Exploiting Translational Correspondences for Pattern-Independent MWE Identification

Sina Zarrieß
Department of Linguistics
University of Potsdam, Germany
sina@ling.uni-potsdam.de

Jonas Kuhn
Department of Linguistics
University of Potsdam, Germany
kuhn@ling.uni-potsdam.de

Abstract

Based on a study of verb translations in the Europarl corpus, we argue that a wide range of MWE patterns can be identified in translations that exhibit a correspondence between a single lexical item in the source language and a group of lexical items in the target language. We show that these correspondences can be reliably detected on dependency-parsed, word-aligned sentences. We propose an extraction method that combines word alignment with syntactic filters and is independent of the structural pattern of the translation.

1 Introduction

Parallel corpora have proved to be a valuable resource not only for statistical machine translation, but also for crosslingual induction of morphological, syntactic and semantic analyses (Yarowsky et al., 2001; Dyvik, 2004). In this paper, we propose an approach to the identification of multiword expressions (MWEs) that exploits translational correspondences in a parallel corpus. We will consider in translations of the following type:

(1) Der Rat sollte unsere Position **berücksichtigen**.
The Council should our position consider.

(2) The Council should **take account of** our position.

This sentence pair has been taken from the German - English section of the Europarl corpus (Koehn, 2005). It exemplifies a translational correspondence between an English MWE *take account of* and a German simplex verb *berücksichtigen*. In the following, we refer to such correspondences as **one-to-many translations**. Based on a study of verb translations in Europarl, we will explore to what extent one-to-many translations provide evidence for MWE realization in the target language. It will turn out that crosslingual correspondences realize a wide range of different linguistic patterns that are relevant for MWE identification, but that they pose problems to automatic word alignment. We propose an extraction method that combines distributional word alignment with syntactic filters. We will show that these correspondences can be reliably detected on dependency-parsed, wordaligned sentences and are able to identify various MWE patterns.

In a monolingual setting, the task of MWE extraction is usually conceived of as a lexical association problem where distributional measures model the syntactic and semantic idiosyncracy exhibited by MWEs, e.g. (Pecina, 2008). This approach generally involves two main steps: 1) the extraction of a candidate list of potential MWEs, often constrained by a particular target pattern of the detection method, like verb particle constructions (Baldwin and Villavicencio, 2002) or verb PP combinations (Villada Moirón and Tiedemann, 2006), 2) the ranking of this candidate list by an appropriate assocation measure.

The crosslingual MWE identification we present in this paper is, a priori, independent of any specific association measure or syntactic pattern. The translation scenario allows us to adopt a completely data-driven definition of what constitutes an MWE: Given a parallel corpus, we propose to consider those tokens in a target language as MWEs which correspond to a single lexical item in the source language. The intuition is that if a group of lexical items in one language can be realized as a single item in another language, it can be considered as some kind of lexically fixed entity. By this means, we will not approach the MWE identification problem by asking for a given list of candidates whether these are MWEs or not. Instead, we will ask for a given list of lexical items in a source language whether there exists a one-to-many translation for this item in a target language (and whether these

Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, pages 23–30,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

one-to-many translations correspond to MWEs). This strategy offers a straightforward solution to the interpretation problem: As the translation can be related to the meaning of the source item and to its other translations in the target language, the interpretation is independent of the expression's transparency. This solution has its limitations compared to other approaches that need to automatically establish the degree of compositionality of a given MWE candidate. However, for many NLP applications, coarse-grained knowledge about the semantic relation between a wide range of MWEs and their corresponding atomic realization is already very useful.

In this work, we therefore focus on a general method of MWE identification that captures the various patterns of translational correspondences that can be found in parallel corpora. Our experiments described in section 3 show that one-to-many translations should be extracted from syntactic configurations rather than from unstructured sets of aligned words. This syntax-driven method is less dependent on frequency distributions in a given corpus, but is based on the intuition that monolingual idiosyncracies like MWE realization of an entity are not likely to be mirrored in another language (see section 4 for discussion).

