Corpus-derived data on German multiword expressions for lexicography

Ulrich Heid and Marion Weller
Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

We show a parsing-based architecture for the extraction of German verbal multiword expressions. It uses dependency parsing as a preprocessing step, allows us to extract syntactic patterns of arbitrary form from the parsed data, and comprises a relational database where each extracted multiword occurrence is stored along with the sentence it is extracted from, and with a number of morphosyntactic and syntactic features. These features serve (i) for an automatic decision about the likely idiomatization of the candidate under review, and (ii) in later lexicographic work to get a clear picture of lexicographically relevant linguistic properties of the selected candidates.
We use dependency-parsed text, because this allows us to find non-adjacent multiwords and to use subcategorization knowledge to identify e.g. verb + object pairs more reliably than on the basis of ourface patterns.
The extraction results illustrate the potential of the tools; we can identify morphosyntactic preferences in collocations (these often indicate idiomatization), longer collocational or idiomatic structures (where e.g. the core elements and possible modifies can be clearly distinguished), lexical variation in idioms, as well as certain specific features of collocations or idioms (e.g. preferences for negation).
As all data are stored in a database, which supports a variety of generalization steps, it is in principle possible to prepare different layouts (i.e. presentations and selections) of dictionary entries, for different user groups and user needs.

1. Introduction

In this paper, we report about work on the extraction of German multiword expressions from text corpora. The multiwords are extracted along with their morphosyntactic preferences and with corpus sentences illustrating these preferences. The extracted data include collocations, but also verbal idioms and idiomatic predicative prepositional phrases.

We first give lexicographic motivation for the need to extract linguistic preferences of multiwords (section 2). We then discuss the extraction architecture used (section 3): it relies on syntactically analyzed corpora, on the extraction of word pairs or word tuples and their morphological and syntactic features, as well as on the storage of all extracted data in a database. We can then sort, group, and interpret the stored data. In section 4, we present examples of the extraction results which illustrate morphosyntatic preferences of collocations and lexical variation in idioms. We finally discuss lexicographic applications of the extracted data (section 5) and avenues for future work (section 6).

The examples presented here have been derived from ca. 270 million words of German newspaper text (1987–1999). At the time of writing, we are in the process of parsing a substantial fragment of Baroni/Kilgarriff's (2006) *German Web as Corpus* data (DeWaC): we intend to run our tools on ca. 880 million words from this corpus and to combine the results with those obtained from ca. 400 milion words of newspaper text. In one of the applications, we used ca. 75 million words of texts from a journal from the field of trademark legislation.

2. Collocations and Idioms: lexicographic needs

For the purpose of this paper, we follow Bartsch (2004: 76), for a working definition of collocations; she writes: [collocations are] '*lexically and/or pragmatically constrained recurrent cooccurrences of at least two lexical items which are in a direct syntactic relation with each other*'. As far as the presentation of collocations in dictionaries is concerned, we essentially follow Tarp (2008) and start from the assumption that for production-oriented

dictionaries, a detailed lexicographic description of collocations is necessary; ideally, this description should be as detailed with respect to morphosyntactic, syntactic, semantic and diasystematic classification (markings for register, region etc.) as the lexicographic description of single words, thus conferring collocations the status of secondary treatment units (cf. Heid/Gouws 2006).

According to Tarp (2008), idioms are mostly relevant for reception-oriented dictionaries, where their form and their meaning are to be described. As there is no clearcut boundary between more collocational and more idiomatic multiwords (cf. Grossmann/Tutin's (2003) classification of collocations into regular, transparent and opaque (= idiom-like) ones), a dictionary aimed at several user types, usage situations and dictionary functions may need to cover both, collocations and idioms, in (almost) equal depth.
The fact that many dictionaries lack sufficient descriptive detail on multiword items has been frequently deplored: cf. Jesenšek (2009: 65ff), for a recent statement, and for proposals concerning descriptive categories needed.

2.1. Descriptive categories for collocations

Collocations should be described, in our view, with respect toproperties of their bases and their collocates (cf. Hausmann 2004), but also with respect to properties they have as a whole (cf. Heid/Gouws 2006). We summarize descriptive proposals by Heid (1998), Bartsch (2004), Heid/Gouws (2006), Jesenšek (2009) and others in Table 1.

