UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Chapitre d'actes 2007

Published version Open Access

This is the published version of the publication, made available in accordance with the publisher's policy.

Collocation translation based on sentence alignment and parsing

Seretan, Violeta; Wehrli, Eric

How to cite

SERETAN, Violeta, WEHRLI, Eric. Collocation translation based on sentence alignment and parsing. In: Actes de la 14e conference sur le Traitement Automatique des Langues Naturelles (TALN 2007). Toulouse, France. [s.l.] : IRIT Press, 2007. p. 401–410.

This publication URL: https://archive-ouverte.unige.ch/unige:80051

© This document is protected by copyright. Please refer to copyright holder(s) for terms of use.

Collocation translation based on sentence alignment and parsing

Violeta SERETAN, Eric WEHRLI
Language Technology Laboratory (LATL) - University of Geneva,
2 Rue de Candolle, 1211 Geneva, Switzerland
{Violeta.Seretan,Eric.Wehrli}@lettres.unige.ch

Résumé. Bien que de nombreux efforts aient été déployés pour extraire des collocations à partir de corpus de textes, seule une minorité de travaux se préoccupent aussi de rendre le résultat de l'extraction prêt à être utilisé dans les applications TAL qui pourraient en bénéficier, telles que la traduction automatique. Cet article décrit une méthode précise d'identification de la traduction des collocations dans un corpus parallèle, qui présente les avantages suivants : elle peut traiter des collocation flexibles (et pas seulement figées) ; elle a besoin de ressources limitées et d'un pouvoir de calcul raisonnable (pas d'alignement complet, pas d'entraînement) ; elle peut être appliquée à plusieurs paires des langues et fonctionne même en l'absence de dictionnaires bilingues. La méthode est basée sur l'information syntaxique provenant du parseur multilingue Fips. L'évaluation effectuée sur 4000 collocations de type verbe-objet correspondant à plusieurs paires de langues a montré une précision moyenne de 89.8% et une couverture satisfaisante (70.9%). Ces résultats sont supérieurs à ceux enregistrés dans l'évaluation d'autres méthodes de traduction de collocations.

Abstract. To date, substantial efforts have been devoted to the extraction of collocations from text corpora. However, only a few works deal with the subsequent processing of results in order for these to be successfully integrated into the NLP applications that could benefit from them (e.g., machine translation). This paper presents an accurate method for identifying translation equivalents of collocations in parallel text, whose main strengths are that : it can handle flexible (not only rigid) collocations ; it only requires limited resources and computation (no full alignment, no training needed) ; it deals with several language pairs, and it can even work when no bilingual dictionary is available. The method relies heavily on syntactic information provided by the Fips multilingual parser. Evaluation performed on 4000 verb-object collocations for different language pairs showed an average accuracy of 89.8% and a reasonable coverage (70.9%). These figures are higher that those reported in the evaluation of related work in collocation translation.

Mots-clés : traduction de collocations, extraction de collocations, parsing, alignement de textes.

Keywords: collocation translation, collocation extraction, parsing, text alignment.

1 Introduction

Collocations constitute a subclass of phraseological units (or *multi-word expressions*) that received particular attention in several research fields—e.g., second language learning, lexicography and natural language processing—both because of their massive presence in language and because of their specific features : although they look similar to regular constructions, they are unpredictable for non-native speakers and usually do not have a literal translation. Consider, for instance, the collocation *to break a record*. A non-native speaker of English would hardly choose *break* as the support verb for the noun *record*. Moreover, this collocation does not translate in a word-for-word fashion into French (**casser un record*), but as *battre un record* (lit., *to beat a record*).

For several decades already, sustained efforts have been put into developing methods for the automatic extraction of collocations from text corpora, as well as into the evaluation of extraction results ; see (Seretan & Wehrli, 2006) for a thorough review. Considerably less work deals instead with the post-processing of extracted collocations and with their further integration into other NLP applications, like machine translation, natural language generation, parsing, word sense disambiguation, or text classification. Among the few exceptions, we can mention works dealing with the semantic classification of collocations (Wanner *et al.*, 2006), the extraction of synonymous collocations (Wu & Zhou, 2003), the translation of collocations (Smadja *et al.*, 1996; Lü & Zhou, 2004), or the use of collocations in machine translation (Orliac & Dillinger, 2003), in natural language generation (Heid & Raab, 1989), and in text classification (Williams, 2002). Unfortunately, these efforts remained generally isolated and at an early stage of development, despite the largely acknowledged critical role played by such expressions in many NLP tasks (Sag *et al.*, 2002) and the continuous improvement of extraction techniques.

