Contexts, Patterns, Interrelations - New Ways of Presenting Multi-word Expressions

Kathrin Steyer
Institute for the German Language
R 5, 6-13
D-68161 Mannheim, Germany
steyer@ids-mannheim.de

Annelen Brunner
Institute for the German Language
R 5, 6-13
D-68161 Mannheim, Germany
brunner@ids-mannheim.de

Abstract

This contribution presents the newest version of our 'Wortverbindungsfelder' (fields of multi-word expressions), an experimental lexicographic resource that focusses on aspects of MWEs that are rarely addressed in traditional descriptions: Contexts, patterns and interrelations. The MWE fields use data from a very large corpus of written German (over 6 billion word forms) and are created in a strictly corpus-based way. In addition to traditional lexicographic descriptions, they include quantitative corpus data which is structured in new ways in order to show the usage specifics. This way of looking at MWEs gives insight in the structure of language and is especially interesting for foreign language learners.

1 Our concept of MWEs

We study MWEs from a linguistic perspective and are mainly interested in two questions: What can we learn about the nature of MWEs and their status in language by studying large corpora? And how can we present MWEs in novel lexicographic ways that reflect our findings? The MWE field presented in this contribution is a prototype that reflects our current ideas regarding these questions. It can be explored online free of charge at http://wvonline.ids-mannheim.de/wvfelder-v3/index.html.

Our approach is based on the concept 'Usuelle Wortverbindungen' (UWV, Steyer 2000; Steyer 2004; Steyer 2013), which defines MWEs as conventionalized patterns of language use that manifest themselves in recurrent syntagmatic structures. This includes not only idioms and idiosyncratic structures, but all multi-word units which have acquired a distinct function in communication. Our focus is on real-life usage, pragmatics and context. We work bottom-up in detecting and describing MWE units in a strongly corpus-driven way (Sinclair 1991; Tognini-Bonelli 2001; Hanks 2013), taking iterative steps to arrive at conclusions about language use. Methologically, our approach bears some similarities to Stefanowitsch/Gries' 'collostructions' (Stefanowitsch/Gries 2003) though we are less interested in syntactic and grammatical structures - as it is common in construction grammar approaches - but see MWEs primarily as parts of the lexicon and feel closer to phraseology.

The basis of our research is DeReKo (Deutsches Referenzkorpus, Institut für Deutsche Sprache 2012), the largest collection of written German available today which has over six billion word tokens and is located at the Institute for the German Language (IDS). In the current stage of our work, which is mainly explorative, we use DeReKo as it is. This means our text basis is dominated by newspaper texts from the last 10-15 years. Though this is surely not a 'balanced' corpus, we argue that it still reflects much of contemporary written language use, as newspaper texts are a medium that is widely disseminated.

Though the interpretation and main analysis is done manually, automatic methods form an important basis to our work. We use a sophisticated method of collocation analysis developed at the IDS (Belica 1995) to get indications which word combinations constitute MWEs and to explore contexts in which an MWE is commonly used. In addition to that, we use a pattern matching tool developed in our project to explore and structure corpus evidence and gain further insight into the behavior and variations of MWE candidates.

Our special interest lies in the fact that MWEs are not as fixed as is often assumed, but often behave as patterns and show multiple interrelations. Therefore, we also describe MWE patterns - a

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 82–88,
Gothenburg, Sweden, 26-27 April 2014.

