Corpus-driven study of multi-word expressions based on collocations from a very large corpus

Annelen Brunner and Dr Kathrin Steyer
Project "Usuelle Wortverbindungen"
Institute for the German Language, Mannheim
brunner@ids-mannheim.de, steyer@ids-mannheim.de

Abstract

We present a corpus-driven approach to the study of multi-word expressions, which constitute a significant part of language. As a data basis, we use collocation profiles computed from DeReKo (Deutsches Referenzkorpus), the largest available collection of written German which has approximately two billion word tokens and is located at the Institute for the German Language (IDS).

We employ a strongly usage-based approach to multi-word expressions, which we think of as conventionalised patterns in language use that manifest themselves in recurrent syntagmatic patterns of words. They are defined by their distinct function in language.

To find multi-word expressions, we allow ourselves to be guided by corpus data and statistical evidence as much as possible, making interpretative steps carefully and in a monitored fashion. We develop a procedure of interpretation that leads us from the evidence of collocation profiles to a collection of recurrent word patterns and finally to multi-word expressions.

When building up a collection of multi-word expressions in this fashion, it becomes clear that the expressions can be defined on different levels of generalisation and are interrelated in various ways. This will be reflected in the documentation and presentation of the findings. We are planning to add annotation in a way that allows grouping the multi-word expressions according to different features and to add links between them to reflect their relationships, thus constructing a network of multi-word expressions.

1. What do we study?

The availability of large corpora has changed the possibilities for linguistic research considerably: They give access to a large quantity of real life language data. It has also changed the perspective on language itself. As Sinclair and many other recent researchers have pointed out, language does not solely work by applying grammatical rules to a set of lexical items. Conventionalised chunks play an important role in the everyday usage of language and greatly

shape its structure (*cf.* Sinclair, 1991; Hausmann 2004).

In many instances, the context of a word form, i.e., its collocations, have been studied to arrive at a better understanding of the meaning of the single word form (e.g., Hanks, 2004). On the other hand, corpus data is used to derive abstract, grammatical patterns for the usage of a word (e.g., Hunston/Francis, 2000).

Our approach is slightly different. We are interested not in single word forms but in multi-word expressions. We aim to find out which ones are common in contemporary German and to capture their behaviour and meaning. By following the evidence of corpora as opposed to intuition, it is possible to discover multi-word expressions that are not yet listed in classical handbooks of phraseology and to detect new usages and forms of known multi-word expressions (e.g. Moon, 1998). We try to make the most out of the evidence available by using the objective results of statistical collocation analysis for pre-structuring and taking a careful, corpus-driven approach to interpretation.

Terminology in the field of phraseology is quite diverse (for an up-to-date overview of the field *cf.* Burger *et al.*, 2007), so we begin by clarifying what we consider the object of our research. Compared to phraseological approaches, which require deviation from grammatical or semantic norms as a necessary criterion, we have a broad concept of multi-word expressions, which is heavily influenced by experience with empirical language data and centres around usage. In this respect, we adhere to Firth's contextual theory of meaning, here summarised by Tognini-Bonelli: "In the Firthian framework the typical cannot be severed from actual usage, and 'repeated events' are the central evidence of what people do, how language functions and what language is about." (Tognini-Bonelli, 2001: 89). In the context of early first language acquisition, Tomasello explains that patterns of language use are generalised to different degrees of abstraction when people use 'similar' expressions in 'similar' situations. Consequently, there are no elements of language that do not have a communicative meaning, as they are all derived from language use (*cf.* Tomasello, 2006: 21). We keep this view in mind when dealing with multi-word expressions.

The German name we use for our object of research, "Usuelle Wortverbindungen" (*cf.* Steyer, 2000), reflects this usage-based perspective as it can be paraphrased as 'multi-word patterns that are common in usage'. These are defined as conventionalised patterns in language use that manifest themselves in recurrent syntagmatic patterns of words (Steyer, forthcoming). Like Feilke, we believe that multi-word expressions become frozen by usage and are pragmatically bound to conventionalised contexts (*cf.* Feilke, 2004: 47).

To be of interest to us, multi-word expressions must neither be completely frozen nor deviate from the grammatical norm. A certain degree of fixedness, however, is important to our model, since structural stability is necessary in order for the unit to become a recognizable chunk with a distinct meaning or function in the language use attached to it.

This function is attached to the multi-word expression as a whole as opposed to its parts, but is not to be confused with idiomaticity. When studying corpus data, it quickly becomes clear that idiomaticity can hardly be an objective criterion for defining multi-word expressions. Whether the whole has a different meaning than the sum of its parts often depends on the meaning assigned to the parts. However, dictionaries show that the number of meanings assigned to a single word often differ significantly. Therefore, it would be necessary to first pin down the meaning of each component of the multi-word expression in a corpus-based way, a very difficult task, especially because, as has been pointed out by Hanks, "there are no literal meanings, only varying degrees of probability" (Hanks, 2004: 247).

