An integrated environment for extracting and translating collocations

Seretan, Violeta

An integrated environment for extracting and translating collocations

Violeta Seretan
Language Technology Laboratory
University of Geneva
violeta.seretan@unige.ch

Abstract

This paper describes the way collocations, which constitute an important part of the multi-word lexicon of a language, are integrated into a multilingual parser and into a machine translation system. Different processing modules concur to ensure an appropriate treatment for collocations, from their automatic acquisition to their actual use in parsing and translation. The main concerns are, first, to cope with the syntactic flexibility characterising collocations, and second, to make sure that the collocation phenomenon is modelled in a rather comprehensive manner. The paper discusses, in particular, issues such as the necessity to extract collocations from syntactically parsed text (rather than from raw text), the identification of collocations consisting of more than two words, the detection of translation equivalents in parallel texts, and the issue of representing collocational information in a lexical database. The processing framework built represents an unprecedented environment that provides an advanced and comprehensive treatment for collocations.

1. Introduction

One of the most studied phenomena in corpus linguistics is that of word collocation, concerned with the specific relations that settle between words systematically occurring together. The particular interest in this topic was not only theoretically, but also, to a large extent, practically motivated. Thus, the study of cooccurrence patterns in different theoretical frameworks, such as the contextualist theory of meaning (Firth 1957), the text cohesion theory (Halliday and Hasan 1976), the meaning-text theory (Mel'čuk 1998, Polguère 1998), or theories on the lexis-grammar interface (Hunston and Francis 2000), was backed up by pedagogical, lexicographic, and natural language processing (NLP) investigation driven by practical goals (see, among others, Cowie 1981, Sinclair 1991, Fontenelle 1997, Lewis 2000).

Collocation is a key phenomenon impacting the linguistic activity, be it in a second language learning or in a computer application context. From the perspective of text production, collocational knowledge is crucial for language students, translators, and NLP systems alike. Compare, for instance, collocations like *ride a bike* and *warm greetings* with their non-collocational counterparts **drive a bike* and **hot greetings*. Although the latter phrases, when generated by non-native speakers or by collocationally-uninformed systems, are still comprehensible and conform to the general syntactic and semantic rules of English, they do not represent the conventional way of expressing the desired meaning; such phases, called *anti-collocations* (Pearce 2001), are felt by the native speakers as either unnatural or completely inadequate. Collocational knowledge is crucial for ensuring native-like selection, i.e., the choice of the appropriate (or preferred) way to convey a given meaning.

From the opposed perspective of text analysis, collocations are very useful in a wide range of applications, thanks to their disambiguation power. For instance, as Church and Hanks (1990) indicate, homographic or homophonic words which are problematic for OCR and speech recognition (like *form* and *farm*), can be decided between by checking their respective collocates. In the case of word sense disambiguation, collocations help discriminate between senses of polysemous word in virtue of the "one sense per collocation" hypothesis, according to which words have a strong tendency to exhibit only one sense in a given collocation (Yarowsky 1993). Moreover, as shown by Fontenelle (1997) and Hindle and Rooth (1993), collocations can be used to guide attachment in parsing, in case of structural ambiguity.

Collocation extraction—the task of automatically acquiring collocations from text corpora—became in the past decades an important NLP application, and an area of intensive research. To date, substantial efforts have been devoted to devising extraction methods and to evaluating them (e.g., Lafon 1984, Church and Hanks 1990, Krenn 2000, Smadja 1993, Daille 1994, Kilgarriff *et al*. 2004, Evert 2004, Seretan and Wehrli 2006, Charest *et al*. 2007; for an inclusive review, see Seretan 2008). Candidate collocations automatically obtained from corpora were used as raw material in a number of lexicographic projects, including, notably, COBUILD (Sinclair 1995).

A more limited number of NLP studies were in contrast concerned with the post-processing of the raw extraction results, in order to made them more usable by humans or machines. Such work addressed, for instance, the issue of providing a semantic classification for collocations (Wanner *et al*. 2006), of identifying synonymous collocations (Wu and Zhou 2003), and of automatically finding a translation for collocations (Smadja *et al*. 1996, Lü and Zhou 2004). A few other studies actually dealt with the integration of results in other NLP applications which could benefit from them, e.g., natural language generation (Heid and Raab 1989), text classification (Williams 2002), or machine translation (Orliac and Dillinger 2003).

Machine translation (MT) is one of the NLP applications in which collocational knowledge plays a critical role. According to Orliac and Dillinger (2003), collocations are not only useful for MT, but they constitute the key factor in producing a more acceptable output. In fact, identifying collocations in the source text and using collocations rather than literal counterparts in the target text (e.g., *ride a bike* instead of **drive a bike*, or, in French, *poser une question* instead of **demander une question*, *battre un record* instead of **casser un record*) is of outmost importance for ensuring translation adequacy and fluency.[1]

There are two main reasons why collocations are particularly problematic for machine translation. The first is that amongst all the multi-word expressions in a language, collocations are by far the most numerous (Mel'čuk 1998); unlike other phenomena, like compounds (e.g., *wheel chair*) or idioms (e.g., *to pay lip service*), collocations occur in virtually any sentence (Howarth and Nesi 1996). The second is that collocations have a high syntactic flexibility, which makes their treatment much more difficult than that of other, more rigid types of expressions. For those languages with a relatively free word order, a superficial treatment which does not try to decode the sentence structure is insufficient for dealing with the cases involving long-distance dependencies between the component words of collocations. As many researchers (Smadja 1993, Breidt 1993, Krenn 2000, Pearce 2002, Evert 2004) have indicated, collocation should ideally be identified from syntactically parsed text rather than from plain text, as soon as appropriate tools (i.e., sufficiently robust and precise parsers) become available.