Our goal in this paper is twofold: First, we want to investigate to what extent one-to-many translational correspondences can serve as an empirical basis for MWE identification. To this end, Section 2 presents a corpus-based study of the relation between one-to-many translations and MWEs that we carried out on a translation gold standard. Second, we investigate methods for the automatic detection of complex lexical correspondences for a given parallel corpus. Therefore, Section 3 evaluates automatic word alignments against our gold standard and gives a method for high-precision one-to-many translation detection that relies on syntactic filters, in addition to word-alignments.

2 Multiword Translations as MWEs

The idea to exploit one-to-many translations for the identification of MWE candidates has not received much attention in the literature. Thus, it is not a priori clear what can be expected from translational correspondences with respect to MWE identification. To corroborate the intuitions introduced in the last section, we carried out a corpus-based study that aims to discover linguistic patterns exhibited by one-to-many translations.

Verb	**1-1**	**1-n**	**n-1**	**n-n**	N_o
anheben (v_1)	53.5	21.2	9.2	16	325
bezwecken (v_2)	16.7	51.3	0.6	31.3	150
riskieren (v_3)	46.7	35.7	0.5	17	182
verschlimmern (v_4)	30.2	21.5	28.6	44.5	275

Table 1: Proportions of types of translational correspondences (token-level) in our gold standard.

We constructed a gold standard covering *all* English translations of four German verb lemmas extracted from the Europarl Corpus. These verbs subcategorize for a nominative subject and an accusative object and are in the middle frequency layer (around 200 occurrences). We extracted all sentences in Europarl with occurences of these lemmas and their automatic word alignments produced by GIZA++ (Och and Ney, 2003). These alignments were manually corrected on the basis of the crosslingual word alignment guidelines developped by (Graça et al., 2008).

For each of the German source lemmas, our gold standard records four translation categories: one-to-one, one-to-many, many-to-one, many-to-many translations. Table 1 shows the distribution of these categories for each verb. Strikingly, the four verbs show very different proportions concerning the types of their translational correspondences. Thus, while the German verb *anheben* (en. *increase*) seems to have a frequent parallel realization, the verbs *bezwecken* (en. *intend to*) or *verschlimmern* (en. *aggravate*) tend to be realized by more complex phrasal translations. In any case, the percentage of one-to-many translations is relatively high which corroborates our hypothesis that parallel corpora constitute a very interesting resource for data-driven MWE discovery.

A closer look at the one-to-many translations reveals that these cover a wide spectrum of MWE phenomena traditionally considered in the literature, as well as constructions that one would usually not regard as an MWE. Below, we will shortly illustrate the different classes of one-to-many translations we found in our gold standard.

Morphological variations: This type of one-to-many translations is mainly due to non-parallel realization of tense. It's rather irrelevant from an MWE perspective, but easy to discover and filter automatically.

(3) Sie **verschlimmern** die Übel.
They aggravate the misfortunes.

(4) Their action **is aggravating** the misfortunes.

Verb particle combinations: A typical MWE pattern, treated for instance in (Baldwin and Villavicencio, 2002). It further divides into transparent and non-transparent combinations, the latter is illustrated below.

(5) Der Ausschuss **bezweckt**, den Institutionen ein
The committe intends, the institutions a
politisches Instrument an die Hand zu geben.
political instrument at the hand to give.

(6) The committee **set out** to equip the institutions with a political instrument.

Verb preposition combinations: While this class isn't discussed very often in the MWE literature, it can nevertheless be considered as an idiosyncratic combination of lexical items. Sag et al (2002) propose an analysis within an MWE framework.

(7) Sie werden den Treibhauseffekt **verschlimmern**.
They will the green house effect aggravate.

(8) They will **add to** the green house effect.

Light verb constructions (LVCs): This is the most frequent pattern in our gold standard. It actually subsumes various subpatterns depending on whether the light verbs complement is realized as a noun, adjective or PP. Generally, LVCs are syntactically and semantically more flexible than other MWE types, such that our gold standard contains variants of LVCs with similar, potentially modified adjectives or nouns, as in the example below. However, it can be considered an idiosyncratic combination since the LVCs exhibit specific lexical restrictions (Sag et al., 2002).

(9) Ich werde die Sache nur noch **verschlimmern**.
Ich will the thing only just aggravate.

