Lexis
Journal in English Lexicology

4 | 2010
Corpus Linguistics and the Lexicon

Using parallel text for the extraction of German multiword expressions

Fabienne Fritzinger

Electronic version
URL: http://journals.openedition.org/lexis/564
DOI: 10.4000/lexis.564
ISSN: 1951-6215

Publisher
Université Jean Moulin - Lyon 3

Electronic reference
Fabienne Fritzinger, « Using parallel text for the extraction of German multiword expressions », *Lexis* [Online], 4 | 2010, Online since 14 April 2010, connection on 10 December 2020. URL : http://journals.openedition.org/lexis/564 ; DOI : https://doi.org/10.4000/lexis.564

Using parallel text for the extraction of German multiword expressions[1]

Fabienne Fritzinger[2]

Abstract

A procedure for the identification of semantically opaque (i.e. idiomatic) German multiwords is presented. We focus on verb + PP combinations that are lexicographically relevant (extracted via dependency parsing [Schiehlen 2003]) of the kind ins Leben rufen – "to initiate", lit.: "to call into life". Starting from [Villada Moirón and Tiedemann 2006], the method exploits the fact that opaque combinations are translated as a whole, whereas compositional uses would show regular, individual translations of the words involved. The translations into other languages are obtained by applying GIZA++ [Och and Ney 2003] word alignment to the EUROPARL corpus [Koehn 2005]. Numerous experiments are performed to further optimise the original method: several parameters are analysed individually as well as in combination with each other. This leads to the following results: depending on the actual parameter settings, values between 0.800 and 0.936 (in terms of uninterpolated average precision) are reached amongst the highest scoring 200 multiword candidates, as opposed to a baseline of 0.584, using the 200 most frequent multiwords in decreasing order of their occurrence frequency.

Keywords: multiword expressions – multilingual corpus – dependency parsing – statistical word alignment

[1] This paper is (partly) based on my diploma thesis submitted to the University of Stuttgart in August 2008.

[2] University of Stuttgart, Institute for Natural Language Processing, Azenbergstr.12, 70174 Stuttgart, Germany: fritzife@ims.uni-stuttgart.de

1. Introduction

1.1. Definitions

Roughly speaking, a Multiword Expression (MWE) consists of at least two words and is to be used as a whole. All MWEs have in common that they are idiosyncratic i.e. somehow peculiar, either in terms of their appearance or in their behavior. One example of idiosyncratic behaviour would be non-compositional (or opaque) semantics of an MWE: the meaning of the whole expression is not derivable from the meaning of its component parts, e.g. shoot the breeze. These are often referred to as idioms. Other typical examples of MWEs include multiword nominals, e.g. New York or grocery store, whole phrases like let the cat out of the bag, and constructions like multiword prepositions and adverbs: in spite of, by and large.

There is a wide range of different definitions of the phenomenon of Multiword Expressions. Bannard [2007] describes MWEs as follows:

> A multiword expression is usually taken to be any word combination (adjacent or otherwise) that has some feature (syntactic, semantic, or purely statistic) that cannot be predicted on the basis of its component words and/or the combinatorial process of language.

According to Moon [1998],

> there is no unified phenomenon to describe, but rather a complex of features that interact in various, often untidy ways and represent a broad continuum between non-compositional (or idiomatic) and compositional groups of words.

In Sag et al. [2002] MWEs are defined as

> idiosyncratic interpretations that cross word boundaries.

Additionally, Sag et al. [2002] present an extensive classification into different MWE types. Other such classification attempts can be found in Villada Moirón [2005] and Heid [2008] who both provide valuable insights into the variety of MWEs.

Besides Multiword Expression, there are also other terms like Collocation, Idiom or Fixed Expression that are frequently used to describe the phenomenon. However, these terms do not always denote the same range of MWE characteristics: Moon [1998] defines collocations purely in terms of statistics, i.e. the term is reduced to frequently co-occurring words independent of any semantics. In contrast, Evert [2004] states that

> a collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components,

and in Zinsmeister and Heid [2003]

> collocations are habitual combinations where the collocate cannot easily be substituted, but which are not necessarily non-compositional.

Amongst the characteristics mentioned so far, namely idiosyncrasy, semantic opacity, statistical co-occurrence, unpredictability and habitual usage, our work mainly focusses on the extraction of MWEs that are semantically opaque i.e. the meaning of the whole expression is not derivable from the meaning of its component words. We present a procedure that exploits the fact that an opaque combination is supposed to be translated as a whole, whereas compositional uses would show regular, individual translations of the words involved. The procedure is exemplified for German verb + prepositional phrase combinations (V+PP), but is supposed to be language independent and could furthermore be applied to any kind of MWE construction.

1.2. Relevance for lexicography

Multiword Expressions frequently occur in all kinds of natural language text. It is assumed that there are at least as many MWEs as there are single words (Jackendoff [1997: 156]). The fact that about half of the entries in the semantic online lexicon WORDNET (Fellbaum [1998]) are MWEs supports this assumption and Sag et al. [2002: 2] state that this is even likely to be an underestimate.

Due to the special form and behavior of MWEs, it is important for language learners and users to know them. Especially their opaque semantics might lead to confusion. It is thus reasonable to include the most frequent and most relevant MWE constructions in a lexicon. A corpus-based attempt to extract MWEs has the advantage that the retrieved expressions are part of the language in use. Furthermore, example sentences can easily be extracted from the corpus. They should be included in lexicons in order to provide the language learner/user with context information about the expressions. Corpus retrieval leads to huge amounts of data, that can hardly be processed manually. It is thus useful to apply (semi-) automatic methods to pre-classify the data set into valid vs. non-valid MWEs in order to minimize the costs for manual processing.

1.3. Related Work

In the field of automatic MWE identification procedures, there are three main trends observable in the literature (classification adopted from Heid [2008]): MWE identification based on their i) statistical, ii) syntactical and iii) semantic properties. In the following, some examples for each of the trends are given:

In the early 1990s, Church and Hanks [1990] applied a variant of the mutual information (mi) measure to automatically identify word associations. The underlying idea of this approach is to divide the joint probability of a word pair (x, y) by the probabilities of observing x and y independently from one another. This relatively simple statistical measure works well for word pairs, but unfortunately, it is not applicable for MWEs consisting of more than two words. A few years later, Smadja [1993] presented a tool that overcomes this shortcoming and that at the same time accounts for syntactic consistency of the identified MWEs. First, it identifies highly associated word pairs based on statistical measures. Then, two filters are applied to these word pairs: one to extend them to MWEs of arbitrary length, while the other filter scans for syntactic consistency of the MWEs and automatically rejects invalid candidates. An extensive survey of statistical word association measures, including their mathematical background, can be found in Evert [2004].

