POSTPRINT

Multi-word patterns and networks

How corpus-driven approaches have changed our description of language use

Kathrin Steyer

Institut für Deutsche Sprache

This paper discusses a theoretical and empirical approach to language fixedness that we have developed at the Institut für Deutsche Sprache (IDS) ('Institute for German Language') in Mannheim in the project Usuelle Worterbindungen (UWV) over the last decade. The analysis described is based on the Deutsches Referenzkorpus ('German Reference Corpus'; DeReKo) which is located at the IDS. The corpus analysis tool used for accessing the corpus data is COSMAS II (CII) and – for statistical analysis – the IDS collocation analysis tool (Belica, 1995; CA). For detecting lexical patterns and describing their semantic and pragmatic nature we use the tool lexpan (or 'Lexical Pattern Analyzer') that was developed in our project. We discuss a new corpus-driven pattern dictionary that is relevant not only to the field of phraseology, but also to usage-based linguistics and lexicography as a whole.

Keywords: pattern-based lexicography, German reference corpus, phraseology, language fixedness, multiword expressions

1. Introduction

This paper discusses a theoretical and empirical approach to language fixedness that we have developed at the *Institut für Deutsche Sprache* (IDS) ('Institute for German Language') in Mannheim in my project *Usuelle Worterbindungen* (UWV) over the last decade.[1] Our research has always had two main areas on focus: (i) corpus linguistic exploration of phraseological phenomena, and (ii) new forms of online

---

lexicographic representation. From the beginning our work has been driven by the core question, "How can we interpret the results of quantitative analyses in a qualitative way?". Following in the comprehensive work of John Sinclair (cf. Herbst et al., 2011, Granger and Meunier, 2008), of Elena Tognini-Bonelli (2001) and many other scholars of corpus linguistics, we also distinguished between 'corpus-based' and 'corpus-driven' approaches and considered ourselves "corpus-driven human beings".[2] Drawing such strict lines is undoubtedly important in the initial development of a new research field. But nowadays we know that the distinction corpus-based vs corpus-driven seems to focus solely on the degree by which we allow ourselves to be led by the data – and the borders are fuzzy. We use the term 'corpus-driven' in a broader sense: a bottom-up corpus linguistic approach that allows us to find typical patterns by collecting many similar cases of usage. Looking at many cases of usage does not mean describing what is already known and visible, it means discovering hidden structures. Not only do we find more data, but we also detect new interrelations, unusual cross-connections, and surprising relationships and networks.

The following analysis is based on the *Deutsches Referenzkorpus* ('German Reference Corpus'; DeReKo)[3] which is located at the IDS. The corpus analysis tool used for accessing the corpus data is *COSMAS II* (CII) and – for statistical analysis – the IDS collocation analysis tool (Belica, 1995; CA). For detecting lexical patterns and describing their semantic and pragmatic nature we use the tool *lexpan* (or 'Lexical Pattern Analyzer') that was developed in our project (see Section 4.3).

Despite our focus on German resources, the principles of qualitative interpretation that we discuss should be transferable to other corpora and languages and other tools like the collocation analysis function in Sketch Engine (SkE).

2.   The rocky road of qualitative interpretation

We imagined that with corpora everything would be better, faster, larger and – most importantly – more accurate in regard to understanding and describing language use. We thought that we could compile new corpus-based dictionaries and lexical information systems that would capture all linguistic aspects. Collecting and restructuring mass data with machine-aided methods in the last 30 years allowed linguists to discard their old-fashioned slip boxes and to look for samples of real language use in new quantitative dimensions – an empirical revolution. The early

---

2.   This phrase was coined by Patrick Hanks in his plenary speech at the Malaga EUROPHRAS conference in 2015.

3.   For this paper, I used the largest DeReKo subcorpus of written language, the W archive with

euphoria was short-lived, however, as it soon became clear that the empirical road would be more tricky to navigate than initially expected. The borders between 'langue' and 'parole' become more and more fuzzy. If we look at the data, we find that syntactical phenomena are deeply motivated by pragmatics; we have to distinguish between ad hoc constructions and frequent usage, etc. Computational linguists gave us extensive databases and sophisticated tools. But the human researcher, the linguist and the lexicographer, cannot always keep pace in the face of mass data. If one gets, for instance, a search result list of 100,000 occurrences or more, one will need a navigation system that structures the results and shows the most typical clusters of usage. Another reason for this growing gap is the fact that the results of computer processes are often a black box for the human interpreter. To give an example: When interpreting results of statistical collocation analyses like the IDS CA we should not overestimate the ranking of a specific lexical item in a single collocation profile. Such profiles are always just snapshots at a particular time based on a specific corpus. A second analysis can elevate other lexical items up to the higher ranks while previous results are relegated. Nevertheless, the important point is that the underlying pattern is stable, despite the correlation of a specific lexical unit to another lexical unit.[4] The designation 'underlying pattern' means that collocations have to be interpreted first and foremost as correlations between groups or classes of units, not between specific words. Without knowing this fundamental principle a user could run into danger of misinterpreting the results or questioning the method as a whole.

Patterns are the key to why speakers understand each other in everyday communication in spite of ever-changing lexical material, syntactic variance and strong idiosyncrasies. The idea that language production and learning work by repetition, recurrence, and imitation on the basis of patterns, templates, and schemas is not new. It has been discussed at length, especially in cognitive linguistics and language acquisition research, as well as in research fields like pattern grammar (Hunston and Francis, 2000) and corpus pattern analysis (Hanks, 2013) – I won't go into the long history of terms like 'multiword units', 'fixed expression', 'formulaic language', 'schemes', 'patterns' and so on. For this I recommend the very readable paper by Hans-Jörg Schmid (2014). Of course, automatic pattern retrieval is not a new invention either, e.g. it is used extensively in data mining.

But regarding a *qualitative* reconstruction of hidden patterns in language use and their applications in lexicography and second language teaching, we are just at the beginning.

This chapter discusses this central pattern-based change of perspective in phraseology and beyond from the point of view of corpus linguistics.

---

4.   The only exception is a fixed correlation between specific words that means words cannot be

3. Kinds of lexical fixedness

3.1 From multiword expressions to patterns

Phraseology as a sub-discipline of lexicology has a long tradition.[5] The following central research question runs through the history of phraseology until today: Under which circumstances does a sequence of words become a *holistic unit*, a *lexical item*, a *lexicon entry*? The answers have been varied and depended to a significant degree on the predominant linguistic paradigms.