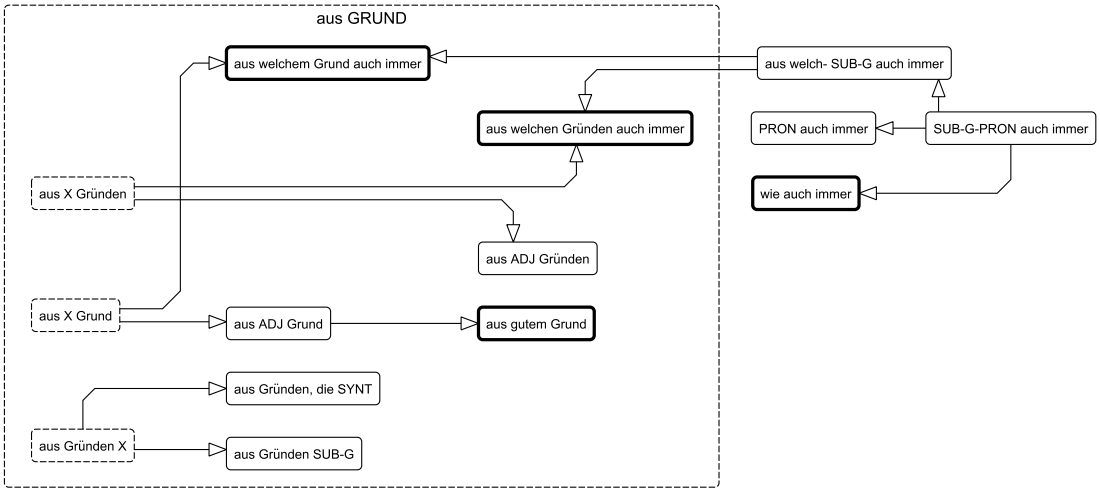

Figure 1: Part of the MWE field centered around *Grund* and preposition *aus*.

more abstract form of MWEs which are only partially fixed. An example for a fixed MWE is *Pi mal Daumen* (*pi times thumb* - 'approximately'), a multi-word expression that is always used in exactly this form. MWE patterns on the other hand consist of fixed lexical components as well as slots that can be filled in different ways. In spite of this variability, the whole pattern has a holistic meaning and function. An example is the expression *wie NOUN in jemandes Ohren klingen* (*to sound like NOUN in someone's ears* - 'to be perceived in a certain way' (specified by NOUN)). The NOUN slot can be filled with different words in order to specify the general meaning of the pattern. In section 2.3 we will go into further detail about how a slot in an MWE pattern can be filled.

The MWE field presented in this contribution centers around the word *Grund* (*reason/basis/foundation*) combined with several prepositions. It is the newest of several versions of MWE fields which have been described elsewhere (cf. Brunner/Steyer 2009; Brunner/Steyer 2010) and are available at our website `http://wvonline.ids-mannheim.de` as well. This newest version focusses more on hierarchies of MWEs and MWE patterns and incorporates additional resources like collocation analyses in its descriptive texts. In the following, we will highlight some features of the MWE field which illustrate our focus on interrelations, contexts and patterns.

2 MWE field *Grund*

2.1 Interrelations

Figure 1 shows a part of the MWE field, centered on the word *Grund* and preposition *aus*. Each node is linked to a lexicographic description. Figure 2 presents a screenshot of one of those articles. In addition to narrative descriptions and manually selected usage examples from our corpus, the articles also include components that are derived from quantitative corpus data. Specifically, these are collocation analyses as well as filler tables for MWE patterns. The function of these components will be explained in more detail in sections 2.2 and 2.3.

In Figure 1, you can observe the relations between MWEs (thick border) and MWE patterns (regular border). The nodes with the dashed border represent repeating surface structures which themselves have no common holistic meaning but show the lexical interconnectedness between the MWEs and MWE patterns.

All nodes enclosed in the square field contain the elements *Grund* and *auf*. The nodes on the far right are extensions which do not belong to the core of the MWE field as it was defined, but are connected lexically and functionally to MWEs that do. We decided to include those 'external nodes' to give a glimpse of how the building blocks of language connect even beyond the artificial borders that where necessary when defining the MWE field.