As to syntactic approaches to MWE extraction, Fazly and Stevenson [2006] investigate the idiomaticity of verb phrases by combining the degree of syntactic fixedness (in terms of article use, singular vs. plural, word order) with the variability that MWEs exhibit with respect to the selection of their lexical components. Another way to use knowledge on syntactic variation to identify verb phrases as MWEs is presented by Bannard [2007]. He distinguishes three different types of syntactic variation, namely: the addition, dropping or variation of a determiner, modification of the noun (e.g. by introducing an adjective), and passivisation of the verb. The variation is determined by counting the presence or absence of the respective features. If the phrase as a whole exhibits less flexibility than expected based on the flexibility of its parts, it is assumed to be a valid MWE.

Finally, there exist a number of approaches that make use of special characteristics in terms of semantics in order to identify MWEs. In Lin [1999], mutual information (mi) scores of MWEs as they occurred in the text and various counterparts of these MWEs (where any of the components is replaced by a similar word) are compared. If there is a significant difference between the score of the original MWE and that of its substituted variant, the original is supposed to be a valid MWE candidate. On the other hand, one of the procedures described by Baldwin et al. [2003] makes use of Latent Semantic Analysis to measure the similarity between an MWE and its components. Here, the underlying assumption is that the less similar the whole construction and its parts turn out to be, the more likely it is supposed to be an MWE.

The procedure presented in this paper belongs to the semantics-based line of work. It aims at identfying MWEs using their opaque semantics. To approximate the semantics of an MWE, the procedure makes use of the MWE's translations into another language. It builds on the assumption that semantically opaque constructions are typically translated as wholes, whereas compositional uses would show regular, individual translations of the words involved. Statistical word alignment is applied to get respective translations of MWEs and two different scores are then used to rank MWE candidates in decreasing order of opacity. The first implementation of this procedure was published by Villada Moirón and Tiedemann [2006], who investigated the identification of Dutch MWEs. Fritzinger [2008] contains a detailed description of a re-implementation for German MWEs.

1.4. Outline

Figure 1 shows an architectural sketch of the procedure (taken from Fritzinger [2008]). It can be divided into three parts: the (monolingual) extraction of MWE candidates (section 2), supplying translations for these candidates (section 3) and finally, ranking the candidates in decreasing order of opacity using their translations (section 4).

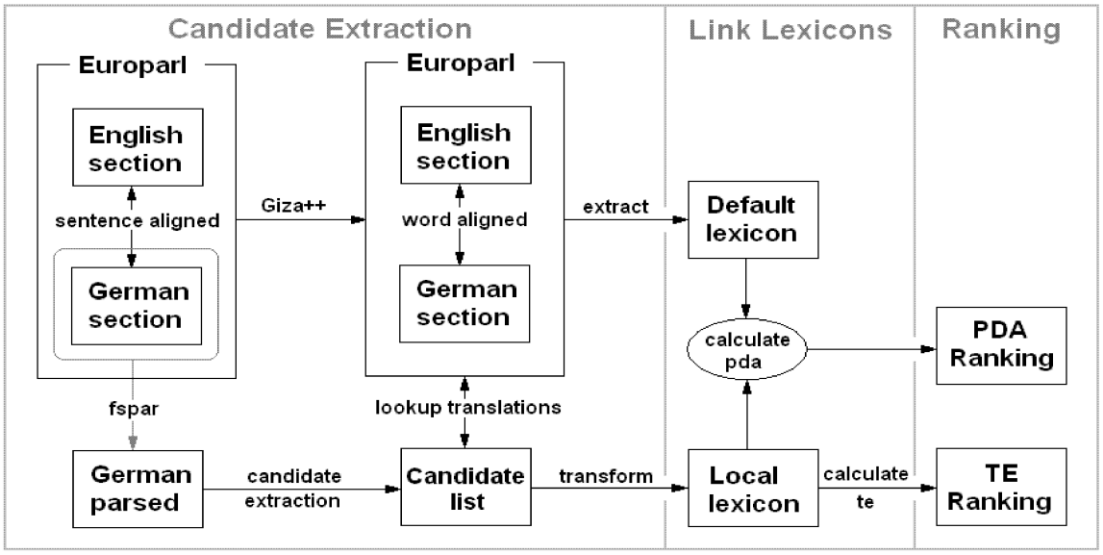


Fig. 1. Architectural sketch of the procedure

Section 5 reports on several experiments that were performed to further optimize the ranking results, before section 6 closes with a conclusion and outlook.

2. Extraction of Multiword Expressions

2.1. Data Material

In order to confirm that the procedure does in fact work to identify semantically opaque MWEs, it is reasonable to apply it to a huge text corpus, ideally available in several different languages. Koehn [2005] used the proceedings of the European Parliament debates of 1996 to 2006 to prepare a parallel corpus named EUROPARL. This corpus is freely available for download in 11 official European languages[3] and contains about 35 million words and roughly 1.5 million sentences per language. Currently, this corpus is widely used for statistical natural language applications of different kinds.

2.2. Candidate Extraction

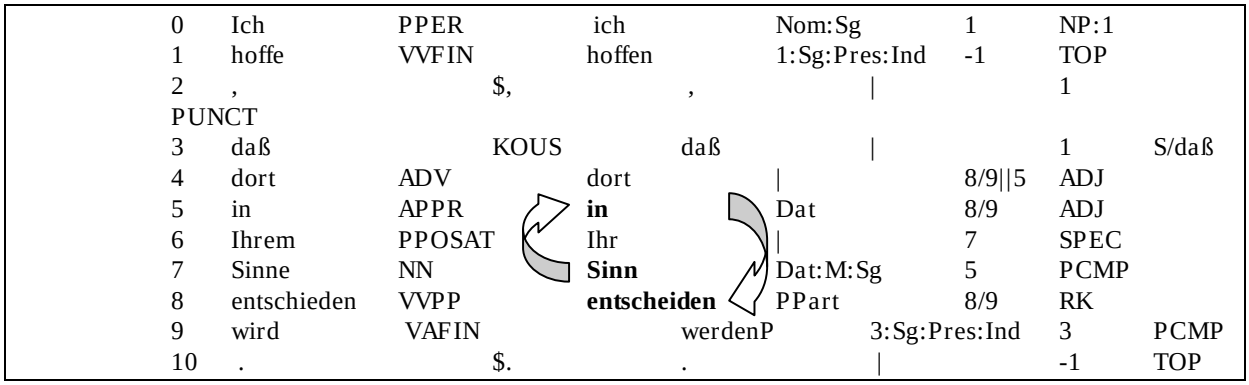
We focus on the extraction of German verb+PP combinations. German is a morphologically rich language and its word order is less strict than e.g. that of English. As a consequence, the verb and its accusative object may be separated by an arbitrary number of intervening words, as shown in Fig. 2.

