

# Wortverbindungsfelder Fields of multi-word expressions

Annelen Brunner<sup>1</sup>, Kathrin Steyer<sup>1</sup>  
Institute for the German Language, Mannheim

## Abstract

In this paper we outline our corpus-driven approach to detecting, describing and presenting multi-word expressions (MWEs). Our goal is to treat MWEs in a way that gives credit to their flexible nature and their role in language use. The bases of our research are a very large corpus and a statistical method of collocation analysis. The rich empirical data is interpreted linguistically in a structured way which captures the interrelations, patterns and types of variances of MWEs. Several levels of abstraction build on each other: surface patterns, lexical realizations (LRs), MWEs and MWE patterns. Generalizations are made in a controlled way and in adherence to corpus evidence. The results are published online in a hypertext format.

**Keywords:** multi-word expression, collocation, corpus-driven, usage-based, corpus linguistics, phraseology, lexicology.

## 1. Methodological approach

We present a structured approach to the study of multi-word expressions (MWEs) which applies a strongly corpus-driven method and results in a novel type of lexicographic description and presentation (*cf.* Steyer and Brunner 2009, Brunner and Steyer 2009).

Based on the concept of *Usuelle Wortverbindungen* (Steyer 2000; Steyer 2004), we regard multi-word expressions as conventionalized patterns of language use that manifest themselves in recurrent syntagmatic structures (*cf.* Feilke 2004). MWEs can comprise fixed lexical components as well as abstract components representing a certain subset of lexical items. Our concept encloses not only idioms and idiosyncratic structures, but all multi-word units which have acquired a distinct function in communication. Real-life usage, pragmatics and context are central to our approach.

In detecting as well as describing these units we work bottom-up in a strongly corpus-driven way (*cf.* Sinclair 1991; Tognini-Bonelli 2001). The following principles, which

---

<sup>1</sup> {brunner,steyer}@ids-mannheim.de

correspond to the definition of corpus-driven work detailed by Tognini-Bonelli, characterize our approach.

We use the empirical basis of a very large corpus. DeReKo (Deutsches Referenzkorpus, KLa2009), located at the Institute for the German Language (IDS), is the largest collection of written German available today and comprises over 3.7 billion word tokens, mostly from modern newspaper articles. At the current stage we use DeReKo as it is, as our focus is on the model of analysis.

The data is pre-structured by statistical collocation analysis. The algorithm we use (“Kookkurrenzanalyse”, Belica 1995) is a sophisticated method which clusters keyword in context (KWIC) lines in several hierarchical levels and also computes the most common order of the surface forms which appear in those clusters (cf. KLa2009, Keibel and Belica 2007). The results are a very good basis for our work, as the statistical method shows regularities in the data in a very objective way by considering only word form surfaces. However, we do not take the clusters as they are but use them as a starting point for human interpretation.

Interpreting this rich empirical data we try to take as few pre-conceived notions of how language works as possible and develop the analysis and presentation of the data to fit corpus evidence. We work bottom up from the language surface structure and take monitored steps of interpretation.

In strong adherence to corpus data, we only describe MWEs and variations of MWEs which are attested in our corpus, so the results are always grounded on empirical evidence. As a result of studying corpus data, we came to consider three characteristics as central to the nature of MWEs:

- Usage and context are crucial when identifying and describing MWE entities.
- Most MWEs are variable and can very often be modified and extended in various ways.
- There are rich interrelations between MWEs such as similarities and contrastive nuances in usage, combinations of MWEs which create rich forms of expression and more abstract groups of structurally similar MWEs, which are no longer completely fixed on the lexical surface.

These characteristics are emphasized in our model for describing MWEs.