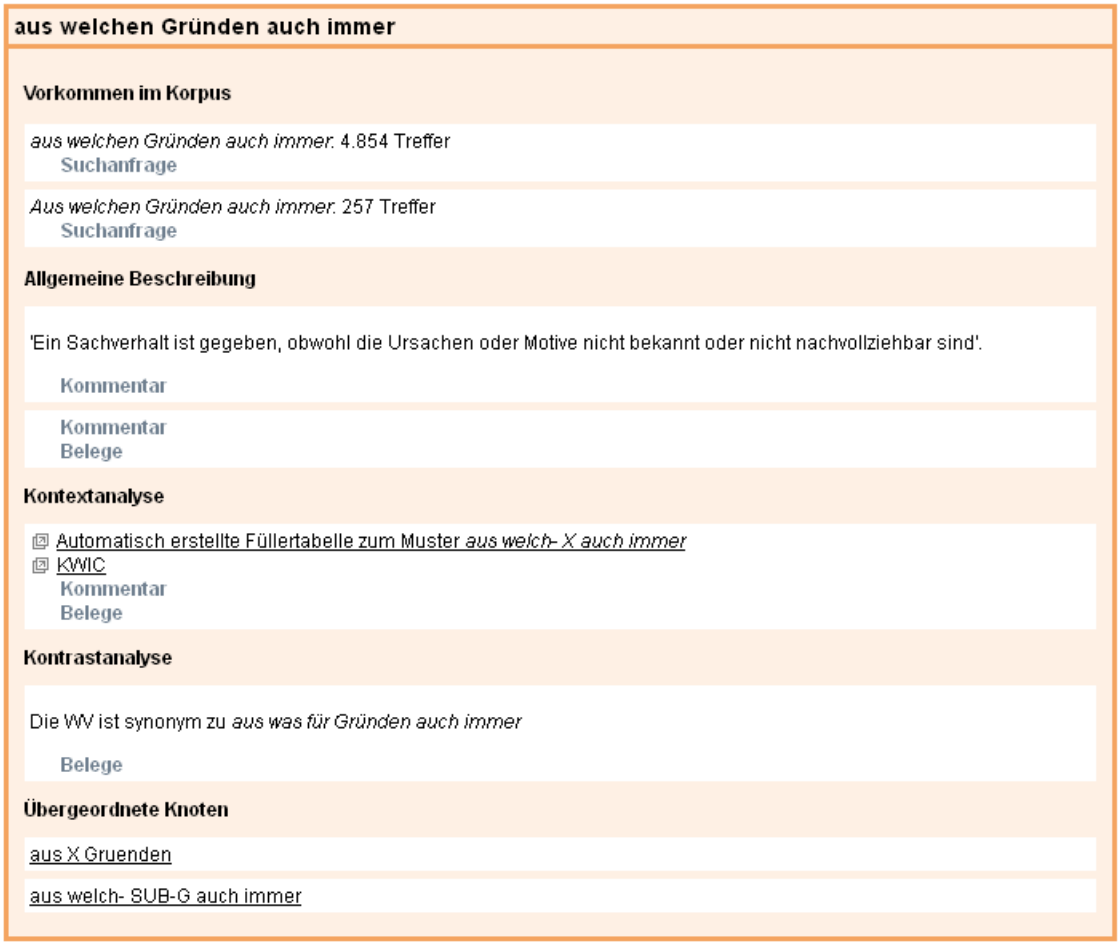

Figure 2: MWE article *Aus welchen Gründen auch immer* from the MWE field *Grund*. The article parts are 'Frequency in the Corpus', 'General description', 'Context Analysis', 'Contrast Analysis' and 'Superordinated Nodes'. The part 'Context Analysis' contains links to a filler table and to the corresponding KWIC lines.

In this example the core field contains the MWEs *aus welchem Grund auch immer* and *aus welchen Gründen auch immer* ('for whatever reason/s'). However, the lexical components *auch immer* are part of more general patterns as well. The word form *Grund* can be substituted by different nouns in the MWE pattern *aus welch- SUB-G auch immer* (e.g. *Motiv (motive), Richtung (direction)*). In the MWE pattern *PRON auch immer* the place is taken by an interrogative pronoun (e.g. *was (what), wo (where), wer (who), warum (why)*). One of those pronoun fillers, *wie* (*how*), is much more frequent than the others, which justifies the definition of a separate MWE *wie auch immer*, which can be translated as 'howsoever' or 'to whatever extent' (see section 2.3 for more details).

The basic structure of the MWE field thus highlights the different degrees of abstraction of MWEs and the functional use of lexical clusters like *auch immer*. The lexicographic descriptions linked to the nodes explain the interrelations and the differences in usage and meaning.

2.2 Contexts

Another important aspect of our approach to MWEs is that we pay close attention to the contexts in which they are commonly used. A good tool to explore this empirically is collocation analysis. In addition to narrative descriptions and manually selected corpus examples we therefore include the results of collocation analysis in our articles.