The rise of corpus linguistics fundamentally expanded the subject area of phraseology and the external perception of this discipline. It became more and more evident how essential multiword expressions (MWEs) are for understanding language use itself. In addition to this it became apparent that all – sometimes even competing – concepts of MWEs are based on the same fundamental principle of language, namely linguistic frozenness and fixedness. Compositional collocations and idioms, for example, differ in their degree of lexical fixedness and semantic opacity, their recognisability and prototypicality. Nevertheless, they all share one important characteristic: they are autonomous units that fill a specific role in communication. There is no core and no periphery. The difference is only in the degree of recognisability for the observer. These word clusters did not become fixed expressions by chance, but because speakers required an economical way to complete communicative tasks. Against this background we proposed the term 'usuelle Wortverbindungen' ('multiword expression in common use'; UWV)[6] in Steyer (2000). UWVs are conventionalised patterns of language use that manifest themselves in recurrent syntagmatic structures. This includes not only idioms and idiosyncratic structures, but all multiword units that have acquired a distinct function in communication. Our focus is on real-life usage, pragmatics and context. The central characteristic is the autonomous status as a communicative and entrenched cognitive unit[7] (see Section 3.2).

---

5.   Concerning phraseology as an independent discipline of linguistics I would like mention the two volumes of the *International Handbook of Phraseology* edited by Burger et al. (2007), the *Einführung in die Phraseologie* ('Introduction to Phraseology') edited by Burger (2015 in its 5th edition) and the two volumes of the De Gruyter International Bibliography edited by Mieder in 2009. For some years, a special journal for phraseology, the *Yearbook of Phraseology*, has been published by De Gruyter Mouton. A reference book for English phraseology is Moon (1998) and – of course – Sinclair (1991). I also mention Herbst et al. (2011), Granger and Meunier (2008) and Gries (2008). Information about the "phraseological community" can be found on the website of the European Society of Phraseology (EUROPHRAS) (www.europhras.org).

6.   English word-for-word translations of German are put in single quotation marks and in brackets.

7.   The term 'entrenchment' is one of the central concepts of cognitive grammar, first and fore-

In recent times, there has been a perceivable shift not only in phraseology and multiword research but also in usage-based linguistics as a whole (cf. Steyer, 2015). Due to the ability to detect invisible structures based on linguistic mass data, it has been shown that phrasemes, idioms, and frozen sentences like proverbs are not as singular and unique as was often assumed in phraseology in the past. Usually, they are specific lexical realisations of templates, more noticeable and more fixed than ad hoc formulations, but not unique. Such templates or patterns emerge from repeated usage and can be instantiated with ever-changing lexical elements, both phraseological and non-phraseological. In addition to corpus linguistics, construction grammar (CxG) also played a part in this paradigm shift in phraseology and it was against this backdrop that Dobrovol'skij introduced, for example, the term 'phraseme construction' in 2011.

Despite the fact that we share many commonalities with the CxG we prefer the term 'Wortverbindungsmuster' ('multiword patterns[8]; MWP) (cf. Steyer, 2013, 2016) as a subtype of the general term 'lexikalisch geprägte Muster 'lexical patterns'; LP) (Steyer 2018).[9] Our pattern concept focuses much more on structures and inter-relations of lexical items and wants to contribute to a usage-based theory of lexis. This approach arose from the tradition of phraseology as a genuine discipline of lexicology. That does not connote the negation of the syntax level. Naturally, our explorations are based on syntactic structures. But the dominance of the syntactic view can induce us to overlook the complexity of lexical phenomena. Probably this is a heuristic problem of analysis: One cannot observe all phenomena with the same intensity but has to fade some of them into the background (e.g. syntactical phenomena) for a much clearer observation of others (lexical structures and networks).[10]

Multiword patterns are conventionalised lexical schemes that are frozen by recurrent use. Recurrence is defined as the repeated appearance of similar linguistic structures in comparable contexts (cf. "geprägte komplexe Ausdrucksmuster" by Feilke, 1996, p. 187). Multiword patterns consist of fixed lexical components as well as obligatory slots that can be filled with specific entities (Renouf and Sinclair, 1991). These fillers have similar semantic and/or pragmatic characteristics, but do not necessarily belong to the same morpho-syntactic category. Sometimes all that they have in common are functional characteristics, which cannot be captured by traditional ontologies. Speakers are able to recall those schemes as lexicon entries

---

8. Biber also use ous the term 'multi-word pattern', but much more in the sense of multiword formulaic sequences like *it should be noted* or *as we have seen* (Biber, 2009).

9. Another type of lexical patterns is the so called 'sentence pattern', primarily proverb patterns like *There is more than one way to* VP (*skin a cat*) or *He who* V V (*He who want reap must sow*).

10. The discussion between phraseologists and construction grammarians to find out a common

and fill the gaps in a specific communicative situation in a functionally adequate way. The MWP concept focuses on the semantic and functional restrictions of the slot fillers much more than any other scheme or pattern theory.

3.2 MWPS as autonomous units

The main criterion for a multiword pattern is that it has a *holistic quality* that gives it a status as an autonomous unit. 'Holistic quality' does not necessarily mean that it is idiomatic. The MWP can just have a specific function, even in a very abstract sense. Example (1) illustrates which components are mandatory for the autonomous status:

(1) [*in* ADJ *Zeit*]
['in ADJ time']
ADJ fillers: *absehbarer* ('foreseeable') / *kurzer* ('short') / *nächster* ('next')

The adjective slot is mandatory for the holistic meaning, 'forthcoming'. In German, we cannot reduce this pattern to a binary MWE *in Zeit ('in time'). By contrast, the MWE *mit Genugtuung* ('with satisfaction') is a binary autonomous lexical item (P+N) with the meaning 'positively perceived'. Its recurrent internal adjective fillers, e.g. *großer* ('great') → *mit großer Genugtuung* ('with great satisfaction') only modify the core meaning as specific context markers (see Section 3.3).

Example (2) shows the functional nature of slot types and the distinct context restriction of MWPs:

(2) [*allen* N *zum Trotz*]
['despite all N']

The most frequent N filler group includes speech acts in a broader sense like the fillers *Vorhersagen* ('predictions'), *Prognosen* ('forecasts') or *Einwänden* ('objections') (pattern: [*allen Vorhersagen / Prognosen / Einwänden zum Trotz*] '*despite all predictions / forecasts / objections*'). Some of these have explicit negative connotations: *Unkenrufen* ('cries of naysayers'), *Getöses* ('hullabaloo') or *Horrormeldungen* ('horror stories'). Sometimes they are extended by adjectives, e.g. *anderslautenden Gerüchte* ('contrary rumours'), *düsteren Prognosen* ('dark predictions') or *vollmundigen Versprechungen* ('overblown promises'), e.g. *despite all overblown promises*. Another filler group includes references to speakers, like *Zweifler* ('sceptics'), *Kritiker* ('critics') or *Pessimisten* ('pessimists'), typically with some degree of negative connotation. Despite the variance of the fillers the MWP [*allen* N *zum Trotz*] has a holistic meaning: 'although something has been expected to go a certain way, it has turned out differently', from positive to negative and vice versa.