Two long-term projects that have been carried out in our laboratory since the 1990s are related to development of a robust multilingual parser, called *Fips* (Laenzlinger and Wehrli 1991, Wehrli 2007), and a large-scale rule-based MT system relying on Fips, called *ITS-2* (Wehrli 1998, Wehrli *et al*. 2009a). As collocations were an important concern in both projects, constant efforts have been directed over the years towards the design of appropriate processing modules for supporting the integration of collocations in these systems, based on the acquisition of collocational resources from text corpora.

The present paper reviews these efforts and describes our processing framework, which constitutes an unprecedented environment providing an advanced and comprehensive treatment for collocations. The rest of the paper is structured as follows. Section 2 briefly introduces the parsing and translation systems, and discusses the importance of collocations for each of them. Section 3 describes the framework itself, and provides relevant results as well as evaluation details, whenever appropriate. Section 4 reviews a couple of additional applications developed at LATL in relation to collocations and multi-word expressions in general. Section 5 concludes the paper by discussing directions for future work.

2. Collocations in a parsing and in a machine translation system

2.1 The *Fips* parser

Fips (Laenzlinger and Wehrli 1991, Wehrli 2007) is a multilingual symbolic parser originally designed for French, and later extended to English, Italian, Spanish, German, and Greek; other languages, among which Romanian, Japanese, Russian and Romansh, are currently under development. Fips can be characterised as a strong lexicalist, bottom-up, left-to-right parser which builds, for an input sentence, a rich structural representation combining:

1. the constituent structure: a parse tree reflecting the hierarchical organisation of the words in the sentence (this is similar to the *c-structure* in Lexical Functional Grammar, LFG, Bresnan 2001);
2. the interpretation of constituents in terms of arguments: a predicate-argument table identifying the grammatical relations between the main constituent of the sentence (similar to the *f-structure* in LFG);
3. the interpretation of elements like clitics, relative and interrogative pronouns in terms of intra-sentential antecedents;
4. co-indexation chains linking extraposed elements (e.g., fronted NPs and *wh* elements) to their canonical positions.

Without being actually bound to a specific theory, Fips relies on generative grammar concepts inspired by the Minimalist Program (Chomsky 1995), LFG (Bresnan 2001), and the Simpler Syntax Model (Culicover and Jackendoff 2005). A constituent is represented as a simplified A-bar structure, with no intermediate level; it has the form $[_{XP}$ L X R], where X, the lexical head, stands for the usual lexical categories (N - noun, V - verb, A - adjective etc.), and L and R denote lists of left and right sub-constituents, possibly empty.

The parsing algorithm proceeds by iteratively performing one of the following three types of operations: creation of constituent structures corresponding to the lexical entries (*Project*), combination of adjacent constituents into larger constituents (*Merge*), and movement of constituents from the canonical position to the surface position (*Move*). The application of these operations is constrained by both language-independent grammar rules (constituting the core parser engine) and language-specific rules (defined for each language supported by the parser).

Alternatives are pursued in parallel and are pruning heuristics are employed for limiting the search space.

The manually-built lexicons are a key component of the parsing system. The lexical entries contain, in fact, rich information that guides the parser (such as subcategorisation information, selectional properties, syntactico-semantic features likely to influence the parsing process). Collocational information, which is constantly added to the lexicons as it is acquired from corpora, is also part of the information which guides the analysis. Each lexical items has associated a list of collocations in which it participates. The parser checks the presence of these collocations in the current sentence; once one of the collocating words is found, it assigns it the correct lexeme reading on the basis of the information listed therein. In case of competing analyses, preference is give to the attachment of constituents whose lexical heads make up a collocation. Currently, the number of collocations in our lexicon is much lower than that of single lexemes; for instance, there are about 14000 entries for French, corresponding to about 35% of the 40000 single lexeme entries. According to theoretical stipulations, however, this number should be one order of magnitude larger (Mel'čuk 1998), hence our constant efforts to acquire more collocations.

Even if we did not yet quantify the impact that collocational entries have on the output of the parser, we expect it to be a positive one (cf. Hindle and Rooth 1993), and we are pursuing the goal of increasing the coverage of collocations in our monolingual lexicons. The need for a large collocational coverage is much more compelling in text production than it is in text analysis, as will be seen below.

2.2 The *ITS-2* machine translation system

ITS-2 (Wehrli 1998, Wehrli *et al*. 2009a) is a large-scale MT system based on syntactic transfer with the parser Fips. It aims to provide automatic translations between all the languages supported by the parser, and it uses, basically, a same generic transfer module which is further refined for each language pair. The language pairs currently supported are English/Italian/Spanish/German to French, and French to English. Like Fips, ITS-2 uses an abstract linguistic level of representation inspired from recent work in generative grammar (Chomsky 1995, Bresnan 2001, Culicover and Jackendoff 2005). This level is both rich enough to express the structural diversity of all the language pairs taken into account, and abstract enough to capture the generalizations hidden behind obvious surface diversity.