This paper describes a method for obtaining collocation equivalents from sentence-aligned texts that is based on the parsing of source and target sentences. The main advantages of this method are that it can deal with flexible (as opposed to rigid) collocations ; it does not require an extensive computation or huge training resources ; and it does not rely crucially on the availability of bilingual dictionaries.

The paper is organized as follows. Section 2 presents a review of previous work on collocation translation. Section 3 introduces our method and briefly describes the processing modules on which it relies (the multilingual parser, the collocation extractor, and the sentence aligner). Experimental results, an evaluation of the method and the error analysis are presented in section 4. Section 5 concludes the paper by discussing the relative merits of the newly introduced method with respect to existing methods, and by pointing out the ways in which this method can be improved in order to attain better performance.

2 Previous Collocation Translation Work

Corpus-based collocation translation has previously been dealt with in a number of works. One of the earliest is (Kupiec, 1993), that identifies noun phrase correspondences between English and French from Hansard parallel corpus. Both source and target corpora are POS-tagged, then NPs are detected with a finite-state recognizer. For mapping correspondences, the author uses Expectation Maximization (EM), an iterative re-estimation algorithm. The method was evaluated on a small set of 100 NPs, and achieved 90% accuracy. Also, Van der Eijk (1993) performed

a similar experiment for Dutch to English, but the reported accuracy was lower, since the evaluation was performed on a larger test set. This method differs from (Kupiec, 1993) in that it uses relative frequencies for computing the mappings between source and target terms.

Pursuing the same goal, Dagan and Church (1994) use a word aligner to propose (multiple) candidate translations for rigid noun phrases. Unlike the previous approaches, their system, TERMIGHT, has the advantage of being able to find translations even for infrequent terms. But like the preceding systems, it deals with rigid constructions only.

Later, the first proper collocation translation system, Champollion, has been implemented by Smadja *et al.* (1996). It relies on Xtract (Smadja, 1993) for detecting source collocations in English, then it applies a statistical correlation metric, namely the Dice coefficient, for identifying their translation equivalents in the aligned French sentences in Hansard corpus. Noticeably, this system can also deal with flexible collocations (e.g., verbal phrases). It requires an additional post-processing step in which the order of words in a flexible collocation is decided, as no syntactic information is available. The system has been evaluated by three human annotators, and showed a precision of 77% and, respectively, 61% on two different test sets of 300 collocations each.

Finally, the work of Lü and Zhou (2004) can deal with flexible collocations as well ; moreover, these are validated syntactic constituents, since extracted with a dependency parser. The syntactic types considered are verb-object, adjective-noun, and adverb-verb. Collocations are extracted from monolingual corpora in English and Chinese by applying the log-likelihood ratios statistical test on the dependency pairs identified. The translation is then performed with a statistical translation model that estimates word translations with EM. The head and the dependent word are assigned uneven probabilities, while the dependency relation is considered to be preserved across languages. The method (whose reported coverage is 83.98%) has been evaluated on a test set of 1000 randomly selected collocations. It achieved between 50.85% and 68.15% accuracy, depending on the syntactic type.

3 Translating Collocations Using Parsing Information

3.1 The method

The translation procedure we developed involves a series of steps relying on other processing modules, shortly described below. The procedure assumes that a parallel corpus is available, and that both the source and target languages are supported by the parser. First, collocations are extracted from the source corpus by using a hybrid extraction procedure (section 3.3) that combines a standard statistical technique with the deep syntactic analysis performed with the Fips parser (section 3.2). In the next step, for each collocation pair extracted, a limited number of sentence contexts is selected amongst all its contexts of occurrence in the source corpus ; in our present experiments, we considered a maximum of 50 contexts for each collocation.