3.3   Extended context patterns (ECPS)

ECPs are recurrent – but not mandatory – context extensions. The context partners can appear inside an autonomous lexical unit as well as in its immediate surroundings (external ECP). Let's take a look at the example from the previous section again: [*mit* X *Genugtuung*] is an ECP of *mit Genugtuung* ('with satisfaction') with the following three frequent extension types (among others)[11]:

(3) *mit großer   Genugtuung*
         tiefer
         grimmiger
    ('with great / deep / grim satisfaction')

(4) *mit Stolz und   Genugtuung*
         Freude und
         Häme und
    ('with pride and / joy and / scorn and satisfaction')

(5) *mit einem Hauch von Genugtuung*
         einem Anflug von
         einem Schuss
         einer Prise
    ('with a hint of / a touch of / a shot of / a pinch of satisfaction')

The extensions in (3) and (4) have two functions: intensification and/or connotation. (5) illustrates a third group of internal extensions, so called syntagmatic connotative quantifiers. These extensions express the observation of a rather hidden emotional reaction from the speaker who perceives something positively.

A recurrent external ECP in postposition of the MWE *mit Genugtuung* is the combination with verbs that refer to communicative or cognitive acts embedded in a '*that*-clause', e.g. [*mit Genugtuung zur Kenntnis nehmen* ('take note') / *feststellen* ('see') / *registrieren* ('notice'), *dass* ('that')…].

All these functional restrictions cannot be predicted *a priori* and do not always follow rules. They can only be discovered by an inductive reconstruction based on large corpora and sophisticated automatic methods.

---

11. The English equivalent behaves very similarly and also has comparable recurrent ECPs [*with* X *satisfaction*].

4. Corpus-linguistic methodology and interpretation

We now explain our iterative corpus linguistic methodology using examples of qualitative interpretation. Our empirical approach includes several steps, each with a specific explanatory potential: (a) complex phrase searches and reciprocal analyses are using *COSMAS II*; (b) IDS CA; (c) slot-filler analysis with *lexpan*.

4.1 Corpus searches

We look at the data, and then we hypothesise about the fixedness, variance and function of the MWE or MWP. Based on these findings we then configure the search strategy and return to the corpus. This can be repeated for several cycles. Thus, we study the nature of MW patterns by exploring KWIC concordances of multiword units. As a result of our focus on syntagmatic word surface structures, our approach is guided by the following two principles: First, we made the decision that searching without grammatical annotations follows our firm conviction that many MWPs cannot be found based on tagged corpora. MWPs often don't conform to syntactical phrases like NPs or PPs and traditional word classes change their function in a MWP. MWPs are primarily *functional* lexical buildings blocks.[12]

Our second principle is that we always use the word form – not the lemma – in our searches. As Sinclair already emphasised:

> 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, p. 8)

We will explain the distinctive use of word forms in the next section.

4.2 Collocation profiles

The IDS CA can be used to detect significant word pairs and MWEs as well as recurrent syntagmatic context patterns. This method enables us to identify typical aspects of meaning and usage of a MWE or a MWP – the extension of the principle of contextualism to multiword phenomena.

Figure 1 shows small snippets of collocation profiles of the word forms *Grund* (1,646,568 hits), *Grunde* (198,390 hits) and *Gründen* (559,751 hits):

---

12. Therefore, the untagged DeReKo corpus is our preferable resource.

Kookkurrenzpartner	LLR	Frequenz	Syntag pattern	Kommentar	Kwics
warum	1569	325	99% ein |der Grund […] warum	why	Kwics
gutem	901	137	100% aus |Aus mit gutem […] Grund	good	Kwics
genug	809	213	97% Grund […] genug für …	enough	Kwics
Feiern	714	107	99% Grund zum Feiern	celebrating	Kwics
Freude Doppelten	516	4	100% Doppelten Grund zur Freude	joy double	Kwics
Freude	516	133	96% Grund zur Freude	joy	Kwics
Boden	481	139	98% in Grund [und] Boden	bottom	Kwics
weshalb	422	89	98% ein |der Grund […] weshalb die …	wherefore	Kwics
Jubeln	315	36	100% Grund zum Jubeln	cheering	Kwics
Kookkurrenzpartner	**LLR**	**Frequenz**	**Syntag pattern**	**Kommentar**	**Kwics**
genommen	8163	1142	100% im |Im Grunde […] genommen	taken	Kwics
ist egal	683	45	51% ist mir |es im Grunde […] egal	doesn't matter	Kwics
ist	683	1406	54% im |Im Grunde […] ist	is	Kwics
Herzens	550	54	92] im |Im Grunde ihres |seines Herzens	of the heart	Kwics
gelegt	392	94	100% zu Grunde […] gelegt	laid	Kwics
ja anderes	377	4	50% ja im Grunde … anderes	indeed different	Kwics
ja	377	136	60% Im |im Grunde […] ja	indeed	Kwics
Kookkurrenzpartner	**LLR**	**Frequenz**	**Syntag pattern**	**Kommentar**	**Kwics**
gesundheitlichen	6697	611	99% aus gesundheitlichen […] Gründen	health	Kwics
beruflichen	4069	445	100% aus beruflichen […] Gründen	professional	Kwics
finanziellen	3130	430	99% aus finanziellen […] Gründen	financial	Kwics
ungeklärten bislang	1142	16	100% Aus |aus bislang […] ungeklärten Gründen	as yet unkown	Kwics
unerfindlichen	1007	70	100% aus unerfindlichen Gründen	mysterious	Kwics

Figure 1. CA profiles *Grund – Grunde – Gründen* (snippets) (CII random sample of 10,000)[13]

As we can see, there are significant differences in the highest LLR-ranks of the CA profiles: The common singular form *Grund* has a wide range of syntagmatic patterns, e.g. *Grund, warum* ('reason why'), *aus gutem Grund* ('with good reason'), *Grund genug* ('reason enough'), [*Grund zum|zur* N (*Feiern / Freude / Jubeln*)] ('cause for N (celebrating / joy / cheering)'). There are also fixed MWEs like the intensifier *in Grund und Boden* ('in ground and bottom') which means 'totally'.

In contrast to this, the collocation profile of the – much less common – singular form *Grunde* is strongly focused on realisations of the preposition noun combination *im* plus *Grunde*: (i) as adverbial MWEs *im Grunde* and *im Grunde genommen*. Both MWEs can be translated as 'basically', (ii) as a MWP with the function of intensification [*im Grunde* PRON *Herzens*] ('at the bottom of PRON heart'). These three entities can also take the meaning 'eigentlich' ('actually').[14] We can see significant context partners like *ist egal* -> *im Grunde ist es egal* ('doesn't matter -> actually it doesn't matter') or *nichts anderes als* ('nothing other than -> actually it is nothing other than').