(10) I am just **making** things **more difficult**.

Idioms: This MWE type is probably the most discussed in the literature due to its semantic and syntactic idiosyncracy. It's not very frequent in our gold standard which may be mainly due to its limited size and the source items we chose.

(11) Sie **bezwecken** die Umgestaltung in eine zivile
They intend the conversion into a civil
Nation.
nation.

(12) They **have in mind** the conversion into a civil nation.

	v_1	v_2	v_3	v_4
N_{type}	22 (26)	41 (47)	26 (35)	17 (24)
V Part	22.7	4.9	0.0	0.0
V Prep	36.4	41.5	3.9	5.9
LVC	18.2	29.3	88.5	88.2
Idiom	0.0	2.4	0.0	0.0
Para	36.4	24.3	11.5	23.5

Table 2: Proportions of MWE types per lemma

Paraphrases: From an MWE perspective, paraphrases are the most problematic and challenging type of translational correspondence in our gold standard. While the MWE literature typically discusses the distinction between collocations and MWEs, the boarderline between paraphrases and MWEs is not really clear. On the hand, paraphrases, as we classified them here, are transparent combinations of lexical items, like in the example below *ensure that something increases*. However, semantically, these transparent combinations can also be rendered by an atomic expression *increase*. A further problem raised by paraphrases is that they often involve translational shifts (Cyrus, 2006). These shifts are hard to identify automatically and present a general challenge for semantic processing of parallel corpora. An example is given below.

(13) Wir brauchen bessere Zusammenarbeit, um die
We need better cooperation to the
Rückzahlungen **anzuheben** .
repayments.OBJ increase.

(14) We need greater cooperation in this respect to **ensure that** repayments **increase** .

Table 2 displays the proportions of the MWE categories for the number of types of one-to-many correspondences in our gold standard. We filtered the types due to morphological variations only (the overall number of types is indicated in brackets). Note that some types in our gold standard fall into several categories, e.g. they combine a verb preposition with a verb particle construction. For all of the verbs, the number of types belonging to core MWE categories largely outweighs the proportion of paraphrases. As we already observed in our analysis of general translation categories, here again, the different verb lemmas show striking differences with respect to their realization in English translations. For instance, *anheben* (en. *increase*) or *bezwecken* (en. *intend*) are frequently translated with verb particle or preposition combinations, while the other verbs are much more often translated by means of LVCs. Also, the more specific LVC patterns differ largely among the verbs. While *verschlimmern* (en. *aggravate*) has many different adjectival LVC correspondences, the translations of *riskieren* (en. *risk*) are predominantly nominal LVCs. The fact that we found very few idioms in our gold standard may be simply related to our arbitrary choice of German source verbs that do not have an English idiom realization (see our experiment on a random set of verbs in Section 3.3).

In general, one-to-many translational correspondences seem to provide a very fruitful ground for the large-scale study of MWE phenomena. However, their reliable detection in parallel corpora is far from trivial, as we will show in the next section. Therefore, we will not further investigate the classification of MWE patterns in the rest of the paper, but concentrate on the high-precision detection of one-to-many translations. Such a pattern-independent identification method is crucial for the further data-driven study of one-to-many translations in parallel corpora.

3 Multiword Translation Detection

This section is devoted to the problem of high-precision detection of one-to-many translations. Section 3.1 describes an evaluation of automatic word alignments against our gold standard. In section 3.2, we describe a method that extracts loosely aligned syntactic configurations which yields much more promising results.

3.1 One-to-many Alignments

To illustrate the problem of purely distributional one-to-many alignment, table 3 presents an evaluation of the automatic one-to-many word alignments produced by GIZA++ that uses the standard heuristics for bidirectional word alignment from phrase-based MT (Och and Ney, 2003). We evaluate the rate of translational correspondences on the type-level that the system discovers against the one-to-many translations in our gold standard. By *type* we mean the set of lemmatized English tokens that makes up the translation of the German source lemma. Generally, automatic word alignment yields a very high FPR if no frequency threshold is used. Increasing the threshold may help in some cases, however the frequency of the translation types is so low, that already at a threshold of 3, almost all types get filtered. This does not mean that the automatic word alignment does not discover any correct correspondences at all, but it means that the detection of the exact set of tokens that correspond to the source token is rare.