Property type	Values (examples)	Example collocations
Lexical combination	• collocate = single word • collocate = idiom	*felsenfest schwören* *Stein und Bein schwören*[1]
Morphosyntactic preferences	• number (sg/pl) (+ %) • determination (+ %) • modifiability • negation (+ %)	*in Dienst[en] stehen*: 36% sg vs. 64 % pl. *In die Kritik geraten*: 84% def. sg. *have high hopes*, preferred adj. *Ask + ADJ + question*, variable adj. *Sich keine Blöße geben*: over 50% negated
Syntactic subcategorization ('Valency patterns')	- roles - grammatical functions - grammatical categories	*In + Dienst + stehen* EMPLOYER EMPLOYEE MOD | PPATTR SUBJECT NP_{Gen} | *bei*-PP NP
Pragmatic annot.	• Style, register, ...	*Dienst schieben* (fam.)

Table 1. Examples of lexicographically relevant properties of collocations

2.2. Descriptive categories for idioms

Obviously, aspects of morphosyntactic fixedness also play a role for the lexicographic description of idiomatic expressions[2]. For receptive purposes, however, dictionaries must primarily provide meaning explanations and diasystematic marks of regional, historical, register, domain variation etc. (Tarp 2008).

[1] This construction could also be classified as 'Teilidiom' (partial idiom).

[2] Fixedness has been used as an indicator of idiomaticity in data extraction, cf. Fazly/Stevenson 2006.

Certain idiomatic expressions are, however, quite variable in their form (cf. Kwaśniak 2006, for a corpus-based study). This variation may even make the identification of idiomatic expressions in a text reception situation rather difficult. Thus such variation should be accounted for even in reception-oriented dictionaries[3]. In table 2, we display the variants of the German negative polarity idiom *auf keine Kuhhaut gehen* ('be just incredible', lit: not fit on any cow's skin), as found in our corpus.

- auf keine Kuhhaut gehen	neg = *kein*
- auf keine Kuhhaut mehr gehen	neg = *kein* + adv (*mehr*)
- (schon) nicht mehr auf die K. gehen	neg = nicht + adv + def. Article
- kaum mehr auf eine Kuhhaut gehen	neg = adv (*kaum* + indef. article)

Table 2. Variants of the German idiom *auf keine Kuhhaut gehen*

2.3. From lexicographic needs to requirements for corpus-based data extraction

The lexicographic needs summarized briefly above translate directly into requirements for data extraction from corpora. The properties discussed for collocations are mostly preferential in nature: thus, data extraction should quantify the observed properties, to identify these preferences. Idiom variation is also preferential, as likely the stable core elements of an idiom can only be defined by means of a quantitative analysis of the uses and variants found in corpus data (cf. also Cignoni/Coffey 1998).

For German, idioms and collocations which contain a verbal element pose a major problem for data extraction. Due to the variability of German constituent order and to case syncretism, surface-based approaches to multiword candidate extraction tend to provide too little recall[4] to really be usable on low frequency data (cf. also Seretan 2008). We have thus opted for data extraction from a parsed German corpus.

3. Multiword data extraction

Our extraction is based on dependency-parsed data. We use the FSPar parser (Schiehlen 2003), which provides as its output dependency trees, encoded in a linearized format, as shown for the sentence *die zweite Studie lieferte ähnliche Ergebnisse* ('the second study produced similar results'), in figure 1 and in table 3, below.

To find collocation and idiom candidates, we check the predicate/argument structures of each parsed sentence, thus roughly following a pattern-based approach, similar to that of Heid (1998), Kilgarriff et al. (2004), etc., but on full parses. Linear FSPar output contains (from left to right, in table 3) the position of each word form in the sentence (`posn.`), the word form itself, its part-of-speech (`pos`), its lemma, its morpho-syntactic features (`morph.prop.`), as well as (in the two rightmost columns), the position number of the word's dependency governor (`gov.`), and the grammatical function (`functn.`) the element has with respect to its governor.

[3] In dictionaries usable for recognizing idioms in running text, such variation is sometimes accounted for.