The third snippet shows the CA profile of the plural form *Gründen*. It is again very different. The context partners also indicate a strong restriction. Adjective collocation partners like *gesundheitlichen* ('health'), *beruflichen* ('professional') or *finanziellen* ('financial') are very dominant and they all come from realisations of the recurrent syntagmatic pattern [*aus* ADJ *Gründen*] ('for ADJ reason') (see Section 4.2).[15] To verify this observation we configure our search query for *Gründen* to yield only those occurrences without the preposition *aus* (in a range of up to five words before *Gründen*). Another useful method is to compare automatically the CA profiles of a MWE with the profile of a single lexeme with a similar meaning and function (see Figure 2). Using this strategy we can figure out the overlapping contexts and the differences in usage between these lexical units. Figure 2 illustrates a comparison of the CA profiles of the MWE *im Grunde* and the single lexeme *eigentlich* which can be translated as 'actually, basically'.

When comparing the CA profile of *eigentlich* with the CA profile of *im Grunde* we can see that although many clusters are similar, some contexts are strongly preferred by *eigentlich* and not highly ranked for *im Grunde*.

For example, there are adjectives in upper case which commonly appear before *eigentlich*, e.g. *Schade* ('Pity') (*Schade eigentlich* 'Pity actually'), *Komisch* ('Odd') (*Komisch eigentlich* 'Odd actually'), *Merkwürdig* ('Strange, Curious') (*Merkwürdig*

---

14. Another significant partner is the form *gelegt* ('laid'). This indicates an inflected form of the functional verb *zu Grunde legen* ('to take as a basis').

15. Most of these adjective collocation partners do not appear in the profiles for the singular forms.

Kookkurrenzpartner	LLR	Frequenz	Syntag pattern	Kommentar	Kwics
genommen	11287	1473	99% im Grunde genommen	taken	Kwics
ist	1189	1974	72% ist […] im Grunde …	is	Kwics
Herzens	719	88	100% im Grunde seines |ihres Herzens	of the heart	Kwics
gar nicht	479	128	89% im Grunde gar […] nicht	not at all	Kwics
ja	456	257	59% ist ja […] im Grunde …	indeed	Kwics
anderes	411	122	96% ist im Grunde nichts anderes als	different	Kwics
egal	280	73	97% ist im Grunde … egal	doesn't matter	Kwics

Kookkurrenzpartner	LLR	Frequenz	Syntag pattern	Kommentar	Kwics
gar nicht	789	212	93% eigentlich […] gar […] nicht	not al all	Kwics
sollte	670	303	67% sollte […] eigentlich	should	Kwics
Warum	543	156	100% Warum […] eigentlich nicht	Why	Kwics
ja	487	265	80% ja […] eigentlich	is indeed	Kwics
wollte	457	200	65% wollte […] eigentlich	would	Kwics
ist egal	395	26	65% ist [mir] eigentlich […] egal	never mind	Kwics
Schade	146	33	96% Schade […] eigentlich	Pity	Kwics
Wieso	75	21	100% Wieso […] eigentlich	Why	Kwics

Figure 2. CA profiles of the MWE *im Grunde* and the single lexeme *eigentlich* (snippets) (CII random sample of 10,000)

eigentlich 'Strange|Curious actually'). The corpus data show that the adverb *eigentlich* cannot be substituted by the MWE *Im Grunde* in those patterns: **Schade im Grunde*; **Komisch im Grunde*; **Merkwürdig im Grunde*. The capitalisation of the adjectives in the *eigentlich* profile indicates that this pattern appears at the beginning of a sentence. The KWIC (see Example (6)) and a selected citation (see Example (7)) show that these ECPs are elliptical constructions with the pragmatic function of an anaphoric comment.

(6) *P14 aber eine Schreibmaschine gibt es nicht mehr. **Schade, eigentlich**.*
    'but there isn't a typewriter anymore. ***Pity actually***.'

(7) *Die Redaktion beherbergt geschätzte 150 Computer, dazu Großdrucker, Scanner und sogar noch Fax-Geräte, aber eine Schreibmaschine gibt es nicht mehr. **Schade, eigentlich**.*
Die Presse, 27.07.2014, S. 36,37; Erika, Olympia und Remington: *Meine Schreibmaschine und ich*:
'The desk hosts approximately 150 computers, plus larger printers, scanners and even fax machines, but there isn't a typewriter anymore. ***Pity actually***.'

This example is a good argument for an analysis without lemmatising the collocation partners. We wouldn't have detected this special use of elliptical constructions in the CA profile if the partner word forms were conflated to the lemma *schade*.

Furthermore, the qualitative comparison of the corpus citations of *im Grunde* versus *eigentlich* shows that speakers use the MWE *im Grunde* (*genommen*) much more indirectly and give the communicative partner the chance to agree or disagree. In contrast, the adverb *eigentlich* directly expresses a truth claim. So, even if a quasi-synonymous single lexeme exists, the MWE shows differences in usage which become apparent when studying large quantities of data.

4.3 KWIC bundles and slot-filler analysis

Our central goal is the corpus-driven detection of fixedness and variance to learn about the nature of multiword and extended context patterns. For this purpose, we have developed the language-independent pattern matching tool *lexpan*[16] that can be used for bundling large quantities of KWICs by search patterns based on qualitative hypotheses or on a specific research question. In addition, it supports the exploration of lexical patterns with variable slots and the qualitative annotation of CA profiles and the fillers of pattern slots.

Since the beginning of our project this tool has not only been developed to support the semi-automatic detection of patterns – as a heuristic analysis instrument – but also to serve as a working environment for the corpus-driven lexicographer. So, it is possible to export and visualise the results (systemised as KWIC bundles, as qualitative annotated CA profiles and filler tables) and to use them as part of a new form of lexicographic representation of MWEs and MWPs (see Section 5).

---

16. We offer *lexpan* to all interested users; the program is available for download on the *lexpan* website (lexpan).
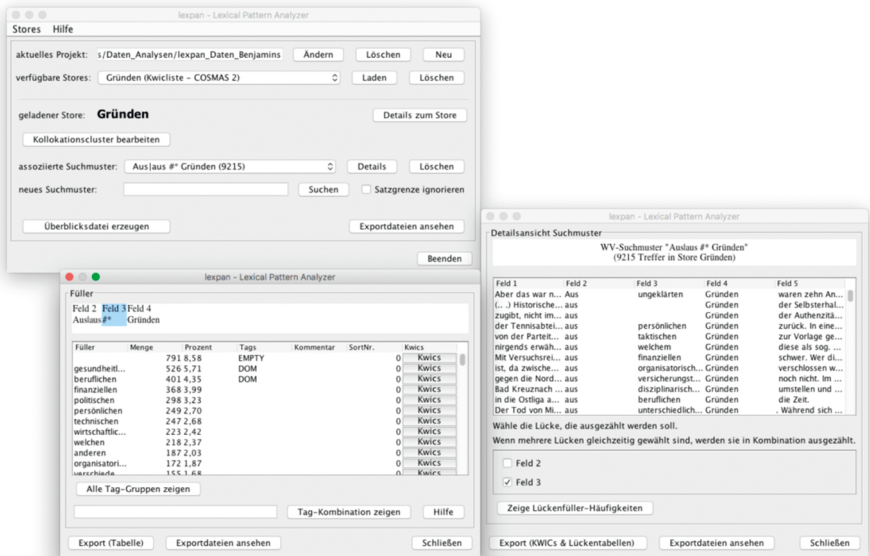
Figure 3. *lexpan* – user interface

