Extracting Multiword Translations from Aligned Comparable Documents

Reinhard Rapp
Aix-Marseille Université, Laboratoire
d'Informatique Fondamentale
F-13288 Marseille, France
reinhardrapp@gmx.de

Serge Sharoff
University of Leeds
Centre for Translation Studies
Leeds, LS2 9JT, UK
S.Sharoff@leeds.ac.uk

Abstract

Most previous attempts to identify translations of multiword expressions using comparable corpora relied on dictionaries of single words. The translation of a multiword was then constructed from the translations of its components. In contrast, in this work we try to determine the translation of a multiword unit by analyzing its contextual behaviour in aligned comparable documents, thereby not presupposing any given dictionary. Whereas with this method translation results for single words are rather good, the results for multiword units are considerably worse. This is an indication that the type of multiword expressions considered here is too infrequent to provide a sufficient amount of contextual information. Thus indirectly it is confirmed that it should make sense to look at the contextual behaviour of the components of a multiword expression individually, and to combine the results.

1 Introduction

The task of identifying word translations from comparable text has received considerable attention. Some early papers include Fung (1995) and Rapp (1995). Fung (1995) utilized a context heterogeneity measure, thereby assuming that words with productive context in one language translate to words with productive context in another language, and words with rigid context translate into words with rigid context. In contrast, the underlying assumption in Rapp (1995) was that words which are translations of each other show similar co-occurrence patterns across languages. This assumption is effectively an extension of Harris' (1954) distributional hypotheses to the multilingual case.

This work was further elaborated in some by now classical papers, such as Fung & Yee (1998) and Rapp (1999). Based on these papers, the standard approach is to start from a dictionary of seed words, and to assume that the words occurring in the context of a source language word have similar meanings as the words occurring in the context of its target language translation.

There have been suggestions to eliminate the need for the seed dictionary. However, most attempts, such as Rapp (1995), Diab & Finch (2000) and Haghighi et al. (2008) did not work to an extent that the results would be useful for practical purposes. Only recently a more promising approach has been investigated: Schafer & Yarowsky (2002), Hassan & Mihalcea (2009), Prochasson & Fung (2011) and Rapp et al. (2012) look at aligned comparable documents and deal with them in analogy to the treatment of aligned parallel sentences, i.e. effectively doing a word alignment in a very noisy environment. This approach has been rather successful and it was possible to improve on previous results. This is therefore the approach which we will pursue in the current paper.

However, in contrast to the above mentioned papers the focus of our work is on multiword expressions, and we will compare the performance of our algorithm when applied to multiword expressions and when applied to single words.

There has been some previous work on identifying the translations of multiword units using comparable corpora, such as Robitaille et al. (2006), Babych et al. (2007), Daille & Morin (2012); Delpech et al. (2012). However, none of this work utilizes aligned comparable documents, and the underlying assumption is that the translation of a multiword unit can be determined by looking at its components individually, and by merging the results.

In contrast, we try to explore whether the translation of a multiword unit can be determined solely by looking at its contextual behavior, i.e. whether it is possible to also apply the standard approach as successfully used for single words. The underlying fundamental question is whether the meaning of a multiword unit is determined by

Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 87–95,
Gothenburg, Sweden, April 27, 2014. ©2014 Association for Computational Linguistics

the contextual behavior of the full unit, or by the contextual behavior of its components (or by a mix of both). But multiword expressions are of complex nature, as expressed e.g. by 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." The current paper is an attempt to systematically approach one aspect of this complexity.

2 Approach

Our approach is based on the usual assumption that there is a correlation between the patterns of word-co-occurrence across languages. However, instead of presupposing a bilingual dictionary it only requires pre-aligned comparable documents, i.e. small or medium sized documents aligned across languages which are known to deal with similar topics. This could be, for example, newspaper articles, scientific papers, contributions to discussion groups, or encyclopaedic articles. As Wikipedia is a large resource and readily available for many languages, we decided to base our study on this encyclopaedia. The Wikipedias have the so-called interlanguage links which are manually inserted by the authors and connect articles referring to the same headword in different languages.

Given that each Wikipedia community contributes in its own language, only occasionally an article connected in this way will be an exact translation of a foreign language article, and in most cases the contents will be rather different. On the positive side, the link structure of the interlanguage links tends to be quite dense. For example, of the 1,114,696 German Wikipedia articles, 603,437 have a link to the corresponding English Wikipedia article.

