

A Survey of Idiomatic Preposition-Noun-Verb Triples on Token Level

Fabienne Fritzing, Marion Weller, Ulrich Heid

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

Abstract

Most of the research on the extraction of idiomatic multiword expressions (MWEs) focused on the acquisition of MWE types. In the present work we investigate whether a text instance of a potentially idiomatic MWE is actually used idiomatically in a given context or not. Inspired by the dataset provided by (Cook et al., 2008), we manually analysed 9,700 instances of potentially idiomatic preposition-noun-verb triples (a frequent pattern among German MWEs) to identify, on token level, idiomatic vs. literal uses. In our dataset, all sentences are provided along with their morpho-syntactic properties. We describe our data extraction and annotation steps, and we discuss quantitative results from both EUROPARL and a German newspaper corpus. We discuss the relationship between idiomaticity and morpho-syntactic fixedness, and we address issues of ambiguity between literal and idiomatic use of MWEs. Our data show that EUROPARL is particularly well suited for MWE extraction, as most MWEs in this corpus are indeed used only idiomatically.

1. Introduction

The phenomenon of multiword expressions (MWEs) has gained considerable attention in NLP research during the past decade¹. The idiosyncratic behaviour of MWEs on different levels of linguistic description may cause severe problems in NLP applications like e.g. parsers or MT-systems if they are not detected and treated adequately (Sag et al., 2002). Due to the diversity of different MWE phenomena and their frequent occurrence in all kinds of texts, MWEs need to be reliably identified and treated.

1.1. Background

Most of the research on the extraction of idiomatic MWEs focused on the acquisition of MWE types. The procedures made use of several corpus-observable idiosyncratic properties of MWEs: they were identified either based on their co-occurrence frequency (Evert, 2004), their morpho-syntactic fixedness – e.g. (Fazly and Stevenson, 2006), (Bannard, 2007) – or their semantics – e.g. (Lin, 1999), (Baldwin et al., 2003), to name only a few examples.

However, most of these approaches operate on lexical type level, stating, e.g. that *spill+beans* is idiomatic, but not on token level. Contrary to this, we intend to take into account whether a text instance of a potentially idiomatic MWE is actually used idiomatically in a given context or not. In fact, there are a number of idiomatic MWEs that can also have a straightforward literal meaning. It is possible to automatically distinguish the idiomatic from the literal use in the way (Katz and Giesbrecht, 2006) did by using latent semantic analysis. In one of their case-studies they found that two thirds of the occurrences of the German idiom *ins Wasser fallen* (lit.: “to fall into the water”,

idiom.: “to be cancelled”) were idiomatic uses, as opposed to one third literal uses. In the case of *ins Wasser fallen*, the two meanings exhibit the same morpho-syntactic surface form. However, sometimes the surface form may help to distinguish the different idiomatic vs. literal uses. Quite often, morpho-syntactic features also support a separation of “homonymous” idioms, which have the same lexical items as components, or of different (idiomatic) readings of a “polysemous” idiom (see Section 5.2. below). An example of homography is the German idiom *in Gang kommen* which means “to be set in motion” when it appears in singular form without determiner, while the same used in plural form with definite article *in die Gänge kommen*, bear the meaning “to get organised”. A literal meaning is also thinkable, e.g. in singular with definite article *in den Gang kommen*, where it would mean something like “to reach the hallway”. These examples show that it is not sufficient to handle MWEs solely on the basis of the lemmas of their components, but that their context and surface form has also to be taken into account.

To our knowledge, (Katz and Giesbrecht, 2006) were so far the only authors who investigated the automatic identification of idiomatic vs. literal uses of German MWEs. For English, however, there has been some more work in this field recently: this includes unsupervised methods like e.g. (Sporleder and Li, 2009) who make use of lexical cohesion in order to recognise different uses of idiomatic MWEs or (Fazly et al., 2009) who use combined knowledge of canonical forms and context information; there have also been supervised methods like (Diab and Bhutata, 2009), who used the MWEs’ context and surface form features in a classification approach based on machine learning.