Let's take a closer look at how to work with *lexpan*. With its help one can formulate search patterns to capture specific surface characteristics and to bundle the KWIC lines accordingly, e.g. *Gründen*:

Figure 4. Example for KWIC bundles for *Gründen* (*lexpan* snippet)
(CII random sample of 10,000)

In the case of the *Gründen* KWICs, only one search pattern covers 92.15% of all occurrences, namely the preposition *aus* plus an arbitrary number of unspecified words plus *Gründen* (aus|Aus #* Gründen) (see KWIC snippet in Figure 4). The other three search patterns capture the MWP [*mit* X *Gründen*] ('with/for X reasons') and the MWEs *ohne Angabe von Gründen* ('without giving reasons') (see KWIC snippet in Figure 4) and *eine Reihe von Gründen* ('a number of reasons'). The remaining KWICs (5.21%) feature occurrences that are not bound to *Gründen* patterns, e.g. *zu den Gründen äußerte er sich nicht* ('he didn't comment on the reasons') (verb pattern: *sich zu* etw. *äußern* 'to comment on sth').

Once the KWIC lines have been captured by a search pattern, *lexpan* can count the slot fillers and present them in filler tables.[17] Such tables show the absolute and relative frequencies of the lexical fillers for one or more slots. Furthermore, *lexpan* allows us to incorporate qualitative annotations (tags and comments) for grouping the fillers according to semantic and pragmatic characteristics, e.g. in this table for the pattern *aus* X *Gründen* ('for X reasons') we use the column "Kommentar" ('comment') for English translations of the fillers:

Lückenfüller	Anzahl	Prozentanteil	Tag	Kommentar	Kwics
	791	8,58	[EMPTY]		Kwics
gesundheitlichen	526	5,71	[DOM]	health	Kwics
beruflichen	401	4,35	[DOM]	professional	Kwics
finanziellen	368	3,99	[DOM]	financial	Kwics
politischen	298	3,23	[DOM]	political	Kwics
persönlichen	249	2,70	[DOM]	private	Kwics
guten	114	1,24	[VAL]	good	Kwics
taktischen	90	0,98	[DOM]	tactical	Kwics
unerfindlichen	49	0,53	[VAL]	mysterious	Kwics
bisher unbekannten	14	0,15	[VAL]	as yet unknown	Kwics

Figure 5. Filler table of *aus|Aus* 0 … n slots *Gründen* (*lexpan* snippet) (CII random sample of 10,000)

The highest ranked slot is annotated as a zero-gap which indicates cases where no lexical item appears between the fixed elements of the pattern: in Figure 5 this zero-gap stands for the lexical core *aus Gründen* ('for reasons'), a MWE with the meaning 'because of' with typical NP extensions like *der Verkehrssicherheit / des*

---

17. Methodologically, our slot-filler approach bears some similarities to Stefanowitsch and Gries' 'collostructions' (Stefanowitsch, 2013). As mentioned in Section 3.1., we feel closer to the holistic perspective of phraseology.

Datenschutzes / ('for reasons of traffic safety / data protection'). The other fillers can be grouped into functional groups, like adjectives referring to a specific domain [DOM], e.g. *politisch* ('political'), *persönlich* ('personal'), *finanziell* ('financial'); *technischen* ('technical') or to the validity of reasons or motifs [VAL] with an evaluative potential of the adjectives, e.g. *nachvollziehbar* ('understandable'), *gut* ('good') or *triftig* ('valid').

Particularly interesting is the first group of adjectives referring to domains like politics, privacy, finance and many others. The ADJ realisations seem to be much more neutral than the other evaluative ADJ fillers like *good* or *understandable*. Speakers say, "The reason is X". But in these cases we can also observe a specific pragmatic aspect: When somebody steps back *aus persönlichen Gründen* ('for personal reasons') the reason is really quite unspecified, we are left to speculate whether there were family problems or possibly internal pressure to step down.

The speakers use the plural form of the pattern to express a certain degree of vagueness and avoid a direct interpretation or evaluation.

It is also very interesting to examine those fillers that occur only once; these are often compound words.

(8) *strömungspolitischen* ('trend-**political**')

(9) *interviewtaktischen* ('interview-**tactical**')

(10) *wettertechnischen* ('weather-**technical**')

(11) *nebulösen* ('nebulous') / *schikanösen* ('vexatious')

Traditionally, as linguists, we would interpret such hapax legomena as ad hoc realisations of the MWPs, as phenomena that can be disregarded for the lexicographic description. But when exploring the filler tables it becomes clear that they as well contribute to forming the emergent pattern and follow the same predispositions: The examples in (8), (9) and (10) are compounds which use words from the highest ranks as their basis. The examples in (11) are evaluative adjectives which are also a dominant filler group in the higher ranks.

lexpan also enables to detect n-grams.[18] This is particularly useful for describing context patterns like typical chunks in the environment of the MWE *aus Gründen* ('for reasons'):

---

18. For the examples in Figure 6 and 7 I extended the random sample to 100,000.

Lückenfüller	Anzahl	Prozentanteil	Kommentar	Kwics
nicht … nur	66	0,87	not only	Kwics
vor … allem	53	0,70	first and foremost	Kwics
und … zwar	36	0,47	in fact	Kwics
aber … auch	23	0,30	but also	Kwics
nicht … zuletzt	22	0,29	last but not least	Kwics

Figure 6. Filler table of ECP two slots + *aus Gründen* (*lexpan* snippet) (CII random sample of 100,000)

Frequent chunks (see Figure 6) like *nicht nur* ('not only'), *vor allem* ('first and foremost'), *und zwar* ('in fact') or *nicht zuletzt* ('last but not least') can be classified as functional chunks that indicate a specific argumentation frame used by the speakers.