One interesting aspect is the difference between

Total	Anzahl	LLR	Kookurrenzen	syntagmatische Muster
36400	36400	44956	Was	99% Was [ist ...] eigentlich
51283	14883	29960	Warum	99% Warum [...] eigentlich
102957	51674	28059	was	99% was [...] eigentlich
113896	10939	24436	müsste	99% müsste [...] eigentlich
123499	9603	11569	obwohl	99% obwohl [...] eigentlich
125982	2483	10271	worum	100% worum [es ...] eigentlich
129066	3084	8837	Wieso	100% Wieso [...] eigentlich
132056	2990	8058	Schade	100% Schade [...] eigentlich
138601	6545	7501	Wo	100% Wo [ist ...] eigentlich
138613	12	6969	müßte Humptata-Musik	100% die unvermeindliche |unvermeidliche Humptata-Musik müßte eigentlich
149934	7750	6891	warum	99% warum [...] eigentlich
163416	13482	6513	wollte	99% |ch wollte [...] eigentlich
168752	5336	4620	müssten	99% müssten [...] eigentlich
168769	17	3625	Woher nimmst	100% Woher nimmst [du |Du] eigentlich
177773	7413	3589	wer	99% wer [...] eigentlich
178666	893	3317	Worum	100% Worum [geht es ...] eigentlich
196190	17524	3296	denn	99% denn [...] eigentlich
198747	2557	3229	Gibt	100% Gibt [es] eigentlich
199781	1034	2754	Wozu	100% Wozu [...] eigentlich

Figure 3: Highest ranking results of the collocation analysis for *eigentlich* (scope: 5 words in front).

MWEs and their single-lexeme quasi-synonyms. For example the meaning of the MWE *im Grunde* is very close to the lexeme *eigentlich* (*actually*). Figures 3 and 4 show the highest ranking results of a collocation analysis that focusses on a window of five words in front of the units *eigentlich* and *im Grunde* respectively and calculates the log likelihood ratio.[1] When comparing the results for these two units you can see that there are some contexts that are strongly preferred by *eigentlich* but are not highly ranked for *im Grunde*. Notable are the combination *schade eigentlich* (*sad actually*) as well as combinations with interrogative adverbs like *wie (how), was (what), warum (why)*. The MWE *im Grunde*, on the other hand, has strong collocation partners that are capitalized conjunctions like *aber (but)* or *denn (because)*. This indicates a clear tendency to appear near the beginning of a sentence in contexts where an argument is made, which is not prominent for *eigentlich*. So even if a quasi-synonymous single lexeme exists, the MWE shows differences in usage which become apparent when studying large quantities of data.

[1]For details on the collocation analysis used here see Perkuhn/Belica 2004. The settings were: *Korpus: W-gesamt - alle Korpora des Archivs W (mit Neuakquisitionen); Archiv-Release: Deutsches Referenzkorpus (DeReKo-2013-II); Analyse-Kontext : 5. Wort links bis 0. Wort rechts; Granularität: grob; Zuverlässigkeit: analytisch; Clusterzuordnung: mehrfach; Auf 1 Satz beschränkt: ja; Lemmatisierung: nein; Funktionswörter: zugelassen; Autofokus: aus*

2.3 Patterns

As mentioned before, MWE patterns are of special interest to us. When exploring MWEs, we use a pattern matching tool that allows us to search large quantities of keyword in context lines (KWICs) for combinations of fixed strings and slots. The lexical fillers of these slots can also be counted and presented in the form of frequency tables. This allows us to explore which kinds of variations are possible and typical for an MWE. The filler tables can show quite different 'profiles' for a slot. In the following, we will give some examples.

For the MWE *aus welchen Gründen auch immer* (*for whatever reasons*) we checked whether the element *Gründen* can be modified by searching for the pattern `aus welchen #* Gründen auch immer` (`#*` stands for a slot that can be filled by any number of words). Table 1 shows the absolute and relative frequencies that where calculated from KWIC lines of our corpus. In the vast majority of cases, the slot is empty, which means that the MWE is used exactly in the form cited above: *aus welchen Gründen auch immer*. It is thus very stable, though not completely inflexible, as there is also evidence of adjectives that are used to further specify the reasons in question, e.g. *persönlichen Gründen (personal reasons)*.

A different example of filler behavior can be observed when studying the pattern `# auch immer` (`#` marks a slot that has to be filled with exactly one word). Table 2 shows that this slot