Hoffentlich zieht die Kommission solche Dinge bei der Planung der Politik in diesem Gebiet ernsthaft in Betracht.

Fig.2. Example of the verb+PP combination in Betracht ziehen – to take into consideration

[3] These languages are the following: Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish.

In order to get reliable information about syntactic relations between words and to simultaneously reduce the words to their lemma forms, the German section of EUROPARL is parsed with FSPAR (Schiehlen [2003]), a finite-state based dependency parser (cf. Figure 3 for an output example). The MWEs are extracted from the parsing output using a PERL script, that works as follows: the script first identifies the nouns of a sentence by means of the third column of the parsing output which contains the part-of-speech information (cf. Fig. 3: POS: NN, Lemma: Sinn). Then, the word(s) which the noun depends on are retrieved by help of the 6th column, containing the dependency structure in terms of sentence positions (cf. Fig. 3: 5 for Sinn). If one of these words turns out to be a preposition (cf. Fig. 3: POS: APPR, Lemma: in), this word's dependency link is followed. Finally, if the preposition depends on a verb or a verb complex (cf. Fig. 3: 8/9, entscheiden werden) the main verb, having a part of speech beginning with VV (here: entscheiden), is retrieved. Thereby, the identification of a verb+PP combination is completed and the lemmas of the actual MWE's constituent words (in this case in Sinn ent-scheiden) are extracted. In this way, about 1.75 million verb+PP combinations (of roughly 950,000 different types) are extracted from the parsed German section of EUROPARL. In the following section, we describe how translations are provided for this MWE candidate list. These translations are later required for ranking them (cf. section 4) according to their semantic opacity.



0	Ich	PPER	ich	Nom:Sg	1	NP:1	
1	hoffe	VVFIN	hoffen	1:Sg:Pres:Ind	-1	TOP	
2	,	$,	,	|	1	PUNCT	
3	daß	KOUS	daß	|	1	S/daß	
4	dort	ADV	dort	|	8/9||5	ADJ	
5	in	APPR	**in**	Dat	8/9	ADJ	
6	Ihrem	PPOSAT	Ihr	|	7	SPEC	
7	Sinne	NN	**Sinn**	Dat:M:Sg	5	PCMP	
8	entschieden	VVPP	**entscheiden**	PPart	8/9	RK	
9	wird	VAFIN	werdenP	3:Sg:Pres:Ind	3	PCMP	
10	.	$.	.	|	-1	TOP	

Fig. 3. Output example of FSPAR

3. Establishing Translation Links

3.1. Word Alignment

For the identification of German MWEs that have an opaque semantics, the MWE's translations into English, French and Swedish are investigated. However, the EUROPARL corpus as it is available for download is only sentence-aligned. To get word equivalences, the corpus first requires an alignment at word level. The statistical word alignment toolkit GIZA++ implemented by Och [2003] is used to provide this alignment. In order to avoid data sparseness and also to improve alignment quality, it is reasonable to run GIZA++ on lemmatized and lowercased texts. The German section of EUROPARL was already parsed in the course of the monolingual extraction of MWE candidates. The parsing result is re-used to extract a lemmatized version of the German section. As lemmatization is often a by-product of tagging, a freely available probabilistic tagger (Schmid [1994]) is applied to lemmatize the English and French sections of EUROPARL. Because we did not have access to a Swedish parser nor tagger, the Swedish section is only stemmed using the Snowball stemmer list, consisting of about 30,000 words (Snowball, [2008]). The lemmatized German, English,

French and Swedish sections of EUROPARL are lowercased with PERL before undergoing word alignment. GIZA++ is then run in its default configuration for each of the translation directions Foreign language to German. Figure 4 contains an example alignment: it can be divided into three parts: the first line contains the unique ID of the sentence pair, as well as the lengths of the participating sentences and an alignment score that indicates the quality of the word alignment. The numbers in brackets behind the French part (which is the source language) refer to positions of words in the German sentence and thereby map e.g. concurrence ({1}) to wettbewerb. Words without equivalence in the target language are mapped to NULL.

```
# Sentence pair (888) source length 18 target length 13 alignment score : 5.20089e-26
wettbewerb ja , einschränkung von beihilfe , wo nötig und wo möglich !
NULL ({3 7 11}) le ({}) concurrence ({1}) : ({}) oui ({2}) . ({}) le ({}) limitation ({4}) du ({5})
aide ({6}) : ({}) là ({}) où ({8}) ce ({}) être ({}) nécessaire ({9}) et ({10}) possible ({12}) . ({13})
```

Fig. 4. GIZA++ output, French to German

Fritzinger [2008] showed that the ranking procedure works best for 1:1 translation links. However, GIZA++ always produces 1:n alignments from source to target language. German is a language with a considerable proportion of morphologically rather complex words, while other languages (as e.g. English or French) tend to use several words to express the same concept. As a consequence, the alignment from Foreign language to German contains more 1:1 links than the respective other direction does, and this alignment direction is thus more suitable for our approach than the opposite one.

3.2. Link Lexicons

Based on the output of GIZA++ word alignment (as shown in Figure 4 above), two types of link lexicons are created: the default lexicon and the local lexicon. These two lexicons are required for the calculation of the two ranking scores (cf. Section 4).

wettbewerb	= concurrence (5443), compétition (195), compétitivité (138), compétitif (99)
einschränkung	= restriction (460), limitation (219), limiter (100), limite (76)
nötig	= nécessaire (889), besoin (166), falloir (144), NULL (104)

Fig. 5. Extract of the default lexicon French to German

For the default lexicon, all established translation links on word level are extracted across the whole text, without any contextual information. Then, the links are counted and the four most frequent links (called default alignments in Villada Moirón and Tiedemann [2006]) for each word are retained. An example is given in Figure 5. The numbers in brackets indicate the frequency of the link: e.g. wettbewerb has been translated by concurrence 5443 times, by compétition 195 times, etc. The default lexicon has to be established once for each of the language pairs under investigation. It will later serve to calculate the proportion of default alignments (cf. section 4.2).