1.2. Objectives

In the present work we do not investigate new methods for MWE classification in context. Instead, we take one step back and present a German resource that could be useful

¹Cf. e.g. the ACL-sponsored workshops on multiword expressions, such as *Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (2009)* or *A Broader Perspective on Multiword Expressions (2007)*.

for future supervised methods, similar to e.g. (Diab and Bhutada, 2009) and for evaluation of extraction tools.

Inspired by the VNC-Tokens Dataset of (Cook et al., 2008), consisting of ca. 3,000 manually annotated corpus sentences of 53 English verb-noun combinations (VNCs), we created a dataset for German: our set contains the manually analysed results of 77 German preposition-noun-verb triples (PNVs: a frequent pattern among German MWEs) in roughly 9,700 sentences. Each instance is provided along with a detailed morpho-syntactic feature description of the MWE and a classification into either literal or idiomatic use. Some cases cannot be decided, as the context given in the sentence is not sufficient to determine the intended reading. Even though we primarily conceived the dataset to serve as a basis for the development of new supervised MWE classification approaches, we will also discuss some examples based on the quantitative distributions of the different readings and their morpho-syntactic feature preferences. We thereby intend to enhance the awareness of literal uses of presumably idiomatic MWEs.

Previous work in this field (for German) includes a corpus-based study (Hümmer, 2007), where the literal vs. idiomatic meaning of 60 German MWEs (of different structural patterns) was investigated from a linguistic and phraseological point of view. Finally, we are aware of one more dataset for English which is (as (Cook et al., 2008) and ours) also conceived to serve future supervised extraction methods, namely the IDIX corpus of (Sporleder et al., 2010). It covers 50 English idioms (mainly V+NP and V+PP) in roughly 5,000 instances and will be available as an add-on to the BNC XML edition.

1.3. Outline

In the following, we first present our preprocessing and extraction procedures (section 2) and then describe the manual filtering and classification into literal vs. idiomatic occurrences (section 3). In section 4, we give a quantitative overview of our results and discuss some cases in more detail in section 5, before we turn to some concluding remarks in section 6.

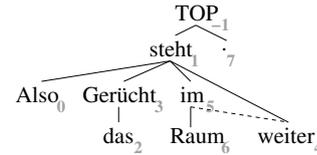
2. Preprocessing

2.1. Data

We use two corpora: a German newspaper (*Frankfurter Allgemeine Zeitung*, FAZ) which we expect to contain both, literal and idiomatic uses of idiomatic PNVs and the proceedings of the European parliament, EUROPARL (Koehn, 2005), which we assume to mainly exhibit idiomatic uses:

description (short name)	size	years
<i>Frankfurter Allgemeine Zeitung</i> (FAZ)	70Mio	97/98
European parliament debates (EUROPARL)	35Mio	96-06

In German, the constituent words of (verbal) multi-word constructions do not always occur adjacently. An example sentence for the PNV *im Raum stehen* (lit. “stand in the room”, to be dealt with) is given below. The whole sentence is translated as “Thus, the rumour is still to be dealt with”.



(a) Tree representation

0	Also	ADV	also		1	ADJ
1	steht	VVFIN	stehen	3:Sg:Pres:Ind*	-1	TOP
2	das	ART	d		3	SPEC
3	Gerücht	NN	Gerücht	Nom:N:Sg	1	NP:1
4	weiter	ADV	weiter		1 5	ADJ
5	im	APPRARTin		Dat:M:Sg	1	ADJ
6	Raum	NN	Raum	Dat:M:Sg	5	PCMP
7	.	.\$.		-1	TOP

(b) FSPAR output

Figure 1: Dependency analysis of the example sentence.

Also steht das Gerücht weiter im Raum .
 Thus stands the rumour still in the room

2.2. Parsing

In order to reliably extract PNVs, a deep syntactic analysis is essential, due to the above mentioned non-adjacency phenomena. Furthermore, as a by-product of parsing, we get a full morpho-syntactic analysis of a PNV’s constituent words.

