Extraction of German multiword expressions from parsed corpora using context features

Marion Weller, Ulrich Heid

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12 D 70174 Stuttgart
[wellermn, heid]@ims.uni-stuttgart.de

Abstract

We report about tools for the extraction of German multiword expressions (MWEs) from text corpora; we extract word pairs, but also longer MWEs of different patterns, e.g. verb-noun structures with an additional prepositional phrase or adjective. Next to standard association-based extraction, we focus on morpho-syntactic, syntactic and lexical-choice features of the MWE candidates. A broad range of such properties (e.g. number and definiteness of nouns, adjacency of the MWE's components and their position in the sentence, preferred lexical modifiers, etc.) along with relevant example sentences, are extracted from dependency-parsed text and stored in a data base. A sample precision evaluation and an analysis of extraction errors are provided along with the discussion of our extraction architecture. We furthermore measure the contribution of the features to the precision of the extraction: by using both morpho-syntactic and syntactic features, we achieve a higher precision in the identification of idiomatic MWEs, than by using only properties of one type.

1. Introduction

Much work has been done on the automatic extraction of multiword expressions (MWEs) from text corpora (cf. e.g. the regular ACL SIGLEX workshops on multiword items). Some of these efforts mainly aim at the identification of semantically non-compositional combinations, leaving their detailed linguistic description for a separate step (e.g. (Fazly and Stevenson, 2006)).

In this article, we describe an approach which allows us to do both steps in one go: identifying idiomatic MWEs, and providing detailed corpus-based material for their lexical description, e.g. in entries of the lexicon of a formal grammar or in a database similar to the Dutch DuELME lexicon (Grégoire, 2007).

Our approach is applied to German data and uses dependency-parsed material (section 2). The parser relies on a substantial amount of lexical and subcategorization information; its underspecified output distinguishes between attachment and label ambiguities. We describe the handling of both ambiguity types. The objective is to maximally exploit the corpus data while keeping precision as high as possible.

We use various forms of syntactic patterns to identify candidates, going beyond simple verb+object, verb+prepositional phrase types: these are only basic patterns that can be expanded by adjectives, further nouns or prepositional phrases (PPs) (cf. section 4). The approach involves a variety of morpho-syntactic properties of the elements of the collocation candidates, insofar as we not only extract MWE-candidates, but also numerous context parameters: these morpho-syntactic properties are mostly preferential in nature, and many of them are idiomaticity indicators (section 3).

We do not differentiate between idioms and collocations; in order to keep our terminology simple, we will call the extracted patterns *collocation candidates*.

2. Extraction methods

2.1. Preprocessing: dependency parsing

As German constituent order is quite flexible, the components of multiword expressions do not always occur adjacently. Using the results of a deep syntactic analysis allows us to easily extract MWEs of different patterns together with their morpho-syntactic features, even if the components are distant.

We work with FSPAR (Schiehlen, 2003), a finite-state based dependency parser which provides an explicitly underspecified (disjunctive) representation of syntactic ambiguities, both at the level of attachment and grammatical function (expressed by means of case). Its output format is one token per line, annotated with POS-tag, lemma, morpho-syntactic features, governors and information about the token's grammatical function in the sentence (cf. the columns in the table in figure 1). Morpho-syntactic features of nouns or adjectives indentified by the parser (and extracted along with the collocation candidates) include gender, number and case. As can be seen in the example (cf. figure 1) reproduced in figure 2, morpho-syntactic annotation as well as dependencies can be ambiguous: The parser lists two possible attachments for the preposition in line 11, referring either to the noun directly above it (*Kopf in den Sand*: '*head in the sand*') or to the verb in line 6 (*stecken [...] in den Sand*: '*hide [...] in the sand*'). Attachment ambiguities (line 11) and in particular case syncretism (ambiguous at the level of grammatical functions, cf. line 17) can be an obstacle for an accurate extraction of collocation candidates.