To get translational equivalences for MWEs, local link lexicons are required. For each MWE and each language pair, a new local link lexicon has to be established. In contrast to the default lexicons, they only contain the translation links of those sentences in which the

MWE's component words actually form that MWE. Figure 6 shows a local link lexicon example for the MWE am Ball bleiben[4] (lit.: to stay on the ball, fig.: to keep at it).

an	= NULL (8), à (1)
ball	= bouillir (1), profil (1), brèche (1), habile (1), cap (1), attendre (1), surveiller (1), mouvement (1), pousser (1)
bleiben	= NULL (2), conserver (1), rester (1), continuer (1), habile (1), maintenir (1), marmite (1), pousser (1)

Fig. 6. Local link lexicon for am Ball bleiben, French to German

4. Ranking According to Opacity

4.1. Translational Entropy Score

In statistics, entropy measures the amount of information in a random variable (Manning and Schütze [1999]). Here, translational entropy (henceforth: te, but $H(T_s|s)$ in the formula given in Fig. 7) is used to express the translational inconsistency of a word (cf. Melamed [1997]). It is calculated using the following formula (taken from Melamed [1997]):

$$H(T_s|s) = -\sum_{t \in T_s} P(t|s) \log P(t|s)$$

Fig. 7. Formula for translational entropy calculation

where T_s is the set of all translation links from the source word s into different target words t. The hypothesis is that MWE constructions having a transparent or compositional semantics exhibit a lower diversity of translations than non-compositional MWEs. Thus, the te score is supposed to be low for compositional MWEs and high for non-compositional ones. An example calculation of the te score for the non-compositional MWE am Ball bleiben is shown in Figure 8, cf. also the local link lexicon in Figure 6 above. In contrast to this MWE, which has a te score of 1.530, a more transparent construction like an Macht bleiben – to remain in power, yields a te score of 1.184.

te_{an}	= - [(8/9 ln 8/9) + (1/9 ln 1/9)]	≈	0.348
te_{ball}	= - [9 * (1/9 ln 1/9)]	≈	2.197
$te_{bleiben}$	= - [(2/9 ln 2/9) + (7 * (1/9 ln 1/9))]	≈	2.043
$te_{an\_ball\_bleiben}$	= $(te_{an} + te_{ball} + te_{bleiben}) / 3$	≈	1.530

Fig. 8: Calculation of te score for am Ball bleiben

4.2. Proportion of Default Alignments

The proportion of default alignments (pda) score indicates how many of an MWE's local translation links are also default alignments. The following formula (taken from Villada Moirón and Tiedemann [2006]) is used for pda calculation:

[4] Note that am is short for an dem. Therefore it appears as an in its lemmatized form.

$$pda(S) = \frac{\sum_{s \in S} \sum_{d \in D_s} align\_freq(s,d)}{\sum_{s \in S} \sum_{t \in T_s} align\_freq(s,t)}$$

Fig. 9. Formula for the calculation of pda

where S is the whole MWE, T_s is the set of all translation links from the source word s into different translated words t, while d stands for their default translations. D_s contains the word's default alignments and aligned_freq(x,y) is the frequency of translation links from word x to word y in the context of the MWE S (here: the verb + prepositional phrase combination). The hypothesis behind this score is the following: the more default alignments are found among the translation links of the local link lexicon, the more likely it is that the MWE at hand has a compositional, rather transparent semantics.

an	= au (17), NULL (15), à (6), respecter (1)
macht	= pouvoir (24), halabja (1), force (1), avatar (1), dictateur (1), panache (1), place (1), prêtes (1)
bleiben	= rester (17), conserver (3), maintenir (3), NULL (3), prêtes (1), accrocher (1), reste (1), avatar (1)

Fig. 10. Local link lexicon for an Macht bleiben, French to German

To get an impression of how the pda score works, consider first the local link lexicons of (i) the non-compositional MWE am Ball bleiben (cf. Figure 6 above) and (ii) the transparent construction an Macht bleiben (cf. Figure 10 above). An extract of the default lexicon, covering all words of both expressions is given below, in Figure 11. Comparing the local link lexicons with the default lexicon, it can be seen that e.g. none of the local links of the word Ball belongs to the set of its default alignments, while the most frequent local entry for Macht, namely pouvoir, is a default alignment. This observation gives a first intuition about the compositionality of the two MWEs.

an	= NULL (66277), à (20302), au (15287), participer (3741)
ball	= balle (64), ballon (19), baller (2), petits-enfants (1)
macht	= pouvoir (1861), puissance (396), force (94), possible (68)
bleiben	= rester (1852), NULL (1305), maintenir (198), demeurer (185)

Fig. 11. Extract of the default lexicon French to German

The detailed pda calculations for both expressions are given in Figures 12 (a)+(b). Roughly speaking, the frequency of all those local alignments that are also default alignments is divided by the total number of local alignments.

pda_{an}	= 8/9 $_{(NULL)}$ + 1/9 $_{(à)}$	= 9/9	= 1
pda_{ball}	= 0/9 $_{(no\ matching\ default\ alignment)}$	= 0/9	= 0
$pda_{bleiben}$	= 2/9 $_{(NULL)}$ + 1 $_{(rester)}$ + 1 $_{(maintenir)}$		= 4/9 ≈ 0.4445
$pda_{an\_ball\_bleiben}$	= (pda_{an} + pda_{ball} + $pda_{bleiben}$) / 3		≈ **0.4815**

Fig. 12 (a). pda calculation for am Ball bleiben, French to German

pda_{an}	= 17/39 $_{(au)}$ + 15/39 $_{(NULL)}$ + 6/39 $_{(à)}$	= 38/39	≈ 0.9744
pda_{macht}	= 24/31 $_{(pouvoir)}$		= 24/31 ≈ 0.7742
$pda_{bleiben}$	= 17/30 $_{(rester)}$ + 3/30 $_{(maintenir)}$ + 3/30 $_{(NULL)}$		= 23/30 ≈ 0.7667
$pda_{an\_macht\_bleiben}$	= (pda_{an} + pda_{macht} + $pda_{bleiben}$) / 3		≈ **0.8492**

Fig. 12 (b). pda calculation for an Macht bleiben, French to German

The examples show that the pda score of 0.4815 for am Ball bleiben is considerably lower than the pda score of 0.8492 for an Macht bleiben and these results confirm our hypothesis.