verb	n > 0		n > 1		n > 3	
	FPR	FNR	FPR	FNR	FPR	FNR
v_1	0.97	0.93	1.0	1.0	1.0	1.0
v_2	0.93	0.9	0.5	0.96	0.0	0.98
v_3	0.88	0.83	0.8	0.97	0.67	0.97
v_4	0.98	0.92	0.8	0.92	0.34	0.92

Table 3: False positive rate and False negative rate of GIZA++ one-to-many alignments

This low precision of one-to-many alignments isn't very surprising. Many types of MWEs consist of items that contribute most of the lexical semantic content, while the other items belong to the class of semantically almost "empty" items (e.g. particles, light verbs). These semantically "light" items have a distribution that doesn't necessarily correlate with the source item. For instance, in the following sentence pair taken from Europarl, GIZA++ was not able to capture the correspondence between the German main verb *behindern* (en. *impede*) and the LVC *constitute an obstacle to*, but only finds an alignment link between the verb and the noun *obstacle*.

(15) Die Korruption **behindert** die Entwicklung.
The corruption impedes the development.

(16) Corruption **constitutes an obstacle to** development.

Another limitation of the word-alignment models is that are independent of whether the sentences are largely parallel or rather free translations. However, parallel corpora like Europarl are know to contain a very large number of free translations. In these cases, direct lexical correspondences are much more unlikely to be found.

3.2 Aligning Syntactic Configurations

High-precision extraction of one-to-many translation detection thus involves two major problems: 1) How to identify sentences or configurations where reliable lexical correspondences can be found? 2) How to align target items that have a low occurrence correlation?

We argue that both of these problems can be adressed by taking syntactic information into account. As an example, consider the pair of parallel configurations in Figure 1 for the sentence pair given in (15) and (16). Although there is no strict one-to-one alignment for the German verb, the basic predicate-argument structure is parallel: The verbs arguments directly correspond to each other and are all dominated by a verbal root node.

Based on these intuitions, we propose a generate-and-filter strategy for our one-to-many translation detection which extracts partial, largely parallel dependency configurations. By admitting target dependency paths to be aligned to source single dependency relations, we admit configurations where the source item is translated by more than one word. For instance, given the configuration in Figure 1, we allow the German verb to be aligned to the path connecting *constitute* and the argument Y_2.

Our one-to-many translation detection consists of the following steps: a) candidate generation of aligned syntactic configurations, b) filtering the configurations c) alignment post-editing, i.e. assembling the target tokens corresponding to the source item. The following paragraphs will briefly caracterize these steps.

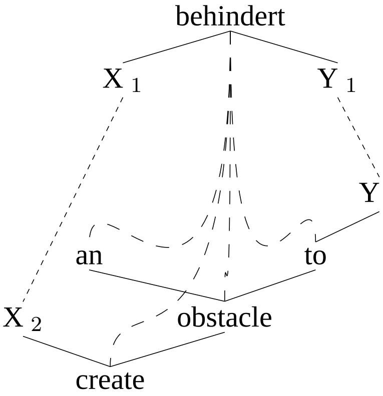

Figure 1: Example of a typical syntactic MWE configuration

Data We word-aligned the German and English portion of the Europarl corpus by means of the GIZA++ tool. Both portions where assigned flat syntactic dependency analyses by means of the MaltParser (Nivre et al., 2006) such that we obtain a parallel resource of word-aligned dependency parses. Each sentence in our resource can be represented by the triple $(D_G, D_E, A_{G,E})$. D_G is the set of dependency triples (s_1, rel, s_2) such that s_2 is a dependent of s_1 of type rel and s_1, s_2 are words of the source language. D_E is the set of dependency triples of the target sentence. $A_{G,E}$ corresponds to the set of pairs (s_1, t_1) such that s_1, t_1 are aligned.

Candidate Generation This step generates a list of source configurations by searching for occurences of the source lexical verb where it is linked to some syntactic dependents (e.g. its arguments). An example input would be the configuration `( (verb,SB,%), (verb,OA,%))` for our German verbs.