The system is intimately linked to the Fips parser, of which it exploits not only the detailed linguistic analysis built for the source sentences, but also the monolingual (source and target language) lexicons. The system's own lexical database contains bilingual equivalences defined over entries in the monolingual lexicons. After parsing the source sentence, ITS-2 transfers the rich structural representation it obtains from Fips into the target language, by recursively processing the parse tree. Lexical transfer (the mapping of lexical items from one language into another) occurs at head level, i.e., when X is processed in each constituent of the form $[_{XP}\ L\ X\ R]$. This process yields a target-language equivalent item, often (but not always) of the same category. Next, following the projection mechanism of the parser, the target structure is build on the basis of the target items obtained. A particular treatment is undergone by those constituents interpreted as predicate arguments, as their structure may in part be determined by the target predicate.

One of the most important aspects of the transfer is the handling of multi-word expressions and, in particular, of collocations. Literal translation yield, most often, to unsatisfactory—if not completely inappropriate—results. For those multi-word

expressions like compounds and some less flexible types of idioms, which undergo little or none morphological variation and which have a lexical status (in the sense that they act like single words, or words-with-spaces, as they do not allow inversion or insertion of additional words in-between), the standard, lexical transfer is enough. Collocation is, on the contrary, a phenomenon at the interface of lexis and syntax, which is subject to all of the grammatical operations that regular lexical combinations undergo. The transfer of collocations is therefore relatively more complex, and consists of the following steps:

1. identification in the source sentence: detecting the presence of a collocation in the source sentence is the first condition which has to be met for its successful translation. In ITS-2, source collocations are identified during the sentence analysis performed by Fips parser. Each item of the collocation is marked as such in the parse tree. The parser may recognise a collocation even if its component items are inverted or far apart in the sentence. The syntactic link may be recovered regardless of the word distance and of the superficial form. In the cases involving complex grammatical transformation, the recovery of the syntactic link is only possible through a fine analysis having to recourse to all of the information the parser provides (constituent structure, argument structure, co-indexation chains, and interpretation of certain pronouns; see Section 3.1).
2. transfer: one the source collocation has been identified and its members marked as such in order to prevent their literal translation, the system looks up the bilingual lexicon for a translation of that collocation. If no translation is found, then it returns a literal translation. Otherwise, it considers the target equivalent (either a simple or a complex lexeme, in particular, a collocation) and proposes it in the target representation.
3. generation: morphological and grammatical transformations apply on the obtained target representation, in order to generate the form of the target sentence. The application may be constrained by collocation-specific restrictions recorded in the lexicon. If no constraints have been stated, the collocation items undergo exactly the same morphosyntactic processes as ordinary lexical combinations.

Using collocational information in a MT system was shown to sensibly augment the quality of the translation output (Orliac and Dillinger 2003). This information is implicitly contained in state-of-the-art (statistical) MT systems, in the form of *n*-grams (sequences of adjacent words identified from text corpora). However, the syntactic flexibility of some collocations may lead to the fragmentation of *n*-gram data for these systems, and consequently to the failure to translate these collocations.

We compared the ability of the ITS-2 system to translate flexible collocations against that of two state-of-the-art competing systems, Google Translate[2] and Systran[3]. The test set consisted of 200 verb-object collocations in two languages, English and Italian, corresponding to 20 distinct pairs occurring in randomly-selected sentences in a corpus, and was translated with each of the three systems into French. The evaluation experiment, reported in Wehrli *et al*. (2009b), found that ITS-2 placed between the two competing systems in terms of precision[4] on the English data, while on the Italian data it ranked first. Moreover, is was found that the performance of ITS-2 is less affected by the increase in distance between the component words, than that of the other systems. This positive result, obtained in spite of the early stage of development of ITS-2, is due to the use of syntactic parsing for the identification of the source collocations in the input text.

3. The collocation processing framework

As discussed in the previous sections, paying special attention to collocations was long since a concern in the work related to the development of the Fips parser and ITS-2 translation systems at LATL. In the last years, the interest in automatically acquiring collocational knowledge from text corpora materialized into a fully-fledged extraction system (Seretan 2008). Its early development was made in collaboration with translators from an international organisation in Geneva in the framework of a joint research project (2002-2004). This project was, basically, aimed at extracting collocations from the translation archives of this organisation, and displaying the results in context by using a monolingual and a bilingual concordancer.

In its present state, the tool developed offers functionalities for:

- extracting collocations from a text corpus by first syntactically analysing it with the Fips parser;
- visualising extracted collocations with the help of a concordancer in the source document and, simultaneously via on-the-fly alignment, in the document that corresponds to its translation in a given target language, if multilingual versions exist for the source document;
- detecting a translation for the collocations extracted, by processing the target sentences corresponding to the source sentences in which collocations occur;
- extracting collocations made up of more than two items, and visualising them by means of a monolingual or bilingual concordancer;
- manually validating collocations by marking them during visualisation and storing them into a monolingual/bilingual collocation database, together with relevant details and usage examples.