In the past, we successfully used the dependency parser FSPAR (Schiehlen, 2003) for several different MWE extraction tasks, e.g. (Fritzinger, 2009), (Weller and Heid, 2010). FSPAR is highly efficient and relies on a large lexicon. Its output, given in Figure 1 (b), is to be read as follows:

- 1st column: position of a word in the sentence
- 2nd column: token
- 3rd column: part of speech²
- 4th column: base form (lemma)
- 5th column: morpho-syntactic information (case, gender, etc.)
- 6th column: dependency relation: position of the word’s governor
- 7th column: grammatical function (subject, object, etc.)

The dependency tree representation in Figure 1(a) is not provided by the parser; we inserted it here in order to enhance readability of the example. As can be seen from Figure 1, the noun *Raum* is dependent on the preposition *im*: the 6th column in the noun’s row (cf. Fig. 1 (b)) points to sentence position 5, where the preposition is located. Analogously, the preposition *im* is dependent on the verb *steht*.

In case of structural or labelling ambiguities, the parser provides an underspecified output, as e.g. for the attachment of *weiter* (1||5) in Figure 1: it is either dependent on *steht* (1) or *im* (5).

²based on the STTS tagset:

<http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html>

2.3. Extraction

PNVs are extracted by applying PERL scripts on the dependency-parsed corpora. Besides the PNV’s base form, numerous morpho-syntactic features (which are contained in the parsing output) are collected and stored in a PostgreSQL database together with the sentence in which the PNV occurred (see (Weller and Heid, 2010) for details). Hence, all information is kept easily accessible. An extract of the features for the example sentence (*Also steht das Gerücht weiter im Raum.*) is given below:

PNV	DET	FUS	NUM	VLAST	NEG
in Raum stehen	def	+	Sg	-	noneg

where DET contains information about the noun’s determiner (e.g. definite, indefinite, possessive, etc.), FUS indicates whether the preposition is conflated with the determiner (“Fusion”, as it is the case here: *in + dem = im*), NUM refers to the noun’s number, VLAST shows if the sentence at hand is a verb-last sentence and finally, NEG indicates whether the PNV occurred in a negative context (values: neg or noneg).

The 10 most frequent PNV types retrieved from FAZ and EUROPARL, respectively, are given in Table 1, idiomatic expressions being bold-faced. All triples in our tables are given as lemma sequences.

(a) FAZ			
PNV			freq.
um	Prozent	steigen	5,902
auf	DM	steigen	3,512
zu	Verfügung	stehen	2,762
bei	Prozent	liegen	2,586
um	Prozent	erhöhen	2,483
um	Prozent	wachsen	2,225
um	Leben	kommen	2,160
zu	Verfügung	stellen	2,061
auf	DM	erhöhen	1,985
um	Prozent	steigern	1,765

(b) EUROPARL			
PNV			freq.
zu	Ausdruck	bringen	4,995
von	Bedeutung	sein	4,962
zu	Kenntnis	nehmen	2,740
um	Uhr	stattfinden	2,725
nach	Tagesordnung	folgen	2,586
zu	Verfügung	stehen	2,042
für	Bericht	stimmen	1,812
zu	Verfügung	stellen	1,784
in	Frage	stellen	1,739
für	Arbeit	danken	1,687

Table 1: Most frequent PNVs, idiomatic ones **bold**.

3. Annotation

3.1. Filtering

In order to identify which of the extracted PNV triples can be idiomatic, (cf. Table 1), we looked up the 1000 most frequent PNV-triples of each corpus in a standard printed dictionary of German idiomatic expressions (Duden, 2002).

There are some idiomatic MWES that contain sequences of a preposition, a noun and a verb, but the triple alone does not bear the idiomatic meaning: an example is *Licht ins Dunkel bringen* (lit.: “to bring light into the dark”, to reveal sth.), containing the sequence *ins Dunkel bringen* which has the form of a PNV. In the present study, we excluded such expressions from our data.

The remaining MWES that we found in (Duden, 2002), were filtered in order to finally contain only idiomatic expressions that, according to our intuition, potentially also can have a literal meaning³. This leads us to the distribution given in Table 2. Most of the expressions we investigated occurred in both corpora. In total, we annotated sentences for 77 different MWES.

