Identifying bilingual Multi-Word Expressions for Statistical Machine Translation

Dhouha Bouamor[1,2,3], Nasredine Semmar[1], Pierre Zweigenbaum[2,3]

[1]CEA-LIST, Vision and Content Engineering Laboratory, F91191 Gif sur Yvette Cedex, France
[2] LIMSI-CNRS, F-91403 Orsay France
[3] Univ. Paris Sud, Orsay, France
dhouha.bouamor@cea.fr, nasredine.semmar@cea.fr, pz@limsi.fr

Abstract

MultiWord Expressions (MWEs) repesent a key issue for numerous applications in Natural Language Processing (NLP) especially for Machine Translation (MT). In this paper, we describe a strategy for detecting translation pairs of MWEs in a French-English parallel corpus. In addition we introduce three methods aiming to integrate extracted bilingual MWEs in MOSES, a phrase based Statistical Machine Translation (SMT) system. We experimentally show that these textual units can improve translation quality.

Keywords: bilingual Multi-Word Expression, Vector Space Model, Statistical Machine Translation

1. Introduction

A Multi-Word Expression (MWE) can be defined as a combination of words for which syntactic or semantic properties of the whole expression can not be obtained from its parts (Sag et al., 2002). Such units are made up of collocations (*cordon bleu*), expressions more or less frozen (*kick the bucket*), named entities (*New York*) etc. (Sag et al., 2002; Constant et al., 2011). They are numerous and constitute a significant portion of the lexicon of any natural language. (Jackendoff, 1997) claims that the frequency of MWEs in a speaker's lexicon is almost equivalent to the frequency of single words. While easily mastered by native speakers, their interpretation poses a major challenge for NLP applications especially those addressing semantic aspects of language.

For Statistical Machine Translation (SMT) systems, various improvements of translation quality were achieved with the emergence of phrase based approaches (Koehn et al., 2003). Phrases are defined as simply arbitrary n-grams with no sophisticated linguistic motivation consistently translated in a parallel corpus. In such systems, the lack of an adequate processing of MWEs could affect the translation quality. In fact, the literal translation of an unrecognized expression is the source of an erroneous and incomprehensible translation. For example, it would suggest *"way of iron"* as a translation of *"chemin de fer "* instead of *"railway"*. It is therefore important to make use a lexicon in which MWEs are handled. But such kind of resource is not necessarily available in all languages, and if they exist, as described (Sagot et al., 2005), they do not cover all MWEs of a given language.

In this paper, we consider any non-compositional contiguous sequence, belonging to one of the three classes defined by (Luka et al., 2006), as a MWE. Classes of MWEs were distinguished on the basis of their categorical properties and their syntactic and semantic congealing degrees and are made up of *compounds*, *idiomatic expressions* and *collocations*. Based on this classification, we present a method combining linguistic and statistical information to extract and align MWEs in a French-English parallel corpus aligned at the sentence level. Then, we introduce three methods aiming to integrate extracted bilingual MWEs into MOSES, the state-of-the-art phrase based SMT system and study in what respect we can improve translation quality by the use of such units.

The remainder of this paper is organized as follows: the next section (section 2) describe in some details previous works addressing the task of semantically equivalent translations extraction and their applications. In section 3, we introduce a method for identifying French and English MWEs and then present, in section 4, the algorithm we implemented to acquire translation pairs of MWEs and report our evaluation results. In section 5 three methods aiming to integrate MWEs in an SMT system are introduced and obtained results are discussed. We, finally, conclude and present our future work, in section 6.

2. Related Work

In recent years, a number of techniques have been applied to the task of bilingual MWEs extraction from parallel corpora. Most works start by identifying monolingual MWE candidates then, apply different alignment methods to acquire bilingual correspondences. Monolingual extraction of MWEs techniques revolve around three approaches: (1) symbolic methods relying on morphosyntactic patterns (Okita et al., 2010; Dagan and Church, 1994); (2) statistical methods which use association measures to rank MWE candidates (Vintar and Fisier, 2008) and (3) Hybrid approaches combining (1) and (2) (Wu and Chang, 2004; Seretan and Wehrli, 2007; Daille, 2001; Boulaknadel et al., 2008). None of the approaches is without limitations. It is difficult to apply symbolic methods to data without syntactic annotations. Furthermore, due to corpus size, statistical measures have mostly been applied to bigrams and trigrams, and it become more problematic to extract MWEs of more than three words.