2.1 Pre-processing and MWE extraction

We used the same versions of Wikipedia as in Rapp et al. (2012) and used the same processing. After download, each Wikipedia was minimally processed to extract the plain text contents of the articles. In this process all templates, e.g. 'infoboxes', as well as tables were removed, and we kept only the webpages with more than 500 characters of running text (including white space). Linguistic processing steps included tokenisation, tagging and lemmatisation using the default UTF-8 versions of the respective TreeTagger resources (Schmid, 1994).

From the pre-processed English and German Wikipedia, we extracted the multiword expressions using two simple principles, a *negative POS filter* and a *containment filter*. The negative POS filter operates in a rule-based fashion on the complete list of n-grams by removing the unlikely candidates according to a set of constraints, such as the presence of determiners or prepositions at the edges of expressions, see a similar method used by (Justeson & Katz, 1995). With some further extensions this was also used to produce the multiword lists for the dictionary of translation equivalents (Babych et al., 2007).

We did not use positive shallow filters. These would need to capture the relatively complex structure of the noun, verb and prepositional phrases, while avoiding noise. This can often lead to a lack of recall when more complex constructions cannot be captured. In contrast, negative shallow filters simply avoid obvious noise, while passing other multiword expressions (MWEs) through, which are very often legitimate syntactic constructions in a language in question. For example, the following English filters[1] rejected personal pronouns (PP) and conjunctions (CC) at the edges of expressions (using the Penn Treebank tagset as implemented by Treetagger):

```
^[^ ]+~~PP |~~PP$
^[^ ]+~~CC |~~CC$
```

Similarly, any MWE candidates including proper nouns (NP) and numerals (CD) were discarded:

```
~~NP
~~CD
```

In the end, this helps in improving the recall rate while using a relatively small number of patterns: 18 patterns were used for English, 11 for German.

The containment filter further rejects MWEs by removing those that regularly occur as a part of a longer acceptable MWE. For example, *graphical user* is an acceptable expression passing through the POS filter, but it is rejected by the containment filter since the overwhelming majority of its uses are in the containing MWE *graphical user interface* (1507 vs 1304 uses in Wikipedia, since MWEs are still possible, e.g., *graphical user environment*).

[1] We use here the standard notation for regular expressions as implemented in Perl (Friedl, 2002). For example, '^' means 'beginning of line' and '$' means 'end of line'.

English keyterms for 'Airbus 320 family'		
Score	*f*	*Keyterm*
34.88	4	final_JJ assembly_NN
31.22	3	firm_NN order_NN
30.73	3	series_NN aircraft_NN
29.07	4	flight_NN control_NN
27.38	3	wing_NN area_NN
23.26	3	final_JJ approach_NN
22.19	2	lose_VV life_NN
20.63	6	passenger_NN and_CC crew_NN
17.54	2	first_JJ derivative_NN
17.34	2	fly-by-wire_NN flight_NN control_NN
16.63	3	flight_NN deck_NN
16.41	2	crew_NN die_VV
15.08	2	pilot_NN error_NN
14.98	2	passenger_NN capacity_NN
14.38	2	turbofan_NN engine_NN
14.03	2	development_NN cost_NN
12.30	2	maiden_JJ flight_NN
11.54	2	direct_JJ competition_NN
10.75	2	overall_JJ length_NN
10.39	2	overrun_VV the_DT runway_NN
9.54	2	flight_NN control_NN system_NN
9.31	2	fuel_NN consumption_NN
8.63	2	roll_VV out_RP
7.86	3	crew_NN member_NN
7.54	2	crew_NN on_IN board_NN
7.33	2	bad_JJ weather_NN
6.63	2	landing_NN gear_NN

German keyterms for 'Airbus-A320-Familie'		
Score	*f*	*Keyterm*
155.25	20	Triebwerk
62.88	4	Fly-by-Wire-System
59.76	8	Erstflug
57.67	8	Absturz
43.79	4	Endmontage
43.70	4	Hauptfahrwerk
41.77	4	Tragflügel
36.52	8	Unfall
35.90	6	Unglück
33.25	3	Abfluggewicht
33.10	5	Auslieferung
30.01	3	Treibstoffverbrauch
29.00	2	Triebwerkstyp
28.51	2	Zwillingsreifen
18.20	2	Absturz_NN verursachen_VV
16.28	3	Passagier_NN Platz_NN
16.23	2	Triebwerk_NN antreiben_VV
13.41	2	Steuerung_NN d_AR Flugzeug_NN
12.52	2	Absturz_NN führen_VV
11.68	2	Rumpf_NN befinden_VV
8.59	2	Insasse_NN ums_AP Leben_NN
8.55	2	Zeitpunkt_NN d_AR Unglück_NN