[4] Ivanova et al. 2008 applied the Sketch Engine approach (Kilgarriff et al. 2004) to German and have shown that pattern-based extraction performs less well on German than on English, unless the equivalent of parsing-based information is available.

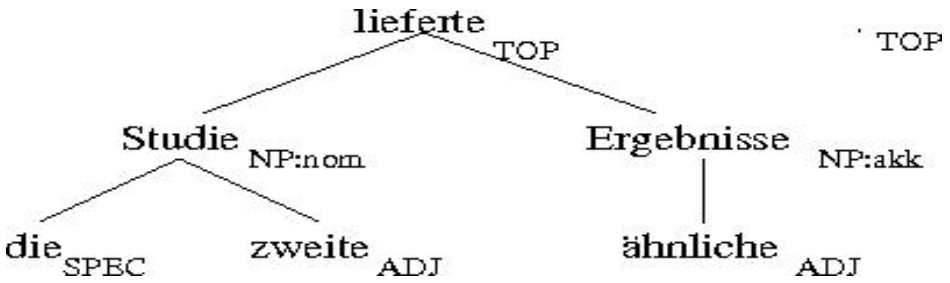

Figure 1. Dependency analysis of *die zweite Studie lieferte ähnliche Ergebnisse*

posn	word form	pos	lemma	morph-prop.	gov.	functn.
0	Die	ART	d	|	2	SPECIFIER
1	zweite	ADJA	2.	|	2	ADJ
2	Studie	NN	Studie	Nom:F:Sg	3	NP:nom
3	lieferte	VVFIN	Liefern	3:Sg:Past:Ind*	-1	TOP
4	ähnliche	ADJA	Ähnlich	|	5	ADJ
5	Ergebnisse	NN	Ergebnis	Akk:N:Pl	3	NP:akk
6	.	$.	.	|	-1	TOP

Table 3. Linearized output of the dependency parse from figure 1

The extraction starts from the finite verb (numbered line 3 in table 3), and searches, for example, for its subject (NP:nom, position 2), its direct object (NP:akk, line 5), etc. In this way, also modifying adjectives (marked as ADJ in table 3, e.g. *ähnlich* for *Ergebnisse*, line 4) and morphosyntactic features (column 5, `morph.prop.`) can be extracted. Each extraction result is a tuple of, e.g. a verb and its object noun, along with values for lemmas, morphosyntactic properties, modifiers, etc. Data about the position of the items in the sentence (e.g. being fronted, extraposed etc.) or about potentially subcategorized complements can also be extracted.

All extracted data sets are stored in a relational database. As these data sets are only candidates, different procedures for subdividing the candidate set into collocations, idioms and trivial combinations are applied. These range from standard association measures for word pairs (e.g. LogLikelihood, cf. Evert 2005) over the analysis of morphosyntactic fixedness as an indicator of idiomaticity to a combination of fixedness-based monolingual features with translation-based semantic opacity indicators[5]. These procedures sort the candidate lists (true positives are then expected to appear at the top of the lists), and the lexicographer has to select, e.g. using an interface like LexiView (Evert et al. 2004) or Kilgarriff et al.'s recent tickbox tools.

The tool setup presented here has the advantage of allowing us to carry out two lexicographically relevant activities in one go: the identification of collocation and idiom candidates in texts and -- provided big enough texts are avaliable -- their morphosyntactic classification in terms of fixed vs. variable properties. The contents of our database can thus be investigated in different ways and with
different objectives.

[5] The latter are extracted from word-aligned parallel text, cf. Fritzinger (2009).

4. Examples of lexicographically relevant results

In this section, we discuss a few examples of the output of our tools; data of this kind may be useful as a basis for descriptive lexicographic work.