Concerning the alignment task, noumerous approaches have already been introduced to deal with this problem.

Some works make use of simple word alignment tools (Dagan and Church, 1994). Others rely on machine learning algorithms such as the *Expectation Maximisation (EM)* algorithm (Kupiec, 1993; Okita et al., 2010). In another direction, (Tufis and Ion, 2007; Seretan and Wehrli, 2007) introduce a linguistic approach in which they claim that MWEs keep in most cases the same morphosyntactic structure in the source and target language, which is not universal. For example the French MWE *insulaire en développement*, aligned with the English MWE *small island developing* does not share the same morphosyntactic structure. Most of methods described above aim at identifiying MWEs in a corpus to construct or extend a bilingual lexicon. Having such type of textual units is useful for a variety of NLP applications such as information retrieval (Vechtomova, 2005), word sense disambiguation (Finlayson and Kulkarni, 2011).

Few works has however focused on the extraction of bilingual MWEs in order to improve an MT system performance. In (Lambert and Banchs, 2005), authors introduce a method in which a bilingual MWEs lexicon was used to modify the word alignment in order to improve the translation quality. In their work, bilingual MWEs were grouped as one unique token before training alignment models. They showed on a small corpus, that both alignment quality and translation accuracy were improved. However, in a further study, a lower BLEU score is reported after grouping MWEs by part-of-speech on a large corpus (Lambert and Banchs, 2006). More recently, (Ren et al., 2009) described a method integrating an in-domain bilingual MWE to Moses and gained $+0.61$ of BLEU score compared to the baseline system. In a preliminary study (Bouamor et al., 2011), we presented a technique for extracting bilingual multi-word expressions. In that study, MWEs identified on a small corpus($10K$) were integrated as a bilingual lexicon in the phrase table of the MOSES system. This method yeilds an improvement of 0.24 points BLEU score. In this paper, we applied the same MWEs extraction technique with a various improvements in a large corpus.

3. MWEs Extraction

In this section, we describe the approach to extract monolingual MWEs from a French-English parallel corpus. Generally, the choice of an effective way to deal with this problem depends on the further use of MWEs, and resources availability. The method we present here is based on a symbolic approach relying on morphosyntactic patterns. Relatively simple, it handles both frequent and infrequent expressions and do not use any dictionary. It only involves a full morphosyntactic analysis of source and target texts. For this, we used the CEA LIST Multilingual Analysis platform (LIMA) (Besançon et al., 2010) which produces a set of part of speech tagged normalized lemmas. Since most MWEs consist of noun, adjectives and prepositions, we adopted a linguistic filter keeping only n-gram ($2 \leq n \leq 4$) units which match with a list of a hand created morphosyntactic patterns. Such process is used to keep only specific *strings* and filter out undesirable ones such as candidate composed mainly of stop words (*"of a, is a, that was"*). Our algorithm operates on lemmas instead of surface forms which can draw on richer statistics and overcome the data sparseness problems. In Table 1 we give an example of MWE produced for each pattern. There exists extraction patterns (or configuration) for which no MWE has been generated (i.e. Noun-Adj). To this list are added some prepositional idiomatic expressions (*in particular, in the light of, as regards...*) and named entities (*Midle East, South Africa, El-Salvador...*) recognized by the morphosyntactic analyzer. Then, we scored all extracted MWEs with their total frequency of occurrence in the corpus. To avoid