Table 1. English and German keyterms for 'Airbus 320 family' (lists truncated). Score = log-likelihood score; f = occurrence frequency of keyterm; NN = noun; VV = verb; AR = article; AP = article+preposition; JJ = adjective; CC = conjunction; RP = preposition.

2.2 Keyterm extraction

As the aligned English and German Wikipedia documents are typically not translations of each other, we cannot apply the usual procedures and tools as available for parallel texts (e.g. the Gale & Church sentence aligner and the Giza++ word alignment tool). Instead we conduct a two step procedure:

1. We first extract salient terms (single word or multiword) from each of the documents.
2. We then align these terms across languages using an approach inspired by a connectionist (Rumelhart & McClelland, 1987) Winner-Takes-It-All Network. The respective algorithm is called WINTIAN and is described in Rapp et al. (2012) and in Rapp (1996).

For term extraction, the occurrence frequency of a term in a particular document is compared to its average occurrence frequency in all Wikipedia documents, whereby a high discrepancy indicates a strong keyness. Following Rayson & Garside (2000), we use the log-likelihood score to measure keyness, since it has been shown to be robust to small numbers of instances. This robustness is important as many Wikipedia articles are rather short.

This procedure leads to multiword keyterms as exemplified in Table 1 for the Wikipedia entry *Airbus A320 family*. Because of compounding in German, many single-word German expressions are translated into multiword expressions in English. So we chose to include single-word expressions into the German candidate list for alignment with English multiwords.

One of the problems in obtaining multiword keyterms from the Wikipedia articles is relative data sparseness. Usually, the frequency of an individual multiword expression within a Wikipedia article is between 2 and 4. Therefore we had to use a less conservative threshold of 6.63 (1% significance level) rather than the more standard 15.13 (0.01% significance level) for the log-likelihood score (see Rayson & Garside, 2000, and http://ucrel. lancs.ac.uk/llwizard.html).

2.3 Term alignment

The WINTIAN algorithm is used for establishing term alignments across languages. As a more detailed technical description is given in Rapp et al. (2012) and in Rapp (1996), we only briefly describe this algorithm here, thereby focusing on the neural network analogy. The algorithm can be considered as an artificial neural network where the nodes are all English and German

terms occurring in the keyterm lists. Each English term has connections to all German terms. The connections are all initialized with values of one when the algorithm is started, but will serve as a measure of the translation probabilities after the completion of the algorithm. One after the other, the network is fed with the pairs of corresponding keyterm lists. Each German term activates the corresponding German node with an activity of one. This activity is then propagated to all English terms occurring in the corresponding list of keyterms. The distribution of the activity is not equal, but in proportion to the connecting weights. This unequal distribution has no effect at the beginning when all weights are one, but later on leads to rapid activity increases for pairs of terms which often occur in corresponding keyterm lists. The assumption is that these are translations of each other. Using Hebbian learning (Rumelhart & McClelland, 1987) the activity changes are stored in the connections. We use a heuristic to avoid the effect that frequent keyterms dominate the network: When more than 50 of the connections to a particular English node have weights higher than one, the weakest 20 of them are reset to one. This way only translations which are frequently confirmed can build up high weights.

It turned out that the algorithm shows a robust behaviour in practice, which is important as the corresponding keyterm lists tend to be very noisy and, especially for multiword expressions, in many cases may contain hardly any terms that are actually translations of each other. Reasons are that corresponding Wikipedia articles are often written from different perspectives, that the variation in length can be considerable across languages, and that multiword expressions tend to show more variability with regard to their translations than single words.

3 Results and evaluation

3.1 Results for single words

In this subsection we report on our previous results for single words (Rapp et al., 2012) as these serve as a baseline for our new results concerning multiword units.