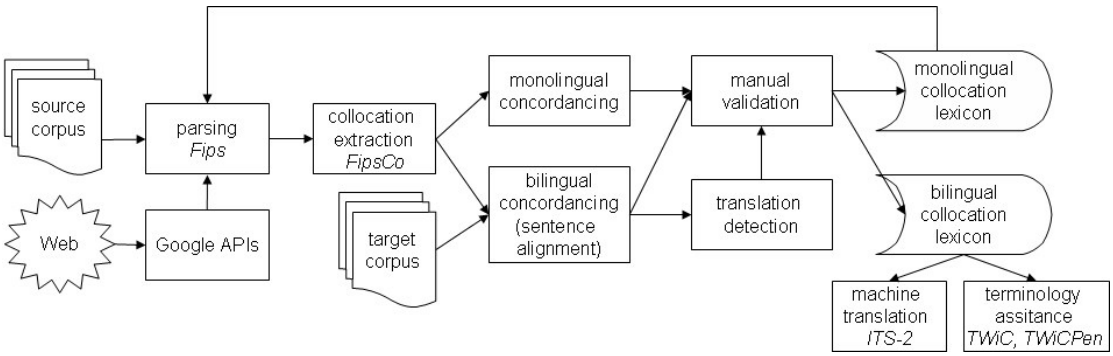

Figure 1. Architecture of the collocation processing framework

The collocation processing framework we describe in this paper is composed of the collocation extraction and visualisation tool, the parser on which it relies, the dedicated interfaces that allow lexicographers to manually validate collocations and enter them into the lexical database of Fips and ITS-2, as well as of several additional tools, including a Web-based collocation extractor and two client applications (that will be presented in Section 4). Figure 1 shows the architecture of the framework.

The processing takes place in a cyclic manner, as follows. Syntactic parsing, with, among other sources of information, the collocations already present in the parser's lexicon, is used to extract new collocations from the source (monolingual) corpus. The collocations extracted are manually added, after validation, into the lexicon of the parser and of the MT system. In the later case, a translation is either manually provided, or it is found from parallel corpora using a procedure of collocation

translation detection. The collocations stored are then used in the ITS-2 translation system, in the two applications of terminology assistance, or again in the parsing system, in a new cycle of parsing and extraction. As an alternative to text corpora, collocations with a given word may also be extracted from the Web.

The rest of this section provides details about the following modules of the framework: the monolingual collocation extractor, the translation detection module, the validation module, and the Web extractor module.

3.1 Monolingual collocation extraction

3.1.1 Method and evaluation

In *FipsCo*, the collocation extraction system implemented at LATL (Goldman *et al*. 2001, Seretan *et al*. 2004a, Seretan and Wehrli 2009), the input text—usually originating from a text corpus—is first syntactically analysed with the parser Fips, then it is processed with standard statistical methods able to grade the strength of association between words. Such methods are called *association measures*; for a comprehensive repertory, see Evert (2004); Pecina (2008).

From the (partial or complete) syntactic analyses built by the parser for each sentence, FipsCo identifies syntactic cooccurrences of words, which will constitute, together with corpus frequency information, the input for the statistical computation. The output consists of a so-called significance list, in which the cooccurrences are ranked according to their chances to constitute collocations. A significant cooccurrence is one that is not merely due to chance,[5] that is, it is not an ordinary association of words, like *(to) buy – (a) house* or *green – apple*.

FipsCo differs from other extractors in that it uses the criterion of syntactic proximity, rather than that of linear proximity, in order to identify collocation candidates in the input text. In fact, FipsCo adopts the lexicographic definition of collocation, seen as a syntactically-bound combination of words as in Benson *et al*. (1986), instead of the more widely-used definition given in purely statistical terms, according to which collocation is the frequent cooccurrence of words in a short space of each other in a text (Sinclair 1991).

A similar approach is taken in the existing extractors based on shallow parsing, e.g., SketchEngine (Kilgarriff *et al*. 2004), and in the few other extractors relying on full parsing which exist, e.g., Lin (1998) or Orliac and Dillinger (2003). Compared to the the latter extractors, FipsCo is more general and more robust. It deals with a large variety of syntactic configurations, it supports multiple languages—namely, French, English, Italian, Spanish, German, and Greek—, and the parser on which it is based has a large grammatical coverage. Indeed, Fips can handle complex constructions, such as passive-, relative-, interrogative- and cleft constructions, enumerations, coordinations, subordinations, or appositions. A few examples are provided below, which contain collocation instances (shown in italics) which have actually been extracted by FipsCo from sentences occurring in our corpora. Example (1) corresponds to a relative construction, (2) to a coordination, and (3) to an apposition.

(1) a very simple *question* which everyone in this country would like to *ask*

(2) the *problem* is therefore, clearly a deeply rooted one and cannot be *solved*

(3) the broad economic policy *guidelines*, the aims of our economic policy, do not *apply* to the euro zone alone

The question whether an extraction approach based on syntactic parsing is an efficient one, given the difficulties of analysing large amounts of raw text, the possible failures of the parser and its inherent errors, was at the heart of our concerns.

In several evaluation experiments (Seretan and Wehrli 2006, Seretan 2008, Seretan and Wehrli 2009) conducted for data in 4 languages (English, French, Spanish and Italian), we compared the performance of FipsCo in terms of precision with that of a standard extraction technique, in which collocation candidates are selected from POS-tagged text using the linear proximity criterion. Our findings were that, as far as the very first results are concerned (top 100 collocations in the significance list), the two methods are equally precise. However, when other levels of the list were investigated (from 1 to 10% of the whole list), we found that FipsCo significantly outperformed the standard syntax-free method. On these levels, the average grammatical precision of FipsCo (the percentage of grammatical results) is 86.9%, much above the baseline set by the syntax-free method, which is only 22.7%.

3.1.2 Extraction of arbitrarily long collocations

An issue which was given particular attention in our work (Seretan *et al*. 2004b) is that of fragments of collocations which are obtained with existing technology, mainly as consequence of the limitation of existing measures to binary associations.[6] An example of such fragment is *mass destruction*. A lexicographer may decide that in addition to this binary combination, it would also be useful to store in the lexicon larger collocations of which it is usually part, for instance:

(4) weapons of mass destruction
(5) proliferation of weapons of mass destruction

Such nested collocations tend to occur more often in larger constructions than independently; some other examples are *defence of rights* – part of *defence of human rights*, *proliferation of weapons* – part of *proliferation of nuclear weapons*), *to abolish a penalty* – part of *to abolish the death penalty*.