4.1. Morphosyntactic preferences in collocations

In table 4, we reproduce the absolute frequency data obtained for different morphosyntactic forms of the (negative polarity) collocation (*kein*) *Wort + verlieren* ((`not to say a word' (about)), as extracted from ca. 270 million words of news text. From left to right, the table contains data about the number of occurrences (`f`), possible modifying adjectives (`modif.`), the noun, the type of determiner observed (`det.`), the lemma of those determiners which are quantifiers (`quantif.`), the number, verb, negation and possible modifying adverbs.

f	modif	noun	n_det	quantif	num	v_lemma	neg	adv
88		Wort	quant	kein	Sg	verlieren		
11		Wort			Sg	verlieren		kaum
11		Wort			Sg	verlieren		auch nur
10	einzig	Wort	quant	kein	Sg	verlieren		
9		Wort	quant	viel	Pl	verlieren		
8		Wort			Sg	verlieren		
8		Wort	quant	kein	Sg	verlieren		aber
8		Wort	quant	viel	Pl	verlieren	+	
7		Wort	quant	kein	Sg	verlieren		mehr
6		Wort	quant	ein_paar	Pl	verlieren		
6	groß	Wort	quant	kein	Pl	verlieren		
5		Wort	quant	kein	Sg	verlieren		auch
5		Wort			Pl	verlieren		

Table 4. Variation in *(kein) Wort verlieren*

Table 4 shows that the form *kein Wort (einziges) verlieren* is most typical, and that there are variants like *kaum ein Wort verlieren* and *..., ohne auch nur ein Wort zu verlieren*; wie also find e.g. (*nicht*) viele Worte verlieren, in the plural.

f	prep	n_lemma	v_lemma	n_det	fusion	num	conjunctn
966	zu	Ergebnis	kommen	def	-	Sg	daß
411	zu	Ergebnis	kommen	indef	-	Sg	
308	zu	Ergebnis	kommen		-	Pl	
238	zu	Ergebnis	kommen	def	-	Sg	
89	zu	Ergebnis	kommen	def	+	Sg	
66	zu	Ergebnis	kommen	def	+	Sg	daß
62	zu	Ergebnis	kommen	dem	-	Sg	
35	zu	Ergebnis	kommen	quant	-	Sg	
34	zu	Ergebnis	kommen		-	Sg	
23	zu	Ergebnis	kommen	def	-	Sg	daß daß
23	zu	Ergebnis	kommen	indef	-	Sg	daß
17	zu	Ergebnis	kommen		-	Pl	daß
14	zu	Ergebnis	kommen	def	-	Pl	
10	zu	Ergebnis	kommen	def	-	Sg	wenn daß

Table 5. *zu + Ergebnis + kommen*: number and determination

In table 5, we show data for the collocation *zu+Ergebnis+kommen* ('come to the conclusion, that...') taking a *dass* (that-) clause. The table shows that this collocation is typically definite, when it has a *dass*-clause (*zu dem Ergebnis kommen, dass*), but not necessarily so without the complement clause. The database entries also show that the article and the preposition are typically not fused (feature `fusion`, 966: *zu dem* vs. 66: *zum*). The fused article+preposition

form occurs more frequently when the construction does not include a complement clause, as in *wir kommen [jetzt] zum Ergebnis*. Whether both usages are part of the same collocation or not, is open for discussion.

The tables summarize similar usage patterns, condensing thereby the output a lexicographer would be presented with, e.g. in a concordance. As all data are represented in a relatinal database, the granularity of presentation can be varied: from a Sketch-Engine-like overview of lexical combinability to the more fine-grained variation patterns shown above.

4.2. Lexical variation in idioms

We have used our data collection also for the analysis of lexical variation in idiomatic expressions. An example is given in table 6, concerning the idiom '*kein+Mucks*+Verb' ('not say/hear a word'): indeed *geben* (give), *machen* and *tun* (make, do) are found, next to *hören* (hear). The data in table 6 also show that the 270 million word corpus is too small to really investigate preferences with respect to the use of verbs in the 'kein+Mucks+Verb'-construction[6]. The different verbs could nonetheless all usefully be given in a dictionary entry.

f	noun	det	quantif.	n_num	v_lemma	adv
6	Mucks	quant	kein	Sg	geben	
5	Mucks	quant	kein	Sg	machen	
4	Mucks	quant	kein	Sg	tun	
3	Mucks	quant	kein	Sg	hören	
2	Mucks	quant		Sg	machen	(kaum mehr)

Table 6. The idiom component *kein + Mucks* with different verbs

4.3. Complex syntactic patterns in idioms

As we extract all multiword data according to syntactic patterns (e.g. verb + direct object, verb + prepositional phrase), one might wonder about more complex patterns in idiomatic multiwords: in a simplistic approach, these would not be found. To avoid this, we also include a set of more complex patterns for multiwords containing verbs.