The WINTIAN algorithm requires as input vocabularies of the source and the target language. For both English and German, we constructed these as follows: Based on the keyword lists for the respective Wikipedia, we counted the number of occurrences of each keyword, and then applied a threshold of five, i.e. all keywords with a lower frequency were eliminated. The reasoning behind this is that rare keywords are of not much use due to data sparseness. This resulted in a vocabulary size of 133,806 for English, and of 144,251 for German.

Using the WINTIAN algorithm, the English translations for all 144,251 words occurring in the German vocabulary were computed. Table 2 shows the results for the German word *Straße* (which means *street*).

For a quantitative evaluation we used the ML1000 test set comprising 1000 English-German translations (see Rapp et al., 2012). We verified in how many cases our algorithm had assigned the expected translation (as provided by the gold standard) the top rank among all 133,806 translation candidates. (Candidates are all words occurring in the English vocabulary.) This was the case for 381 of the 1000 items, which gives us an accuracy of 38.1%. Let us mention that this result refers to exact matches with the word equations in the gold standard. As in reality due to word ambiguity other translations might also be acceptable (e.g. for *Straße* not only *street* but also *road* would be acceptable), these figures are conservative and can be seen as a lower bound of the actual performance.

GIVEN GERMAN WORD	*Straße*	
EXPECTED TRANSLATION	*street*	
	LL-SCORE	TRANSLATION
1	215.3	road
2	148.2	street
3	66.0	traffic
4	46.0	Road
5	42.6	route
6	34.6	building

Table 2. Computed translations for *Straße*.

3.2 Results for multiword expressions

In analogy to the procedure for single words, for the WINTIAN algorithm we also needed to define English and German vocabularies of multiword terms. For English, we selected all multiword terms which occurred at least three times in the lists of English key terms, and for German those which occurred at least four times in the lists of German key terms. This resulted in similar sized vocabularies of 114,796 terms for English, and 131,170 for German. Note that the threshold for German had to be selected higher not because German has more inflectional variants (which does not matter as we are working

with lemmatized data), but because – other than the English – the German vocabulary also includes unigrams. The reason for this is that German is highly compositional, so that English multiword units are often translated by German unigrams.

Using the WINTIAN algorithm, the English translations for all 131,170 words occurring in the German multiword vocabulary were computed, and in another run the German translations for all 114,796 English words. Table 3 shows some sample results.

For a quantitative evaluation, we did not have a gold standard at hand. As multiword expressions show a high degree of variability with regard to their translations, so that it is hard to come up with all possibilities, we first decided not to construct a gold standard, but instead did a manual evaluation. For this purpose, we randomly selected 100 of the German multiword expressions with an occurrence frequency above nine, and verified their computed translations (i.e. the top ranked item for each) manually. We distinguished three categories: 1) Acceptable translation; 2) Associatively related to an acceptable translation; 3) Unrelated to an acceptable translation.

	English → German	
	husband_NN and_CC wife_NN	
Rank	*Aktivity*	*Translation*
1	2.98	Eheleute
2	1.09	Voraussetzung
3	1.08	Kirchenrecht
4	0.76	Trennung
5	0.35	Mann
6	0.24	Kirche
7	0.08	Mischehe
8	0.08	Diakon

	German → English	
	Eheleute	
Rank	*Aktivity*	*Translation*
1	3.01	husband_NN_and_CC_wife_NN
2	1.26	married_JJ_couple_NN
3	1.02	civil_JJ_law_NN
4	1.02	equitable_JJ_distribution_NN
5	1.02	community_NN_property_NN
6	0.52	law_NN_jurisdiction_NN
7	0.05	racing_NN_history_NN
8	0.05	great_JJ_female_JJ

Table 3. Sample results for translation directions EN → DE and DE → EN.

We also did the same computation for the reverse language direction, i.e. for English to German. The results are listed in Table 4. These results indicate that our procedure, although currently state of the art for single words, does not work well for multiword units. We investigated the data and located the following problems:

- The problem of data sparseness is, on average, considerably more severe for multiword expressions than it is for single words.
- Although the English and the German vocabulary each contain more than 100,000 items, their overlap is still limited. The reason is that the number of possible multiword units is very high, far higher than the number of words in a language.
- We considered only multiword units up to length three, but in some cases this may not suffice for an acceptable translation.
- In the aligned keyterm lists, only rarely correct translations of the source language terms occur. Apparently the reason is the high variability of multiword translations.