4.3. Evaluation Methodology

4.3.1. Judgement

The procedure aims at automatically ranking MWEs with the result that on the top of the list there is supposed to be a high density of semantically opaque MWEs. In order to be able to evaluate the procedure, the candidate MWEs first need to be classified according to their semantics into compositional vs. non-compositional expressions. As there is no adequate (electronic) lexicographic resource to look up valid MWEs, we carried out a human judgement on the data. To avoid personal preferences and thereby influencing the ranking results, the highest scoring 200 candidates of all lists were merged into one single list which is then classified according to semantic opacity. This classification was then re-introduced into the result lists of the different parameter settings, before the evaluation metrics were applied.

4.3.2. Baseline

The performance of the procedure is compared to a baseline ranking: it contains all MWE candidates, ranked in decreasing order of their frequency of occurrence. One typical characteristic of MWEs is their conventionality: in the course of time, certain word combinations became established MWEs in everyday language use. Moreover, as valid MWEs usually do not allow for much lexical variation, they are expected to occur more frequently than occasional multiword constructions. Figure 13 shows the 10 most frequent MWEs taken from the baseline ranking (the last column indicates the frequency of occurrence).

+	zu	Ausdruck	bringen		5027
+	von	Bedeutung	sein	4974	
-	nach	Tagesordnung	folgen	2922	
+	in	Lage	sein	2830	
+	zu	Kenntnis	nehmen		2731
+	zu	Verfügung	stehen	2046	
-	um	Uhr	stattfinden	1998	
-	für	Bericht	stimmen		1889
+	zu	Verfügung	stellen	1789	
-	für	Arbeit	danken	1738	

Fig. 13. Top 10 candidate MWEs of the baseline ranking

4.3.3. Uninterpolated Average Precision

The uninterpolated average precision (abbreviated uap) is a metric originating from the field of Information Retrieval. It is used to reflect the quality of a ranking procedure. According to Manning and Schütze [1999: 535], the uap "aggregates many precision numbers into one evaluation figure". It is calculated using the following formula:

$$uap = \frac{\sum_{S_c} P(S_1..S_c)}{|S_c|}$$

Fig. 14. uap formula

where P(S1..Sc) is the precision of true positives. The uap is calculated at each point c where a true positive is found in the ranked list, then, it is averaged over all precision points (Villada Moirón and Tiedemann [2006]). The baseline ranking given in Figure 13 above is used to give an exemplary uap calculation: (*1/1 + 2/2 + 3/4 + 4/5 + 5/6 + 6/9) / 6 ≈ 0.841*. The denominators are the positions where true positives (i.e. valid MWEs) are found in this ranking: 1, 2, 4, 5, 6 and 9. The numerators are the numbers of true positives found up to (and including) this position.

4.4.4. Precision (proportion of true positives)

The uap metric reflects only the quality of the ranking, it does not reflect the number of true positives found. To give an example, imagine a list of 100 words where only the first word is a true positive. The uap for this list amounts to 1.000, which indicates perfect ranking. However, it would be useful to know that only one true positive was found at all. The precision is therefore additionally indicated to evaluate the rankings. It is calculated by simply dividing the number of true positives by the size of the ranked list, and reflects the percentage of true positives in the list. The precision is henceforth abbreviated ptp, proportion of true positves.

5. Experiments

5.1. Parameters

To optimise the results of the procedure, different parameter configurations are explored:

- all: original method (no parameter applied)
- freq25: the MWE candidate list is reduced to those that occurred at least 25 times.
- freq50: the MWE candidate list is reduced to those that occurred at least 50 times.
- wnl (without Null Links): Null Links are removed from the link lexicons.
- wpr (without prepositions): the values of the prepositions are ignored in the calculations.

5.2. Results (1): single parameters

All experiments are performed for each of the three language pairs (FR-DE, EN-DE, SE-DE) and for both ranking metrics (te and pda). The resulting ranked lists are all evaluated using uap and ptp scores. The procedure is first applied to the plain data (all). The comparison of the other parameters to all makes the impact of the individual parameters visible. The results of the rankings with single parameters are given in Figure 15 (a)-(d).