For example, in German there is a multi-word expression *Kapital verspielen*. *Kapital* is commonly translated as 'financial capital, funds' and *verspielen* as 'to gamble away'. However, in the vast majority of cases, corpus data from collocation profiles tells us that what a speaker means when using this multi-word expression is not actually 'to gamble away money', but rather 'to put at risk what you have/to squander opportunities'. So, is the meaning of the whole here different from the meaning of the parts? Not if you take into account that *verspielen* is often used in the sense of 'to squander/to put at risk' and that *Kapital* is used in the sense of 'opportunities/chances/potential' in other contexts, too.

For our definition of a multi-word expression, the question, whether *Kapital verspielen* is idiomatic or not, is secondary. The important fact is that this combination of words is commonly used in German language and works as a functional unit in communication. From corpus evidence, you can see that it is used within a specific pragmatic context to express criticism or admonition. This is confirmed by examining typical contexts of the multi-word expression brought forth by collocation analysis, which include:

*Kapital **darf** nicht verspielt werden*
***Gefahr**, Kapital zu verspielen*
*Kapital **leichtsinnig** verspielen*

Multi-word expressions can even have a specific function in language if they appear to be completely transparent. An example would be the "aus ADJECTIVE Gründen" multi-word expressions discussed in example 2 below.

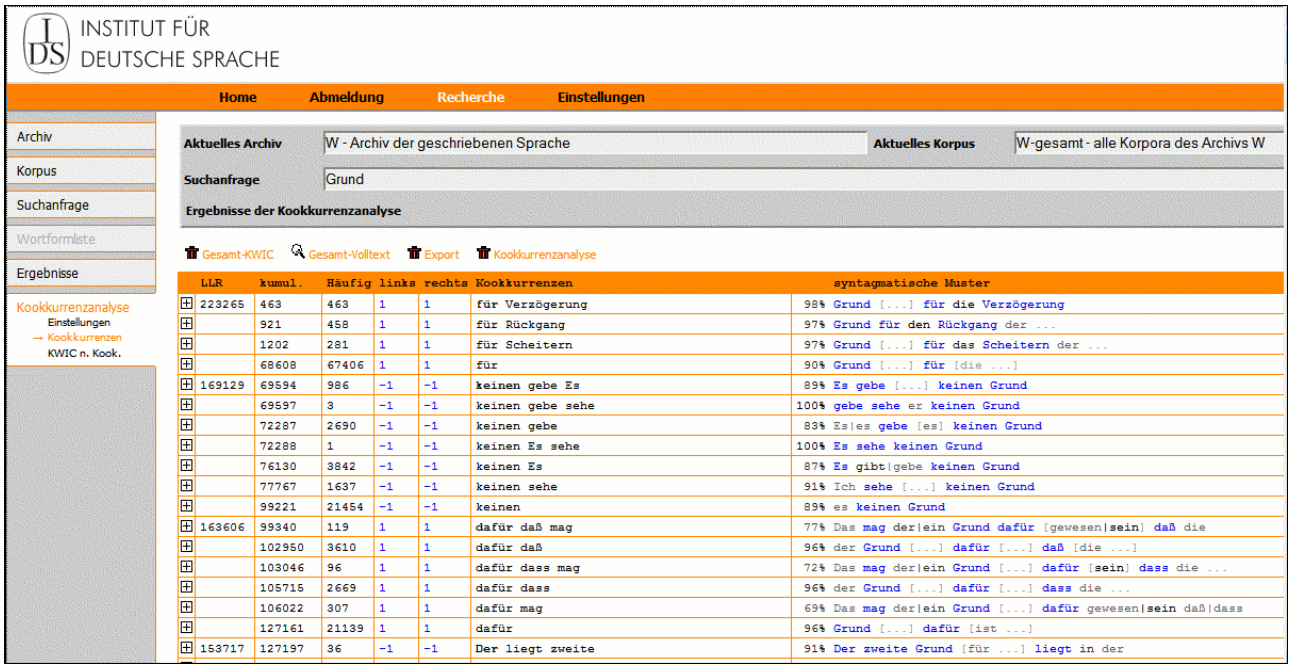
To sum up, usage is for us the key to identifying as well as to describing multi-word expressions. We base our research on collocation profiles computed from a very large corpus, thus, relying on a statistical measure of typicality. For the definition of multi-word expressions, their meaning and/or function in language use is most important.

2. How to find multi-word expressions

As stated above, our goal is to study multi-word expressions with a corpus-driven method. Therefore, we allow ourselves to be guided by corpus data as much as possible, making interpretative steps carefully and in a monitored fashion. We follow the dictum of corpus-driven linguistics:

"In a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence." (Tognini-Bonelli, 2001: 84).

Our corpus basis is DeReKo (Deutsches Referenzkorpus), with over 2 billion tokens the largest corpus of written German available today. It is located at the Institute for the German Language (IDS) and accessible via the corpus research tool COSMAS II (http://www.ids-mannheim.de/cosmas2). In addition, we use a sophisticated analytical method for structuring corpus data, collocation analysis (Kookkurrenzanalyse), as developed by Cyril Belica (Belica 1995), which can also be used via COSMAS II. This method takes a target word and, using the log-likelihood-ratio measurement, identifies the words that appear statistically significantly in a given radius around it, its collocation partners. Belica's method also sub-structures the result by calculating further partner words, which appear with the target word and its primary collocation partner. Thus, a hierarchical structure of collocation partners for a target word is constructed, and the KWICs (keyword-in-context lines from the corpus), which were the basis to the calculations, are clustered accordingly. For each cluster and sub-cluster, the method also calculates a 'syntagmatic pattern' from the assigned KWIC surfaces that reflects the most frequent positioning of the target and partner words as well as other words that appear often in the cluster. For more detailed information on this method see: Homepage of the project 'Methoden der Korpusanalyse und -erschließung', (http://www.ids-mannheim.de/kl/projekte/methoden); Tutorial Kookkurrenzanalyse (http://www.ids-mannheim.de/kl/misc/tutorial.html); Perkuhn, 2007.)