2.2. Extraction

The first step in the extraction process is to identify verbs, and then to basically 'collect' all potential complements referring to the respective verb. In the sentence in table 1, we would find that the prepositional phrases *bei Gefahr*

pos	word form	pos-tag	lemma	morpho-syntactic features	gover-nor	gramm. function
0	”	$)	”	|	-1	TOP
1	Kann	VMFIN	können	3:Sg:Pres|1:Sg:Pres	-1	TOP
2	sehr	ADV	sehr	|	4/1||3	ADJ
3	schnell	ADJD	schnell	|	4/1	ADJ
4	laufen	VVINF	laufen	Inf	4/1	RK
5	und	KON	und	|	-1	TOP
6	steckt	VVFIN	stecken	3:Sg:Pres|2:Pl:Pres	16	ADJ
7	bei	APPR	bei	Dat	6	ADJ
8	Gefahr	NN	Gefahr	Dat:F:Sg	7	PCMP
9	den	ART	d	|	10	SPEC
10	Kopf	NN	Kopf	Akk:M:Sg	6	NP:8
11	in	APPR	in	Akk	6||10	ADJ
12	den	ART	d	|	13	SPEC
13	Sand	NN	Sand	Akk:M:Sg	11	PCMP
14	”	$)	”	|	-1	PUNCT
15	,	$,	,	|	16	PUNCT
16	sagt	VVFIN	sagen	3:Sg:Pres|2:Pl:Pres	-1	TOP
17	G.	NE	G.	Nom:M:Sg|Dat:M:Sg	16	NP1:|NP:8

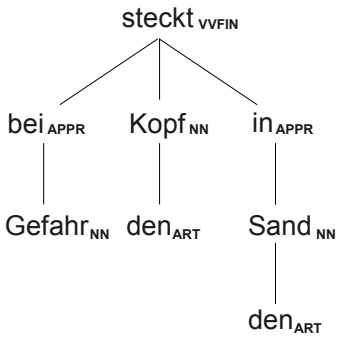

Figure 1: Slightly simplified output of FSPAR. The numbers in the *governor* column indicate to which line the respective word refers. The extracted structure is illustrated on the right.

”	Kann	sehr	schnell	laufen	und	steckt	bei	Gefahr	den	Kopf	in	den	Sand	”	,	sagt	G.
”	Can	very	fast	run	and	hides	with	danger	the	head	in	the	sand	”	,	says	G.

” It can run very fast and hides its head in the sand in case of danger”, says G.

Figure 2: Translation of the sentence used as example in figure 1.

(line 7) and *in den Sand* (line 11) refer to the verb *stecken* in line 6, as well as the direct object *Kopf* (line 10). Information about the number and determiner of noun phrases is extracted as well.

2.3. Ambiguity handling

Due to the difficulty to distinguish between complement and adjunct PPs, pp-attachment is very often ambiguous. Since any PP or a combination of several PPs might be collocationally relevant, we include every dependency listed in the parse output.

We assume that we can statistically even out the noise we get by including every possible pp-attachment listed in the parse-output: when applying statistical association measures (e.g. log-likelihood ratio, (Evert, 2005)), idiomatic and collocational expressions (being highly associated) are better rated than mere free combinations.

However, noun phrases with case label ambiguities are not included in the data collection, in an attempt to keep the data as clean as possible. While this clearly leads to a loss in recall, we will not have to deal with multiple entries, as it would be the case with several noun phrases being annotated as either subject or as direct/indirect object.

2.4. Storage

Extraction results are stored in a PostgreSQL-database. Table 1(a) shows a (partial) entry containing the results obtained from the sentence given in figure 1. An entry in the column V_LEM is obligatory, while all other fields are optional as long as there is at least one subject, object or prepositional phrase.

(a)

V_LEM	SUBJ	ACC OBJ	ACC NUM	ACC DET	DAT obj	PREP PHRASE
stecken	—	Kopf	sg	def	—	bei:Gefahr in:Sand

(b)

V_LEM	ACC OBJ	ACC DET	PREP	N_LEM	N NUM	N DET
stecken	Kopf	def	in	Sand	sg	def
stecken	Kopf	def	bei	Gefahr	sg	—

Table 1: Extraction results for the verb *stecken* from table 1.

Since PPs often occur as adjuncts, there can be an unlimited number of them in PREP_PHRASE. To be able to study each of these prepositional phrases individually, multiple values in this column are split into separate entries, (illustrated in table 1(b)), which are enriched with morphological information about the respective PP. This design allows us to study patterns containing a single prepositional phrase as well as patterns with several combined PPs.