Pattern	English/ French MWEs
Adj-Noun	Plenary meeting / Libre circulation
Noun-Adj	... / Parlement européen
Noun-Noun	Member state / Etat membre
Past_Participle -Noun	Developped country/ ...
Noun-Past_Participle	Parliament adopted/ Pays developpé
Adj-Adj-Noun	European public prosecutor /...
Adj-Noun-Adj	Social market economy / Bon conduite administratif
Adj-Noun-Noun	Renewable energy source / ...
Noun-Noun-Adj	... / Industrie automobile allemand
Noun-Adj-Adj	... / Ministère public européen
Adj-Noun-Adj	... / Important débat politique
Noun-Prep-Noun	Point of view / Chemin de fer
Noun-Prep-Adj-Noun	Court of first instance/ Court de première instance
Noun-Prep-Noun-Adj	.../ Source d'énergie renouvelable
Adj-Noun-Prep-Noun	European court of justice/ ...
Noun-Adj-Prep-Noun	.../ Politique européen de concurrence

Table 1: French and English MWE's morphosyntactic patterns

an over-generation of MWEs and remove irrelevant candidates from the process, a redundancy cleaning approach is introduced. In this approach, if a MWE is nested in another, and they both have the same frequency, we discard the smaller one. Otherwise we keep both of them. We consider also the case in which a MWE appears in a high number of terms and discard all longer ones. Because our approach does not use additional correlations statistics such as Mutual Information or Log Likelihood Ratio, it finds translations for all extracted MWEs (both frequent and infrequent ones), to our knowledge, none of other approaches can make this claim.

4. Vector Space Model for MWEs alignment

We present a method aiming to find for each MWE in a source language its adequate translation in the target one. Traditionally, this task was handled through the use of external linguistic resources such as bilingual dictionaries or simple words alignment tools. We propose a *ressource-independant* method which simply requires a parallel corpus and a list of input MWEs candidates to translate.

Our approach is based on aspects of the distributional semantics (Harris, 1954), where a specific representation is associated to each expression (source and target). We associate to each MWE an N sized vector, where N is the number of sentences in the corpus, indicating whether it appears or not in each sentence of the corpus. Our algorithm

NUMÉROPHRASE	PHRASE
2	... to go before the courts **once more** because the public prosecutor ....
55	I would therefore **once more** ask you to ensure ....
n-1	.... and then, in September, it voted **once more** to approve ...
n	...being amended **once more** in cooperation with the...

⇓

EXPRESSION	1	2	3	4	.....	55	.....	n-1	n
once more	0	1	0	0	.....	1	.....	1	1

Figure 1: Représentation vectorielle de l'expression ≪*once more*≫

is based on the Vector Space Model (VSM). VSM (Salton et al., 1975) is a well-known algebraic model used in information retrieval, indexing and relevance ranking. This vector space representation will serve, eventually, as a basis to establish a translation relation between each pair of MWE. Figure 1 illustrates a vector representing the English MWE *"once more"*.

To extract translation pairs of MWEs, we propose an iterative alignment algorithm operating as follows:

1. Find the most frequent MWE exp in each source sentence.

2. Extract all target translation candidates, appearing in all parallel sentences to those containing exp.

3. Compute a confidence value V_{Conf} for each translation relation between exp and all target translation candidates.

4. Consider that the target MWE maximizing V_{Conf} is the best translation.

5. Discard the translation pair from the process and go back to 1.

To compute the confidence value V_{Conf}, we adopted the *Jaccard Index*, a frequently used measure in information retrieval. It is defined as

$$Jaccard = \frac{I_{st}}{V_s + V_t - I_{st}} \quad (1)$$

and based on the number I_{st} of sentences shared by each target and a source MWE. This is normalized by the sum of the number of sentences where the source and target MWEs appear independently of each other (V_s and V_t) decreased by I_{st}. In table 2, a sample of aligned MWEs by means of the algorithm described above.

By observing some pairs, we noticed that our method has two advantages: (1) It allows the translation of MWE aligned in most previous work (Dagan and Church, 1994; Ren et al., 2009) using simple word alignment tools to establish word-to-word alignment relations. In our work, we capture the semantic equivalence between expressions such as *"insulaire en dévelopement"* and *"small island developing"* in a different way. (2) It also permits the alignment of idioms such as *à nouveau* → *once more*.