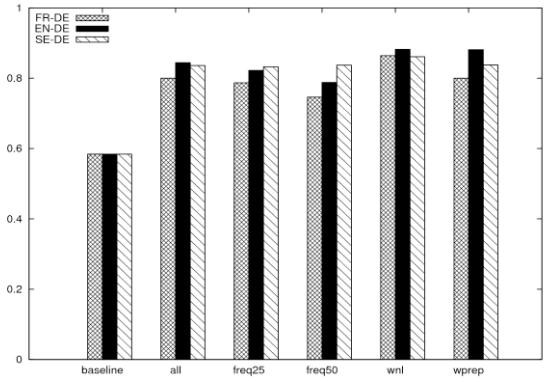


Fig. 15 (a). ranked by te, evaluated with uap

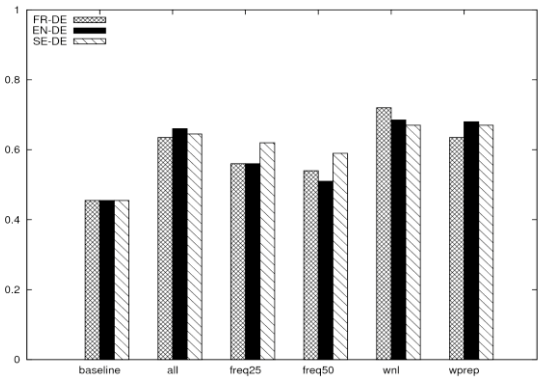


Fig. 15 (b). ranked by te, evaluated with ptp

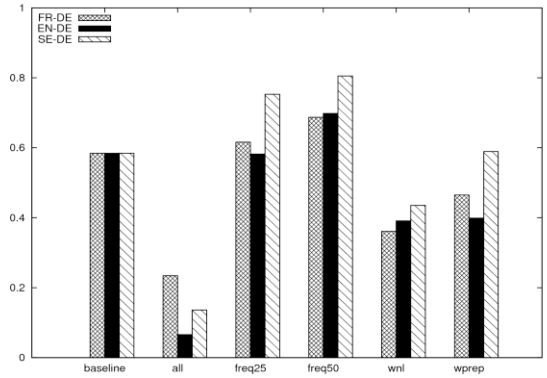


Fig. 15 (c). ranked by pda, evaluated with uap

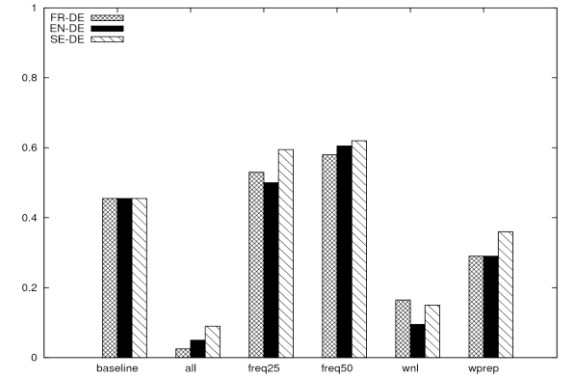


Fig. 15 (d). ranked by pda, evaluated with ptp

Considering these figures, it is obvious that across all parameter settings, the te ranking score outperforms the baseline and all pda ranking attempts. Furthermore, the te ranking exhibits a more stable performance in contrast to pda. The latter yields reasonable results only for settings where a frequency constraint is involved, i.e. it beats the baseline only for the two parameters freq25 and freq50. On the other hand, if the pda rankings perform well, their uap values can get quite close to the respective results for te. Consider e.g. freq50 for the language pair Swedish to German: here, the uap value of te ranking amounts to 0.837, closely followed by the pda ranking's uap value of 0.805. One reason for the bad performance of pda ranking without parameters lies in the fact that MWEs occurring only once or very few times often tend to yield pda scores of 0 and thus reach high positions in the course of the ranking process. For te, this risk of false positives that have a very low frequency and thus could have a negative influence on the result is considerably lower: high entropy scores result from a high diversity of links, and to exhibit a high diversity of links, a certain frequency of occurrence of the MWE is required. As to the comparison of the two evaluation metrics uap and ptp, a closer look at Figures 15(c)+(d) reveals that in the setting all, the two metrics give different impressions on the best performing language pair: in terms of uap, it is French to German that performs best, but at the same time, this pair also performs worst in terms of ptp. However, for most of the other parameter settings, the two metrics express similar trends.

5.3. Results (2): combined parameters

This section reports on the impact of combined parameters on ranking quality. First, the constraints wnl (without Null links) and wpr (without prepositions) are combined with the two

frequency constraints freq25 and freq50. See Figures 16 (a)+(b) for a comparison of te and pda ranking in terms of uap values.

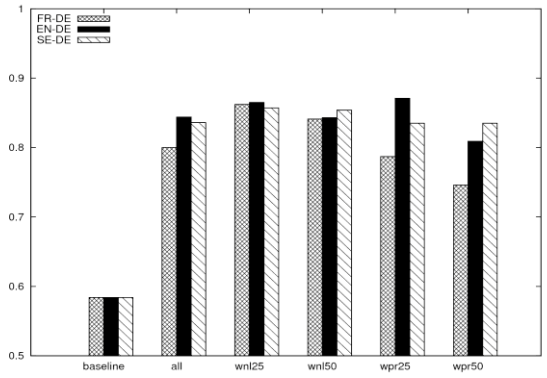


Fig. 16 (a). ranked by te, evaulated with uap

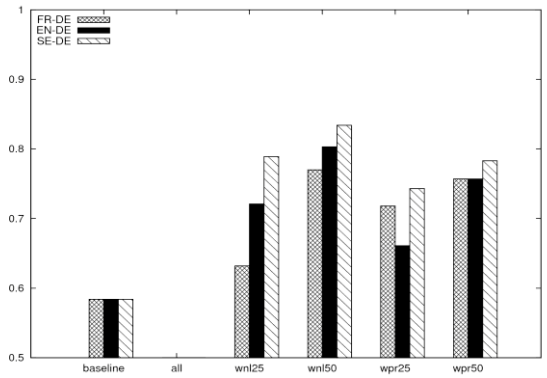


Fig. 16 (b). ranked by pda, evaluated with uap

Even though the te ranking for the combinations is sometimes worse than that for single parameter settings (e.g. in case of wnlf25, where wnl performs better), it still outperforms pda ranking across all settings (either single parameters or combined ones). Nevertheless, pda ranking quality has considerably improved, compared to its performance for single parameter rankings: this time, each of the parameter combinations outperforms the baseline and the results of single parameter settings. This observation is also consistent with the above hypothesis that pda is affected to the frequencies of MWEs.

In a second series of experiments, the combination of the wnl and wpr parameters together with the two frequency constraints is investigated. These results are reported in Figures 17 (a)+(b). It is not surprising that the pda ranking quality is again best when a frequency constraint is involved. As for te ranking, the combination wnlwpr turns out to yield better results than any of its two variants with frequency constraints: for French to German, a uap value of 0.911 is reached for the wnlwpr setting, which is the best overall result so far.

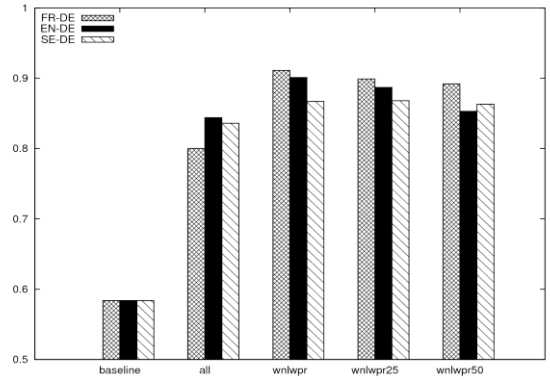


Fig. 17 (a). ranked by te, evaluated with uap

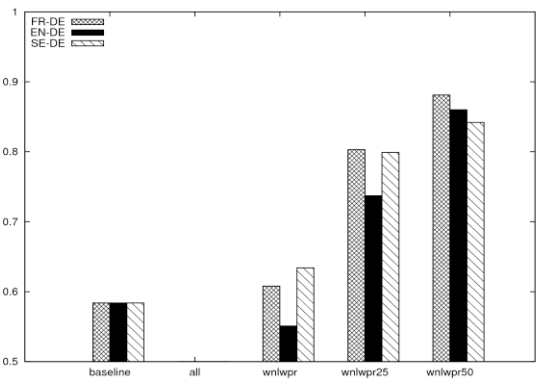


Fig. 17 (b). ranked by pda, evaluated with uap

5.4. Results (3): merging language pairs

Up to now, the outcomes of all experiments have been reported for each of the three language pairs under investigation. Even if there are slight variances observable in the results for the different pairs, there is no explicitly best performing pair. Whether or not the different

language pairs contain similar candidates in the 200 best performing candidates of each list is a question that has not yet been answered. Therefore, the impact of combining several language pairs is investigated, under the hypothesis that if there was a certain similarity of the top scoring candidates across the language pairs, their combination should yield similar or even better results.