Table 2 gives an impression of a more extensive entry: the expression *gute Miene zum üblen Spiel machen* is idiomatic (’*make the best of it*’, lit. ’*make (a) brave face to the bad game*’) and only complete with both adjectives.

3. Context parameters

We extract the following context parameters: reflexivity of the verbal element of the collocation candidate, negation, adverbs and fusion of a preposition and a definite article

V_LEM	SUBJ	ACC OBJ	ACC ADJ	PREP	N LEM	N ADJ
machen	Diplomat	Miene	gut	zu	Spiel	übel

FUS	NEG	ADV	SENTENCE
+	–	wieder mal	Der frühere Diplomat macht wieder mal gute Miene zum üblen Spiel ...

Table 2: Example for a data-base entry. Number and determiner of the nouns are not shown.

(cf. *zur* fused vs. *zu der* unfused).
Each noun has the features *number*, *determiner* and *adjective*. There are 6 possible values for determination: *definite, indefinite, demonstrative, possessive, null* and *quantifying*. In the case of a quantifying determiner, we also extract its lemma like *jeder* ('*every*') or *keiner* ('*no*').
Occurring (nearly) always in a negated context is characteristic of a certain group of MWES, namely negative polarity items. Since there are far more possibilities for negative contexts than just negation of the verb or noun (which can be derived from the negation particle *nicht* (feature NEG) or the quantifier *kein*), we experimented with modeling negative contexts, to specifically extract negative polarity items. For example, an inherently negative verb like *bezweifeln* ('*to doubt*') would indirectly negate a subsequent statement. For a more detailed description of negative polarity items see (Lichte and Soehn, 2007) or (Fritzinger et al., 2010a).
In the case of preposition-noun-verb combinations (PNV-triples), we also applied two syntactically motivated features. As we expect the components of idiomatic MWEs to be (immediate) neighbours, we compute a simple adjacency measure. It is based on the positions of the preposition, the noun and the verb. While trivial preposition-noun-verb combinations can appear in combination with e.g. adjectives and adjuncts like relative clauses (as demonstrated in example 1), idioms and collocations tend to exclude such modifiers and hence are (immediately) adjacent (see example 2). Preposition, noun and verb of a non-trivial combination cannot be immediate neighbours if a determiner or adjective is part of the idiom/collocation (shown in example 3). However, the relative position of verb, noun and preposition is still fixed.

1 **Auf** kleinen **Zetteln**, die an Bäume geklebt worden waren , **stand**: "Wilson kommt".

On small **notes**, that to trees glued been had, **stood**: "Wilson comes".

On small notes that had been glued to trees, it read: "Wilson comes".

2 Sie glauben, daß dadurch die Wirtschaft wieder **in Fahrt kommt**.

They believe that thereby the economy again **in run comes**.

They believe that thereby the economy gets going again.

3 Den Grünen werden vorschnelle und unüberlegte Politiksprünge **in** die **Schuhe geschoben**.

The Greens are overhasty and unreflected political moves **in** the **shoes shoved**.

Thes Greens are made responsible for overhasty and unreflected political moves.

Furthermore, idiomatic MWEs rarely occur at the very beginning of a sentence (in *vorfeld* position), except in contrastive contexts, e.g. together with adverbs like *jedoch* ('*however*').
We distinguish different types of *vorfeld* occurrences: All the MWE components (preposition, noun, verb) can be in the *vorfeld* with the verb being either finite or infinite followed by an auxiliary verb. Idiomatic MWEs can appear in this structure ((partial) VP-fronting, cf. example 4), although usually only in contrastive or otherwise special contexts. In the second type, only the preposition and the noun are in *vorfeld* position, followed by an auxiliary in the subsequent position (the *linke Satzklammer*, left verbal position) with the main verb later in the sentence. This structure does not work for idiomatic MWEs (example 5), but it can be used without problems for trivial MWEs (see example 6). We can thus use the restrictions on this particular context as one syntactic feature (among others) to tell idiomatic PNVs apart from non-idiomatic ones.
The non-trivial triple *in Stellung bringen* ('*to bring into position*') (total frequency 188), occurs once in the *vorfeld* (example 4). Moving the verb out of the *vorfeld*, (as demonstrated in example 5), leads to an ungrammatical sentence as opposed to the sentence in example 6 with the trivial expression *in Klinik bringen* ('*to bring into hospital*').