The source sentences are then aligned using a sentence-aligner (section 3.4) and the equivalent target sentences are gathered into a small corpus specific to each source collocation. As the whole translation procedure is automatic, no manual validation is performed on the resulting sentence alignments. The corpus of target sentences is subsequently processed with Fips, and candidate collocation pairs are extracted using the same method as in the case of source collo-

cations. Finally, to find out the translation of the source collocation given these candidate pairs, we apply a matching procedure, which is described in section 3.5.

3.2 The Fips parser

Fips is a deep symbolic parser based on generative grammar concepts that was developed at LATL over the last decade (Wehrli, 2006). It is written in Component Pascal and adopts an object-oriented implementation design allowing to couple language-specific processing with a generic core module. The parsing algorithm proceeds in a bottom-up fashion, by applying (general or language-specific) licensing rules, by treating alternatives in parallel, and by using pruning heuristics. The parser currently supports the following languages: English, French, Spanish, Italian, German (other languages are under development as well).

In Fips, each syntactic constituent is represented as a simplified X-bar structure of the form $[_{XP}$ L X R] with no intermediate level, where X is a variable ranging over the set of lexical categories[1]. L and R stand for (possibly empty) lists of, respectively, left and right subconstituents. The lexical level contains detailed morphosyntactic and semantic information available from the manually-built lexicons. In the structures returned by the parser, extraposed elements (interrogative phrases, relative pronouns, clitics, etc.) are coindexed with empty constituents in canonical positions (i.e., typical argument or adjunct positions).

3.3 Collocation extraction with Fips

The first step in the collocation extraction process is the identification of collocation candidates. Once a sentence has been parsed by Fips, the resulting structure is checked for potential collocational pairs, by recursively examining all the pairs consisting of the head of a phrase and an element of one of its left or right subconstituents. Those pairs that satisfy certain constraints are retained as valid collocation candidates. The constraints may refer both to the lexical items individually, and to the combination as a whole. For instance, proper nouns and auxiliary verbs are ruled out, and combinations are considered valid if in a configuration like the following[2]: A-N: *wide range*, N-A: *work concerned*, N-N: *food chain*, N(subject)-V: *rule applies*, V-N(object): *strike balance*, V-P: *reflect upon*, V-P-N(argument or adjunct): *comply with rule*, N-P-N: *fight against terrorism*, V-A: *steer clear*, V-Adv: *desperately need*, Adv-A: *highly controversial*, A-P: *favourable to*, coordinated A-A: *nice and warm*, coordinated N-N: *part and parcel*. It is worth noting that each lexical item may in turn be a complex lexeme (e.g., a compound or a collocation), like *death penalty* in *abolish the death penalty* ; such a lexeme can be recognized as a single lexical item by Fips, if is is present in its lexicon.

In the second extraction step, the candidate pairs that have been identified in step one are partitioned into syntactically-homogeneous classes, then log-likelihood ratios test (Dunning, 1993) is applied on each class. *Log-likelihood ratios* (LLR) is a statistical hypothesis test that can be used to identify statistically-significant pairs among candidates (i.e., collocations) based on lexical co-occurrence evidence organized in a so-called contingency table, for each two lexical items making up a candidate pair. This table lists, essentially, the joint frequency of the two

[1]The lexical categories are N(oun), Adj(ective), V(erb), P(reposition), Adv(erb), C(onjonction), Inter(jection), to which we add the two functional categories T(ense) and F(unctional).

[2]The list of configurations is not exhaustive. It is, in fact, growing as more corpus evidence is considered.

items, the marginal frequency of each item, and the total number of pairs in the corresponding class. The result of extraction is represented by the initial list of candidate pairs ranked according to the LLR score ; the higher the score, the more likely that the pair constitutes a collocation. More details about the extraction procedure can be found in (Nerima *et al.*, 2003).

3.4 Sentence alignment

Given a parallel corpus, a sentence alignment tool finds, for each source sentence, the corresponding sentence in the target corpus (i.e., the translation equivalent of that sentence, or the target sentence). State-of-the-art sentence aligners are based on the char-length of words in sentences, on lexical clues (e.g., numbers, cognate words) and possibly exploit the macro-structure of documents (titles, sections, paragraphs)[3].