Filtering Given our source candidates, a valid parallel configuration $(D_G, D_E, A_{G,E})$ is then defined by the following conditions:
1. The source configuration D_G is the set of tuples (s_1, rel, s_n) where s_1 is our source item and s_n some dependent.
2. For each $s_n \in D_G$, there is a tuple $(s_n, t_n) \in A_{G,E}$, i.e. every dependent has an alignment.
3. There is a target item $t_1 \in D_E$ such that for each t_n, there is a $p \subset D_E$ such that p is a path $(t_1, rel, t_x), (t_x, rel, t_y)...(t_z, rel, t_n)$ that connects t_1 and t_n. Thus, the target dependents have a common root.

To filter noise due to parsing or alignment errors, we further introduce a filter on the length of the path that connects the target root and its dependents and w exclude paths cross contain sentence boundaries. Moreover, the above candidate filtering doesn't exclude configurations which exhibit paraphrases involving head-switching or complex coordination. Head-switching can be detected with the help of alignment information: if there is a item in our target configuration that has an reliable alignment with an item not contained in our source configuration, our target configuration is likely to contain such a structural paraphrases and is excluded from our candidate set. Coordination can be discarded by imposing the condition on the configuration not to contain a coordination relation. This Generate-and-Filter strategy now extracts a set of sentences where we are likely to find a good one-to-one or one-to-many translation for the source verb.

Alignment Post-editing In the final alignment step, one now needs to figure out which lexical material in the aligned syntactic configurations actually corresponds to the translation of the source item. The intuition discussed in 3.2 was that all the items lying on a path between the root item and the terminals belong to the translation of the source item. However, these items may have other syntactic dependents that may also be part of the one-to-many translation. As an example, consider the configuration in figure 1 where the article *an* which is part of the LVC *create an obstacle to* has to be aligned to the German source verb.

Thus, for a set of items t_i for which there is a dependency relation $(t_x, rel, t_i) \in D_E$ such that t_x is an element of our target configuration, we need to decide whether $(s_1, t_i) \in A_{G,E}$. This translation problem now largely parallels collocation translation problems discussed in the literature, as in (Smadja and McKeown, 1994). But, crucially, our syntactic filtering strategy has substantially narrowed down the number of items that are possible parts of the one-to-many translation. Thus, a straightforward way to assemble the translational correspondence is to compute the correlation or association of the possibly missing items with the given translation pair as proposed in (Smadja and McKeown, 1994). Therefore, we propose the following alignment post-editing algorithm:
Given the source item s_1 and the set of target items T, where each $t_i \in T$ is an element of our target configuration,

1. Compute $corr(s_1, T)$, the correlation between s_1 and T.

2. For each t_i, t_x such that there is a $(t_i, rel, t_x) \in D_E$, compute $corr(s_1, T + \{t_x\})$

3. if $corr(s_1, T + \{t_x\}) \geq corr(s_1, T)$, add t_x to T.

As the Dice coefficient is often to give the best results, e.g. in (Smadja and McKeown, 1994), we also chose Dice as our correlation measure. In future work, we will experiment with other association measures. Our correlation scores are thus defined by the formula:

$$corr(s_1, T) = \frac{2(freq(s_1 \wedge T))}{freq(s_1) + freq(T)}$$

We define $freq(T)$ as the number of sentence pairs whose target sentence contains occurrences of all $t_i \in T$, and $freq(s_1)$ accordingly. The observation frequency $freq(s_1 \wedge T)$ is the number of sentence pairs that where s_1 occurs in the source sentence, and T in the target sentence.

The output translation can then be represented as a dependency configuration of the following kind :*((of,PMOD,%), (risk,NMOD,of),(risk,NMOD,the), (run,OBJ,risk), (run,SBJ,%))* which is the syntactic representation for the English MWE *run the risk of*.

3.3 Evaluation

Our translational approach to MWE extraction bears the advantage that evaluation is not exclusively bound to the manual judgement of candidate lists. Instead, we can first evaluate the system output against translation gold standards which are easier to obtain. The linguistic classification of the candidates according to their compositionality can then be treated as a separate problem.