4 **In Stellung gebracht** worden seien Al-Samoud-Raketen mit einer Reichweite von 200 Kilometern

In position brought been have Al-Samoud missiles with a range of 200 kilometres

Al-Samoud missiles with a range of 200 kilometres have been positioned

5 *__In Stellung__ seien Al-Samoud-Raketen mit einer Reichweite von 200 Kilometern __gebracht__ worden

6 __In__ die __Klinik__ hatten die Eltern sie gegen ihren Willen __gebracht__.

__Into__ the __hospital__ had the parents her against her will __brought__.

The parents took her into a hospital against her will.

The restrictions on the use in *vorfeld* can thus be used as indicators of idiomatized MWE candidates. They are a symptom of a more general syntactic-semantic phenomenon, and even if they are not relevant for human-oriented lexicography, they are indeed relevant for the lexicon of text generation tools; provided an adequate information structure, it may be useful for such a system to have access to data about the possibility of VP- or PP-fronting in multiwords.
Since many idiomatic MWEs are morpho-syntactically fixed in number and determination, and also may be special in terms of their syntactic behaviour, these parameters can be used for sorting syntactically defined word co-occurrences (PNV-triples in table 3) into idiom/collocation candidates vs. non-fixed non-idiomatic combinations. Some of these

	MWE	f	NUM	DET	ADJ
-	in Jahr aussehen	271	Pl 150 Sg 121	def 129 – 85 dem 31	– 208
+	auf Barrikade gehen	167	Pl 165 Sg 2	def 165 – 2	– 165

Table 3: Comparison between morphologically fixed and non-fixed preposition-verb-noun triples.

parameters are also an input for a detailed lexicographic description of the specific properties of a given MWE.
The parser also marks the head of compound nouns, allowing us to generalize across transparent compounds and, alternatively, to identify those compounds which do not inherit the collocation preferences of their heads (which are mostly lexicalized, non-transparent compounds; (Zinsmeister and Heid, 2004)).
A further step towards more general predicate-argument-structure extraction, e.g. for texts from specialized language, could be to replace the actual nouns by an abstract term, such as the categories provided by GERMANET[1].

4. Experiments and results

4.1. Patterns

As illustrated in table 4, idiomatic expressions can be of different length and form. While components like certain adjectives or nouns can be obligatory within a given MWE, there might be more variation in other cases. It is also possible that an idiomatic MWE occurring mostly with a certain (group of) adjective(s) can also felicitously occur without an adjective. The adjectives in lines 2, 5 and 8 of table 4 are integral parts of the respective MWE. Looking at the expression '*auf* ADJ *Ohren stoßen*' ('*(not) find a good listener*', table 5), it becomes evident that there is a clear preference for the two most frequent adjectives. However, to be idiomatic at all, the expression must occur with an adjective, while e.g. the combination *Wert legen* ('*to insist'*) occurs (roughly) equally often with or without adjective.
The fact that longer patterns contain less complex ones leads to a problem of overlapping results: when extracting data covered by short patterns, we do not want to find incomplete expressions. Similarly, we are not interested in finding valid, short MWEs combined with varying adjuncts. For more details, see section 4.4.

4.2. Data and evaluation

The corpus collection we worked with (269 million tokens) consists of newspaper articles (1987-99) and the proceedings of European parliament debates (EUROPARL[2]). Table 6 gives an impression of the number of extracted syntactic structures and the results of a simple evaluation we carried out for this selection of patterns.

4.3. Extraction errors

Generally, we can distunguish between two kinds of undesired extraction results: the first ones are trivial Mwes, that

	Syntactic pattern	example
1	**noun verb**	Wert legen
2	**adj noun verb**	grün Licht geben
3	NOUN NOUN VERB	(dem) Fass (den) Boden ausschlagen
4	**prep noun verb**	in Sand setzen
5	**prep adj noun verb**	für bar Münze nehmen
6	**noun prep noun verb**	Kopf in Sand stecken
7	ADJ NOUN PREP NOUN VERB	dick Strich durch Rechnung machen
8	ADJ NOUN PREP ADJ NOUN PREP	gut Miene zu böse Spiel machen
9	PREP NOUN PREP NOUN VERB	mit Kanone auf Spatz schießen
10	PREP NOUN PREP NOUN	wie Sand an Meer
11	PREP NOUN VERB VERB	mit Angst (zu) tun bekommen

Table 4: Different syntactic patterns with entries of idiomatic expressions (examples are lemmatized).