Corpus	in Duden	lit.&idiom.	only idiom.
FAZ	155	69	86
EUROPARL	108	35	73
Total	196	77	119

Table 2: Idiomatic MWES amongst the most frequent 1000.

3.2. Classification

For each of the 69 MWES from FAZ and 35 MWES from EUROPARL that can have a literal and an idiomatic interpretation (cf. Table 2), we randomly extracted 100 sentences together with several morpho-syntactic features from our database. For MWES that occurred less than 100 times, all available sentences are extracted. This leads to a total of 9,740 sentences: 6,690 sentences for the 69 MWES from FAZ and 3,050 sentences for the 35 MWES from EUROPARL.

Then, these sentences were annotated independently by two native speakers of German⁴ with respect to the literal (L) vs. idiomatic (I) use of their MWES. In some cases, the actual reading could not clearly be determined. These were marked as ambiguous (A). Furthermore, we also observed a number of extraction errors which we marked separately (X). Two examples of such extraction errors are given in Table 3. The extracted PNV triple is highlighted in **bold** face, while the correct dependency structure is underlined. These errors appear due to the fact that FSPAR leaves structural and label ambiguities unresolved in its output. In favour of recall we decided to extract all existing triples. For example, the sentence given in Table 3(b) was wrongly assigned to the extraction results for (*etwas*) *auf Weg bringen* (lit.: “to bring (sth.) on the way”, to get sth. started). However, it contains another MWE *zu Fall bringen* (lit.: “to bring down s.o./sth.”), which would be the correct MWE to extract from this sentence. Some more error examples observed in the output of our tools can be found in (Weller and Heid, 2010). Both annotators analysed all of the 9,740 extracted MWE instances independently from one another. Out of the 6,690 sentences from FAZ, they agreed in their annotation in 6,550 cases (97.9%). The remaining 140 cases of disagreement are distributed over literal vs. idiomatic use (69

³There are numerous expressions that can never bear a literal meaning (e.g. *in Betracht ziehen*, to take sth. into consideration), that are not relevant for the present study.

⁴Both being computational linguists and authors of this paper.

(a) PP-adjunct

Es gibt wenig Berichte aus erster Hand, aber viele Spekulationen.
 “There are few first-hand reports, but many speculations.”

(b) Wrong MWE detection

...soll auf rechtlichem Wege zu Fall gebracht werden.
 “...should on the juridical way to fall brought be”
 ...should be caused to fall by means of juridical procedures

Table 3: Examples for typical extraction errors.

cases), literal vs. ambiguous use (17 cases) and finally, 54 cases of disagreement concerning idiomatic vs. ambiguous use. These problematic cases were carefully discussed with other German native speakers in order to find a reliable final annotation.

4. Quantitative Results

Table 4 contains the total numbers of literal vs. idiomatic uses across all PNVs investigated. As expected, the overall proportion of literal uses is considerably higher in the newspaper corpus FAZ (5.15%: 345 of 6,690) than in EUROPARL (1.02%: 31 of 3,050).

corpus	all	idiom.	lit.	amb.	extr.err.
FAZ	6,690	6,176	345	75	94
EUROPARL	3,050	2,937	31	14	68

Table 4: Distribution of literal and idiomatic uses.

Table 5 gives an overview of the quantitative distribution of the readings of all 77 MWEs we considered. Items marked “-” did not occur amongst the 1000 most frequent MWEs of the respective corpus.

PNV	FAZ			EP		
	L	I	A	L	I	A
an Tagesordnung sein	0	88	0	-	-	-
an Tag legen	0	95	0	0	99	0
auf Bank sitzen	32	59	8	-	-	-
auf Bein stellen	1	98	0	-	-	-
auf Fahne schreiben	0	87	1	-	-	-
auf Hand liegen	0	99	0	0	99	0
auf Kopf stellen	3	93	3	-	-	-
auf Platz verweisen	3	93	0	-	-	-
auf Programm stehen	0	98	0	-	-	-
auf Seite haben	10	81	2	-	-	-
auf Seite sein	15	53	7	3	40	0
auf Seite stehen	16	75	3	12	86	1
auf Spiel stehen	0	99	1	0	100	0
auf Straße gehen	1	94	0	4	54	0
auf Spur kommen	1	97	0	-	-	-
auf Strecke bleiben	0	94	2	0	80	2
auf Weg bringen	3	85	2	0	97	2
auf Weg machen	3	87	0	0	38	0
aus Auge verlieren	0	100	0	-	-	-
aus Hand geben	0	85	0	-	-	-
in Auge behalten	0	100	0	0	99	0
in Auge fassen	0	100	0	0	100	0
in Auge haben	9	83	0	0	85	0
in Auge sehen	-	-	-	0	78	3