LLR	kumul.	Häufig	links	rechts	Kookkurrenzen	syntagmatische Muster
223265	463	463	1	1	für Verzögerung	98% Grund [...] für die Verzögerung
	921	458	1	1	für Rückgang	97% Grund für den Rückgang der ...
	1202	281	1	1	für Scheitern	97% Grund [...] für das Scheitern der ...
	68608	67406	1	1	für	90% Grund [...] für [die ...]
169129	69594	986	-1	-1	keinen gebe Es	89% Es gebe [...] keinen Grund
	69597	3	-1	-1	keinen gebe sehe	100% gebe sehe er keinen Grund
	72287	2690	-1	-1	keinen gebe	83% Es|es gebe [es] keinen Grund
	72288	1	-1	-1	keinen Es sehe	100% Es sehe keinen Grund
	76130	3842	-1	-1	keinen Es	87% Es gibt|gebe keinen Grund
	77767	1637	-1	-1	keinen sehe	91% Ich sehe [...] keinen Grund
	99221	21454	-1	-1	keinen	89% es keinen Grund
163606	99340	119	1	1	dafür daß mag	77% Das mag der|ein Grund dafür [gewesen|sein] daß die
	102950	3610	1	1	dafür daß	96% der Grund [...] dafür [...] daß [die ...]
	103046	96	1	1	dafür dass mag	72% Das mag der|ein Grund [...] dafür [sein] dass die ...
	105715	2669	1	1	dafür dass	96% der Grund [...] dafür [...] dass die ...
	106022	307	1	1	dafür mag	69% Das mag der|ein Grund [...] dafür gewesen|sein daß|dass
	127161	21139	1	1	dafür	96% Grund [...] dafür [ist ...]
153717	127197	36	-1	-1	Der liegt zweite	91% Der zweite Grund [für ...] liegt in der

Figure 1: A clipping from the results of collocation analysis for the word form *Grund* as presented by the COSMAS II web interface.[1]

This automated pre-structuring of the corpus evidence is an excellent starting point for our research, since the fact that word forms appear together in a statistically significant way often provides evidence for a recurrent syntactical and semantic connection between them (*cf.* Belica/Steyer forthcoming). We prefer using collocation analysis without lemmatisation (which would also be available using Belica's method), as we agree with Sinclair's statement: "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) Starting our work at the surface level of language is a logical step since language users, too, are only confronted with the surfaces, and all categories are secondary interpretations. As we want to treat the data given to us by the objective means of corpus evidence and statistical calculation as carefully as possible and avoid jumping to conclusions based on our intuition too early, we propose several separate steps of interpretation to arrive at a multi-word expression.

Our first interpretative step is to define search patterns that are matched to the KWICs of a collocation cluster in order to group together those that have a similar, stable structure. This step is

[1] This profile was calculated 28 June 2007 with the following settings: Korpusbasis: W-gesamt - alle Korpora des Archivs W; Anfrage: "Grund"; Kontext: -5, 5; Zuverlässigkeit: analytisch, Granularität: fein; Autofokus: ja; Lemmatisierung: nein; Funktionswörter ignorieren: nein; höchstens 1 Satz: ja; Zuordnung: mehrfach. Note that the collocation profile is statistical evidence and may change as weak partners may disappear and be replaced by others.

necessary, as collocation analysis calculates statistical affinity between word forms in a given radius around a target word, without taking order or filler word forms into account. Because of this, several different multi-word expressions often occur in the same cluster (if they are made up of the same word forms), or instances of the same expression are sorted into different clusters (e.g., if there is morphological or orthographic variance in the elements that make up the multi-word expression).

For the definition of our search patterns, we take clues from the syntagmatic patterns provided by Belica's method and use mainly word forms that have appeared as significant collocation partners of a target word. However, the patterns may also include other word forms if those contribute to the structure of the detected unit or serve to distinguish between different units.

These search patterns and their results are then bundled together into what we call 'recurrent word patterns'. Recurrent word patterns serve to collect information about a particular structure, applying a limited set of basic generalisations, currently: orthographic variance, order of the components (especially important for multi-word expressions with verbal components), and variance in the grammatical form of the components. Note that these generalizations do not have to be applied. They should not if it turns out that a particular order or grammatical form is typical for a multi-word unit. Defining recurrent word patterns on the basis of the search patterns is, therefore, again an interpretive step.

The resulting recurrent word patterns are purely surface-based constructs, snippets of language that have a strong indication of occurring in this specific form. They are influenced by particularities of the corpus and the settings of collocation analysis and can often be seen as a set of typical realisations of more abstract multi-word expressions.