In other cases, the slot-filler analysis reveals a high degree of restriction, e.g. for the internal extension of the syntagma *mit … Grund* ('with … reason'):

Lückenfüller	Anzahl	Prozentanteil	Kommentar	Kwics
ein	10912	49,04	a	Kwics
gutem	7365	33,10	good	Kwics
	1856	8,34	'zero gap'	Kwics
dem	317	1,42	the	Kwics
einigem	118	0,53	some	Kwics
festem	4	0,02	concrete	Kwics
goldenem	4	0,02	golden	Kwics

Figure 7. Filler table *mit* 0–1 slots *Grund* (*lexpan* snippet) (CII random sample of 100,000)

In Figure 7, we can observe that the first two fillers *ein* ('a') and *gutem* ('good') have a disproportionately high frequency (prototypical fillers). The third rank is taken by a zero gap. The large number of remaining fillers are distributed with low frequencies. The zero gap indicates the MWE *mit Grund* ('with reason') with the meaning 'legitimately'. The frequent adjective filler *gutem* ('good') points to the extended pattern of this MWE and has the function of intensification. Due to its frequency, the ECP *mit gutem Grund* can also be considered an autonomous MWE that also means 'legitimately'. The other frequent filler, the indefinite article *ein* ('a'), constitutes the interesting syntagma *mit ein Grund* ('with **a** reason'); *with* in this case means 'among others' so the pattern can be paraphrased as 'a reason among others'. At first glance it seems that the components of this syntagma are the word

form *mit* ('with') and the NP *ein Grund* ('a reason'). But the NP slot is filled in a very systematic manner with nouns that refer to causative concepts. Beside *Grund* there are e.g. *Ursache* ('cause') -> **mit eine Ursache** ('a cause among others') or *Motiv* ('motive') -> **mit ein Motiv** ('a motive among others'). From the pattern perspective, the fixed lexical element is *mit ein/eine* ('with a'). That seems to be non-grammatical (the regular grammatical unit is the NP (indefinite article + noun)), but *mit ein/eine* ('with a') can be regarded as a functional chunk as a component of the LP [*mit ein/eine* NOUN].

5. A new type of corpus-driven, pattern-based MW dictionaries

Parallel to the development of the MWP model and the corpus-driven methodology, we have been investigating how these results could lead to new forms of description of common language use, e.g. for a change of perspective in foreign language acquisition and teaching (Steyer, 2009; Steyer and Brunner, 2014). The fundamental insight that only a small number of expressions are saved as isolated entries in the mental lexicon whereas the majority bases on lexical patterns as part of a network calls for new forms of lexicographical representation. A learner needs information about the status of a lexical unit: Is it a strongly fixed MWE? If so, then the item must be memorised like a word. Or: Is this MWE only a typical realisation of a pattern? If so, then it is necessary to understand the semantic or functional principle underlying this pattern and to learn which realisations are typical in common use.

Another innovation related to developments in language technology is that mass data not only can used as a basis for empirical analyses but also as a form of lexicographic information.[19]

A prototype is our corpus-driven, pattern-based pilot study "Wortverbindungsfelder" ('Multiword fields'), especially Version 3, "Grund" (http://wvonline. ids-mannheim.de/wvfelder-v3/index.html) that contains the following four data types:

---

19. Pioneering work in this area is also done by Hanks and his colleagues in their Pattern Dictionary of English Verbs (PDEV). This online dictionary presents usage patterns of English verbs with their KWIC lines, systematised according to semantic properties. Thus, users can understand and learn these patterns not on the basis of idealised example sentences but in their authentic usage contexts.

a. automatically compiled corpus data (KWICs, CA profiles, filler tables)
b. semi-automatically compiled corpus data that illustrates specific aspects of us-age; this is original corpus data that has been selected and/or annotated man-ually, including
   – qualitatively selected KWIC bundles and citations
   – qualitatively annotated and systematised collocation fields
   – qualitatively annotated and systematised filler groups
c. narrative lexicographic comments
d. hierarchies of MWE and MWP nodes on several levels of abstraction which contain several types of cross connections.

The structures of MWEs and MW patterns, the distinctive characteristics of their usage, their interrelations, and the regularities of the underlying patterns are pre-sented primarily on the basis of the arrangement and hierarchical ordering of KWIC lines. In this way one can learn about language use by studying numerous real-word usage examples with related characteristics. Filler tables like the ones cited in this article, collocation profiles and other data contribute to the understanding of the MWE and its place in networks and hierarchies.

We want to explain the main ideas of this concept using the example of the node [*aus Gründen* SUB-G] ('for reasons NP') (as mentioned before) and the MWE node *aus welchen Gründen auch immer* ('for whatever reasons'). Figure 8 shows a snippet of the pattern node [*aus Gründen* SUB-G] ('for reasons NP'):

Figure 8. "Multiword fields" online: Node *aus Gründen* SUB-G ('for reasons NP') (snippet)

This node contains narrative comments, e.g. on the abstract meaning of the pattern *aus Gründen* SUB-G ('for reasons NP') (see "Allgemeine Beschreibung") and on pragmatic features (see "Kommentar"). Furthermore, there are links to the CA

profile of this pattern (see snippet in Figure 9) and the filler table of the NP slot. This table is complemented by a linguistic comment about pattern restriction based on the word building phenomena of the noun fillers.

In addition to this, the section "Kontrastanalyse" ('contrast analysis') provides information on a quasi-equivalent of the chunk *aus Gründen* ('for reasons), the prepositional single lexeme *wegen* ('because of') (pattern: [*wegen* NP] ['because of NP']) by showing its CA profile. Our online platform allows the user to compare those two profiles (Figure 9).
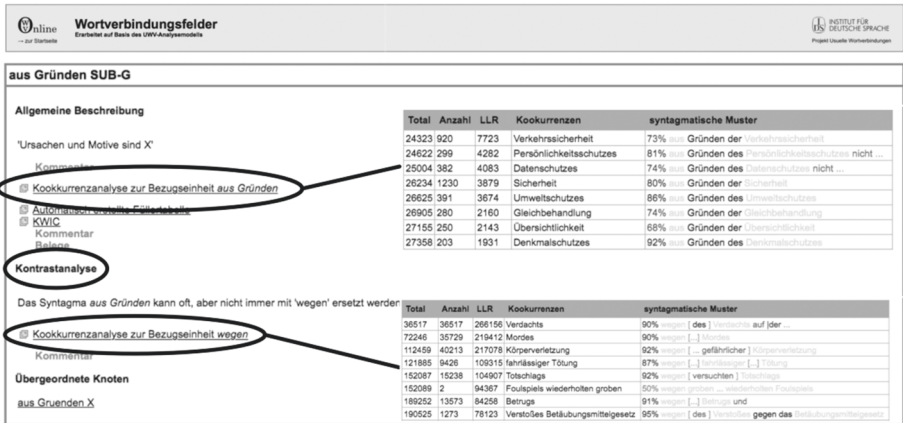
Wortverbindungsfelder
Erarbeitet auf Basis des UWV-Analysemodells

INSTITUT FÜR DEUTSCHE SPRACHE
Projekt Usuelle Wortverbindungen

aus Gründen SUB-G

Allgemeine Beschreibung

'Ursachen und Motive sind X'