We employed our own sentence-aligner based on lexical clues and on context-size matching for paragraph detection, followed by a one-by-one sentence alignment within the aligned paragraphs. The method, which is fully described in (Nerima *et al.*, 2003), has the advantage of computing a partial, on-the-fly sentence alignment for a given source sentence identified by the file position of a word inside that sentence. This aligner is best suited for our purpose, as it allows the rapid identification of the target sentence given an item of the source collocation, without us being forced to align the whole source and target documents. Although the aligner's accuracy is not perfect (between 88% and 93.5%), the translation results obtained with our procedure suggest that it is nonetheless satisfactory for this specific task.

3.5 The matching procedure

To actually translate a collocation, we try to match it against the collocation candidates extracted from the associated target sentences. Like in (Lü & Zhou, 2004), we assume that the mapping between the source collocation and the target collocation preserves the syntactical relation involved, meaning, for instance, that a verb-object collocation in French translates into a verb-object collocation in English[4]. Therefore, we first perform a syntactic filter on the target candidate pairs that retains only the appropriate pairs, i.e., those that involve the same syntactic relation as the source collocation.

We then (optionally) apply a 'semantic' filter as follows: first, we derive from the syntactic type information about the semantic head of a collocation (usually called *base*). For instance, the base of a verb-object collocation is the object, that of an N-A collocation is the noun N, etc. Collocations are known to preserve the translation of the base, while the translation of collocate can vary across languages. For instance, in translating *break a record* into French, the noun *record* is preserved, while the verbal collocate *break* (*casser*) is transformed into *battre*. Whenever translation information for the base of the source collocation is available in our bilingual dictionaries, we consider all its possible translations and we apply a filter on the target candidate pairs accordingly ; otherwise, this filter is skipped. Finally, we select as target collocation (i.e., as translation of the source collocation) the most frequent pair among the remaining pairs, after the filters described above have been applied.

[3]Lack of space prevents us from providing more details here.

[4]This is obviously not always the case. Yet, this (simplifying) assumption was shown by Lü and Zhou (2004) to hold in the majority of cases.

4 Results and Evaluation

4.1 Translation experiment

The translation experiment described in this paper was performed on collocations extracted from a parallel corpus in four languages (English, French, Italian and Spanish), which is a subpart of Europarl parallel corpus of European Parliament proceedings (Koehn, 2005). It contains 62 files in each of the four languages that correspond to the complete proceedings for the year 2001. Table 1 displays several statistics on the corpus (rows 1–4) and on the collocations extracted with the method presented in section 3.3 (rows 5–6).

Row	Statistics	English	French	Italian	Spanish
1	Size (MB)	21.4	23.7	22.9	22.7
2	Tokens	4158622	4770835	4134549	4307360
3	Sentences	161802	162671	160906	172121
4	Average sentence length (tokens)	25.7	29.3	25.7	25
5	Total collocation pairs extracted	851500	988918	880608	901224
6	Distinct collocation pairs extracted	333428	327366	333848	315532
7	V-O pairs in translation set (500 distinct)	28005	27058	25787	23003
8	Frequency range for pairs in translation set	5–852	6–784	7–1085	6–480

TAB. 1 – Experimental statistics: corpus size, collocations extracted, translation sets size.

From the whole set of collocations extracted, we have chosen for our translation experiment the top 500 verb-object collocations obtained in each language. These 500 collocation types correspond to many more instances occurring in the corpus ; row 7 of Table 1 displays the total number of instances in each translation set, and row 8 shows the frequency range for the collocation types in each set.

The translation method described in section 3 has been applied on these translation sets in each of the possible directions. Therefore, for the 4 languages considered, there are 12 language pairs on which the method was applied. Several (randomly chosen) translations obtained are listed in Table 2.

4.2 Evaluation of results and error analysis

The random examples of translations shown in Table 2 suggest that the accuracy of our method is quite high. In fact, the evaluation performed until now shows that surprisingly good results can be achieved with this rather simple method.

The results obtained for a couple of language pairs in the translation experiment presented here have been thoroughly checked by a human judge. Whenever necessary, the contexts of the source collocation in the original documents have been inspected and confronted against the target sentences with the help of a concordancer connected to our collocation extractor and sentence aligner (Seretan *et al.*, 2004). The accuracy results for the language pairs evaluated until now are shown in the second column of Table 3. As it can be seen, comparable accuracy is achieved for the language pairs for which a bilingual (mono-lexeme) dictionary is available (90.9% to 94.1%). When such a dictionary is not available, results are worse, but still satisfactory (82.4% to 85.8%). The average accuracy obtained is 89.8%.