French → English MWEs
european parliament /parlement européen
military coup / coup d'état
in favour of /en faveur de
no smoking area/ zone non fumeur
small island developing / insulaire en développement
good faith / de bonne foi
competition policy / politique de concurrence
process of consultation / processus de consultation
railway sector / chemin de fer
with regard to / en ce qui concerne
cut in forestation / coupe forestier

Table 2: Sample of aligned MWEs

5. Bilingual MWEs in MOSES

In the previous section, we described the approach we followed to extract translation pairs of MWEs. Because of the lack of a common benchmark data sets for evaluation in MWE extraction and alignment research, we decided to study in what respect these units are useful to improve the performance of phrase based SMT systems. In such systems, *phrase tables* are the main knowledge source for the machine translation decoder. The decoder consults these tables to figure out how to translate an input candidate in a source language in the target one. However, due to the errors in automatic word alignment, extracted phrases could be meaningless. To alleviate this problem, we propose three techniques to make use of bilingual MWEs in an SMT system and compare their performances.

5.1. Methods

5.1.1. Retraining model with MWE

In this method (noted "BASELINE+TRAIN"), we add the extracted bilingual MWE as a parallel corpus and retrain the model. By increasing the occurences of bilingual MWEs, considered as good phrases, we expect a modification of alignment and probability estimation.

5.1.2. MWEs in the phrase table

Here we attempt to extend an SMT system's phrase table by integrating the found bilingual MWEs candidates[1] as in (Bouamor et al., 2011). We, then use the Jaccard Index (proposed for each pairs of MWE) to define the two

[1]the MWEs extracted following the approach we present in section 4

directions translation probabilites and set the lexical probabilities to 1 for simplicity. So, for each phrase in a given input sentence, the decoder will search all candidate translation phrases by taking into account bilingual MWEs. This method is denoted "BASELINE+TABLE" in the remaining part of this paper.

5.1.3. New feature for MWE

(Lopez and Resnik, 2006) pointed out that better feature mining can lead to substantial gain in translation qualtity. We followed this claim and extend "BASELINE+TABLE" by adding a new feature indicating whether a phrase is a MWE or not. The aim of this method ("BASELINE+FEAT") is to guide the system to choose bilingual MWEs instead of its phrases.

5.2. Baseline system

We use the factored translation model of the Moses[2] SMT system as our baseline system. It is an extention of the phrase based models which are limited to the mappings of phrases without any explicit use of linguistic information. The factored model enables the use of additional annotations at the word level. We present a model that operates on lemmas instead of surface forms, in which the translation process is broken up into a sequence of mapping steps that either :

- Translate source lemmas into target's ones.
- Generate surface forms given the lemma.

The features used in baseline system include:(1) four translation probability features, (2) two language models, (3) one generation model and (4) word penalty. For the "BASELINE+TRAIN" method, bilingual MWEs are added into the training corpus, as result, new alignment and phrase table are obtained. For "BASELINE+TABLE" method, bilingual units are incorporated in the Baseline system's phrase table. For "BASELINE+FEAT" method, an additional $1/0$ feature is introduced to each entry of the phrase table.

5.3. Data and tools

To train the SMT system's translation model, we used a training set of 100000 parallel sentences extracted from the French-English Europarl Corpus (Koehn, 2005). This corpus groups a set of parallel sentences extracted from the Proceedings of the European Parliament.

	French	English
Training Sentences	100000	
Words	2656209	2537762
Test Sentences	1000	
Words	26862	24389

Table 3: Caracteristics of Training and Test data

First, we tokenized, cleaned up the training corpus and kept only sentences containing at most 50 words. Since we use the factored translation model, we annotated training data with lemmas by mean of the TreeTagger Toolkit[3]. Next, word-alignment for all the sentences in the parallel training corpus is established and use the same methodology as in phrase-based models (symmetrized GIZA++ alignments) to form a phrase table. We also specified two language models using the IRST Language Modeling Toolkit [4] to train two tri-gram models on the total size of the Europarl corpus ($1.8K$ sentences). Besides the regular language model based on surface forms, we have a second language model which is trained on lemmas. Afterwards, we extracted bilingual MWEs from the training corpus and applied the three methods described above.