Total	Anzahl	LLR	Kookurrenzen	syntagmatische Muster
1580	1580	396	Aber	100% Aber [...] im
2783	1203	278	ja	99% ja [...] im
3644	861	251	Denn	100% Denn [...] im
3667	23	147	einchecken	100% einchecken müssen dann starten Sie im
3697	30	121	Dr	100% Herr Dr ... im
3704	7	100	Besitzverteidigung	100% und Besitzverteidigung eingesetzt wird - im
3721	17	96	&	100% & [...] im
3757	36	87	bzw	100% bzw [...] im
3781	24	86	usw	100% usw [... ist] im
3791	10	73	Whishaw Dustin	100% Whishaw Dustin Hoffman"Das Parfüm ist im
4080	279	63	obwohl	100% obwohl [sie] im
4083	3	48	produktionsethischer	100% Sache produktionsethischer Bravheit des Eigensinns im
4086	3	43	Undelicatesse	100% Undelicatesse gegen uns Denker , im
4098	12	43	Kriminalkomödie	100% eine makabre Kriminalkomödie im
4101	3	41	Akku-Beleuchtung	100% heute mit Akku-Beleuchtung unterwegs - im
4106	5	40	Netanjahu-Regierung	100% etwas schwächere Netanjahu-Regierung die aber im
4126	20	40	hinwegtäuschen	100% darüber hinwegtäuschen daß |dass Jodie Foster im
4129	3	40	Bio-Mischung	100% Bio-Mischung seien im
4132	3	38	Zivi-Jobs	100% die Zivi-Jobs bei denen es im

Figure 4: Highest ranking results of the collocation analysis for *im Grunde* (scope: 5 words in front).

Filler	Freq	Rel Freq
	1239	98.33
unerfindlichen	3	0.24
persönlichen	2	0.16
legitimen	1	0.08
durchsichtigen	1	0.08
politischen	1	0.08
rätselhaften	1	0.08
psychologisch-persönlichen	1	0.08
mir nicht verständlichen	1	0.08
besagten	1	0.08
(PR-)	1	0.08
psychologischen	1	0.08
(un)berechtigten	1	0.08
"	1	0.08
(oft ökonomischen)	1	0.08
...	...	...

Table 1: Fillers of the pattern `aus welchen #* Gründen auch immer`.

Filler	Freq	Rel Freq
Wie	9611	10.08
wie	7389	7.75
was	5289	5.55
aber	3397	3.56
Gründen	3157	3.31
es	2288	2.40
Was	1953	2.05
Wer	1825	1.91
sich	1677	1.76
warum	1529	1.60
wo	1486	1.56
wer	1446	1.52
ja	1333	1.40
wem	1292	1.35
ist	1276	1.34
...	...	...

Table 2: Fillers of the pattern `# auch immer`.

is filled by *wie* (capitalized or non-capitalized) in nearly 18 percent of the matches. In this case, a single lexical filler is very dominant. This was a strong indication for us that the pattern *wie auch immer* functions as an MWE while at the same time being a prototypical realization of the pattern *PRON auch immer*. Also quite frequent is the filler *Gründen*, which indicates the pattern *[aus welchen] Gründen auch immer*, and other interrogative pronouns and adverbs like *was (what), wer (who), wem (whom)* etc. This lead us to define the MWE hierarchies as shown in figure 1 and explained in section 2.1.

A different filler profile (Table 3) can be observed for the pattern `aus # Gründen` (*for # reasons*). This is a true MWE pattern, as it has a specific communicative function tied to the plural form of Grund: reasons are mentioned, but left intentionally vague. Table 3 shows that there is a large number of adjectives that can fill the gap. In contrast to the example `X auch immer` above,

Label		Aus|aus	#	Gründen	
SOZ07_10	weshalb das Oratorium	aus	akustischen	Gründen	auch nicht in einer Kirche aufgeführt
WPD11_4133	werden, deren Ausbau	aus	unerfindlichen	Gründen	gestoppt wurde, die Brutalität
BRZ11_258	dem sie sich bisher	aus	finanziellen	Gründen	immer zurückhielten. Um sich auch
M07_208	Oliver Kahn	aus	disziplinarischen	Gründen	für das Hertha-Spiel an Schärfe
E98_409	möglicherweise	aus	wirtschaftlichen	Gründen	zurückgehalten. Schliesslich ist Epo
WDD11_305	schlage diesen Artikel	aus	folgenden	Gründen	als lesenswert vor: fachlich
NUN11_144	die Polizei	aus	ermittlungstaktischen	Gründen	nicht mitteilen.
...	...	...	...	...	...

Table 4: KWIC lines of the pattern `aus # Gründen`.

Filler	Freq	Rel Freq
gesundheitlichen	7355	10.03
beruflichen	6311	8.60
finanziellen	4708	6.42
persönlichen	2660	3.63
organisatorischen	2585	3.52
politischen	2499	3.41
wirtschaftlichen	2180	2.97
privaten	1941	2.65
welchen	1849	2.52
verschiedenen	1779	2.43
diesen	1494	2.04
anderen	1381	1.88
technischen	1260	1.72
zwei	1237	1.69
familiären	1219	1.66
...	...	...