	Source collocation	Translation	Source collocation	Translation
En-Fr	express satisfaction	exprimer satisfaction	accroître transparence	increase transparency
	create condition	créer condition	corriger erreur	*make mistake
	improve safety	améliorer sécurité	perdre vie	lose life
	transpose directive	transposer directive	devenir réalité	become reality
	draw conclusion	tirer conclusion	remercier collègue	thank colleague
En-It	ask question	porre domanda	soddisfare requisito	meet requirements
	have opportunity	avere occasione	modificare direttiva	amend directive
	vote reason	*votare relazione	creare situazione	create situation
	thank presidency	ringraziare presidenza	apportare contributo	make contribution
	congratulate Mrs.	*svolgere lavoro	garantire livello	ensure level
En-Es	achieve goal	alcanzar objetivo	ser placer	be pleasure
	address question	abordar cuestión	recibir respuesta	receive reply
	draw list	hacer lista	ocupar lugar	take place
	play role	desempeñar papel	suspender sesión	suspend sitting
	find way	encontrar salida	*sobrar base	*draw inspiration
Fr-It	déployer effort	compiere sforzo	avere compito	avoir tâche
	transposer directive	recepire direttiva	commettere reato	commettre délit
	demander parole	chiedere parola	approvare risoluzione	adopter résolution
	vacciner animal	vaccinare animale	prendere impegno	prendre engagement
	ménager effort	lesinare sforzo	effettuare studio	mener étude
Fr-Es	poursuivre effort	continuar esfuerzo	emitir dictamen	donner avis
	éradiquer terrorisme	erradicar terrorismo	examinar cuestión	examiner question
	produire électricité	generar electricidad	hacer distinción	faire distinction
	jouer rôle	desempeñar papel	marcar hito	représenter étape
	lever obstacle	eliminar obstáculo	traspasar frontera	passer frontière
It-Es	rispettare principio	respetar principio	promover desarrollo	promuovere sviluppo
	avere impressione	tener impresión	manifestar gratitud	*ringraziare relatore
	approvare posizione	aprobar posición	tener intención	avere intenzione
	rispettare impegno	respetar compromiso	acumular retraso	accumulare ritardo
	affrontare problema	abordar problema	hacer observación	fare osservazione

TAB. 2 – Randomly chosen translation results (incorrect translations or invalid source collocations are marked with an asterisk).

The third column in Table 3 shows the method's coverage. This corresponds, in our case, to the ratio of collocation pairs for which a translation was proposed (70.9% on average). Our method does not propose a translation for a collocation when there are several translation candidates with the same frequency (previous examination of results indicated that taking all candidates in a tie introduces more noise than good translation alternatives), or when there are no candidates left after the two filters have been applied. This situation might occur for the lower frequency collocations ; our translation sets contain collocations whose frequency is as low as 5–7.

Table 3 also reports the impact of frequency on our method's performance. Accuracy and coverage have been computed separately for three frequency intervals (we distinguished between high-, medium-, and low-frequency data, corresponding to the following frequency ranges: 31–50, 16–30, and 1–15). The results obtained suggest that only a minor decrease in accuracy is observed as the frequency decreases, while the coverage is more drastically affected.

Error analysis performed on the evaluated collocations highlighted a series of issues that affect the performance of our method. Since we apply no restriction on the collocate other than the

Language pair	Accuracy				Coverage				Dictionary
	All	31–50	16–30	1–15	All	31–50	16–30	1–15	
English-French	94.1	95.6	93.3	89.8	71.4	75.8	70.7	58.3	+
English-Italian	85.8	86.2	89.3	75.7	64.8	75.5	57.1	44.0	-
French-English	92.8	94.7	89.3	92.7	72.2	80.0	65.5	59.4	+
French-Italian	92.8	91.8	96.5	87.8	72.2	79.6	66.1	59.4	+
French-Spanish	90.9	92.0	90.9	85.7	75.0	81.5	70.8	60.9	+
Italian-English	82.4	87.6	75.2	74.1	63.6	72.9	58.3	41.5	-
Italian-French	94.1	97.0	88.9	93.1	67.8	79.2	60.0	44.6	+
Italian-Spanish	85.3	89.5	80.0	77.8	80.0	89.8	75.0	55.4	-
Average	89.8	91.8	87.9	84.6	70.9	79.3	65.4	53.0	