ADJ	taub	offen	geneigt	verschlossen	—
f	222	92	1	1	1

Table 5: Adjective distribution for the expression *auf* ADJ *Ohren stoßen*.

can, at least partially, be filtered out by applying association measures. But there can also be 'false positive' results, i.e. expressions that appear to be an idiom but are not idiomatic in all observed instances. Here again, we can distinguish two types: literal use of an idiomatic Mwe, vs. a wrong analysis, e.g. of pp-attachment making a combination look like a valid idiom. For more information, see Cook et al. (2008) or Sporleder et al. (2010) for English and Fritzinger et al. (2010b) for German, who investigate the actual rate of literal use of idioms in different corpora.
At this point, we ignore the possibility of literal usage and focus only on wrong pp-attachment in PNV-triples leading to false positives. We carried out an analysis of 69 valid idiomatic PNV-triples with 100 randomly chosen occurrences for each[3]. The sentence length was restricted to 40 tokens.

Example 7 shows the false positive triple *in Betrieb sein* (lit. '*to be in operation*': '*to operate*') found in a sentence with *in Lohn und Brot stehen* (lit. '*to be in pay and bread*': '*to be in so.'s pay*'), where the latter is an idiomatic expression and *in Betrieb* ('*in business*') a mere adjunct.
If the verbal elements are *sein* ('*to be*') or *haben* ('*to have*'), our system might confuse auxiliary verb readings with full verb readings.
In the above examples, extracted triples are printed in **boldface** while correct structures are underlined.
We found that MWEs containing certain noun-preposition combinations are prone to extraction errors; these noun-preposition combinations are commonly used as (idiomatic) adjuncts: *in+Augen* (lit. '*in (his) eyes*': '*in his*

[1] http://www.sfs.uni-tuebingen.de/lsd
[2] http://www.statmt.org/europarl
[3] This set of sentences is smaller if the given idiom has a frequency below 100.

	PNV		NPNV		PANV		NV		ANV	
	EP	NEWS	EP	NEWS	EP	NEWS	EP	NEWS	EP	NEWS
tokens	1.576.220	10.165.415	572.523	3.139.948	404.813	2.580.320	978.083	4.913.847	260.621	1.216.437
n=25	16	23	0	5	3	16	14	19	7	10
n=50	28	39	3	14	8	31	27	37	10	19
n=100	55	82	4	31	16	46	47	71	17	37
n=500	153	290	29	86	57	131	165	247	59	113

Table 6: Number of extracted database-entries ('tokens') for the syntactic patterns highlighted in table 4 and number of non-trivial items in the top *n* when sorted by frequency. Results are split by corpus: Europarl (EP) and newspaper text (NEWS). Triples including nouns like *Mark*, *Dollar* and *Prozent* ('*percent*') have been excluded.

(a)

idiom	false positives	valid occurrences
auf Weg machen	10	90
auf Weg bringen	10	90
in Raum stehen	9	89
in Auge haben	8	92
an Tag legen	5	93
in Betrieb sein	2	98
zu Schau stellen	0	100
unter Lupe nehmen	0	100

(b)

false pos.	0	1	2	3	5	6	7	8	9	10	22
idioms	51	5	3	2	1	1	1	1	1	2	1

Table 7: Number of false positives for a a few selected idioms (a) and distribution of false positives for the examined 69 idioms (b). In total, we found 94 false positives in 6690 sentences.

opinion'), can be part of the MWE *in+Auge+haben* (lit. '*to have in eye*': '*to have sth. in mind*', cf. example 8), but it can also be used as an adjunct. In example 9, *hat* is the verb that belongs to *schlechtes Ansehen haben* ('*to have a bad reputation*'), but not the verb of the preposition-noun-verb triple as in example 8.

7 **waren in** 192 **Betrieben** knapp 20000 Mitarbeiter in Lohn und Brot.

were in 192 **companies** almost 20000 employees in pay and bread.

in 192 companies, almost 20000 members of staff were employed.