5.4. Results and discussion

We test translation quality on a test set of 1000 parallel sentences extracted from the Europarl corpus against one reference per sentence, with respect to BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2011) scores. The difference between these two metrics is that, in contrast with BLEU, which uses only precision based features, METEOR uses recall and precision, with recall weighted higher than precision. Using such measures gives us information about both precision and recall of all systems. In table 4, we report obtained results.

METHOD	BLEU	METEOR
BASELINE	25.64	29.11
BASELINE+TRAIN	**25.94**	**29.26**
BASELINE+TABLE	25.67	29.15
BASELINE+FEAT	22.91	27.18

Table 4: Translation results in terms of BLEU score and METEOR scores

The first notable observation is that the two scores vary similarly for all methods. The best improvement is achieved using the BASELINE+TRAIN method which exploit MWEs as additional parallel "sentences". Compared to the BASELINE system, this method reports a gain of $+0.30$ and $+0.15$ points in respectively BLEU and METEOR scores. The BASELINE+TABLE comes next with a slightly higher BLEU and METEOR scores with respectively an improvement of $+0.03$ and $+0.04$ points. However, the BASELINE+FEAT method have lower scores compared to the BASELINE system. We assume that we obtain such lower scores because, while adding a feature to guide the SMT system in choosing the best translation with preference to MWEs, it neglects other units and consequently fail to propose a good translation.
In order to know in what respect using bilingual MWEs improve translation quality, we manually analyzed the test sentence presented in Figure 2. From observing bilingual data sets, it become evident that in some cases, it is just impossible to perform a word to word alignment between two phrases that are translation of each other. For example, certain combination of words might convey a meaning which

[2] http://www.statmt.org/moses

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[4] http://hlt.fbk.eu/en/irstlm

Source Sentence	Ce n'est que ces dernières années que la plupart des états membres ont investi dans l'amélioration des ***chemins de fer*** et parfois également dans la navigation intérieure.
Reference	Only in the last few years have most member states invested in improving the ***railways*** and sometimes inland shipping too.
BASELINE	They will be that this last year that most member states have invested in improving the ***way to go to fer*** and sometimes also in the navigation internal.
BASELINE+TABLE	They will be that this last year that most member states have invested in improving the ***railways sector*** and sometimes also in the internal navigation.

Figure 2: Translation example

is somehow independent from the words in contains. This is the case of bilingual pairs such as *"Chemin de fer"* and *"Railways"*. We can notice from the example presented above that a word to word alignment stratedy is adopted in the BASELINE system. It provides the following alignments for words contained in the previous example:

- "chemin"="way to go to"
- "de"=Not Translated
- "fer"=Not translated

Here, the French word *"chemin"* was translated into the English phrase *"way to go to"* and the word *"fer"* was not translated since there is no entry in the baseline system's phrase table to which we can associate it. While it is aligned to the target MWE *"railways sector"* in BASELINE+TABLE. We can consider that this is a correctly translated phrase as much as it keeps the same meaning.

6. Conclusion and Future work

We described, in this paper a method aiming to extract and align MWEs in a French-English parallel corpus. The alignment algorithm we propose works only on many to many correspondences and deal with both frequent and infrequent MWEs in a given sentence pair.

We also investigated the performance of three different application strategies by integrating bilingual MWEs in the MOSES SMT system. Results show that, using MWEs as additional parallel sentences to train the translation model improves the best.

Although our initial experiments are positive, we believe that they can be improved in a number of ways. We first plan to extend the morphosyntactic patterns to handle with other forms of MWEs, e.g. starting with a verb. We will also try to develop and evaluate other statistical based methods to align MWEs. In addition to their application in a phrase based SMT system, bilingual MWEs may also be integrated into other MT models such as rule-based translation ones. We also expect to extract such textual units from more available but less parallel data sources: *comparable corpora*.

Acknowledgments

This research work is supported by FINANCIALWATCH (QNRF NPRP: 08-583-1-101) project. This publication was made possible by a grant from the Qatar National Research Fund NPRP 08–583–1–101. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the QNRF.