Hereby he last point seems to have a particularly severe negative effect on translation quality. However, all of these findings are of fundamental nature and contribute to the insight that at least for our set of multiword expressions compositionality seems to be more important than contextuality.

German → English		
Judgment	*Number*	*Example taken from actual data*
Acceptable	5	Jugendherberge → youth_NN hostel_NN
Association	38	Maischegärung → oak_NN barrel_NN
Unacceptable	57	Stachelbeere → horror_NN film_NN

English → German		
Judgment	*Number*	*Example taken from actual data*
Acceptable	6	amino_NN acid_NN → Aminosäure
Association	52	iron_NN mine_NN → Eisenerz
Unacceptable	42	kill_VV more_JJ → Weltmeistertitel_NN im_AP Schwergewicht_NN

Table 4. Quantitative results involving MWEs.

3.3 Large scale evaluation

As a manual evaluation like the one described above is time consuming and subjective, we thought about how we could efficiently come up with a gold standard for multiword expressions with the aim of conducting a large scale automatic evaluation. We had the idea to determine the correspondences between our English and German MWEs via translation information as extracted from a word-aligned parallel corpus.

Such data we had readily at hand from a previous project called COMTRANS. During this project we had constructed a large bilingual dictionary of bigrams, i.e. of pairs of adjacent words in the source language. For constructing the dictionary, we word-aligned the English and German parts of the Europarl corpus. For this purpose, using Moses default settings, we combined two symmetric runs of Giza++, which considerably improves alignment quality. Then we determined and extracted for each English bigram the German word or word sequence which had been used for its translation. Discontinuities of one or several word positions were allowed and were indicated by the wildcard '*'. As the above method for word alignment produces many unjustified empty assignments (i.e. assignments where a source language word pair is erroneously assumed to have no equivalent in the target language sentence), so that the majority of these is incorrect, all empty assignments were removed from the dictionary.

In the dictionary, for each source language word pair its absolute frequency and the absolute and relative frequencies of its translation(s) are given. To filter out spurious assignments, thresholds of 2 for the absolute and 10% for the relative frequency of a translation were used. The resulting dictionary is available online.[2] Table 5 shows a small extract of the altogether 371,590 dictionary entries. Alternatively, we could have started from a Moses phrase table, but it was easier for us to use our own data.

Although the quality of our bigram dictionary seems reasonably good, it contains a lot of items which are not really interesting multiword expressions (e.g. arbitrary word sequences such as *credible if* or the discontinuous word sequences on the target language side). For this reason we filtered the dictionary using the lists of Wikipedia-derived multiword expressions as described in section 2.1. These contained 418,627 items for English and 1,212,341 candidate items for German (the latter included unigram compounds). That is, in the dictionary those items were removed where either the English side did not match any of the English MWEs, or where the German side did not match any of the German candidates.

This intersection resulted in a reduction of our bigram dictionary from 371,590 items to 137,701 items. Table 6 shows the results after filtering the items listed in Table 5. Note that occasionally reasonable MWEs are eliminated if they happen not to occur in Wikipedia, or if the algorithm for extracting the MWEs does not identify them.

The reduced dictionary we considered as an appropriate gold standard for the automatic evaluation of our system.

ENGLISH BIGRAM	GERMAN TRANSLATION
credible if	dann glaubwürdig * wenn
credible if	glaubhaft * wenn
credible if	glaubwürdig * wenn
credible in	in * Glaubwürdigkeit
credible in	in * glaubwürdig
credible investigation	glaubwürdige Untersuchung
credible labelling	glaubwürdige Kennzeichnung
credible manner	glaubwürdig
credible military	glaubwürdige militärische
credible military	glaubwürdigen militärischen
credible only	nur dann glaubwürdig
credible partner	glaubwürdiger Partner
credible policy	Politik * glaubwürdig
credible policy	glaubwürdige Politik
credible reports	glaubwürdige Berichte
credible response	glaubwürdige Antwort
credible solution	glaubwürdige Lösung
credible system	glaubwürdiges System
credible threat	glaubhafte Androhung
credible to	für * glaubwürdig
credible to	glaubwürdig

Table 5. Extract from the COMTRANS bigram dictionary.