In a final step, the recurrent word patterns serve as evidence to postulate multi-word expressions. According to our usage-based approach, the main indicator for a multi-word expression is that a distinct function or meaning in language use can be assigned. Thus, multi-word expressions may (and often will) subsume recurrent word patterns from several collocation clusters and even several profiles.

When building up a collection of multi-word expressions in this fashion, it becomes clear that the expressions can be defined on different levels of generalisation and are interrelated in various ways. This has to be reflected in documentation and presentation of the findings. We are planning to add annotation that allows grouping of the multi-word expressions according to

different features. In addition, we are going to add links between them to reflect their relationships, thus constructing a network of multi-word expressions.

We will illustrate our approach with two examples: One to show which interpretative steps we propose to arrive at the definition of a multi-word expression, and one to shed more light on our concept of multi-word expressions and their interrelations.

2.1 Example 1: From corpus data to multi-word expressions

The following example illustrates how we arrive at a multi-word expression from a collocation profile. Search patterns will be represented in italics and parentheses *(search pattern),* recurrent word patterns in squared brackets [recurrent word pattern], and multi-word expressions in quotation marks "multi-word expression".

We look at the collocation profile of the word form *Grund* ('ground/soil'; 'reason'). For this example, we focus on the cluster of the primary collocation partner *Boden* ('ground/floor'). *Boden* ranks among the strongest collocation partners of *Grund*. The cluster has been sub-structured by Belica's collocation analysis method in the following way:

- **Boden** *„in Grund [und] Boden" (98%)*
  - **Boden von** *„von Grund [und] Boden" (66%)*
    - **Boden von Ausverkauf** *„den Ausverkauf [...] von Grund und Boden" (92%)*
  - **Boden und** *„in Grund [...] und [...] Boden" (91%)*
    - **Boden und von** *„von Grund und [...] Boden" (62%)*
    - **Boden und Ausverkauf** *„den Ausverkauf von Grund und Boden" (93%)*
      - **Boden und von Ausverkauf** *„den Ausverkauf [...] von Grund und Boden [...] und"* (86%)

Figure 2: Collocation partners and syntagmatic patterns from the cluster *Boden* in the collocation profile for *Grund.*[2]

This is an indicator that the word forms *Boden*, *und*, *von* and *Ausverkauf* are significant in their relationship to *Grund*. They should be included in the search patterns we define in order to structure the KWICs into recurring structures. Further hints for ordering, gaps, and other word

[2] The profile cited in this example was calculated 17 May 2007 with the following settings: Korpusbasis: W-gesamt - alle Korpora des Archivs W; Anfrage: "Grund"; Kontext: -5, 5; Zuverlässigkeit: analytisch, Granularität: fein; Autofokus: ja; Lemmatisierung: nein; Funktionswörter ignorieren: nein; höchstens 1 Satz: ja; Zuordnung: mehrfach. Note that the collocation profile is statistical evidence and may change as weak partners may disappear and be replaced by others.

forms that are frequent in the KWICs are given by the syntagmatic patterns provided by the collocation analysis method.

An obvious choice for a search pattern is (*Grund und Boden*), which matches almost all KWICs of the cluster. This pattern is very stable, with nearly no other word forms entering between the components. Those that appear do not contribute as modifiers or complements to the observed structure and can, thus, be neglected in a study of multi-word expressions. With this search pattern, the first recurrent word pattern can be defined, also called [Grund und Boden].

Now we observe the KWICs to further differentiate. Making use of the indicators Belica's collocation analysis method has given us, we now look for the search pattern (*von Grund und Boden)* and notice two interesting facts: First, the meaning of *Grund und Boden* is always the same in this structure: It means 'land, property'. Second, when inserting a gap (indicated here by the #-sign) between the components *von* and *Grund und Boden*, a regularity can be observed. In many instances, the gap contains modifiers to the 'land' the text deals with:

sich häufenden Forderungen aus dem Westen nach Rückgabe	von	enteignetem	Grund	und	Boden	sagte Lothar Pawliczak, in der DDR werde "nicht
Bis jetzt kamen 30.000 Familien in den Genuß	von	enteignetem	Grund	und	Boden	. Die Umverteilung gilt als letzte Möglichkeit, den Sprengsatz
der DDR Heute läuft die Antragsfrist zur Rückerstattung	von	enteignetem	Grund	und	Boden	und Vermögen in der DDR ab/ Hunderttausende WestbürgerInnen
zwei Jahren das totale Chaos", fürchtet er. Die Sicherung	von	erschwinglichem	Grund	und	Boden	für Sozial- und Gemeindewohnungen sei eine der wichtigsten
Die meisten	von	fast einem Dutzend libanesischer Camps sind auf	Grund	und	Boden	gebaut, der vom UN-Hilfswerk für Palästinaflüchtlinge (UNRWA) gemietet
übrigen erfolge die Wahrung des Eigentumsrechtes bei Inanspruchnahme	von	fremdem	Grund	und	Boden	, insbesondere auch des Luftraumes, grundsätzlich nicht rechtsmißbräuchlich, weil
Entscheidung an, die	von	großer Tragweite für Enteignungen und andere	Grund	und	Boden	betreffende Verfahren in Österreich sein könnte. Die Europäische
aus dem Publikum, wie denn der drohende Ausverkauf	von	heimischem	Grund	und	Boden	verhindert werden könne, wurden von Gmachl und Firlei
binden. INNSBRUCK (schra). Bleibt Tirols Politik beim Verkauf	von	heimischem	Grund	und	Boden	nur mehr die Zuschauerrolle? Diese Befürchtung wird seit
"Das Instrument des Grundverkehrs funktioniert gut. Der Ausverkauf	von	heimischem	Grund	und	Boden	ist nicht eingetreten", begründete Streiter seinen Vorschlag, Nöbl
es um eine EG-konforme Regelung für den Erwerb	von	heimischem	Grund	und	Boden	durch "Ausländer". Denn als solche gelten die EG-Bürger
sollen die Länder, so wird argumentiert, den (Aus-)Verkauf	von	heimischen	Grund	und	Boden	an reiche Ausländer besser verhindern können. Nach der derzeit
- etwa die Erlaubnis zum privaten Besitz	von	hundert Quadratmetern	Grund	und	Boden	oder die Einrichtung von Bauernmärkten - interpretierten Optimisten

Figure 3: KWICs from the example cluster, structured by the search pattern *(von # Grund und Boden)*.

To capture this information, we use the gapped search pattern *(von # Grund und Boden*) to define another recurrent word pattern: [von ... Grund und Boden].

A third search pattern can be defined by considering the last statistically significant partner *Ausverkauf* ('sellout'): *(Ausverkauf von # Grund und Boden)*. *Ausverkauf* is a complement to *Grund und Boden* in the meaning of 'land/property'. Other complements, e.g., *Erwerb* ('purchase'), *Verkauf* ('sale'), *Nutzung* ('use') can be observed by looking at the KWICs, though those are not significant partners to *Grund* in this profile. The word *Ausverkauf* is so prominent because DeReKo is dominated by newspaper texts and the 'sellout' of property in the former DDR during German Reunification as well as the concerns citizens of other countries have about foreign investors were

important issues. This shows that by using real life data, real life events also shape the findings. However, this does not deter from the fact that the unit is frequent in language use and should, thus, be noted as a recurrent word pattern [Ausverkauf von ... Grund und Boden].

Following indication from the syntagmatic patterns provided by Belica's method, we now try a fourth search pattern *(in Grund und Boden)*. When examining the KWICs it captures, a very interesting fact occurs: *in Grund und Boden* has radically different semantics than the *Grund und Boden*-patterns described above. It appears as a verb modifier and indicates that the action described leads to a negative state.

geäußert werden. Entweder reden gleich drei Redner einen hinterher	in	Grund	und	Boden	, oder alle sehen sich peinlich berührt an und
schon immer gesagt", erklärt Joe Gordon, "Maggies gesamte Gewerkschaftsgesetze	in	Grund	und	Boden	zu verdammen, war schon immer gewerkschaftlicher Übereifer. Die
in der Minderheit. 15.000 IrInnen pfeifen die englische Nationalhymne	in	Grund	und	Boden	. Danach gelingt es der deutschen Kapelle, aus der
Lande Lenins dazu animiert hatte, die Söldner italienischer Fußballmagnaten	in	Grund	und	Boden	zu stampfen. Marx und Engels fehlten in München, ebenso
gemeint, durch den Nachweis einiger Ungenauigkeiten das ganze Buch	in	Grund	und	Boden	verdammen zu dürfen. Wird der diesjährige Historikertag in
Demokratie" daherredet. Mit dieser Kritik will ich die taz keineswegs	in	Grund	und	Boden	verdammen; sie bringt oft hervorragende Artikel und äußerst
Musik etwas Positives: Die alte Musik wird nicht mehr	in	Grund	und	Boden	verflucht, sondern teilweise original verwendet, neu überarbeitet oder
Michael Holm, wird zwar mit realen Pauken und Trompeten	in	Grund	und	Boden	gespielt, aber der unverändete Text ist schon witzig
wieder zu ersetzen. So klimperten dann "Blechreiz" die Tobenden	in	Grund	und	Boden	. Die Berliner Band bot zwar auch den typischen
zu Tieren hatte und mich wegen meiner mangelnden Standfestigkeit	in	Grund	und	Boden	schämte. Ich frage Euch, was eine derartige Arbeit mit
In ihrer Abwesenheit wurde ihr Heim von einer Kamelherde	in	Grund	und	Boden	getrampelt und zum Glück kam kein Menschenleben zu
von Würde, von korrekter Haltung, von Schneidigkeit, von Stehkragen	in	Grund	und	Boden	. Wer Angst davor hat, sich lächerlich zu machen,
Chorrillo, wo ihre Häuser während der Invasion vom 20.Dezember	in	Grund	und	Boden	gebombt wurden. Guillermo Cochez, ein Abgeordneter der Christdemokratischen Partei,
Niederlage im Januar 1987 kritisierte Lafontaine Raus schlappen Bundestagswahlkampf	in	Grund	und	Boden	. Jochen Vogels Klarsichthüllen-Attribute gelten schon länger nicht als
der Bundesrepublik tat er später dasselbe und stampfte es	in	Grund	und	Boden	. Jeweils mit 1a unerschrockenen Argumenten und Feuerschutz aus
Polemik: den realen Sozialismus - zu Recht natürlich -	in	Grund	und	Boden	stampfen und der Hoffnung der Menschen auf Wiedervereinigung
Fell retten. "Unsere Leute sind in der Lage, uns	in	Grund	und	Boden	zu wirtschaften. Auch wir sind Eigentümer des Betriebs.
von dem eigentlich als Fallobst vorgesehenen James "Buster" Douglas	in	Grund	und	Boden	gehauen wurde, hielt sich King nicht lange mit
In nicht enden wollenden Dialogen sprechen sich die beiden	in	Grund	und	Boden	. Aus der Groteske ist ein plump psychologisierendes Dialogmarathon
damit die rutschfesten Gastgeber ihre wasserscheuen Gegner nach Belieben	in	Grund	und	Boden	, rsp. Schlamm und Matsch spielen können. Die Münchner
bin auch kein Engel." Immer, wenn sie ihren Freund	in	Grund	und	Boden	geredet habe, habe er sie an die Wand
dem Gutachten, das auf Antrag Mompers bei der Senatssitzung	in	Grund	und	Boden	gestimmt und als "unzureichend" gegeißelt worden war. Damit
aus Reinickendorf bis zu den Arminen aus Bielfeld alles	in	Grund	und	Boden	gekickt wurde und am Sonntag auch noch der
Mittagssonne von Verona spielten wir den Gegner mit 9:2	in	Grund	und	Boden	. Das Feuerwerk wurde schon in der dritten Minute
gehoben, nach dem selben Ergebnis gegen die USA aber	in	Grund	und	Boden	verdammt wurden, und nun erwischte es die Mannschaft
personeller Gleichheit hätten sie die wacklige argentinische Abwehr vermutlich	in	Grund	und	Boden	gespielt. Als sich der Schatten über immer größere Teile
für Leute, die überall dabeisein wollen. d.Korr.) und Trauergästen	in	Grund	und	Boden	kritisierten. Nicht ganz zu unrecht natürlich, doch darf