It is a well known fact that collocations may combine to yield more complex collocations, of—theoretically—unrestricted length. Researchers like Heid (1994) have long since remarked the recursive nature of collocations; yet, the practical work deals almost exclusively with collocations made up of two words. Besides the absence of association measures of higher arity (which apply to candidates longer than two items), it is the combinatorial explosion when considering all possible word combinations as candidates that hinders the extraction of longer collocations. Existing methods are therefore confined to adjacent sequences of words (*n*-grams), and are clearly inappropriate to account for the syntactic flexibility of collocations.

As the syntactic configurations which are appropriate for long collocations are not known in advance, it was impossible to follow an extraction procedure similar to the one we used for binary collocations. Instead, we found a different solution, in which we identify long collocations by relying on previously extracted binary collocations. We extend the notion of collocation from cooccurrence of words to cooccurrence of collocations. Thus, by exploiting the recursive character of collocation, we can apply the same association measures as in the case of two-word collocations.

Cooccurrence of two collocations means, more precisely, that they combine syntactically by sharing a common term in the input sentence. For instance, *contrast* is shared by both of the binary collocations *stand in contrast* and *stark contrast* identified in the sentence in (6); their combination yield the longer collocation *stand in stark contrast*.

(6) The apparent remoteness and peacefulness of the area *stand in stark contrast* to the bustling city.

Long collocations are also (partly) obtained in our framework as a side-effect of the standard extraction procedure. More precisely, when a collocation which is present in the parser's lexicon is identified in the source text by Fips, it is treated by the extractor as a single unit, and is further considered as a term of a new (binary) collocation. Thus, once *stark contrast* is added to the lexicon, the parser will recognise it in future analyses, for instance, when it processes the sentence in (6), and will consider the whole combination as the argument of the verb *stand*. The longer combination *stand – in – stark contrast* will therefore be proposed as a binary collocation candidate of verb-preposition-noun type.

3.2 Bilingual collocation extraction

The translation detection module (Seretan and Wehrli 2007) was designed to assist the work of lexicographers who enter collocations into our bilingual lexicons while consulting the extraction results using the bilingual concordancing module. Figure 2 shows the interface of the bilingual concordancer, in which extraction results (obtained from a French corpus and filtered according to different criteria—here, the syntactic type selected is verb-object and the first word of the collocation is *atteindre*) are displayed in the list on the left hand side, the original context is shown in the top text area, and the corresponding text in the translated document is found with an on-the-fly sentence alignment procedure and shown in the bottom text area.