Table 3: Fillers of the pattern `aus # Gründen`.

none of these is so dominant and striking that a separate MWE needs to be considered. However, the fillers can be grouped into functional groups, like type of the reasons (e.g. *politisch (political), persönlich (personal), finanziell (financial)*), validity of the reasons (e.g. *nachvollziehbar (understandable), gut (good), triftig (valid)*) or relevance of the reasons (e.g. *wichtig (important), zwingend (imperative)*).

You can see that filler tables are very useful for different purposes: To confirm the fixedness of an MWE and explore occasional variations, to conceptualize lexical units in order to build up hierarchies, and to further describe and understand the behavior of MWE patterns. Not only do we work with such patterns and filler tables when building the MWE field, we also include them in our descriptions - another way to give a user access to original corpus data structured in an informative way.

Additionally, we provide access to the KWIC lines that were used to calculate the filler tables. Table 4 shows some of the lines that match the pattern `aus # Gründen`. These lines are structured in fields according to the search pattern and the different columns can be sorted. In this way, you can explore the use of specific MWE structures yourself.

3 Conclusion

We believe that our MWE fields allow a different way to look at MWEs which is very useful to understand the structure of language. As they are strictly based on data from a large modern language corpus, our findings also reflect real, contemporary language use. This is especially useful for foreign language learners who struggle to navigate the complexities of fixedness and variability in the German language. In continuing our MWE research, we strive to refine our strategies for description and visualization and also plan to add contrastive studies in the future.

References

Belica, Cyril: Statistische Kollokationsanalyse und Clustering. Korpusanalytische Analysemethode, 1995 ⟨URL: `http://www1.ids-mannheim.de/kl/projekte/methoden/ur.html`⟩ – visited on 28.01.2014.

Brunner, Annelen/Steyer, Kathrin: A Model for Corpus-Driven Exploration and Presentation of Multi-Word Expressions, in:

Levická, Jana/Garabík, Radovan, editors: NLP, Corpus Linguistics, Corpus Based Grammar Research (= Proceedings of SLOVKO 2009, held 25-27.11.2009 in Smolenice, Slovakia), 2009, pp. 54–64.

Brunner, Annelen/Steyer, Kathrin: Wortverbindungsfelder: Fields of Multi-Word Expressions, in: **Granger, Silviane/ Paquot, Magali, editors:** eLexicography in the 21st century: New challenges, new applications. Proceedings of the eLex 2009. Louvaine-la-Neuve: Presses universitaires de Louvain, 2010, Cahiers du CENTAL, pp. 23–31.

Hanks, Patrick: Lexical Analysis: norms and exploitations, Cambridge [u.a.]: MIT Press, 2013.

Institut für Deutsche Sprache: Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache (DeReKo 2012-II), Webseite, 2012 ⟨URL: `http://www.ids-mannheim.de/DeReKo`⟩ – visited on 28.01.2014.

Perkuhn, Rainer/Belica, Cyril: Eine kurze Einführung in die Kookkurrenzanalyse und syntagmatische Muster. Institut für Deutsche Sprache, Mannheim, 2004 ⟨URL: `http://www1.ids-mannheim.de/kl/misc/tutorial.html`⟩ – visited on 28.01.2014.

Sinclair, John: Corpus, Concordance, Collocation, Oxford: Oxford University Press, 1991.

Stefanowitsch, Anatol/Gries, Stephan Th.: Collostructions: Investigating the interaction of words and constructions, in: International Journal of Corpus Linguistics, 8 2003, Nr. 2, pp. 209–243.

Steyer, Kathrin: Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten, in: Deutsche Sprache, 28 2000, Nr. 2, pp. 101–125.

Steyer, Kathrin: Kookkurenz. Korpusmethodik, linguistisches Modell, lexikographische Persepektiven, in: **Steyer, Kathrin, editor:** Wortverbindungen - mehr oder weniger fest, Berlin/New York: de Gruyter, 2004, Jahrbuch des Instituts für Deutsche Sprache, pp. 87–116.

Steyer, Kathrin: Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht, Tübingen: Narr, 2013.

Tognini-Bonelli, Elena: Corpus Linguistics at Work, Amsterdam/Philadelphia: J. Benjamins, 2001.