TAB. 3 – Evaluation results (for the whole translation sets and for different frequency intervals).

syntactic filter[5], our method could propose a wrong candidate if this happens to occur systematically in the context of the right collocate and has the same syntactic type. For instance, a sentence like the one in example 1 below occurs a lot in the corpus. Our method proposes an incorrect translation for *reprendre séance*, namely **suspend a sitting*, since *suspend a sitting* occurs systematically in the context of the right candidate *resume a sitting*, and it has the same syntactic type, verb-object. Moreover, it is easier to analyse than the right candidate, which is in turn more susceptible to be missed by the parser.

1. *The sitting was suspended at 1 p.m. and resumed at 3 p.m.*
2. *This compromise formula broke the deadlock in Council and opened the door to the approval of the negotiating directives.*
   Tale formula di compromesso riuscì a sbloccare la situazione di stallo nel Consiglio e spianò la strada all'approvazione delle direttive negoziali.
3. *En tant qu'homme de science, je voudrais faire une autre remarque, Monsieur le Commissaire.*
   As a scientist, I would like to make another point, dear Commissioner.

A more recurrent situation is that in which one of the items in a collocation is lexicalized across languages, or the whole collocation is lexicalized, i.e., paraphrased as a single word. Example 2 shows the item *situazione di stallo* in the target collocation *sbloccare la situazione di stallo*, which in English is lexicalized as the single word *deadlock*. Our method incorrectly translates *break the deadlock* into *sbloccare la situazione* instead of *sbloccare la situazione di stallo*[6], since the parser does not recognize *situazione di stallo* as a lexical unit. Once this unit is added in the parser's lexicon, our method could find the good translation. An example in which the whole source collocation is lexicalized is *manifestar gratitud* shown in Table 2, whose Italian equivalent is a single word, *ringraziare*. Our method cannot handle such situations, and wrongly adds an object (**ringraziare relatore*) to the otherwise good verbal translation identified.

Quite frequent are also the situations in which the translation of a collocation is difficult to find due to the free human translations the parallel corpus contains: one can find too vague paraphrases: *hold any necessary debates*/ *participer à tous les débats nécessaires* ; omissions of a collocation item: *is to hold a debate*/ *avec le débat* ; complete omission of the collocation: *Once we have held the debate*/ *À ce moment-là*, etc. Similar problems are posed by the syntactic

[5]That is, we do not apply a semantic filter (as in the case of the base word) or finer syntactic constraints, such as imposing a syntactic structure matching between the source and target contexts.

[6]Although this kind of translations can be interpreted as partly correct, we marked them as incorrect as we did not use a gradual scale in our evaluation.

structure changes across languages (e.g., V-N vs. V-P-N: *attend meeting/ participer à reunion*, V-N vs. V-A: *pay attention/ être attentif*), or, interestingly, by negation: *It is no easy task/ Il s'agit d'une rude tâche*.

Clearly, the parsing and alignment errors as well as the coverage of the bilingual lexicons also affect our method's accuracy. If parsing and alignment errors do not influence much (as long as they can be compensated by looking at other contexts[7]), dictionary coverage problems have more drastic consequences: if a translation for a base word is missing from the dictionary and the corpus systematically contains exactly that translation, the method cannot propose a translation for the source collocation. For instance, our French-English dictionary lists, for the entry *remarque*, the following translations: *remark, comment, note*. However, the translation *point* is also needed in order for our method to identify the translation of *faire remarque* from contexts like in example 3 above, that involves the pair *make point*.

5 Conclusion

Thanks to the methodology used, the method we presented has several advantages over existing collocation translation techniques. Unlike (Kupiec, 1993; van der Eijk, 1993; Dagan & Church, 1994), it can handle flexible collocations. Unlike (Smadja *et al.*, 1996), it does not require the postprocessing of results (lexical re-ordering), since target collocations are extracted with a parser. With respect to (Lü & Zhou, 2004), it deals with more syntactic types and more languages ; it does not depend crucially on a bilingual dictionary ; it only uses mono-lexeme translations for the base word (since most of the times the collocate cannot be translated literally) ; and it is considerably simpler. In addition, it only requires several sentence contexts for a collocation, as opposed to the huge textual resources and the expensive training required by state-of-the-art phrase aligners developed in relation with statistical translation[8].