ENGLISH BIGRAM	GERMAN TRANSLATION
credible investigation	glaubwürdige Untersuchung
credible only	nur dann glaubwürdig
credible policy	glaubwürdige Politik
credible response	glaubwürdige Antwort
credible solution	glaubwürdige Lösung
credible system	glaubwürdiges System
credible threat	glaubhafte Androhung
credible to	glaubwürdig

Table 6. Extract from the bigram dictionary after filtering.

[2] http://www.ftsk.uni-mainz.de/user/rapp/comtrans/ There click on "Dictionaries of word pairs" and then download "English – German".

As in section 3.2, the next step was to apply the keyword extraction algorithm to the English and the German Wikipedia documents. Hereby only terms occurring in the gold standard dictionary were taken into account. But it turned out that, when using the same log-likelihood threshold as in section 3.2, only few keyterms were assigned: on average less than one per document. This had already been a problem in 3.2, but it was now considerably more severe as this time the MWE lists had been filtered, and as the filtering had been on the basis of another type of corpus (Europarl rather than Wikipedia).

This is why, after some preliminary experiments with various thresholds, we finally decided to disable the log-likelihood threshold. Instead, on the English side, all keyterms from the gold standard were used if they occurred at least once in the respective Wikipedia document. On the German side, as here we had many unigram compounds which tend to be more stable and therefore more repetitive than MWEs, we used the keyterms if the occurred at least twice. This way for most documents we obtained at least a few keyterms.

When running the WINTIAN algorithm on the parallel keyword lists, in some cases reasonable results were obtained. For example, for the direction English to German, the system translates *information society* with *Informationsgesellschaft*, and *education policy* with *Bildungspolitik*. As WINTIAN is symmetric and can likewise produce a dictionary in the opposite direction, we also generated the results for German to English. Here, among the good examples, are *Telekommunikationsmarkt*, which is translated as *telecommunications market*, and *Werbekampagne*, which is translated as *advertising campaign*. However, these are selected examples showing that the algorithm works in principle.

Of more interest is the quantitative evaluation which is based on thousands of test words and uses the gold standard dictionary. For English to German we obtained an accuracy of 0.77% if only the top ranked word is taken into account, i.e. if this word matches the expected translation. This improves to 1.6% if it suffices that the expected translation is ranked among the top ten words. The respective figures for German to English are 1.41% and 2.04%.

The finding that German to English performs better can be explained by the fact that other than English German is a highly inflectional language. That is, when generating translations it is more likely for German that an inflectional variant not matching the gold standard translation is ranked first, thus adversely affecting performance.

A question more difficult to answer is why the results based on the gold standard are considerably worse than the ones reported in section 3.2 which were based on human judgment. We see the following reasons:

- The evaluation in section 3.2 used only a small sample so might be not very reliable. Also, other than here, it considered only source language words with frequencies above nine.
- Unlike the candidate expressions, the gold standard data is not lemmatized on the target language side.
- The hard string matching used for the gold-standard-based evaluation does not allow for inflectional variants.
- The gold-standard-based evaluation used terms resulting from the intersection of term lists based on Wikipedia and Europarl. It is clear that this led to a reduction of average term frequency (if measured on the basis of Wikipedia), thus increasing the problem of data sparseness.
- As for the same reason the log-likelihood threshold had to be abandoned, on average less salient terms had to be used. This is likely to additionally reduce accuracy.
- For many terms the gold standard lists several possible translations. In the current implementation of the evaluation algorithm only one of them is counted as correct.[3] However, in the human evaluation any reasonable translation was accepted.
- Some reasonable MWE candidates extracted from Wikipedia are not present in the gold standard, for example *credible evidence*, *credible source*, and *credible witness* are not frequent enough in Europarl to be selected for alignment.

We should perhaps mention that it would be possible to come up with better looking accuracies by presenting results for selected subsets of the source language terms. For example, one could concentrate on terms with particularly good cov-

[3] This can be justified because an optimal algorithm should provide all possible translations of a term. If only some translations are provided, only partial credit should be given. But this is likely to average out over large numbers, so the simple version seems acceptable.

erage. Another possibility would be to consider MWEs consisting of nouns only. This we actually did by limiting source and target language vocabulary (of MWEs) to compound nouns. The results were as follows:

English to German (top 1):	1.81%
English to German (top 10):	3.75%
German to English (top 1):	2.03%
German to English (top 10):	3.16%

As can be seen, these results look somewhat better. But this is only for the reason that translating compound nouns appears to be a comparatively easier task on average.