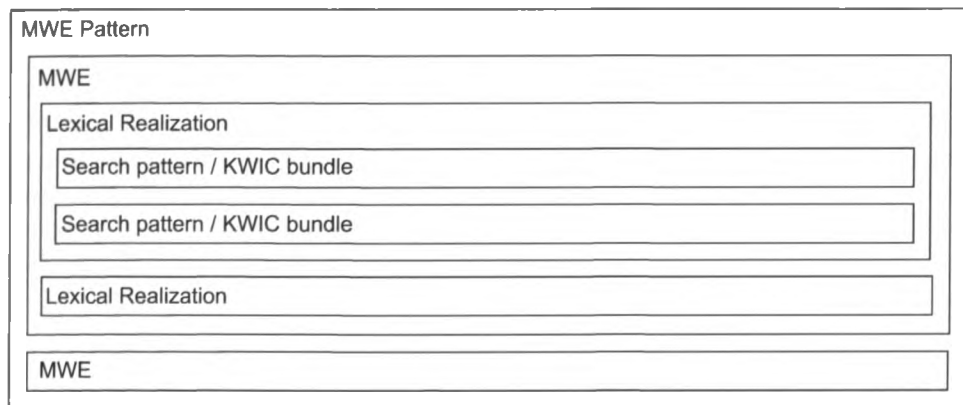
## 2. Model of analysis

Our model of analysis has some similarities to that of Hanks detailed in the description of his *Corpus Pattern Analysis* (CPA):

Concordance lines are grouped into semantically motivated syntagmatic patterns. Associating a ‘meaning’ with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns. The identification of a syntagmatic pattern is not an automatic procedure: it calls for a great deal of lexicographic art. Among the most difficult of all lexicographic decisions is the

selection of an appropriate level of generalization on the basis of which senses are to be distinguished.” (CPA2009)

We, too, have to tackle the task of assigning meaning to syntagmatic patterns and to find the right level of abstraction. CPA aims at describing single words (*cf.* Hanks 2008), while we are interested in MWEs, which adds an additional level of complexity as identifying the surface form itself requires an interpretative effort. To handle the difficulties of generalization, our model has several hierarchical levels which build upon each other. Figure 1 gives an overview of its structure.



*Figure 1. Hierarchical model of analysis*

In this paper, we will mostly focus on the example MWE “in den Ohren klingen” [to sound in the ears].

As a starting point, we conduct a collocation analysis of the target word form “Ohren” [ears]. We decided to only use non-lemmatized word form surfaces as targets for the algorithm, as our model of analysis is strongly surface based. The collocation analysis of different inflectional forms of a lemma can result in quite different profiles and we do not want to gloss over these differences too quickly. In this respect, we adhere to Sinclair’s claim:

“There is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity.” (Sinclair 1991: 8)

Collocation analysis outputs several clusters which are relevant for the MWE “in den Ohren klingen”, mainly those forming around inflected forms of “klingen” [to sound]. The KWIC lines which comprise these clusters will be the basis of our analysis.

### 2.1. Search patterns

On the first level, the KWIC lines which have been clustered by collocation analysis are explored and subjected to further structuring. For this task, we use search patterns based on regular expressions.

This step is necessary because collocation analysis shows the relationships between word surfaces, but does not consider the underlying syntactic structure nor can it recognize similarities in meaning and usage. For example, the cluster “Ohren - klingen” [ears - to sound] also contains realizations of other MWEs like “die Ohren klingen” [the ears resound].

At this point, human interpretation builds on the pre-structuring done by statistics. Search patterns can be defined flexibly to capture the structures we are interested in. They serve an analytical purpose, as they allow us to explore possible surface variations, *e.g.* common fillers of slots between fixed elements which can be examined as well.

For example, we find that though the surface form “in den Ohren klingen” [to sound in the ears] is indeed the most common realization of the MWE, the element “den” [the] is quite often replaced in the actual realizations. With the help of the search patterns we explore the fillers for the slot between “in” and “Ohren” and find that three kinds of fillers are most dominant: possessive pronouns, genitive phrases referring to a person and adjectives denoting groups of people, most often referring to their nationalities.

These are example KWIC lines for the three different kinds of surface realizations:

A98/SEP.60063 und aus früheren Tagen **in unsern Ohren klingen.**

P92/FEB.04217 Wie Hohn mußte **in Strolz' Ohren** der Beifall der Tausenden **klingen**

A01/OKT.36053 Die Erklärungen des saudischen Diplomaten mögen **in westlichen Ohren** hohl und feindselig **klingen**

Search patterns allow us to group instances which have similar surface characteristics so that these groups can serve as the basis of further analysis.