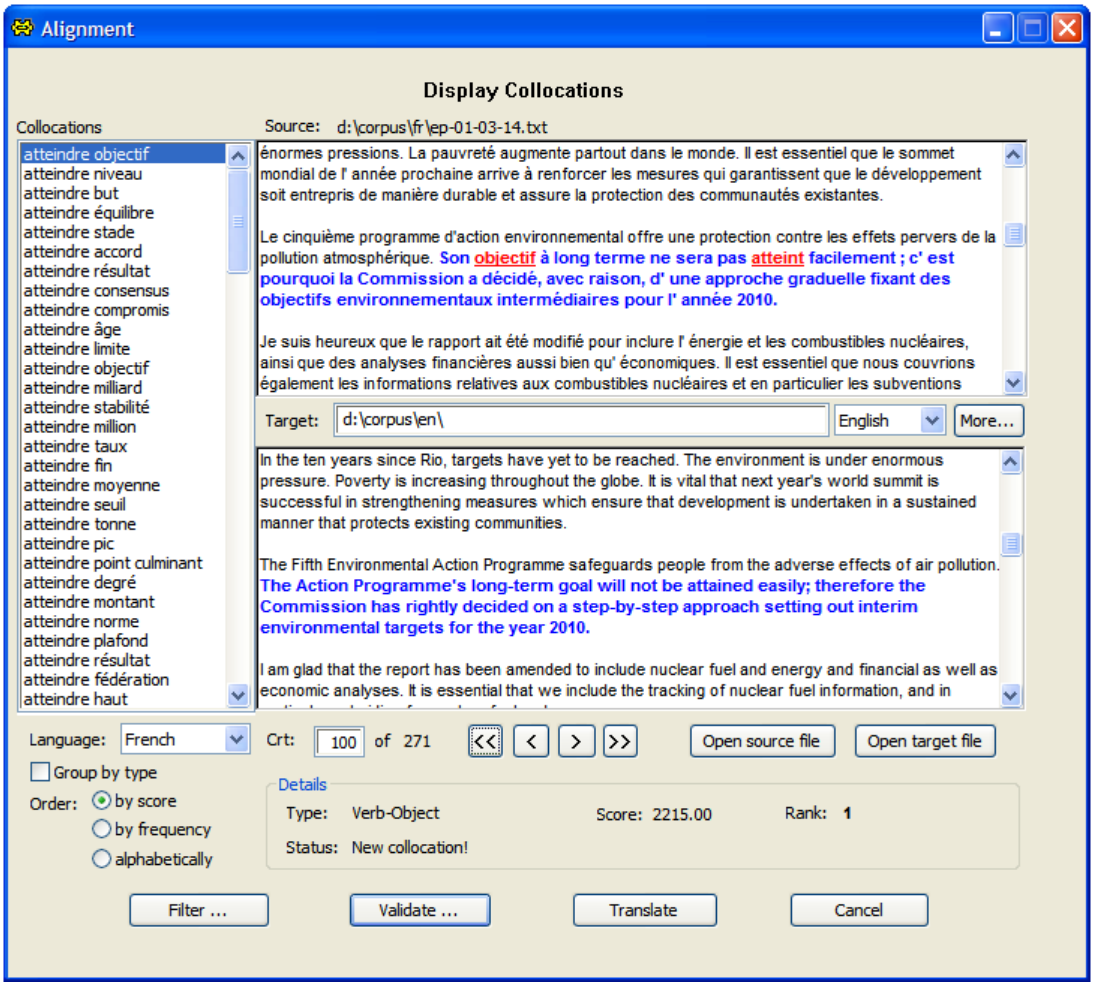

Figure 2. Parallel concordancer (screen capture)

In Figure 2, the collocation consulted is *atteindre – objectif,* and its 100th instance in the source corpus is currently shown, which occurs in the sentence *Son objectif (...)*. The corresponding target sentence is *The Action Programmes's (...)*. Lexicographers may then read the target sentence and identify the translation of the collocation in the given context, namely, the verb-object pair *(to) attain – goal*.

The goal of the translation detection module is to automate this process. The translation procedure scans a limited number of target sentences corresponding to source contexts, from which it tries to identify a potential translation for the given collocation. In doing so, it uses information such as the syntactic type of the source collocation, and the possible mappings of this type from the source language to the target language (for instance, a verb-object French collocation like *relever – défi* might be translated as a verb-preposition-argument English pair, *respond – to – challenge*). It also uses, optionally, translation information for single words. In fact, most collocations allow for a literal translation of the base word (i.e., the word that carries the meaning of the collocation and that is selected in a regular way), but not for the collocate (the other word in the collocation, whose selection is dependent on the base, cf. Mel'čuk 1998). The procedure therefore makes use of translations found for the base word in our bilingual lexicons, if any, in order to find a potential translation for the whole collocation in the target sentences.

The identification of a potential translation relies on the syntactic analysis of the target sentences; therefore, the translation procedure is functional for those target languages that are supported by Fips. The frequency of the target syntactic combinations and their association strength are also taken into account for selecting the combination that is most likely to be the translation of the source collocation.

The evaluation of the method was performed on parallel corpora in four languages, more precisely on English, French, Spanish and Italian texts from the Europarl corpus (Keohn 2005), for 8 of the 12 possible source-target language pairs. It showed an average precision of 89.8% and a recall (percentage of translated pairs) of 70.9%. Further improvements of the method (Seretan, to appear) led to an increase in recall (79.2%), while maintaining a good level of precision (84.1%).

3.3. Representation and storage

The extraction of collocations and their translation equivalents from text corpora is meant to support the lexicographers' work, but by no means to replace it. The automatic acquisition results represent raw material for inclusion in our lexical database. A collocation is only added into the collocation lexicon after manual validation. An entry in this lexicon contains detailed information, including:

- collocation key, i.e., the character string representing the 'normalized' form of the collocation (base word form, canonical word order);
- references to the lexical entries for the two composing lexemes; in case of long collocations, the composing lexemes can be collocations themselves (complex lexemes);
- preposition, for some of the collocations including a nominal component (this preposition does not count as a separate collocation component);
- syntactic type of the collocation; some of the typical types are: adjective-noun, noun-adjective, noun-noun, noun-preposition-noun, subject-verb, verb-object, verb-preposition-noun, verb-adverb, and adverb-adjective;
- frequency information, derived from corpora;
- morphosyntactic features (e.g., plural collocation, plural argument, determinerless argument, possessive noun).

In many cases, the decision of whether a word combination is a collocation, a free combination or a different type of multi-word expression (for instance, compounds or idioms) is hard to take; it is well-known that collocations are situated on a grey area of a continuum between ordinary word associations, which are fully compositional in meaning, and idiomatic expressions, which are fully non-compositional; the first are uninteresting from a lexicographic point of view, whereas the later are very important.

The approach taken at LATL is driven by practical considerations. A lexical combination is selected for inclusion in the lexicon as far as it is difficult to generate by a non-native speaker. The only distinction that is further made is between collocations and compounds. Compounds have lexical category, e.g., noun, adverb, etc.; they act like single words (or words-with-spaces) and undergo little morphologic variation (for instance, in number); all their forms can be listed in the lexicon. Collocations, on the other hand, resemble free word combination as they allow intervening material, inversion, and may, basically, undergo the same grammatical transformations as free combinations; it is not convenient to list all their forms in a lexicon.

No further classification is made between collocations and idioms or other types of multi-word expressions (like phrasal verbs, light verbs, etc). Despite the theoretical attractiveness of such a classification, this is less relevant from the practical point of view of parsing and translation, because all these complex lexical items pose the same challenge: the necessity to correctly identify and translate them, in order to prevent a word-by-word treatment.