We present two experiments in this evaluation section: We will first evaluate the translation detection on our gold standard to assess the general quality of the extraction method. Since this gold standard is to small to draw conclusions about the quality of MWE patterns that the system detects, we further evaluate the translational correspondences for a larger set of verbs.

Translation evaluation: In the first experiment, we extracted all types of translational correspondences for the verbs we annotated in the gold standard. We converted the output dependency configurations to the lemmatized bag-of-word form we already applied for the alignment evaluation and calculated the FPR and FNR of the translation types. The evaluation is displayed in table 4. Nearly all translation types that our system detected are correct. This confirms our hypothesis that syntactic filtering yields more reliable translations that just coocurrence-based alignments. However, the false negative rate is also very high. This low recall is due to the fact that our syntactic filters are very restrictive such that a major part of the occurrences of the source lemma don't figure in the prototypical syntactic configuration. Column two and three of the evaluation table present the FPR and FNR for experiments with a relaxed syntactic filter that doesn't constrain the syntactic type of the parallel argument relations. While not decreasing the FNR, the FPR decreases significantly. This means that the syntactic filters mainly fire on noisy configurations and don't decrease the recall. A manual error analysis has also shown that the relatively flat annotation scheme of our dependency parses significantly narrows down the number of candidate configurations that our algorithm detects. As the dependency parses don't provide deep analyses for tense or control phenomena, very often, a verb's arguments don't figure as its syntactic dependents and no configuration is found. Future work will explore the impact of deep syntactic analysis for the detection of translational correspondences.

MWE evaluation: In a second experiment, we evaluated the patterns of correspondences found by our extraction method for use in an MWE context. Therefore, we selected 50 random verbs occurring in the Europarl corpus and extracted their respective translational correspondences. This set of 50 verbs yields a set of 1592 one-to-many types of translational correspondences. We filtered the types wich display only morphological variation, such that the set of potential MWE types comprises 1302 types. Out of these, we evaluated a random sample of 300 types by labelling the types with the MWE categories we established for the analysis of our gold standard. During the classification, we encountered a further category of oneto- many correspondence which cannot be considered an MWE, the category of alternation. For instance, we found a translational correspondence between the active realization of the German verb *begrüßen* (en. *appreciate*) and the English passive *be pleased by*.

The classification is displayed in table 5. Almost 83% of the translational correspondences that our system extracted are perfect translation types. Almost 60% of the extracted types can be considered MWEs that exhibit some kind of semantic idiosyncrasy. The other translations could be classified as paraphrases or alternations. In our random sample, the portions of idioms is significantly higher than in our gold standard which confirms our intuition that the MWE pattern of the one-to-many translations for a given verb are related to language-specific, semantic properties of the verbs and the lexical concepts they realize.

4 Related Work

The problem sketched in this paper has clear concetions to statistical MT. So-called phrase-based translation models generally target whole sentence alignment and do not necessarily recur to linguistically motivated phrase correspondences (Koehn et al., 2003). Syntax-based translation that specifies formal relations between bilingual parses was established by (Wu, 1997). Our way to use syntactic configurations can be seen as a heuristic to check relaxed structural parallelism.

	Strict Filter		Relaxed Filter	
	FPR	FNR	FPR	FNR
v_1	0.0	0.96	0.5	0.96
v_2	0.25	0.88	0.47	0.79
v_3	0.25	0.74	0.56	0.63
v_4	0.0	0.875	0.56	0.84

Table 4: False positive and false negative rate of one-to-many translations.

Trans. type	Proportion		
		MWE type	Proportion
MWEs	57.5%	V Part	8.2%
		V Prep	51.8%
		LVC	32.4%
		Idiom	10.6%
Paraphrases	24.4%		
Alternations	1.0%		
Noise	17.1%		

Table 5: Classification of 300 types sampled from the set of one-to-many translations for 50 verbs

Work on MWEs in a crosslingual context has almost exclusively focussed on MWE translation (Smadja and McKeown, 1994; Anastasiou, 2008). In (Villada Moirón and Tiedemann, 2006), the authors make use of alignment information in a parallel corpus to rank MWE candidates. These approaches don't rely on the lexical semantic knowledge about MWEs in form of one-to-many translations.