## 2.2. Lexical realizations

Lexical realizations (LRs) are an intermittent step between the hard language surface, as captured by the search patterns and the MWEs. Corpus research clearly shows that the surface form of an MWE is nearly always subject to variation. When generalizing quickly to a single form, many of these nuances are lost. LRs allow us to focus on different typical forms an MWE can take, to show their relationships and to comment on them. An MWE in our model is thus represented by a collection of LRs organized in a tree-like structure.

We distinguish between different kinds of LRs according to a basic set of types which was developed from empirical experience.

For each MWE, a *Core LR* is defined which represents the minimal surface structure necessary to recognize the MWE in its communicative function. Alternative core

realizations can exist, called *Core Variant LRs* in our model. In addition, we define *Extension LRs*, extensions to the core, which can be internal as well as external modifications and additions, e.g. prepositional phrases, verbs, modifying adjectives or adverbs. The last type of LR is *Context LR*, defined to highlight word forms which typically appear close to the MWE realization without being part of its structure.

The *LR Group* represents a container which contains all realizations of the MWE. It also serves as a root element of an LR tree. The other LRs can be arranged in several levels.

The LR structure of the MWE “in den Ohren klingen” is shown in Figure 2.

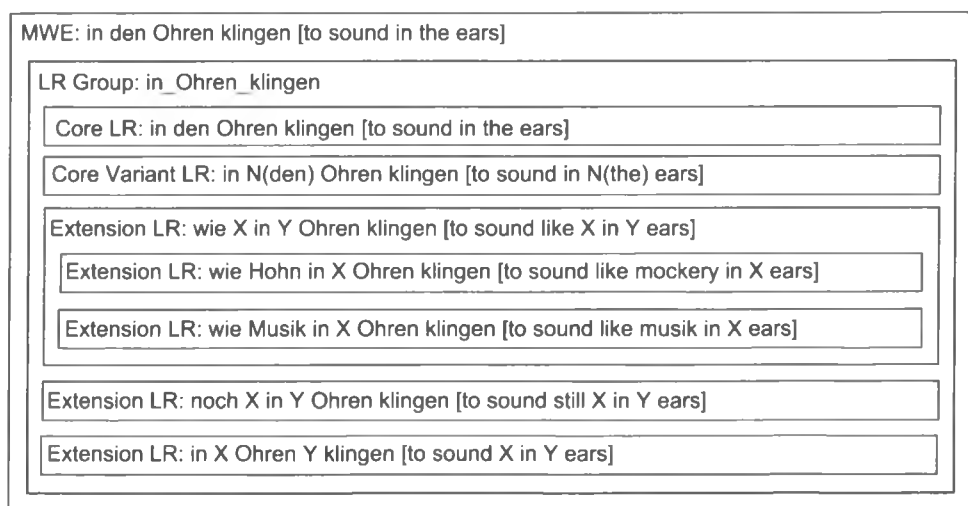


Figure 2. LR tree of the MWE “in den Ohren klingen”<sup>2</sup>

Each LR subsumes the appropriate KWIC lines which are bundled by search patterns and contains more information about the specifics of this realization, for example nuances in meaning and usage and relative frequency information.

LRs can contain slots, which are represented by capital letters in the LR’s name. For example, the Extension LR “wie X in Y Ohren klingen” [to sound like X in Y ears] has two slots: Slot Y further specifies the word form “Ohren” [ears] and its fillers are of the kind we found by studying the search patterns in the previous section – most often: “den”, possessive pronouns, adjectives.

<sup>2</sup> This example does not contain a Context LR. A typical Context LR would be for example “Knopf im Ohr ... Steiff” [button in the ear ... Steiff] which belongs to the MWE “Knopf im Ohr” [button in the ear]. This Context LR highlights a word form, “Steiff” (the name of a toy company), which appears very frequently in the vicinity of the MWE’s Core LR. This is an indicator that the MWE is often used to refer to a characteristic of stuffed animals manufactured by the company Steiff, which have a metal button punched into their ear as a brand label.

Slot X is filled by nouns which serve as a simile for how something is received or experienced. Two fillers for this slot are extremely frequent: “Hohn” [mockery] and “Musik” [music]. Because of their typicality the realizations with these fillers are presented as separate LRs, which are dependent on the LR “wie X in Y Ohren klingen”.