8 weil alle nur die kurzfristigen Chancen **im Auge haben**.

because everybody only the short-term chances **in eye has**.

because everybody has only the short-term chances in mind.

9 daß ein Dienstleistungsunternehmen **in** den **Augen** der Kunden ein so schlechtes Ansehen **habe**.

that a service provider **in** the **eyes** of the clients a such bad reputation **has**.

that a service provider has such a bad reputation in the eyes of the clients.

Table 7(a) shows how many false positives were among 100 randomly chosen occurrences for some of the examined idioms, while table 7(b) gives an overall impression of the evaluation. Note that the first entry in table 7(a) needs a reflexive pronoun to be complete; although information about reflexivity of verbs is available, we chose to exlcude this feature from this evaluation in order to keep the evaluated data as simple as possible: we do not want the same PNV-triple to occur twice, with and without a reflexive pronoun. As can be seen in the examples, the idiomatic version and the adjunct have different preferences for determiner and number: If the morphological preferences of an idiom are known, false positives can be discarded in many cases.

4.4. Detailed analysis of the extracted morpho-syntactic features

As most of the above examples illustrate, collocations and idioms are morphologically and/or syntactically restricted. In the following experiment, we want to examine the usefulness of the extracted morhosyntactic features for the identification of non-trivial MWEs.

The test set consists of the 1013 most frequent PNV-triples ($f \geq 210$) extracted from newspaper text. Nouns like *Mark*, *Franc*, *Prozent* ('*percent*'), being very frequent in newspaper text, but not interesting for MWE-extraction, have been excluded in advance. Two native speakers manually annotated the data deciding about the idiomaticity of each candidate. Cases with conflicting annotation were discussed in detail with a third native speaker. Incomplete patterns, e.g. idiomatic triples with a missing adjective, were marked as valid idioms since we expect the PNV-part under rewiew to be morpho-syntactically restricted in the same way as the complete idiom would be. Overall, 513 candidates were labelled as non-trivial MWEs. To test the indicator value of morpho-syntactic features, we intend to identify these cases by searching for strong preferences (see section 4.5).

For each triple, we computed a *fixedness-score* derived from the averaged or most prominent features of this triple. This score is intended to represent the morpho-syntactic preferences of an MWE. All PNV-triples are then sorted according to their scores.

The quality of a sorted list can be measured with the *uninterpolated average precision* (UAP) score, peaking at value 1 for a perfectly sorted list (Manning and Schütze, 1999). We opt for this measure as we want to split candidate expressions into trivial and non-trivial combinations. Since degrees of fixedness vary, we attempt to sort them ranging from very idiomatic at top to less idiomatic towards the end of the list.

In table 8(a) we present the UAP-scores for the list sorted seaparately by each of the enumerated features. The respec-

(a)

feature	num	det	neg	adjacency	vorfeld
uap	0.607	0.650	0.643	0.694	0.566

(b)

grouping	M_1 det+num	M_2 det+num+neg	S adja+vorf.	M_2+S
uap	0.635	0.681	0.664	0.830

(c)

weighted	M_2+2S	M_2+3S	M_2+4S	M_2+5S	M_2+6S
uap	0.845	0.853	0.857	0.858	0.859

Table 8: UAP-values for the morpho-syntactic features computed separately (a), grouped (b), and weighted (c). Candidates with the same score are ordered alphabetically. Ordering according to frequency results in an UAP-value of 0.651.

tive scores are either based on the number of occurrences of the features in relation to the triple's overall frequency (*vorfeld*, *neg*) or the percentage of the most prominent value (*num*, *det*), i.e. *singular* or *plural* and the determiner occurring most often with a given PNV-triple. The adjacency measure was averaged over all occurrences of a triple and averaged to be within 0 and 1.

Table 8(b) shows the resulting UAP-values when the features are grouped into morphologically and syntactically motivated categories. The features within the groups are weighted equally. Adding the *neg* feature to M_1 improves the sorting quality, while combining morphological (M_2) and syntactic (S) features results in an even bigger improvement. The morphologically and syntactically motivated features are independent from each other and complementary: this can be exploited, for example, to identify a morphologically fixed trivial combination because of its lower S-score for being syntactically unrestricted.