PNV	FAZ			EP		
	L	I	A	L	I	A
in Aussicht stellen	0	100	0	0	100	0
in Betrieb gehen	1	99	0	-	-	-
in Betrieb nehmen	0	100	0	-	-	-
in Betrieb sein	7	91	0	-	-	-
in Bild passen	0	99	0	-	-	-
in Bild setzen	1	88	1	-	-	-
in Buch stehen	32	65	0	-	-	-
in Gang sein	0	100	0	0	100	0
in Gang setzen	-	-	-	0	100	0
in Gespräch sein	6	94	0	-	-	-
in Grenze halten	0	100	0	-	-	-
in Griff bekommen	0	100	0	-	-	-
in Griff haben	0	100	0	-	-	-
in Hand fallen	1	99	0	-	-	-
in Hand haben	18	75	6	7	80	3
in Hand halten	55	41	3	-	-	-
in Hand legen	3	91	0	-	-	-
in Hand nehmen	22	73	4	2	96	0
in Hintergrund treten	0	99	1	-	-	-
in Kopf haben	1	98	0	-	-	-
in Lage sein	2	97	0	-	-	-
in Raum stehen	33	53	2	-	-	-
in Schatten stehen	5	92	3	-	-	-
in Schatten stellen	0	98	1	-	-	-
in Spiel sein	0	90	6	0	62	0
in Szene setzen	0	99	0	-	-	-
in Vordergrund rücken	0	100	0	0	100	0
in Vordergrund stehen	0	97	3	0	100	0
in Vordergrund stellen	0	100	0	0	100	0
in Weg leiten	-	-	-	0	100	0
mit Fuß treten	-	-	-	0	94	1
über Bühne gehen	3	96	1	-	-	-
unter Arm greifen	1	84	0	-	-	-
unter Druck setzen	3	96	1	0	100	0
unter Druck stehen	2	98	0	0	76	0
unter Lupe nehmen	0	100	0	0	75	0
unter Teppich kehren	-	-	-	0	54	0
vor Auge führen	0	100	0	-	-	-
vor Auge halten	-	-	-	0	100	0
vor Tür stehen	21	71	8	2	60	0
zu Ausdruck bringen	0	100	0	-	-	-
zu Ausdruck kommen	0	100	0	0	100	0
zu Einsatz kommen	1	97	1	1	98	0
zu Fall bringen	17	82	1	0	47	0
zu Kasse bitten	0	86	0	-	-	-
zu Schau stellen	0	100	0	-	-	-
zu Tragen kommen	0	100	0	-	-	-
zu Verkauf stehen	0	100	0	-	-	-
zu Wehr setzen	0	98	2	-	-	-
zu Zug kommen	0	98	1	-	-	-

Table 5: Literal (L), idiomatic (I), and ambiguous (A) uses.

Not all of the MWEs in Table 5 have 100 annotated instances; this happens for two reasons: on the one hand, there are MWEs that occurred less than 100 times in the respective corpus⁵, on the other hand, this concerns the extraction errors that we briefly addressed in the previous section. For certain MWEs there are more extraction errors than for others: e.g. *(sich) auf Weg machen*

⁵Note that we restricted the sentence length to 40 words in order to get more reliable parsing results.

(lit.: "make (oneself) on the way", to hit the road) is often extracted in the context of *Fortschritte auf diesem Weg machen* ("make progress on this way") where *auf diesem Weg* ("on this way") is an adjunct of *Fortschritte machen* ("make progress").