The data obtained from ranking the candidate lists of each of the three different language pairs was merged into one list for each of the best performing parameter combinations, as reported in (Figures 17 (a)+(b) above). Out of this merged list, the te and pda values of repeatedly occurring MWE candidates were summed up and averaged. Based on these new te and pda scores, the lists were re-ranked and finally their top 200 MWE candidates were evaluated. Figures 18 (a)+(b) report on the results of this procedure. Most te rankings exhibit a uap of at least 0.9, and even pda reaches the 0.9 mark for the two combinations where frequency constraints are involved. Again, the best overall score (0.936) is reached for te ranking in the parameter combination wnlwpr, i.e. without any frequency constraint. It can thus be stated that translational entropy is less sensitive to frequencies of occurrence than pda. The latter however, outperforms the te ranking results, when frequency constraints are active: wnlwpr50 yields an uap of 0.916 for te and an uap of 0.927 for pda. In terms of ptp, even wnlwpr25 with a ptp of 0.815 for pda is able to surpass te ranking which has a ptp of 0.760. Note that a ptp of 0.815 reflects that 81.5% of the 200 candidates were true positives, i.e. 163 out of 200 MWE candidates had a non-compositional (i.e. opaque) semantics.

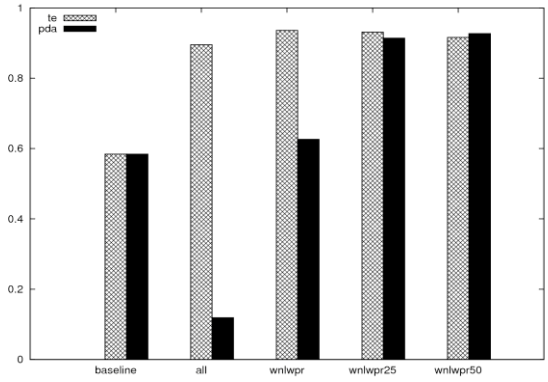


Fig. 18 (a). language combination, eval. with uap

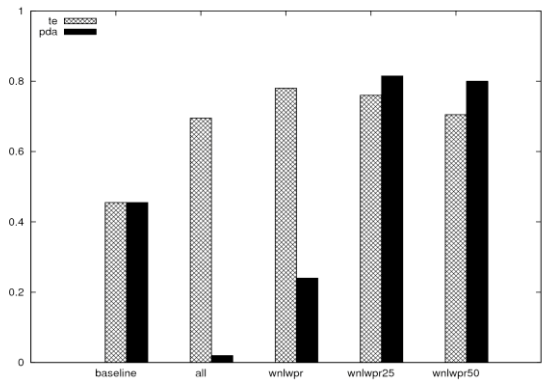


Fig.18 (b): language combination, eval. with ptp

5.5. Recall

In terms of precision, the procedure has proved to produce satisfying results. However, there is so far no intuition about the recall efficiency of the method. Obviously, it would be too time-consuming to manually account for a classification of the 950,000 MWE candidates that are extracted from the German section of the EUROPARL corpus. To break down the data to a reasonable size, only those candidates that occurred at least 25 times were thus manually classified. The resulting list contains 4,566 MWE candidates, whereof 545 are classified to be valid MWEs, i.e. having an opaque semantics. If the frequency constraint is set to 50, there remain 343 candidates from this list. Figures 19 (a)+(b) show the absolute numbers of retrieved true positives amongst the top 200 candidates obtained with those parameter settings where one of the frequency constraints was used. These figures reflect the results of the combination of all three language pairs.

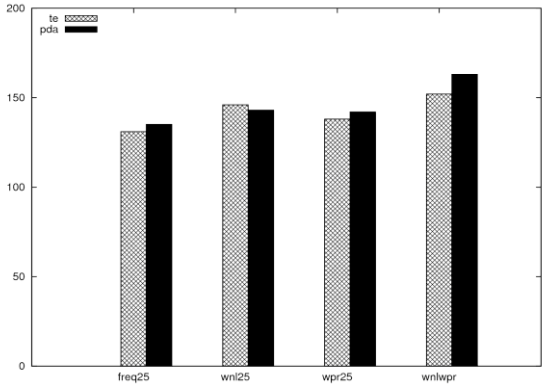


Fig. 19 (a). absolute recall of freq25.

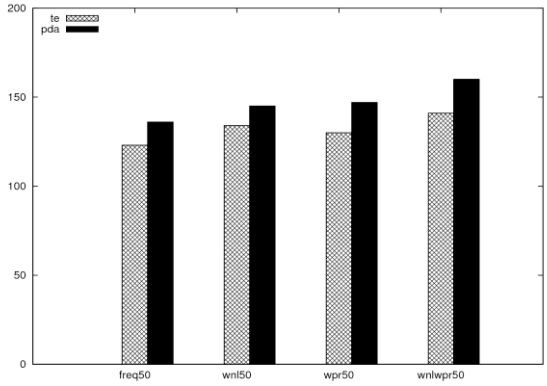


Fig. 19 (b). absolute recall of freq50

These results show that across all settings, the pda rankings always (except once for wnl25) outperform te ranking quality in terms of recall. In terms of the different frequency thresholds of 25 and 50, respectively, there is no significant difference observable in the number of retrieved MWEs.

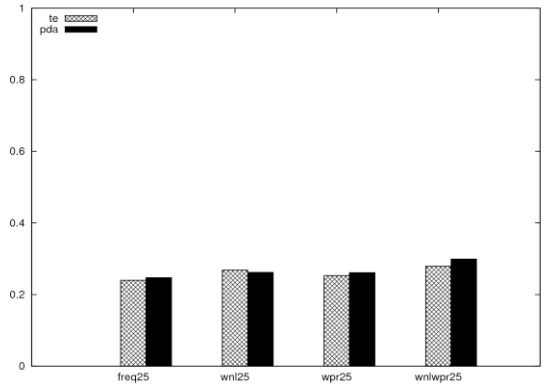


Fig. 20 (a). relative recall of freq25

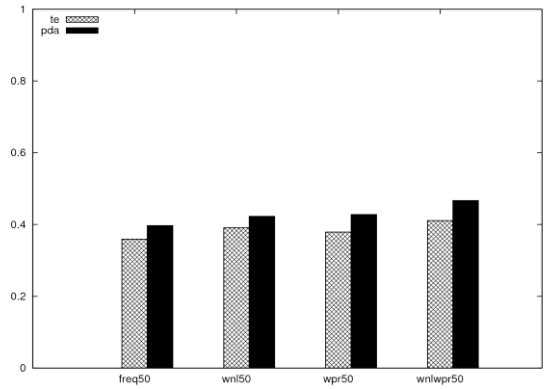


Fig. 20 (b). relative recall of freq50

However, if the relative recall results are compared, the setting with the higher frequency restriction (freq50) performs much better. Out of the 343 MWE candidates that occurred more than 50 times in EUROPARL, 160 are found amongst the top 200 of the best pda ranking (wnlwprf50, cf. Fig. 20 (b)), which amounts to about 46.6% (i.e.160 of 343). In contrast, the best pda ranking (wnlwprf25,) for MWE candidates occurring at least 25 times contained 163 candidates out of a total of 545 in the top 200 (≈ 29.9%).