4 Conclusions and future work

We have presented a method for identifying term translations using aligned comparable documents. Although it is based on a knowledge poor approach and does not presuppose a seed lexicon, it delivers competitive results for single words.

A disadvantage of our method is that it presupposes that the alignments of the comparable documents are known. On the other hand, there are methods for finding such alignments automatically not only in special cases such as Wikipedia and newspaper texts, but also in the case of unstructured texts (although these methods may require a seed lexicon).

Concerning the question from the introduction, namely whether the translation (and consequently also the meaning) of a multiword unit is determined compositionally or contextually, our answer is as follows: For the type of multiword units we were investigating, namely automatically extracted collocations, our results indicate that looking at their contextual behavior usually does not suffice. The reasons seem to be that their contextual behavior shows a high degree of variability, that their translations tend to be less salient than those of single words, and that the problem of data sparseness is considerably more severe.

It must be seen, however, that there are many types of multiword expressions, such as idioms, metaphorical expressions, named entities, fixed phrases, noun compounds, compound verbs, compound adjectives, and so on, so that our results are not automatically applicable to all of them. Therefore, in future work we intend to compare the behavior of different types of multiword expressions (e.g. multiword named entities and short phrases such as those used in phrase-based machine translations) and to quantify in how far their behavior is compositional or contextual.

Acknowledgment

This research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme.

References

Babych, B., Sharoff, S., Hartley, A., and Mudraya, O. (2007). Assisting Translators in Indirect Lexical Transfer. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics ACL 2007*, Prague, Czech Republic.

Daille, B.; Morin, E. (2012). Revising the compositional method for terminology acquisition from comparable corpora. Proceedings of Coling 2012, Mumbai.

Delpech, E.; Daille, B.; Morin, E., Lemaire, C. (2012). Extraction of domain-specific bilingual lexicon from comparable corpora: compositional translation and ranking. Proceedings of Coling 2012, Mumbai.

Diab, M., Finch, S. (2000): A statistical wordlevel translation model for comparable corpora. In: *Proceedings of the Conference on Content-Based Multimedia Information Access* (RIAO).

Friedl, J. (2002). *Mastering Regular Expressions.* O'Reilly.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In: *Proceedings of the Third Annual Workshop on Very Large Corpora*, Boston, Massachusetts. 173-183.

Fung, P.; Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. *Proceedings of COLING/ACL 1998, Montreal, Canada.* 414-420.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., Klein, D. (2008): Learning bilingual lexicons from monolingual corpora. In: Proceedings of ACL-HLT 2008, Columbus, Ohio. 771-779.

Harris, Z.S. (1954). Distributional structure. *Word*, 10(23), 146–162.

Hassan, S., Mihalcea, R. (2009): Cross-lingual semantic relatedness using encyclopedic knowledge. In: *Proceedings of EMNLP*.

Justeson, J.S.; Katz, S.M. (1995). Techninal terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1): 9–27.

Moon, R.E. 1998. Fixed Expressions and Idioms in English: A Corpus-based Approach. Oxford: Clarendon Press.

Prochasson, E., Fung, P. (2011). Rare word translation extraction from aligned comparable documents. In: *Proceedings of ACL-HLT*. Portland .

Rapp, R. (1995). Identifying word translations in non-parallel texts. In: Proceedings *of the 33rd Annual Meeting of the ACL*. Cambridge, MA, 320-322.

Rapp, R. (1996). *Die Berechnung von Assoziationen*. Hildesheim: Olms.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland.* 519–526.

Rapp, R., Sharoff, S., Babych, B. (2012). Identifying word translations from comparable documents without a seed lexicon. In: Proceedings of the 8th Language Resources and Evaluation Conference, LREC 2012, Istanbul.

Rayson, P.; Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora (WCC '00), Volume 9*, 1–6.

Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., Utsuro, T. (2006). Compiling French-Japanese terminologies from the web. In: Proceedings of the 11th Conference of EACL, Trento, Italy, 225-232.

Rumelhart, D.E.; McClelland, J.L. (1987). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 1: Foundations*. MIT Press.

Schafer, C., Yarowsky, D (2002).: Inducing translation lexicons via diverse similarity measures and bridge languages. In: *Proceedings of CoNLL.*

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing*, 44–49.