The MWE *vor Tür stehen* can intuitively have a literal meaning "to stand outside the door" as well as an idiomatic meaning "be imminent". We claim that in everyday speech native speakers have no clear preference for one of the two readings if the MWE is presented to them without context. Not surprisingly, though, the idiomatic meaning is clearly prominent in EUROPARL (60 idiomatic vs. 2 literal occurrences), whereas for FAZ the instances can also often have literal meanings: one fifth (21 of 100) of all occurrences are literal, as opposed to 71 idiomatic instances and 8 ambiguous ones (cf. Table 5).

Consider also e.g. the MWE *zu Fall bringen*: in this case, German native speakers may first think of the idiomatic meaning "to cause a regime (or politician etc.) to fall" which is the only one found in EUROPARL, and then realise that the literal one "to literally bring down sb." (E.g. in boxing, or by tripping someone up) is just as thinkable. This latter meaning in fact shows up several times in the newspaper corpus FAZ, mostly in the context of soccer. Another example that frequently occurs in the domain of soccer (or any other sports), though mostly in its idiomatic meaning, is *auf Bank sitzen* (lit.: "sit+on+bench", idiom.: to sit on the bench = not to be allowed to join the match).

Furthermore, it can be seen from Table 5 that there are several MWEs that never occur in their literal use. The number of such MWE types for each of the corpora is given in Table 6. There are two explanations to that phenomenon: firstly, it is possible that due to chance no literal meaning is amongst the randomly selected 100 instances, but on the other hand, we also have to admit that we were quite generous in the filtering process (see Section 3.1. above) in that we kept all MWEs in which at least the noun can have a literal meaning. Maybe we will remove MWEs appearing only in their idiomatic use in a future version of our dataset.

	all	excl. idiomatic	other readings
FAZ	69	26	43
EP	35	20	15

Table 6: MWE types never occurring literally.

5. Discussion

5.1. Morpho-syntactic preferences

Table 7 shows the distributions of morpho-syntactic features over literal vs. idiomatic uses of the PNVs *in Raum stehen* (lit. "to stand in the room", to be dealt with) and *in Buch stehen* (lit. "to stand in the book", to be a text-book example). The numbers refer to occurrences in FAZ. The idiomatic use of *in Raum stehen* always occurs with a definite article, conflated with the preposition (cf. FUS = +), while the literal uses allow for variations of the determiner (different determiners, conflated or not). Similarly,

(a) in Raum stehen

	#PNVs	DET			FUS		NUM	
		no	def	quant	+	-	sg	pl
lit.	33	2	24	1	19	14	28	5
idiom.	53	0	53	0	53	0	53	0
amb.	2	0	2	0	2	0	2	0

(b) in Buch stehen

	#PNVs	DET			NUM		VLAST	
		no	def	pos	sg	pl	+	-
lit.	32	7	16	3	21	11	6	26
idiom.	65	0	58	7	62	3	18	47

Table 7: Distribution of morpho-syntactic features in FAZ.

(a) Literal use

In den besseren Räumen standen Öfen, an denen er sich wärmte.
 "In the superior rooms stood stoves, at which he himself warmed."
 In the superior rooms were stoves at which he warmed himself.

(b) Idiomatic use

Widersprüche stehen ungelöst im Raum.
 "Contradictions stand unresolved in the room"
 There are unresolved contradictions to be dealt with.

Table 8: Examples taken from FAZ for *in Raum stehen*.

the number of the PNV's noun can be either singular or plural in its literal use, while in the idiomatic use, only singular is possible. See Table 8 for example sentences of both uses. A polysemous MWE that has more than one idiomatic meaning is *in+Buch+stehen*. As can be seen from Table 7(b), almost any combination of number and determiner settings is possible in its literal use. Table 9 contains two examples for literal uses. While the first one is straightforwardly literal (Table 9(a): "text written in a book"), the second one is metaphoric (Table 9(b): "book of nature"), where the metaphor is based on the literal meaning.

(a) Literal use

Was nicht in den Büchern stehe, spiele im Unterricht keine Rolle.
 "What not in the books stands, plays in the class no role"
 What is not written in the books, doesn't matter in class.