Examples are coordinated noun phrases or prepositional phrases (e.g. *Hand und Fuß haben*}, *in Angst und Schrecken versetzen*, cf. table 7) or constructions embedded under modals (*nicht riechen können* ('[to] hate')) or under *lassen* (*sich (nicht) aus der Ruhe bringen lassen* ('[to] keep cool'), cf. table 8). Tables 7 and 8 are semi-formalized and do not correspond to word order. For all items, not only frequency, but also the above mentioned morphosyntactic properties, and variation data are available.

freq.	prep.	coordinated nouns	verb
120	in	Angst und Schrecken	versetzen
69	um	Leben und Tod	gehen
62	mit	Händen und Füßen	sich wehren
51	an	Ecken und Enden	fehlen
45	in	Lohn und Brot	stehen
42	auf	Herz und Nieren	prüfen
35	in	Saus und Braus	leben
20	ausser	Rand und Band	geraten
18	in	Sack und Asche	gehen

Table 7. Preposition-noun-verb collocations with coordinated nouns

[6] Cf. our plans for a similar extraction on the basis of much larger texts, section 1.

freq.	prep.	noun	verb1	verb2
122	zu	Wort	kommen	lassen
70	im	Hals	stecken	bleiben
68	aus	Ruhe	bringen	lassen
56	im	Regen	stehen	lassen
35	mit	Angst	tun	bekommen
14	ins	Bockshorn	sich jagen	lassen

Table 8. Preposition-noun-verb collocations with an additional verb

5. Lexicographic application examples

The data extraction setup described above has been used for the provision of raw material for different lexicographic projects. In one case, data for high-frequency words of German, as well as English and French, have been used as an input to the creation of core vocabularies for learners.

In a recent cooperation with C.H. Beck publishers, Munich, we carried out the same type of data extraction on German specialized texts from the field of intellectual property and trademark legislation, to provide raw material for a dictionary of this field (cf. Heid et al. 2008). As the language of law shows much variation within morphological word families, we have grouped the collocation data according to such families, bringing verb + object collocations, participle + noun collocations and noun + noun collocations or compounds together, cf. table 9 (and Fritzinger/Heid 2009).

verb + object	participle + noun	noun + genitive	compound
Patent anmelden (395)	angemeldetes P. (559)	Anmeldung eines P. (598)	Patentanmeldung (23700)
Interesse abwägen (34)	abgewägtes I. (11)	Abwägung von I. (1150)	Interessenabwägung (2833)
Begriff auslegen (533)	ausgelegter B. (28)	Auslegung eines B. (1229)	Begriffsauslegung (17)
einlegen (1049)	eingelegte B. (451)	Einlegung einer B. (363)	Beschwerdeeinlegung (150)

Table 9. Number of occurrences of word-formation variants of a few specialised collocations

The interest of this exercise, which is carried out automatically, by use of a German morphology system (SMOR, cf. Schmid et al. 2004), lies in the following: not all morphological variants of collocations from juridical phraseology are equally frequent. Thus, it makes sense, in a dictionary, (i) to establish cross-links between variants, and (ii) to inform users about the respective frequency data; for example, *Interessen abwägen* and *Abwägung der/von Interessen* are frequent, whereas the pertaining verb+object combination is not.

Let us now come back to Tarp's (2008) proposals for the lexicographic presentation of collocation data in (printed and electronic) dictionaries. Tarp suggests that very different data need to be presented, as far as collocations are concerned, to users in need of support for text production, vs. users wanting to understand a text. The differences between production-oriented and reception-oriented dictionaries not only concern the layout of collocational entries, they also concern the selection of lexicographic data types (i.e. types of indications) and their ordering, i.e. the typical access paths users follow to the respective indications. We claim that a rich and richly structured database (as it is created by our extraction tools) provides a flexible starting point for deriving such widely varying entries from one single source of data. The differences in the presentation of collocational data for both reception and production are illustrated in figure 2.