However, the syntactically motivated features have a greater impact on sorting quality than the morphological ones as illustrated in table 8(c), where the weight for S has been gradually increased. This might be due to the fact that *det* and *num* are related features while, *vorfeld* and the adjacency measure are independent. In fact, the adjacency measure has only been computed for verb-final sentences and therefore there is no overlap with the *vorfeld* feature.

4.5. Expanding basic patterns

By detecting a strong preference for a certain object and/or adjective, basic, incomplete patterns could be expanded into more complex idioms.

In the following experiment, we want to focus only on adjectives in PNV-triples. There are two aspects we need to consider: Does a given PNV-triple occur always or nearly always with an adjective, and if so, is there lexical variation in adjectives or are they always the same?

As a first step, triples occurring mostly without adjectives are excluded: Out of 1013 PNV-triples,133 never appear with an adjective and 610 occur in at least 90% of all cases without an adjective. When sorting the list of candidate-triples by a score based on the combination of morpho-logical and syntactic features (cf. table 8(b)) and the percentage of adjective occurrences, an UAP-value of 0.833 is achieved, a slight improvement compared to the sorting result without the adjective feature (0.830). By reducing the entire list of 1013 candidates to the 610 supposedly adjective-free candidates, sorting quality is improved to a UAP-value of 0.892. A further reduction to the 133 entirely adjective-free candidates results in an UAP-value of 0.937 (cf. table 9).

size of test set	1013 [all]	610 [ADJ$\leq$0.1]	133 [ADJ=0]
idioms	512	390	99
UAP	0.833	0.892	0.937

Table 9: Results when sorting according to a score based on morpho-syntactically features combined with the precantage of adjectives per candidate.

Using a threshold of 90%, we still allow for a few adjectives to occur occasionally with a candidate triple. By including the percentage of adjective appearances into the score, very few or no adjectives at all are rewarded compared to more adjectives indicating no clear preference.

Creative use of language definitely is an obstacle to our approach: Some idioms a native speaker would intuitively judge as not being able to be used with an adjective appear nevertheless in such combinations. In example 10, the idiom *zu Sache gehen* (lit. '*to go to thing*': '*there is a great ambiance*') is used with the adjective *fröhlich* ('*funny*') to describe the atmosphere at the event, resulting in a sentence on the verge to ungrammaticality.

10 Dort **geht** es bei Schunkelmusik des Musikcorps Stierstadt **zur** fröhlichen **Sache**.

There **goes** it with beer tent music of the music club 'Stierstadt' **to the** funny **thing**.

With beer tent music played by the music club 'Stierstadt', there is a great ambiance.

With supposedly adjective-free triples now identified and exluded from the test set, the remaining 403 candidates have to be divided into a set of idioms where a specific adjective is an obligatory part of the pattern vs. one where the presence of adjectives is common and not restricted.

At this point, the distribution of adjectives co-occurring with a cendidate expression is used to indicate preferences. Table 10 shows candidate triples with their respective most frequent adejctive and the precentage of occurrence in all instances with an adjective. While there are idiomatic expressions with a clear preference, there are also (morpho-syntactically fixed) frequent non-idiomatic formulae with a specific, obligatory adjective. In the case of *mit sofortig Wirkung bestellen* ('*to order with immediate effect*'), only the preposition-adjective-noun construct is fixed: it can be used with different verbs. To overcome this problem, a more detailed study including the association between individual parts of a pattern would be necessary (e.g. (Zinsmeister and Heid, 2004)).

Table 11 shows PNV-triples sorted by a score based on morphological and syntactic features in combination with the percentage of adjective occurrences and adjective distribu-

	PNV-triple	adjective	ADJ
+	auf Bank schieben	lang	1
-	mit Wirkung bestellen	sofortig	1
-	zu Fixing verbilligen	frankfurter	1
+	auf Fuß setzen	frei	0.997
+	in Gang sein	voll	0.992

Table 10: Candidate triples sorted by their preference for a specific adjective.

	PNV-triple	adjective	position
+	auf Bank schieben	lang	1
+	mit Dingen zugehen	recht	2
+	mit Beispiel vorangehen	gut	3
-	auf Welt geben	ganz	4
+	in Ordnung sein	beste	5
+	auf Fuß setzen	frei	15
+	in Gang sein	voll	64
-	mit Wirkung bestellen	sofortig	106
+	zu Fixing verbilligen	frankfurter	140

Table 11: The top-5 candidate triples sorted by morpho-syntactical fixedness in combination with the percentage of adjective occurrences and adjective distribution. 'Position' refers to the position of the triple in the sorted list.

tion. Although some of the trivial items with a strong preference for adjectives (cf. table 10) could be relegated to lower ranks, its UAP-value of 0.566 is very low.