6. Conclusion and Future Work

We presented a procedure for the automatic extraction of Multiword Expressions from parallel text. It proved to successfully rank semantically opaque constructions at the top of the resulting lists. Already when applied without any parameters, the te ranking is able to clearly outperform the baseline ranking (cf. Fig. 15 (a)). It reaches a uap score of 0.800 for the language pair English to German, compared to a baseline of 0.584 when ranked according to frequency of occurrence. If the three language pairs are combined, this value is even higher,

namely 0.895 (cf. Fig. 18 (a)). The pda score, on the other hand, is very sensitive to low frequency data and it yields reasonable results only when frequency constraints are applied.

We conducted a series of different experiments in order to further improve the procedure's outcomes: e.g. if NULL links and the values of the prepositions are left out of the calculations, better results are observable across all settings. Moreover, all experiments were performed for each of the three language pairs under investigation. We used two Germanic languages, namely English and Swedish, as well as one Romance language, namely French. As the procedure builds on the contrast of different translations of MWEs into another language, we would have expected that the French-German language pair turns out to perform best, as these two languages are the most contrasting ones (out of our set). In fact, this was sometimes the case, cf. Fig. 17 (a)+(b), but not always as Fig. 16 (a)+(b) show. It was only the combination of the three language pairs that showed a stable improved performance.

In terms of precision, the procedure reaches values of up to 0.936 uap for te ranking ignoring NULL-links and the values of the prepositions (cf. Fig. 18(a), wnlwpr). This best result is reached using the combined translations from all three languages (English, French and Swedish) when calculating the ranking scores. If the lists are ranked according to pda, which is a relatively simple scoring metric, its best performance comes very close to the best one for te: a uap of 0.927 is reached, when all three language pairs are used, the candidates' frequencies of occurrence is restricted to at least 50, and furthermore, the NULL-links and the values of the prepositions are left out in the calculation (cf. Fig. 18 (a), wnlwpr50).

The results prove that in terms of precision, the procedure works reasonably well. As to recall, however, we have (due to the time consuming manual evaluation methodology) so far no satisfactory answer. Even though the recall was investigated for frequently occurring MWE candidates, this issue shall be further addressed in the future.

Bibliography

BALDWIN Timothy, **BANNARD** Colin, **TANAKA** Takaaki and **WIDDOWS** Dominic, "A statistical approach to the semantics of verb-particles", Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, Japan, 2003: 89-96.

BANNARD Colin, "A measure of syntactic flexibility for automatically identifying multiword expressions in corpora", Proceedings of the ACL Workshop on a Broader Perspective on Multiword Expressions, Prague, Czech Republic, 2007: 1-8.

CHURCH Kenneth and **HANKS** Patrick, "Word Association Norms, Mutual Information, and Lexicography", Computational Linguistics, Vol. 16.1, 1990: 22-29.

EVERT Stefan, The Statistics of Word Cooccurrences: Word Pairs and Collocations, Ph.D. thesis, University of Stuttgart, 2004.

FAZLY Afsaneh and **STEVENSON** Suzanne, "Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations", Proceedings of the EACL, Trento, Italy, 2006: 337-344.

FELLBAUM Christiane (ed.), WordNet: "An Electronic Lexical Database", the MIT Press, 1998

FRITZINGER Fabienne, Extracting Multiword Expressions from Parallel Text, Diplomarbeit, Universität Stuttgart, 2008.

HEID Ulrich, "Computational phraseology: an overview", GRANGER Sylviane and MEUNIER Fanny (eds.), Phraseology – an Interdisciplinary Perspective, John Benjamins Publishing Company, 2008: 337-360.

JACKENDOFF Ray, The Architecture of the Language Faculty, Cambridge, MA: the MIT Press, 1997.

KOEHN Philipp, "Europarl: A parallel corpus for statistical machine translation", Proceedings of 10th MT Summit, Phuket, Thailand, 2005: 79-86.

LIN Dekang, "Automatic Identification of Non-compositional Phrases", Proceedings of the 27th ACL, Maryland, USA, 1999: 317-324.

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

MELAMED I. Dan, "Measuring Semantic Entropy", ACL-Siglex Workshop Tagging with Lexical Semantics: Why, What and How, Washington, D.C., USA, 1997: 41-46.

MOON Rosamund, "Fixed Expressions and Idioms in English" Clarendon Press, Oxford, 1998

OCH Franz Josef and **NEY** Hermann, "A systematic comparison of various statistical alignment models", Computational Linguistics, Vol. 29.1, 2003: 19-51.

SAG Ivan A., BALDWIN Timothy, BOND Francis, COPESTAKE Ann A. and FLICKINGER Dan, "Multiword expressions: a pain in the neck for NLP", Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002), Mexico City, Mexico, 2002: 1-15.

SCHIEHLEN Michael, "A cascaded finite-state parser for German", Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), Budapest, Hungary, 2003: 163-166.

SCHMID Helmut, "Probabilistic part-of-speech tagging using decision trees", Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994: 44-49.

SMADJA Frank, "Retrieving Collocations from Text: Xtract", Computational Linguistics, Vol. 19.3, 1993: 143-177.

Snowball stemming algorithm, http://www.snowball.tartarus.org/index.php, retrieved in August 2008.

VILLADA MOIRÓN Begoña, Data-driven identification of fixed expressions and their modifiability, Ph.D. thesis, University of Groningen, 2005.

VILLADA MOIRÓN Begoña and **TIEDEMANN** Jörg, "Identifying idiomatic expressions using automatic word alignment", Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006): Workshop on multiword expressions in a multilingual context, Trento, Italy, 2006: 33-40.

ZINSMEISTER Heike and **HEID** Ulrich, "Significant triples: Adjective+noun+verb combinations", Proceedings of COMPLEX, Conference on Computational Lexicography and Text Research, Budapest, Hungary, 2003: 92-101.