Figure 4: KWICs from the example cluster, structured by the search pattern *(in Grund und Boden).*

This different meaning and the fact that the pattern occurs frequently and in a stable fashion warrants that another recurrent word pattern is defined: [in Grund und Boden].

The difference between search pattern and recurrent word pattern is not very striking in this example, as all the recurrent word patterns subsume exactly one search pattern. However, the concept becomes clearer when more clusters and perhaps more collocation profiles are examined.

It turns out that in this profile *Bodens* (genitive to *Boden*) is also a significant collocation partner of *Grund* (though much weaker than *Boden*). The cluster of *Bodens* contains primarily a good number of matches for the search pattern *(Grund und Bodens)*, a genitive variant of *Grund und Boden* in the 'land/property' sense. This search pattern would then be added to the recurrent word pattern [Grund und Boden], as variance in grammatical form is one of the 'allowed'

generalisations for recurrent word patterns. [Grund und Boden] is thus a generalisation over search patterns *(Grund und Boden)* and *(Grund und Bodens).* This mechanism is especially helpful when collecting instances of patterns with verbal components that can appear in a lot of different forms.

The definition of the actual multi-word expressions happens on the basis of recurrent word patterns. Only now, an interpretation beyond the surface and syntagmatic particulars is made and the multi-word expressions are defined according to the communicative value of the observed structures. In our example, two multi-word expressions would be defined:

Multi-word expression	**Subsumed recurrent word patterns**	**Explanation**
Grund und Boden	[Grund und Boden], [von ... Grund und Boden], [Ausverkauf von ... Grund und Boden]	The two more specialised recurrent word patterns are subsumed, because they are extensions to *Grund und Boden* and do not serve as functional chunks of their own. In the description of the multi-word expression, other frequent partners to *Grund und Boden* can be mentioned, e.g., *Erwerb von* ('purchase of'), *Umgang mit* ('treatment of').
in Grund und Boden	[in Grund und Boden]	Though it appears analogue to [von Grund und Boden] on the surface level, the structure captured by the recurrent word pattern has a different function and, thus, constitutes a separate multi-word expression. In the description of the multi-word expression, it is noted that it appears very frequently as a verb modifier, and frequent verbal partners are mentioned such as *sich schämen* ('to be ashamed'), *stampfen* ('to stomp'), *reden* ('to talk').

Figure 5: Overview of the defined multi-word expressions.

The reference to the recurrent word patterns from which they are derived is a key element in the description of multi-word expressions, as those link back to search patterns, which in turn point to the actual corpus data and, thus, make the process of generalisation retraceable.

Each multi-word expression will be assigned a paraphrase and be enriched by more information about its particular structure and its contexts of usage. In the following example, the nature of multi-word expressions and their interrelations will be elaborated upon.

2.2 Example 2: Partially lexicalised and multi-levelled multi-word expressions

The multi-word expressions from the example above, “Grund und Boden” and “in Grund und Boden”, both belong to the group of fully lexicalised multi-word expressions. However, our method also captures partially lexicalised multi-word expressions, especially when combining the evidence from several collocation clusters and profiles. An example is given here.