Kookkurrenzanalyse zur Bezugseinheit *aus Gründen*

Automatisch erstellte Fillertabelle

KWIC

Kommentar

Belege

Total	Anzahl	LLR	Kookkurrenzen	syntagmatische Muster
24323	920	7723	Verkehrssicherheit	73% aus Gründen der Verkehrsicherheit
24622	299	4282	Persönlichkeitsschutzes	81% aus Gründen des Persönlichkeitsschutzes nicht ...
25004	382	4083	Datenschutzes	74% aus Gründen des Datenschutzes nicht ...
26234	1230	3879	Sicherheit	80% aus Gründen der Sicherheit
26625	391	3674	Umweltschutzes	86% aus Gründen des Umweltschutzes
26905	280	2160	Gleichbehandlung	74% aus Gründen der Gleichbehandlung
27155	250	2143	Übersichtlichkeit	68% aus Gründen der Übersichtlichkeit
27358	203	1931	Denkmalschutzes	92% aus Gründen des Denkmalschutzes

Kontrastanalyse

Das Syntagma *aus Gründen* kann oft, aber nicht immer mit 'wegen' ersetzt werden

Kookkurrenzanalyse zur Bezugseinheit *wegen*

Kommentar

Total	Anzahl	LLR	Kookkurrenzen	syntagmatische Muster
36517	36517	266156	Verdachts	90% wegen [des] Verdachts auf |der ...
72246	35729	219412	Mordes	90% wegen [...] Mordes
112459	40213	217078	Körperverletzung	92% wegen [... gefährlicher] Körperverletzung
121885	9426	109315	fahrlässiger Tötung	87% wegen [... fahrlässiger] [...] Tötung
152087	15238	104907	Totschlags	92% wegen [versuchten] Totschlags
152089	2	94367	Foulspiels wiederholten groben	50% wegen groben ... wiederholten Foulspiels
189252	13573	84258	Betrugs	91% wegen [...] Betrugs und
190525	1273	78123	Verstoßes Betäubungsmittelgesetz	95% wegen [des] Verstoßes gegen das Betäubungsmittelgesetz

Übergeordnete Knoten

aus Gruenden X

Figure 9. "Multiword fields" online: Contrasting CA profiles of *aus Gründen* und *wegen* (snippets)

In this way, the user gets an idea of the significant usage differences between *wegen* and *aus Gründen*.

The MWE node *aus welchen Gründen auch immer* 'for whatever reasons' is interesting because several levels of abstraction are at play: the word form *Gründen* can be substituted with different nouns in the MWP *aus welch-* SUB-G *auch immer*, e.g. *Motiv* ('motive') or *Richtung* ('direction'). But the lexical components *auch immer* are part of another, more abstract pattern as well: [PRON *auch immer*]. In the MWP, the place of PRON is taken by interrogative pronouns e.g. *was* ('what'), *wo* ('where'), *wer* ('who'), and *warum* ('why').

One of those pronoun fillers, *wie* ('how'), is much more frequent than the others, and this is why there is a separate MWE node *wie auch immer*. This MWE can be translated as 'howsoever' or 'to whatever extent'. The complicated abstractions and relations are visualised by the MWE network (Figure 10). By means of this network the user can understand the progressing abstraction from the lexical surface to the pattern meaning ('sth is not known or not comprehensible') and the
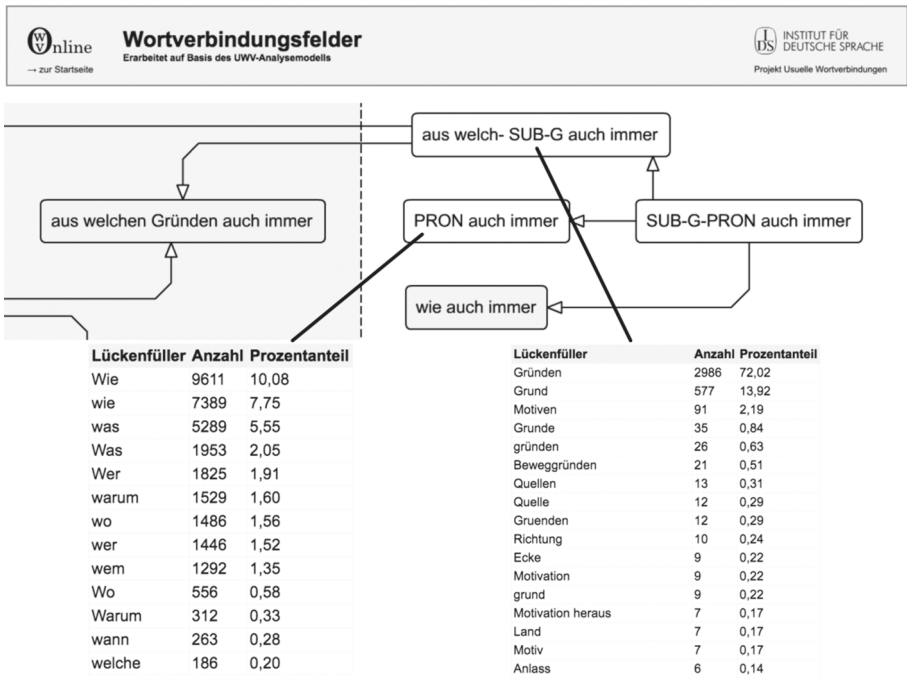
Figure 10. "Multiword fields" online: Pattern hierarchy and filler tables (snippets)

communicative function that is stable on every level (Speakers express a certain degree of scepticism, doubt or criticism).

Currently, the central ideas and components of this lexicographical concept are implemented in the cooperation project *Preposition-Noun-Combinations in Context* (PREPCON).[20]

PREPCON explores preposition-noun-combinations (PNCs) with a recurrent zero gap as autonomous lexical items (e.g. *durch Zufall* ('by chance') or *über Jahre* ('over years')). PREPCON has three different forms of data representation: (i) a fully automated database of 80,000 German PNCs (autonomous and as part of other constructions) with KWICs and frequencies; (ii) a semi-automatic database of temporal PNCs including a collection of 150 autonomous temporal MWEs that also fulfil modal or discourse marker functions; (iii) a contrastive pilot study based on the UWV methodology for a trilingual contrastive description of lexical patterns.

---

20. The cooperation partners are: IDS UWV group (head: Kathrin Steyer, Mannheim), FRASESPAL (head: Carmen Mellado Blanco, Santiago de Compostela/Spain), WICOL (head: Peter Ďurčo, Trnava/Slovakia).