Reception-oriented version

Dienst [...]
∼ *antreten* [duty: come to work]; ∼ *antreten* [job: take up + job] • *zum* ∼ *gehen* [duty: go to work] • ∼ *haben* [duty: be on duty] • ∼ *machen* [duty: be on duty] • ∼ *quittieren* [job: resign] • *aus dem* ∼ *scheiden* [job: resign] • ∼ *schieben* (fam.) [duty: be on duty] • *in jds.* ∼*(en) sein* [job: be employed by so.] • *in jds.* ∼*(en) stehen* [job: be employed by so.] • *in jds.* ∼*(e) treten* [job: take up + job] • *seinen* ∼ *versehen* (formal) [duty: do one's job].

Production-oriented version

Dienst [...]
(a) work, duty [...]
VERB + ∼. be on duty: ∼ *haben/machen/tun*; fam.: ∼ *schieben*; *im* ∼ *sein*; do one's job: *seinen* ∼ *(gewissenhaft ...) versehen*; go/come to work: *zum* ∼ *gehen/kommen*, *seinen* ∼ *antreten*.
(b) job, employment [...]
VERB+∼. be employed by s.o.: *in jmds* ∼*[en] sein/stehen*; take up a job: *seinen* ∼ *antreten*, *in jmds* ∼*e treten*; resign: *den* ∼ *quittieren*, *aus dem* ∼ *[aus]scheiden*.

Figure 2. Collocation data with the German noun *Dienst* in hypothetical entries for respective vs. productive purposes

The example deals with some collocations of the German noun *Dienst*. Instead of semantic paraphrases of the collocations, we use English near equivalents. *Dienst* is polysemous: it has a reading which is roughly equivalent to 'duty' and one which is roughly equivalent to 'job'. If a dictionary user in a text reception situation does not know the collocations, likely he or she is also not able to distinguish the two readings. Thus, an appropriate device seems to be an alphabetic listing of all relevant collocations, along with an indication of the reading underlying each collocation. Obviously, lexical and morphosyntactic variation only need to be mentioned in so far as they might affect the user's access to the collocation description.

To support text production, the dictionary entry should rather be organized in the same way as, e.g. the *Oxford Collocations Dictionary for Students of English*, i.e. by the readings of the base noun ('duty' vs. 'employment' in the right part of figure 2), the grammatical construction type of the collocations (here only: verb+object), as well as the meaning paraphrases (here: English equivalents) of the individual collocations listed. For each collocation, morphosyntactic preferences (e.g. the possibility to have *in jmds. Diensten sein/stehen* in the plural), or diasystematic marks (e.g. the fact that *Dienst schieben* is colloquial) should be mentioned, as these indications are relevant for text production.

One and the same underlying source of lexicographical data should feed both presentational variants. Enabling this is not only a matter of the internal representation of these data, but, crucially, also of data provision.

6. Conclusion

We have shown a parsing-based architecture for the extraction of German multiword expressions. It uses dependency parsing as a preprocessing step, allows us to extract syntactic patterns of arbitrary form from the parsed data, and comprises a relational database where each extracted multiword occurrence is stored along with the sentence it is extracted from, and with a number of morphosyntactic and syntactic features. These features serve (i) for an automatic decision about the likely idiomatization of the candidate under review, and (ii) in later lexicographic work to get a clear picture of lexicographically relevant linguistic properties of the selected candidates.

We still do not know enough about the linguistic properties of German collocations and idioms. The availability of large syntactically annotated corpora now starts to open ways to analyze multiword behaviour in context in sufficient detail. For example, some of our future

research is aimed at the automatic identification of detailed evidence for subcategorization properties of German multiword expressions, and at an analysis of their preferences with respect to word order (which ones allow for the topicalization of their noun phrase or prepositional phrase elements: ??*Gebrauch gemacht hat davon niemand*, 'usage made has of this nobody' -- 'nobody has made use of this'?).

We are also developing an application where regional specificities of German collocations from Austrian, Swiss and German (newspaper) texts are compared, to gain more experience on region-specific collocation variation; this also may lay the foundations of register-sensitive corpus-based lexicographic work on multiword expressions, as the techniques to distinguish regional and register-specific data should be the same.

Future applications will be lexicographic and/or directed towards language technology; we intend to further investigate ways of providing prototype entries of dictionaries based on our data collection.

References

Baroni, M.; Kilgarriff, A. (2006). 'Large linguistically-processed web corpora for multiple languages'. In *Conference Companion of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*. 87-90.