The collocation profile of *Gründen* (dative plural to *Grund*) contains many adjectival collocation partners. Several recurrent word patterns can be defined that share the stable syntagmatic structure [aus … Gründen] (‘for ... reasons’), e.g., [aus politischen Gründen] (‘for political reasons’), [aus zwei Gründen] (‘for two reasons’), [aus unerfindlichen Gründen] (‘for incomprehensible reasons’). Postulating a different multi-word expression for every significant adjective is not only problematic from a methodological point of view, as it is hard to make a clear cut which adjectives to include, but would also gloss over an important abstraction, the fact of syntactic as well as pragmatic similarity of the instances.

Therefore, we define a partially lexicalised multi-word expression with a slot. On a high level of abstraction this would be “aus ADJECTIVE Gründen”. The filler is only specified grammatically here. This multi-word expression can be assigned the general meaning of “giving reasons”.

However, we are interested whether there are restrictions on the adjectives that are used as fillers and arrive at the following sub-categorisation:

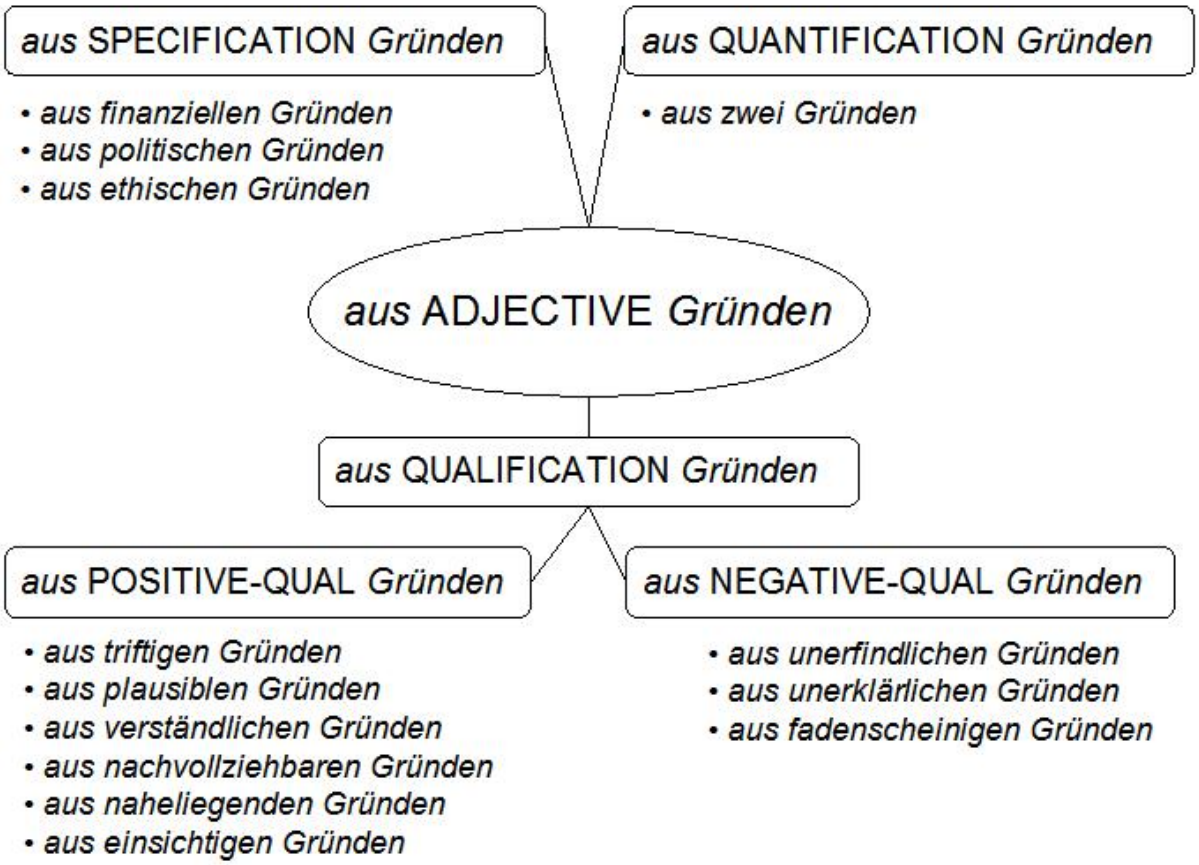


Figure 6: The multi-word expression "aus ADJECTIVE Gründen" and its specialisations.

These sub-categories are chosen not so much because the adjectives themselves can be assigned the abstract labels used here but because the meaning and communicative function of each of the more specific multi-word expressions differ. For example, the multi-word expression "aus SPECIFICATION Gründen" could be paraphrased as follows: "Using this multi-word expression gives an official character to the actions that are explained and at the same time allows the speaker to be vague about the reasons for these actions by using the less-specific plural form that is mandatory for its structure. It typically appears in combination with verbs like *ablehnen*, *absagen*, and *zurücktreten*."

This paraphrase clearly would not be appropriate for the parent "aus ADJECTIVE Gründen" nor for any other of the more specific multi-word expressions.