6. Conclusion

This paper discussed new pattern perspectives that are relevant not only to the field of phraseology, but also to usage-based linguistics and lexicography as a whole. We argued that large corpora and sophisticated automatic methods can only be fruitful for these fields if they are used in combination with qualitative interpretation. Qualitative interpretation requires deep reflection on the explanatory potential of each corpus linguistic method and also on its limitations. The UWV methodology described in this paper is an example of an integrative approach that is focused on discovering hidden structures of language use, especially frozen lexical building blocks and patterns and their functions in communication. Beside the contextualistic approach to multiword units based on collocation profiles, this methodology is focused on the detection and interpretation of patterns that have holistic meanings or functions. Using the UWV tool *lexpan* we illustrated several characteristics of frozenness: fixed lexical multiword expressions, multiword patterns with fixed lexical components and mandatory slots and extended multiword expressions with facultative, but also recurrent contextual extensions of the core. All these types have the potential for cognitive entrenchment. We showed the complex nature of filler tables, from the disproportional frequency of a few lexical units (indicating the prototypical realisation of the pattern) to a continuous range of frequencies. In the first case we can assume more than one entry in the mental lexicon: as an MWE and as a realisation of a pattern. Another interesting observation is that even fillers that occur only once (indicating occasional ad hoc realisations) follow the same predispositions that are apparent for the highest ranks. One of the main results of this research is the insight that there is logic and purpose behind the seemingly endless linguistic creativity. Nothing is arbitrary in language production, but it is not only grammar that guides language production and reception. The genesis of patterns and their functional restrictions cannot be deduced *a priori* based on rules but only *a posteriori* based on an inductive quantitative and qualitative reconstruction.

With regard to a new pattern-based lexicography several important questions emerge: How can we visualise these relationships? Which kind of representation is appropriate for which user group, for example for foreign language learners? The major challenge will therefore be to adapt the concept of the corpus-driven pattern dictionary didactically. We have to find an answer to the very complex question of how to guide a user through the universe of corpus data and networks.

References

Belica, C. (1995). *Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analyse-methode*. Mannheim: Institut für Deutsche Sprache.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3), 275–311. https://doi.org/10.1075/ijcl.14.3.08bib

Burger, H. (2015). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. 5., neu bearb. Aufl. Berlin: Schmidt.

Burger, H., Dobrovol'skij, D., Kühn, P., & Norrick, N. R. (Eds.) (2007). *Phraseology. An International Handbook of Contemporary Research*. 2 vols. (HSK 28, 1/2). Berlin/New York: de Gruyter.

Dobrovol'skij, D. (2011). Phraseologie und Konstruktionsgrammatik. In A. Lasch, & A. Ziem (Eds.), *Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze* (pp. 111–130). Tübingen: Stauffenburg.

Feilke, H. (1996). *Sprache als soziale Gestalt: Ausdruck, Prägung und die Ordnung der sprachlichen Typik*. Frankfurt am Main: Suhrkamp.

Granger, S., & Meunier, F. (Eds.) (2008). *Phraseology. An interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/z.139

Gries, S. Th. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger, & F. Meunier (Eds.), *Phraseology and linguistic theory: A brief survey* (pp. 3–25).

Hanks, P. (2013). *Lexical Analysis. Norms and Exploitations*. Cambridge, MA/London: MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001

Herbst, T., Faulhaber, S., & Uhrig, P. (Eds.) (2011). *The Phraseological View of Language. A Tribute to John Sinclair*. Berlin/Boston: de Gruyter. https://doi.org/10.1515/9783110257014

Hunston, S., & Gill, F. (2000). *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.4

Langacker, R. W. (1987). *Foundations of Cognitive Grammar. Vol 1. Theoretical Prerequisites*. Stanford: Stanford University Press.

Mieder, W. (2009). *International Bibliography of Paremiology and Phraseology*. 2 vols. Berlin: de Gruyter. https://doi.org/10.1515/9783110215595

Moon, R. (1998). *Fixed Expressions and Idioms in English*. A Corpus-Based Approach. Oxford: Clarendon Press.

Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. In K. Aijmer, & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 128–143). London: Longman.

Schmid, H.-J. (2014). Lexico-grammatical patterns, pragmatic associations and discourse frequency. In T. Herbst, H.-J. Schmid, & S. Faulhaber (Eds.), *Constructions, Collocations, Patterns* (pp. 239–293). Berlin/Boston: Mouton de Gruyter.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stefanowitsch, A. (2013). In G. Trousdale, & T. Hoffmann (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 290–306). Oxford/New York: Oxford University Press.

Steyer, K. (2018). Lexikalisch geprägte Muster. Modell, Methoden und Formen der Onlinepräsentation. In K. Steyer (Ed.), *Sprachliche Verfestigung. Chunks, Muster, Phrasem-Konstruktionen* (pp. 227–264). Tübingen: Narr.

Steyer, K. (2016). Corpus-driven Description of Multi-word Patterns. In G. Corpas Pastor, J. Monti, V. Seretan, & R. Mitkov (Eds.), *Workshop Proceedings Multi-Word Units in Machine Translation and Translation Technologies (MUMTTT2015)* (pp. 13–18). Geneva: Editions Tradulex.

Steyer, K. (2015). Patterns. Phraseology in a state of flux. In D. Dobrovol'ski (Ed.), *Phraseology and Dictionaries (Special Issue). International Journal of Lexicography*, 28(3), 279–298. https://doi.org/10.1093/ijl/ecv021

Steyer, K. (2013). *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr.

Steyer, K. (2009). Zwischen theoretischer Modellierung und praxisnaher Anwendung. Zur Korpusgesteuerten Beschreibung useuller Wortverbindungen. In C. Mellado Blanco (Ed.), *Theorie und Praxis idiomatischer Wörterbücher* (pp. 119–145). Niemeyer: Tübingen.

Steyer, K. (2000). Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und Lexikografische Möglichkeiten. *Deutsche Sprache*, 28(2), 101–125.

Steyer, K., & Brunner, A. (2014). Contexts, Patterns, Interrelations – New Ways of Presenting Multiword Expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014)* (pp. 82–88). Gothenburg, Sweden.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6

Internet sources

(Accessed on 2 Juin 2018)

CII: COSMAS II. Korpusrecherche- und analysesystem. Institut für Deutsche Sprache. Mannheim. http://www.ids-mannheim.de/cosmas2/

Institut für Deutsche Sprache: Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2015-II (Release vom 28.09.2015). Mannheim: Institut für Deutsche Sprache. http://www1.ids-mannheim.de/kl/projekte/korpora/

lexpan: Lexical Pattern Analyzer. http://uwv.ids-mannheim.de/lexpan/

PDEV: Pattern Dictionary of English Verbs. http://pdev.org.uk/#about_cpa

PREPCON: Präposition-Nomen-Verbindungen im Kontext. http://uwv.ids-mannheim.de/prepcon/

SkE: Sketch Engine. http://www.sketchengine.co.uk/

Abbreviations

ADJ	adjective		N	noun
CA	IDS collocation analysis		NP	noun phrase
ECP	extended multiword pattern		P	preposition
LP	lexical pattern		PP	prepositional phrase
MWE	multiword expression		UWV	usuelle Wortverbindungen
MWP	multiword pattern			