Such slots are represented as tables in the LR’s body which list the abstract types and/or concrete lexical items that serve as fillers. These tables are created manually as a result of the study and categorization of the KWIC lines. Only systematic slots, *i.e.* slots with fillers which show some regularity, are represented in this way. They give important insight into the paradigmatic variability of MWEs.

In addition to that, each LR gives direct access to the KWIC lines captured by the search patterns it subsumes and to automatically generated lists of the surface realizations of every underspecified element in these patterns – an unrevised slot-filler list. So it is also possible to take a look at the hard corpus data and see the raw frequencies.

### 2.3. MWEs, MWE patterns and relationships between them

MWEs are represented by an LR tree as shown in Figure 2 above. In addition, each MWE is assigned a description which contains a paraphrase that is true for all LRs and represents the core meaning of the MWE. For our example “in den Ohren klingen” the general paraphrase would be: “sth is experienced intensely in a certain way and remembered”. Depending on the realization of this rather complex MWE different aspects of this general meaning are emphasized.

In addition to that, an MWE can also contain information about its typical genre, its phrasal structure and its relative frequency in the collocation profile of the target word form.

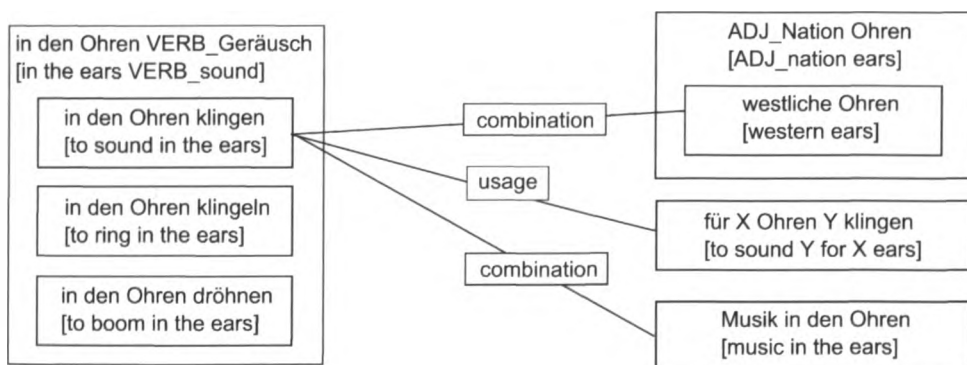
MWE patterns are an additional step of abstraction which is not obligatory for all MWEs. The patterns are generalizations over structurally similar MWEs and contain at least one underspecified component. Two types of MWE patterns can be distinguished:

1. The MWEs which comprise the pattern are near synonyms and the same meaning can be assigned to all of them. In this case, the meaning paraphrase is assigned to the MWE pattern instead of the separate MWEs.
2. The realizations of the underspecified components are all different in meaning. This results in a group of MWEs which each have a distinct meaning but still have a meaning component in common. The MWE pattern is assigned the most general meaning paraphrase, but each MWE still carries its own meaning paraphrase detailing its specifics.

The example MWE “in den Ohren klingen” can be considered part of an MWE pattern “in den Ohren VERB\_Geräusch” [in the ears VERB\_sound] and is grouped together

with similar MWEs. These MWEs are not completely identical in meaning – “in den Ohren klingen”, which is also the most frequent of the three, has a much richer meaning than the other MWEs. However, in one aspect, they are indeed very similar: They can all express the meaning “sth is experienced intensely (most often acoustically)”.

Another important aspect of our model is that interrelations can be defined between MWEs or MWE patterns. These interrelations can be of different kinds, but often involve a similarity in usage or a frequent combination of MWEs or MWE patterns. The interrelation structure of “in den Ohren klingen” is represented in Figure 3.



*Figure 3. Interrelations between MWE “in den Ohren klingen” and other MWEs and MWE patterns*

“In den Ohren klingen” is often combined with “Musik in den Ohren”, resulting in the form “wie Musik in den Ohren klingen” [to sound like music in the ears]. With the MWE pattern “ADJ\_Nation Ohren” it is combined to form the realization “in ADJ\_Nation Ohren klingen” [to sound in ADJ\_nation ears]. The MWE “für X Ohren Y klingen” [to sound Y of X cars] is very similar to one meaning aspect of “in den Ohren klingen”: “to be experienced in a certain way by a certain group of people”.