By contrast, previous approaches to paraphrase extraction made more explicit use of crosslingual semantic information. In (Bannard and Callison-Burch, 2005), the authors use the target language as a pivot providing contextual features for identifying semantically similar expressions. Paraphrasing is however only partially comparable to the crosslingual MWE detection we propose in this paper. Recently, the very pronounced context dependence of monolingual pairs of semantically similar expressions has been recognized as a major challenge in modelling word meaning (Erk and Pado, 2009).

The idea that parallel corpora can be used as a linguistic resource that provides empirical evidence for monolingual idiosyncrasies has already been exploited in, e.g. morphology projection (Yarowsky et al., 2001) or word sense disambiguation (Dyvik, 2004). While in a monolingual setting, it is quite tricky to come up with theoretical or empirical definitions of sense discriminations, the crosslingual scenario offers a theory-neutral, data-driven solution: Since ambiguity is an idiosyncratic property of a lexical item in a given language, it is not likely to be mirrored in a target language. Similarly, our approach can also be seen as a projection idea: we project the semantic information of simplex realization in a source language to an idiosyncratic, multiword realization in the target language.

5 Conclusion

We have explored the phenomenon of one-to-many translations in parallel corpora from the perspective of MWE identification. Our manual study on a translation gold standard as well as our experiments in automatic translation extraction have shown that one-to-many correspondences provide a rich resource and fruitful basis of study for data-driven MWE identification. The crosslingual perspective raises new research questions about the identification and interpretation of MWEs. It challenges the distinction between paraphrases and MWEs, a problem that does not arise at all in the context of monolingual MWE extraction. It also allows for the study of the relation between the semantics of lexical concepts and their MWE realization. Further research in this direction should investigate translational correspondences on a larger scale and further explore these for monolingual interpretation of MWEs.

Our extraction method that is based on syntactic filters identifies MWE types with a much higher precision than purely cooccurence-based word alignment and captures the various patterns we found in our gold standard. Future work on the extraction method will have to focus on the generalization of these filters and the generalization to other items than verbs. The experiments presented in this paper also suggest that the MWE realization of certain lexical items in a target language is subject to certain linguistic patterns. Moreover, the method we propose is completely languageindependent such that further research has to study the impact of the relatedness of the considered languages on the patterns of one-to-many translational correspondences.

References

Dimitra Anastasiou. 2008. Identification of idioms by mt's hybrid research system vs. three commercial system. In *Proceedings of the EAMT*, pp. 12–20.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: a case study on verb-particles. In *Proceedings of the COLING-02*, pp. 1–7.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the ACL*, pp. 597–604 .

Lea Cyrus. 2006. Building a resource for studying translation shifts. In *Proceedings of the 5th LREC*, pp. 1240–1245.

Helge Dyvik. 2004. Translations as semantic mirrors. From parallel corpus to WordNet. *Language and Computers*, 1:311 – 326.

Katrin Erk and Sebastian Pado. 2009. Paraphrase assessment in structured vector space: Exploring parameters and datasets. In *Proc. of the EACL GEMS Workshop*, pp. 57–65.

João de Almeida Varelas Graça, Joana Paulo Pardal, Luísa Coheur, and Diamantino António Caseiro. 2008. Multi-language word alignments annotation guidelines. Technical report, Tech. Rep. 38 / 2008 INESC-ID Lisboa.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the NAACL '03*, pp. 48–54.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*, pp. 79–86.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Malt-parser: A data driven parser-generator for dependency parsing. In *Proc. of LREC-2006*, pp. 2216–2219.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pavel Pecina. 2008. A machine learning approach to multi-word expression extraction. In *Proceedings of the LREC MWE 2008 Workshop*, pp. 54–57.

Ivan A. Sag, Timothy Baldwin, Francis Bond, and Ann Copestake. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the CICLing-2002*, pp. 1–15.

Frank Smadja and Kathleen McKeown. 1994. Translating collocations for use in bilingual lexicons. In *Proceedings of the HLT '94 workshop*, pp. 152–156.

Begoña Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proc. of the EACL MWE 2006 Workshop*, pp. 33–40.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.*, 23(3):377–403.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT 2001*, pp. 1–8.