A limitation of our method is that it relies on a parallel corpus ; on the contrary, (Lü & Zhou, 2004) does not. However, in this setting our method was shown to produce quite accurate results, which suggest that adding parsing information is at least as helpful as using sophisticated statistical techniques. The method can be improved by defining syntactic configuration mappings between languages (in order to account for structure changes across languages, as those mentioned in section 4.2), by increasing the dictionaries coverage, and by including multi-word units in the parser's lexicon. Furthermore, its evaluation must be extended to other syntactic types, preferably once the syntactic mappings will be defined.

Acknowledgements

Part of the research described in this paper has been supported by a grant from the Swiss National Science Foundation (No. 101412-103999).

[7] We measured the performance of our method on low-frequency data in a separate evaluation experiment, and found that the accuracy is still acceptable for collocations occurring exactly 10 and 5 times in the corpus (85.7% and, respectively, 72.0%), but it drops to 39.1% when frequency is equal to 3 (a number of 100 English-French translation pairs have been investigated for each frequency level). The coverage is drastically affected (42%, 25% and 23%). One important reason for this degradation are the unsystematic translations found in the parallel corpus.

[8] Note that the phrase translations produced in PBSMT (Phrase-Based Statistical Machine Translation) do not have a linguistic interpretation/motivation, since a phrase simply means there any sequence of words.

Références

DAGAN I. & CHURCH K. (1994). TERMIGHT: Identifying and translating technical terminology. In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing*, Stuttgart, Germany.

DUNNING T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

HEID U. & RAAB S. (1989). Collocations in multilingual generation. In *European Chapter of the Association for Computational Linguistics (EACL'89)*, p. 130–136, Manchester, England.

KOEHN P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of The Tenth Machine Translation Summit (MT Summit X)*, p. 79–86, Phuket, Thailand.

KUPIEC J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *31st Annual Meeting of the Association for Computational Linguistics*, p. 17–22, Columbus, Ohio, U.S.A.

LÜ Y. & ZHOU M. (2004). Collocation translation acquisition using monolingual corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, p. 167–174, Barcelona, Spain.

NERIMA L., SERETAN V. & WEHRLI E. (2003). Creating a multilingual collocation dictionary from large text corpora. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, p. 131–134, Budapest, Hungary.

ORLIAC B. & DILLINGER M. (2003). Collocation extraction for machine translation. In *Proceedings of Machine Translation Summit IX*, p. 292–298, New Orleans, Lousiana, U.S.A.

SAG I. A., BALDWIN T., BOND F., COPESTAKE A. & FLICKINGER D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, p. 1–15, Mexico City.

SERETAN V., NERIMA L. & WEHRLI E. (2004). A tool for multi-word collocation extraction and visualization in multilingual corpora. In *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*, p. 755–766, Lorient, France.

SERETAN V. & WEHRLI E. (2006). Multilingual collocation extraction: Issues and solutions. In *Proceedings of COLING/ACL Workshop on Multilingual Language Resources and Interoperability*, p. 40–49, Sydney, Australia. 2006.

SMADJA F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, **19**(1), 143–177.

SMADJA F., MCKEOWN K. & HATZIVASSILOGLOU V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, **22**(1), 1–38.

VAN DER EIJK P. (1993). Automating the acquisition of bilingual terminology. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, p. 113–119, Utrecht, The Netherlands.

WANNER L., BOHNET B. & GIERETH M. (2006). Making sense of collocations. *Computer Speech & Language*, **20**(4), 609–624.

WEHRLI E. (2006). TwicPen: Hand-held scanner and translation software for non-native readers. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, p. 61–64, Sydney, Australia: Association for Computational Linguistics.

WILLIAMS G. (2002). In search of representativity in specialised corpora: Categorisation through collocation. *International Journal of Corpus Linguistics*, **7**(1), 43–64.

WU H. & ZHOU M. (2003). Synonymous collocation extraction using translation information. In *Proceeding of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, p. 120–127, Sapporo, Japan.