7. References

R. Besançon, G. De Chalendar, O. Ferret, F. Gara, M. Laib, O. Mesnard, and N. Semmar. 2010. Lima: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. In *Proceedings of LREC*, Malta.

D. Bouamor, N. Semmar, and P. Zweigenbaum. 2011. Improved statistical machine translation using multi-word expressions. In *Proceedings of MT-LIHMT*, Barcelona, Spain.

S. Boulaknadel, B. Daille, and A. Driss. 2008. A multi-term extraction program for arabic language. In *Proceedings of LREC*, Marrakech, Morocco.

M. Constant, I. Tellier, D. Duchier, Y. Dupont, A. Sigogne, S. Billot, et al. 2011. Intégrer des connaissances linguistiques dans un crf: application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Actes de TALN*, Montpellier, France.

I. Dagan and K. Church. 1994. Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on ANLP*, pages 34–40, Stuttgart, Germany.

B. Daille. 2001. Extraction de collocation à partir de textes. In Denis Maurel, editor, *Actes de TALN 2001 (Traitement automatique des langues naturelles)*, Tours, July. ATALA, Université de Tours.

M. Denkowski and A. Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

M. Finlayson and N. Kulkarni. 2011. Detecting multiword expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 20–24, Portland, Oregon, USA.

Z.S. Harris. 1954. Distributional structure. *Word*.

R. Jackendoff. 1997. The architecture of the language faculty. *MIT Press*.

P. Koehn, F. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Lan-*

guage Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 115–124, Edmonton, Canada.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT-SUMMIT*.

J. Kupiec. 1993. An algorithm for finding noun phrases correspondences in bilingual corpora. In *Proceedings of the 31st annual Meeting of the Association for Computational Linguistics*, pages 17–22, Columbus, Ohio, USA.

P. Lambert and R. Banchs. 2005. Data inferred multi-word expressions for statistical machine translation. In *Proceedings of MT SUMMIT*.

P. Lambert and R. Banchs. 2006. Grouping multi-word expressions according to part-of-speech in statistical machine translation. In *Proceedings of the Workshop on Multi-word Expressions in a multilingual context*.

A. Lopez and P Resnik. 2006. Word-based alignment, phrase based translation:what's the link? In *Proceedings of the association for machine translation in the Americas: visions for the future of machine translation*, pages 90–99.

N. Luka, V. Seretan, and E. Wehrli. 2006. Le problème de collocation en tal. In *Nouveaux cahiers de linguistiques Française*, pages 95–115.

T. Okita, M. Guerra, Y. Alfredo Graham, and A. Way. 2010. Multi-word expression sensitive word alignment. In *Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010*, pages 26–34, Beijing.

k. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th Annual meeting of the Association for Computational Linguistics*.

Z. Ren, Y. Lu, Q. Liu, and Y. Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications*, pages 47–57.

I. Sag, T. Baldwin, F. Francis Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions:a pain in the neck for nlp. In *CICLing 2002*, Mexico City, Mexico.

B. Sagot, L. Clément, É. De La Clergerie, P. Boullier, et al. 2005. Vers un méta-lexique pour le français: architecture, acquisition, utilisation. In *Actes de TALN*.

G. Salton, A. Wong, and C. Yang. 1975. A vector space model for automatic indexing. In *Communications of the ACM*, pages 61–620.

V. Seretan and E. Wehrli. 2007. Collocation translation based on sentence alignment and parsing. In Farah Benarmara, Nabil Hatout, Philippe Muller, and Sylwia Ozdowska, editors, *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse, June. ATALA, IRIT.

I. Tufis and R. Ion. 2007. Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure. In *Proceedings of the 4th International Conference on Speech and Dialogue Systems*, pages 183–195.

O. Vechtomova. 2005. The role of multi-word units in interactive information retrieval. In *ECIR2005*, pages 403–420, Berlin.

S. Vintar and D. Fisier. 2008. Harvesting multi-word expressions from parallel corpora. In *Proceedings of LREC*, Marrakech, Morocco.

C. Wu and S. Jason. Chang. 2004. Bilingual collocation extraction based on syntactic and statistical analyses. In *Computational Linguistics*, pages 1–20.