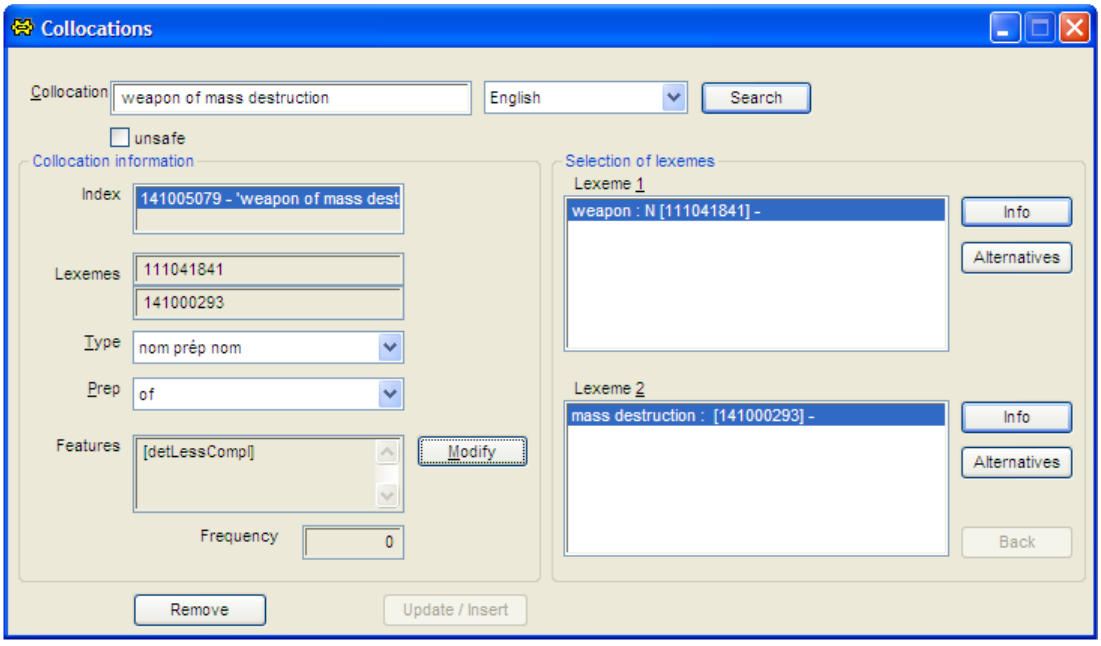

Figure 3. Insertion of collocation in the lexicon – user interface

The interface used for entering collocations into our lexical database is shown in Figure 3. When the user enters the collocation string and clicks on the search button, the parser is activated that detects the syntactic type of the collocation and assigns references (indexes) to the component lexemes. In the example shown for *weapons of mass destruction*, the second item is a complex lexeme, the collocation *mass destruction*.

3.4. Web-based extraction

Another component module of our collocation processing framework is the Web-based extraction module (see Figure 1). It provides an interface between our collocation extractor, FipsCo, and Google's search software, Google APIs.[7] This interface enables the detection of the collocates of a given word in Web documents rather than in a static pre-compiled corpus.

Suppose that a user (lexicographer, translator, language student) wants to know what are the collocations with a given word. Rather than consulting a paper dictionary or collocations pre-extracted from a given corpus, the user has the possibility to use world's largest textual resource, the Wold Wide Web, and powerful search engines to find occurrences of that specific word in Web documents; once the search results in the form of snippets (contexts of the given word in the HTML pages found) are downloaded, they are pre-processed—for instance, for detecting sentence boundaries—and then submitted to FipsCo. Collocation extraction from a relatively limited amount of text, as the one represented by a few hundreds of sentence contexts, was found to be feasible and to provide interesting results (Seretan *et al*. 2004c).

This solution relying on Web search allows us to overcome the data sparseness problem which is characteristic of text corpora.

4. Client applications: *TWiC* and *TWiCPen*

This section briefly describes two applications which use the output of the collocation processing framework described in Section 3. These have been developed in close relation to the ITS-2 translation system, and are both focused on the translation of words in context.

With the advent of the digital era and of Internet, the way users access bilingual dictionary has radically changed. The traditional keyword-based search is being increasingly replaced with an intelligent context-sensitive search, where the match with the relevant dictionary entry (or subentry) is perfected by a text analysis procedure which interprets the context in which the word sought occur. This procedure disambiguates the word in context (in case of part-of-speech ambiguity, as, for instance, between a verb and a noun, and in case of semantic ambiguity, for polysemous words), finds the base word form given the inflected form encountered in the input text, and enables the match with the most appropriate dictionary entry/subentry.

The most powerful feature of the context-sensitive dictionary look-up is, however, the ability to detect whether the word sought is part of a multi-word expression, and in this case, to return a translation for the whole expression, instead of (or in addition to) the translation of the word considered in isolation. The user might thus learn that a verb like *wreak* has a specific sense when it co-occurs with the noun *havoc* than with other words, and that *wreak havoc* is a typical language expression, i.e., a collocation.

TWiC (Wehrli 2004) is an online terminology assistance tool that offers the functionalities of a context-sensitive dictionary. It is a Web-browser plug-in that is activated when the user selects a word on a Web page. It isolates the sentence in which the word selected occur, automatically detects its language, performs its syntactic analysis using Fips, and opens a pop-up window that displays the translation of the word (in a language of user's choice) which is compatible with the context. The translation is taken from the bilingual lexicons of the ITS-2 system; in case the parser detects that the selected word is part of a collocation, TWiC accesses the bilingual collocation lexicons, cf. Figure 1.

TWiCPen (Wehrli 2006) is a similar terminology assistance tool, intended for readers of printed (off-line, rather than on-line) material. The readers select a text span (sentence, sentence fragment, paragraph, etc.) by using a hand-held scanner connected to their PDA or personal computer. The *TWiCPen* interface allows them to navigate word by word in the text, to see the translation of each word in context, similarly to TWiC, and it also provides a translation for the whole text span.

5. Conclusions

Collocational knowledge (information about which words combines with a certain word in order to make up native-like word combinations) is badly needed in any activity concerned with language synthesis, including in a learning, translation, or a natural language processing setting.