Bartsch, S. (2004). *Structural and functional properties of collocations in English, A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Tübingen: Narr.

Cignoni, L.; Coffey, S. (1998). 'A corpus-based study of Italian idiomatic phrases: from citation forms to 'real-life' occurrences'. In *Proceedings of the Euralex International Congress 1998*. Liege. 291-300.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, available from http://www.collocations.de/phd.html, published 2005.

Evert, S.; Heid, U.; Säuberlich, B.; Debus-Gregor, E.; Scholze-Stubenrecht, W. (2004). 'Supporting corpus-based dictionary updating'. In *Proceedings of the eleventh EURALEX International Congress*. I, 255-264.

Fritzinger, F. (forthcoming). 'Using parallel text for the extraction of German multiword expressions'. In *Lexis - E-Journal in English Lexicology* (2010).

Fritzinger, F.; Heid, U. (2009). 'Automatic grouping of morphologically related collocations'. In *Corpus Linguistics Conference 2009 (Proceedings)*, electronic publication.

Fazly, A.; Stevenson, S. (2006). 'Automatically constructing a lexicon of verb phrase idiomatic combinations'. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL-2006*. Trento/New Brunswick: ACL. 337-344.

Grossmann, F.; Tutin, A. (2003). 'Quelques pistes pour le traitement des collocations'. In *Les collocations - analyse et traitement*. Amsterdam: De Werelt. 5-21.

Hausmann, F.J. (2004). 'Was sind eigentlich Kollokationen?. In Steyer, K. (ed.). *Wortverbindungen -- mehr oder weniger fest*. *Institut für Deutsche Sprache, Jahrbuch 2003*. published 2004. 309-334.

Heid, U.; Gouws, R.H. (2006). 'A model for a multifunctional electronic dictionary of collocations', in: *Proceedings of the XIIth Euralex International Congress*. Torino. 979-988.

Heid, U. (1998). 'Towards a corpus-based dictionary of German noun-verb collocations'. In *Proceedings of the Euralex International Congress 1998*. Liège. 301-312.

Heid, U.; Fritzinger, F.; Hauptmann, S.; Weidenkaff, J.; Weller, M. (2008). 'Providing Corpus Data for a Dictionary for German Juridical Phraseology'. In Storrer, A; Geyken, A.; Siebert, A.; Würzner, K.M. (eds., 2008). *Text Resources and Lexical Knowledge -- Selected Papers from the 9th Conference on Natural Language Processing, KONVENS 2008*. Berlin: Mouton de Gruyter. 131-144.

Ivanova, K.; Heid, U.; Schulte im Walde, S.; Kilgarriff, A.; Pomikálek, J. (2008). 'Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case'. In *Proceedings of LREC-2008, Linguistic Resources and Evaluation Conference*. Marrakesh. CD-ROM.

Jesenšek, V. (2009): 'Phraseologische Wörterbücher auf dem Weg zu Phraseologiedatenbanken'. In Mellado Blanco, C. (ed., 2009). *Theorie und Praxis der idiomatische Wörterbücher*. Tübingen: Niemeyer. 65-81.

Kilgariff, A.; Rychly, P.; Smrz, P.; Tugwell, D. (2004). 'The Sketch Engine'. In *Proceedings of the eleventh EURALEX International Congress*. I, 105-117.

Kwasniak, R. (2006): 'Wer hat nun den Salat? - Now who's got the mess? Reflections on phraseological derivation: from sentential to verb phrase idiom'. In *IJL* 19 (4). 459-478.

Schiehlen, M. (2003): 'Combining Deep and Shallow Approaches in Parsing German'. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*. Sapporo: ACL. 112-119.

Schmid, H.; Fitschen, A.; Heid, U. (2004). 'SMOR: A German computational morphology covering derivation, composition, and inflection'. In *Proceedings of LREC-2004*. Lisboa.

Seretan, V. (2008). *Collocation Extraction based on syntactic Parsing*. Geneva: Université de Genève.

Tarp, S. (2008): *Lexicography in the borderland between knowledge and non-knowledge: general lexicographic theory with particular focus on learner's lexicography*. Tübingen: Niemeyer.