### 3. Implementation and presentation

For our analysis, we use a specially developed software tool, which takes collocation clusters as input and is used to match, group and annotate the KWIC lines according to the model described above. The analyzed data are stored in an XML format which allows different modes of visualization.

Currently, our results are presented as fields of MWEs (“Wortverbindungsfelder”), each centered on a specific word form. The hierarchical structures and interrelations between the different units are realized in a hypertext format and direct access to structured corpus data is provided. All levels of description are enriched by lexicographic comments like the description of meaning and usage in the corpus. Thus

the results can be viewed in two different ways. On the one hand, the structure allows for the reconstruction of the typical usage of MWEs from the corpus data and provides a complete documentation of our interpretative method. On the other hand, the narrative comments allow an access more similar to that of traditional lexicographical products. The first version of fields of MWEs, one centred on forms of the word “Grund” [ground/reason] and one centered on forms of the word “Ohr” [ear] are available on the internet, accessible from our site “Wortverbindungen online”: <http://wvonline.ids-mannheim.de/>

Though developed in an experimental research context, we believe that our approach can give valuable impulses to lexicographic practice: Working with real-life data helps revising common misapprehensions about the structure and meaning of MWEs and results in a new form of presentation, highlighting the importance of variability, context and usage. In addition to that, our model presents a novel approach in including corpus data not only as illustration, but as a basis of description, and offers structured access to real-life data, taking advantage of the options of the electronic hypertext format.

## References<sup>3</sup>

- BELICA, C. (1995). *Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode*. Mannheim: Institut für Deutsche Sprache: <http://www.ids-mannheim.de/kl/projekte/methoden/ur.html>.
- BRUNNER, A. and STEYER, K. (2009). A Model for Corpus-Driven Exploration and Presentation of Multi-Word Expressions. In J. Levická and R. Garabík (eds). *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice, Slovakia, 25-27 November 2009. Proceedings*. Bratislava: Tribun: 54-64.
- CPA2009. *Corpus Pattern analysis*. Internet: <http://nlp.fi.muni.cz/projekty/cpa/>.
- FEILKE, H. (2004). Kontext – Zeichen – Kompetenz. Wortverbindungen unter sprachtheoretischem Aspekt. In K. Steyer (ed.). *Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003*. Berlin/New York.
- HANKS, P. (2008). Lexical Patterns. From Hornby to Hunston and Beyond. In E. Bernal and J. DeCesaris (eds). *Proceedings of the XIII Euralex International Congress, Barcelona, 15-19 July 2008*. Barcelona: Institute for Applied Linguistics, Pompeu Fabra University: 89-129.
- KEIBEL, H. and BELICA, C. (2007). CCDB: A Corpus-Linguistic Research and Development Workbench. In *Proceedings of Corpus Linguistics 2007*, Birmingham. [http://corpus.bham.ac.uk/corplingproceedings07/paper/134\\_Paper.pdf](http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf).
- KL2009. *Ausbau und Pflege der Korpora geschriebener Gegenwartssprache. Das Deutsche Referenzkorpus – DeReKo*. Mannheim: Institut für Deutsche Sprache: <http://www.ids-mannheim.de/kl/projekte/korpora/>.
- KL2009: *Methoden der Korpusanalyse- und erschließung*. Mannheim: Institut für Deutsche Sprache: <http://www.ids-mannheim.de/kl/projekte/methoden/>.
- SINCLAIR, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

---

<sup>3</sup> All hyperlinks checked on 14 December 2009.



- STEYER, K. (2000). Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten. *Deutsche Sprache*, 28(2): 101-125.
- STEYER, K. (2004). Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In K. Steyer (ed.). *Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003*. Berlin/New York: DeGruyter: 87-116.
- STEYER, K. and BRUNNER, A. (2009). *Das UWV-Analysemodell. Eine korpusgesteuerte Methode zur linguistischen Systematisierung von Wortverbindungen*. Mannheim: Institut für Deutsche Sprache: <http://www.ids-mannheim.de/pub/laufend/opal/privat/opal09-1.html>.
- TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins Publishing Company (*Studies in Corpus Linguistics* 6).