(b) Metaphorical literal use

...die Wahrheit über ihn steht im Buch der Natur;
 "...the truth about him stands in the book of nature;"
 ...the truth about him is written in the book of nature;

Table 9: Examples taken from FAZ for *in Buch stehen*.

Besides its literal meaning, *im Buch stehen* also has two idiomatic meanings which not only differ in terms of their semantics, but also show clearly diverging morpho-syntactic preferences. In the first reading (cf. Table 10(a)), the construction requires the word *wie* ("how"). Furthermore, the noun often occurs in the older German dative form with "-

(a) Idiomatic use (i)

...ist ein deutscher Bildungsbürger, wie er **im Buche steht**.
“...is a German educated citizen, as he in the book stands.”
...is a textbook example of a German educated citizen.

(b) Idiomatic use (ii)

In den Büchern der Banken stehen riesige Summen...
“In the books of the banks stand huge amounts...”
The accounts of the banks hold huge amounts...

(a) Idiomatic use (i)

Wir müssen der Wahrheit ins Auge sehen.
“We must the truth in the eyes look.”
We have to face the truth.

(b) Idiomatic use (ii)

...den europäischen Verbrauchern **in die Augen sehen** wollen
“...the European consumers in the eyes look want”
... want to have a pure conscience vis-a-vis
the European consumers’

(c) Ambiguous use

...ich sitze nah genug,
um Ihnen in die Augen sehen zu können.
“...I sit close enough, to you in the eyes look can.”
...I sit close enough, to be able to look into your eyes.

Table 10: Examples taken from FAZ for *in Buch stehen*.

e” at its end (*im Buche*) and the number of the noun in this reading is only singular (cf. Table 7(b)). In contrast to this, the noun occurs mostly in plural with a definite article in the second reading (Table 10(b)). Here, the meaning of the MWE is closely related to the domain of accounting.

5.2. Ambiguous Cases

The distinction into literal vs. idiomatic use is not always as clear as shown for the examples of Table 7 in the previous section. We give three examples of the MWE *in Auge sehen* (lit.: “to look in the eye(s)”, idiom.: to face, to have a pure conscience⁶) in Table 11.

The first reading 11(a) “to face sth.” typically occurs with a closed set of abstract nouns like e.g. *Wahrheit* (truth), *Tatsache* (fact), *Problem* (problem) and *Realität* (reality).

The second reading, 11(b) “to have a pure conscience” is still idiomatic. Example 11(b) is however less clear, in that the individuals forming the group of European consumers in fact do have eyes, as opposed to the abstract “truth” in (a). However, it is not likely that this speaker of the European parliament in fact stands in front of European consumers and literally looks into their eyes. A similar borderline case between metaphoric use and idiomaticity is e.g. *in Gespräch sein* (lit.: “to talk”, idiom.: to be a candidate, to negotiate). Here, the second idiomatic meaning “to negotiate” implies the literal meaning “to talk”. It is not clear where the distinction between the two readings is to be made, and the annotation heavily depends on the annotators intuition. For these cases, we thus found annotations after debates with other German native speakers.

In contrast to 11(b), we leave the classification of *in die Augen sehen* unanswered in the third interpretation (given in (c)). Here, the literal meaning “to look into s.o.’s eyes” and the idiomatic one “to have a pure conscience” are both equally possible. As the speaker is close enough to the person he is talking to, he might in fact be able to look into his eyes, and at the same time the addressee should better not lie to him.

Similar conflicts arise for the polysemous MWE *in+Hand+nehmen* (lit.: “to take sth. into the hand”, idiom.: to take over control, to make use of sth.). As for the first idiomatic use of *ins Auge sehen* (Table 11(a)), *in die Hand nehmen* exhibits the idiomatic reading “to take over control” only if used together with a closed set

⁶This meaning is derived from the fact that if someone looks into your eyes, he/she might be able to identify whether you are telling the truth or not. Having a pure conscience you do not have to fear that someone looks into your eyes.

