacke* (‘jacket’) and *Hemd* (‘shirt’) but also *Muster* (‘texture’) is a shared collocate. With the Wortprofil “diff” function, we can tell apart both words by their distinctive meaning(s). *kleinkariert* co-occurs with character traits such as *engstirnig* (‘narrow-minded’) and speech acts such as *Mäkelei* (‘fault-finding’), the word *kariert* typically co-occurs with color-combinations such as *rot-weiß* (‘red-white’).

### 6. Summary and Further Perspectives

In this final section of the paper we would like to come back to our leading questions and summarize which answers the data provide so far. The leading questions are: a) concerning the collocates: are we able to classify these into lexical-semantic classes and group them accordingly; b) concerning the bases: are we able to find significant numbers of shared collocates for lexical-semantically related bases and thus reach some form of generalization and regular patterns?

Ad a) the tool of Lexical Functions that the Meaning-Text Theory provides serves a good starting point for a systematic and theoretically grounded grouping of collocations. The complex example *Bau* also showed that the set of Standard Lexical Function (and combinations thereof) do not suffice to cover all facets that the collocates represent. We have two options: either we introduce “non standard Lexical Functions”, at the risk of producing an extended vocabulary of LFs which is hard to control, or we refer to other lexical-semantic theories that cover these facets more appropriately. The qualia roles in the framework of the “Generative Lexicons” seem to come in handy, at least for our example of a concrete artifact. We want to look further, and with more data, into this issue and try to elaborate a descriptive vocabulary that draws on both theoretical frameworks.

Ad b) again, for our most complex example of *Bau* we could show that a sequence of pairwise intersections of word profiles provides a promising
tool for generalizations over a full set of lexical-semantically related words as well as over subsets of it and by this means to detect and describe (sub-) regularities. GermaNet provides a large and well-structured resource for a large part of the German vocabulary. However, while for some relations such as hyponymy / hypernymy, the “intersect”-functions yields interesting results, this is not the case for other relations. For the relations of synonymy and antonymy, it seems to be more worthwhile to look at the differences and to emphasize these in a vocabulary learning situation. The Wortprofil tool provides both functions.

Concerning the relation between the two theoretical frameworks: Meaning text theory seems to work better with combinations where the collocate relates to arbitrary and language specific aspects of the base word (while the Lexical Functions are universal). The adjectival attributes with evaluative meaning are a good example. The “Generative Lexicon” theory seems to work better with collocates that denote essential aspects (i.e. the qualia) of the base words. While overlapping in part, both theories seem to lend themselves to a combination for practical purposes.

In the context of our large-scale lexicographical project, we want to devote some effort to the identification and description of collocations for a significant amount of headwords. It turns out that a mere alphabetical order of collocates under a specific syntactic relation is not helpful for the user if the number of collocates exceeds a certain limit. We need a more fine-grained ordering of collocates. The experiments which we presented in this paper are conducted in order to find a solid and stable descriptive framework. The most promising path seems to be a combination of more than one theoretical framework. Before we can start with the practical work (beyond some small-scale experiments), we want to fix such a vocabulary and provide to our lexicographers guidelines about the proper use of it. We are confident that such guidelines will be of practical relevance beyond our project, e.g. for other lexicographical projects which will lead to general language dictionaries or to specialized dictionaries of collocations. Furthermore, such a lexicographical view of both theoretical frameworks will be also interesting for these frameworks and their respective development.
Acknowledgements

We would like to thank Jörg Didakowski for implementing the set functions of the Wortprofil, Reinhild Barkey from the GermaNet team in Tübingen to have a closer look on our collocations for Bau from a “Generative Lexicon” point of view, and the two anonymous reviewers for the helpful comments.

References


Linguistics 17(1991) 4, 409-441

PID of the accompanying dataset: hdl.handle.net/11858/00-203Z-0000-0028-2A76-7