While it is relatively easy to find idiomatic MWEs without adjectives, it is difficult to divide triples commonly containing adjectives into idioms with a restricted range of possible adjectives (cf. table 5) and idiomatic expressions with unrestricted combination possibilities.

5. Conclusion

We showed methods to extract German idiomatic multiword expressions along with their morpho-syntactic features from dependency-parsed text. By storing all features, as well as all typical lexical modifiers, complements, etc. of a given MWE, as extracted from the parsing output, we can identify MWE candidates of different complexity. In addition, the stored features and a manually created small gold standard list of ca. 1000 idiomatic MWEs allowed us to start assessing the usefulness of individual features or feature combinations for the identification of idiomatic MWEs. As morpho-syntactic fixedness and syntactic preferences are neither necessary nor sufficient condidtions for idiomaticity, an approach that combines both types of features leads to better accuracy in idiom identification. The context parameters identified for individual MWEs can directly be reused in the lexical description of the expressions: among others, they are inserted into the electronic collocation dictionary under construction in the work described by (Spohr, 2008).

Monolingual context parameters are one type of indicators of idiomaticity. Another one is semantic transparency vs. opaqueness, as it is identified, for example, in work by (Villada Moirón and Tiedemann, 2006), (Fritzinger, 2009), by use of translational behaviour. In the same spirit as with combining morpho-syntactic and structural syntactic features, we are experimenting with a combined use of both types of indicators and find results quite promising.

Another strand of future work concerns the separation of incomplete shorter and complete longer versions of multiword expressions. We intend to carry out this work on larger chunks of text from the web.

6. References

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vcn-tokens dataset. In *Proceedings of the LREC Workshop: Towards a shared task for multiword expressions*, pages 19–22, Marrakech, Morocco.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut fü maschinelle Sprachverarbeitung, University of Stuttgart.

Afsaneh Fazly and Suzanne Stevenson. 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*, pages 337–344, Trento/New Brunswick: ACL.

Fabienne Fritzinger, Frank Richter, and Marion Weller. 2010a. Pattern-based extraction of negative polarity items from dependency-parsed text. In *Proceedings of LREC 2010*, Valletta, Malta.

Fabienne Fritzinger, Marion Weller, and Ulrich Heid. 2010b. A survey of idiomatic preposition-noun-verb triples on token level. In *Proceedings of LREC 2010*, Valletta, Malta.

Fabienne Fritzinger. 2009. Using parallel text for the extraction of german multiword expressions. In *Lexis – E-Journal in English Lexicology*. Issue 4: Corpus Linguistics and the Lexicon.

Nicole Grégoire. 2007. Design and implementation of a lexicon of Dutch multiword expressions. In Nicole Grégoire, Stefan Evert, and Brigitte Krenn, editors, *Proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions*, pages 17–24, Prague, Czech Republic.

Timm Lichte and Jan.-Philipp Soehn. 2007. The retrieval and classification of negative polarity items using statistical profiles. In Sam Featherston and Wolfgang Sternefeld, editors, *Roots: Linguistics in Search of its Evidential Base*, pages 249–266. Mouton de Gruyter, Berlin.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Michael Schiehlen. 2003. A cascaded finite-state parser for german. In *Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, April.

Dennis Spohr. 2008. Requirements for the design of electronic dictionaries and a proposal for their formalisation. In *Proceedings of the XIIIth Euralex International Congress*, Barcelona, Spain.

Caroline Sporleder, Linlin Li, Philip John Gorinski, and Xaver Koch. 2010. Idioms in context: The idix corpus. In *Proceedings of LREC 2010*, Valletta, Malta.

Begoña Villada Moirón and Joerg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context*, Trento, Italy.

Heike Zinsmeister and Ulrich Heid. 2004. Collocations of complex nouns: Evidence for lexicalization. In *Proceedings of KONVENS-2004*, Heidelberg. Springer.