Since multi-word expressions like "aus SPECIFICATION Gründen" share structural as well as functional traits with their parent, but at the same time, have distinct functional traits of their own, it seems legitimate to propose that there are several levels of multi-word expressions in different degrees of abstraction. The number of these levels and the relationships among them is

subject to further research.

3. Prospects

At the moment, we are working on a network of multi-word expressions based on words for body parts like *Ohr* ('ear'), *Kopf* ('head'), *Auge* ('eye'), etc., as well as on examining causative multi-word expressions, starting out at the word forms of the lemmas *Grund* and *warum* ('why').

We strive to build up a collection of multi-word expressions common in the German language according to the usage-based criteria explained above. Important issues are to find out what constitutes the core of a multi-word expression and how slots can be specified.

We also want to study more deeply the links and relationships, surface-based as well as functional, that exist between multi-word expressions. For this, we use linguistic annotation, including structural criteria (e.g., grammatical status of a multi-word expression) as well as features that capture the typical use in the corpus like domain, situation, or genre. An important annotation will be the pragmatic function of a multi-word expression.

However, the set of possible annotations is not fixed yet and will be developed as research continues. This is typical for the way corpus-driven linguistics works: "As the main lines of description become clear, it is to be expected that a descriptive apparatus will take shape in response to the descriptive needs." (Tognini-Bonelli, 2001: 179).

We plan to present our findings in a network structure that illustrates the interrelations of multi-word expressions and can also be linked to electronic dictionaries (for more thoughts about the presentation of multi-word expressions *cf.* Steyer, forthcoming). A network like this can be both helpful to foreign language learners and interesting for linguists.

References

Belica, C. (1995) Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode. Institut für Deutsche Sprache. Mannheim. COSMAS II available on-line from http://www.ids-mannheim.de/cosmas2/ (accessed: 29 June 2007).

Belica, C. and K. Steyer (forthcoming): Korpusanalytische Zugänge zu sprachlichem Usus, in AUC (Acta Universitatis Carolinae), Germanistica Pragensia XX. Prag: Karolinum. Preprint available from http://www.ids-mannheim.de/kl/projekte/uwv (accessed: 29 June 2007).

Burger, H., D. Dobrovolskij, P. Kühn and N. R. Norrick (eds) (2007) Phraseologie/Phraseology. Ein internationales Handbuch zeitgenössischer Forschung/An International Handbook of Contemporary Research. 2 Halbbände (= HSK 28.1/2). Berlin/New York: de Gruyter.

Feilke, H. (2004) Kontext – Zeichen – Kompetenz. Wortverbindungen unter sprachtheoretischem Aspekt, in K. Steyer (ed.) (2004a), pp. 41–64.

Hanks, P. (2004) 'The Syntagmatics of Metaphor and Idiom'. *International Journal of Lexicography* 17 (3), 245–274.

Hausmann, F. J. (2004) Was sind eigentlich Kollokationen? in K. Steyer (ed.) (2004a), pp. 309–334.

Hunston, S. and G. Francis (2000) Pattern Grammar. A corpus-driven approach to the lexical grammar of English. (= Studies in Corpus Linguistics 4). Amsterdam/Philadelphia: Benjamins.

Kämper, H. and L. M. Eichinger (eds) (2007) Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache. (= Studien zur Deutschen Sprache 40). Tübingen: Narr.

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

Perkuhn, R. (2007) "Corpus-driven": Systematische Auswertung automatisch ermittelter sprachlicher Muster, in H. Kämper and L. M. Eichinger (eds) (2007), pp. 465–491.

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

Steyer, K. (2000) ‘Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten’. *Deutsche Sprache*. Jahrgang 28, 2/2000, 101–125.

Steyer, K. (ed.) (2004a) Wortverbindungen – mehr oder weniger fest. (= Jahrbuch des Instituts für Deutsche Sprache 2003). Berlin/New York: de Gruyter.

Steyer, K. (2004b) Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven, in K. Steyer (ed.) (2004a), pp. 87–116.

Steyer, K. and M. Lauer (2007) “Corpus-Driven”: Linguistische Interpretation von Kookkurrenzbeziehungen, in: H. Kämper and L. M. Eichinger (eds) (2007), pp. 493–509.

Steyer, K. (forthcoming) Zwischen theoretischer Modellierung und praxisnaher Anwendung. Zur korpusgesteuerten Beschreibung usueller Wortverbindungen, in C. Mellado Blanco (ed.) Studien zur Phraseologie aus lexikografischer Sicht. Theorie und Praxis der Erstellung von idiomatischen Wörterbüchern. Tübingen.

Tognini-Bonelli, E. (2001) Corpus Linguistics at Work. (= Studies in Corpus Linguistics 6). Amsterdam/Philadelphia: Benjamins.

Tomasello, M. (2006) Konstruktionsgrammatik und früher Erstspracherwerb (translated by Stefanie Wulff and Arne Zeschel), in K. Fischer and A. Stefanowitsch (eds) Konstruktionsgrammatik. Von der Anwendung zur Theorie. (= Stauffenburg Linguistik 40), pp. 19–3. Tübingen: Stauffenburg.