Table 11: Examples from EUROPARL for *in Auge sehen*.

of nouns, like e.g. *Steuer/Ruder* (steering wheel/oar), *Heft* (haft), *Zepter* (sceptre), *Zügel* (rein), *Fäden* (wires) etc. However, even here ambiguities might arise, e.g. in case of a pilot “taking the steering wheel into his hands” while approaching for landing (*Der Pilot nimmt beim Landeanflug das Steuer in die Hand*). Here, the pilot might take over control (i.e. the autopilot is switched off for landing) and at the same time might (literally) grasp the steering wheel.

Instances of the second idiomatic meaning of *in+Hand+nehmen* showed that in a sentence like *Kaum ein Buch habe er je ein zweites Mal in die Hand genommen...* (there is almost no book that he ever [took into his hand/read] a second time...), again the idiomatic meaning “to make use” (here: to read) coincides with the literal meaning “to take into the hand”.

6. Conclusion and Future Work

The analysis of the ca. 9,700 sentences provides interesting evidence: for 26 of the analysed 69 MWEs in the newspaper FAZ (37%) we only found idiomatic uses, while for EUROPARL, 20 of 35 (57%) MWEs were exclusively used idiomatically. These differences are attributed to the different text types and communication situations. Given the lower density of non-idiomatic uses in EUROPARL, this corpus is perhaps an “easier” source for tools extracting idiomatic MWEs than newspaper data.

The correlation between morpho-syntactic fixedness and idiomaticity seems to hold, but a simple listing of MWE lemmas is not always enough to clearly identify idiomatic expressions, as some PNV-triples belong to “homographous” idioms (*in+Buch+stehen*) or have several (some times non-trivially linked) idiomatic readings (polysemy). These cases complicate manual annotation and would require more research into semantic processing, for a clear automatic separation. Finally, PP-attachment errors indeed reduce the precision of a parsing based extraction that is aware of morpho-syntactic features. It will be a task for future work to enhance the extraction tools to be able to remove such cases from the extraction result.

The annotated dataset, we believe, can be used for training machine-learning approaches to idiom identification,

for the evaluation of idiom extraction tools, but also for a more detailed analysis of the properties of idiomatic MWES and for an investigation into the conditions of ambiguity between idiomatic and non-idiomatic interpretation.

7. References

- Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (ACL 2003)*, pages 89–96.
- Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a broader Perspective on Multiword Expressions (ACL 2007)*, pages 1–8, Prague, Czech Republic.
- Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the Workshop: Towards a shared task for multiword expressions (LREC 2008)*, pages 19–22, Marrakech, Morocco.
- Mona T. Diab and Pravin Bhutada. 2009. Verb noun construction mwe token supervised classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (ACL-JICNLP 2009)*, pages 17–22, Singapore.
- Duden. 2002. *Nr. 11 - Redewendungen: Wörterbuch der deutschen Idiomatik*. Bibliographisches Institut & F.A. Brockhaus AG, Mannheim.
- Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. University of Stuttgart, PhD dissertation.
- Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11st Conference of the European Chapter of the ACL (EACL 2006)*, pages 337–344, Trento, Italy.
- Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1).
- Fabienne Fritzingier. 2009. Using parallel text for the extraction of german multiword expressions. *Lexis - E-journal in English Lexicology*, 4.
- Christiane Hümmer. 2007. Meaning and use: a corpus-based study of idiomatic mwus. In Christiane Fellbaum, editor, *Idioms and Collocations*, pages 138–151. Continuum International Publishing Group Ltd.
- Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties (ACL 2006)*, pages 12–19, Sydney, Australia.
- Phillip Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceedings of the 10th MT Summit 2005*, Phuket, Thailand.
- Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguists (ACL 1999)*, pages 317–324, Maryland, USA.
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, pages 1–15, Mexico City, Mexico.
- Michael Schiehlen. 2003. A cascaded finite-state parser for german. In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL 2003)*, Budapest, Hungary.
- Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece.
- Caroline Sporleder, Linlin Li, Philip John Gorinski, and Xaver Koch. 2010. Idioms in context: The idix corpus. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.
- Marion Weller and Ulrich Heid. 2010. Multi-parametric extraction of german multiword expressions from parsed corpora. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.