In this paper we reviewed the framework that was developed over the years at LATL, the Language Technology Laboratory of the University of Geneva, for processing collocations in order to integrate them into a multilingual syntactic analysis system and into a large-scale rule-based machine translation system. The topics addressed ranged from the automatic extraction of collocations from text corpora or from the Web, to the automatic translation of collocations based on parallel corpora, and the use of collocational knowledge in machine translation.

In light of the evaluation experiments ran for a couple of languages, we believe that the high quality of the obtained results is attributable in the first place to the availability of syntactic information for these languages, provided by the Fips parser as well as, indirectly, by its large manually-built lexical database. An important advantage of relying on syntactic information is, in particular, that collocation instances that are different at a surface level can be identified as identical at a deeper level (in the 'normalized' sentence form, in which the words are assigned their base form and their canonical order, e.g., the object following the verb in a SVO language). This led to fewer problems of data fragmentation than are otherwise encountered, especially in those languages like French, Spanish and Italian with a relatively rich morphology and flexible word order.

In future work, we will pursue our efforts acquiring collocational information from text resources for our lexical database. We also plan to improve the treatment of collocations made up or more than two words in our translation system, particularly in the generation step. Future processing of collocations that goes beyond syntactic level and looks into semantic issues (for instance, the representation of meaning and the classification according to meaning) is likely to alleviate the current processing and to enhance the usability of the collocation resources created, for the large public. This is therefore another line of research we intend to pursue in our future work.

Acknowledgements

This work has been supported in part by the Swiss National Science Foundation (grant no. 100012-117944). The author is grateful to Eric Wehrli for fruitful discussions.

Notes

[1] *Adequacy* and *fluency* are the main criteria currently used for judging the quality of a translated sentence: *adequacy* refers to the preservation of the meaning of the source sentence, and *fluency* to the naturalness of the target sentence.

[2] http://www.google.com/language tools (accessed: June 2009).

[3] http://www.systran.co.uk/ (accessed: June 2009).

[4] *Precision* refers to the percentage of correct results among the results produced by a system.

[5] In statistics, an event is called *significant* if it does not occur due to chance alone.
[6] This issue was been also discussed in the related field of terminology (Frantzi *et al*. 2000).
[7] http://www.google.com/apis/ (accessed: June 2009).

References

Benson, M., E. Benson and R. Ilson (1986). *The BBI Dictionary of English Word Combinations*. Amsterdam/Philadelphia: John Benjamins.

Breidt, E. (1993). "Extraction of V-N-Collocations from Text Corpora: A Feasibility Study for German". In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, p. 74–83, Columbus, U.S.A.

Bresnan, J. (2001). *Lexical Functional Syntax*. Oxford: Blackwell.

Charest, S., É. Brunelle, J. Fontaine and B. Pelletier (2007). "Élaboration automatique d'un dictionnaire de cooccurrences grand public". In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, pages 283–292, Toulouse, France.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, Mass.: MIT Press.

Church, K. and P. Hanks (1990). "Word association norms, mutual information, and lexicography". *Computational Linguistics*, 16, 22–29.

Cowie, A. P. (1981). "The Treatment of Collocations and Idioms in Learner's Dictionaries". *Applied Linguistics*, 2, 223–235.

Culicover, P. and R. Jackendoff (2005). *Simpler Syntax*. Oxford: Oxford University Press.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Université Paris 7.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press.

Fontenelle, T. (1997). *Turning a bilingual dictionary into a lexical-semantic database*. Tübingen: Max Niemeyer Verlag.

Frantzi, K. T., S. Ananiadou and H. Mima (2000). "Automatic recognition of multi-word terms: the C-value/NC-value method". *International Journal on Digital Libraries*, 2, 115–130.

Goldman, J.-P., L. Nerima and E. Wehrli (2001). "Collocation Extraction Using a Syntactic Parser". In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, p. 61–66, Toulouse, France.

Halliday, M. A. K. and R. Hasan (1976). *Cohesion in English*. London: Longman.

Heid, U. (1994). “On Ways Words Work Together – Research Topics in Lexical Combinatorics”. In *Proceedings of the 6th Euralex International Congress on Lexicography (EURALEX ’94)*, p. 226–257, Amsterdam, The Netherlands.

Heid, U. and S. Raab (1989). “Collocations in multilingual generation”. In *Proceeding of the Fourth Conference of the European Chapter of the Association for Computational Linguistics (EACL’89)*, p. 130–136, Manchester, England.

Hindle, D. and M. Rooth (1993). “Structural ambiguity and lexical relations”. *Computational Linguistics*, 19, 103–120.

Howarth, P. and H. Nesi (1996). “The teaching of collocations in EAP”. Technical report, University of Leeds.

Hunston, S. and G. Francis (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Kilgarriff, A., P. Rychly, P. Smrz and D. Tugwell (2004). “The Sketch Engine”. In *Proceedings of the Eleventh EURALEX International Congress*, p. 105–116, Lorient, France.

Kjellmer, G. (1994). *A Dictionary of English Collocations*. Oxford: Claredon Press.

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.

Krenn, B. (2000). *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*, volume 7. German Research Center for Artificial Intelligence and Saarland University Dissertations in Computational Linguistics and Language Technology, Saarbrücken, Germany.

Laenzlinger, C. and E. Wehrli (1991). “Fips, un analyseur interactif pour le français”. *TA informations*, 32, 35–49.

Lafon, P. (1984). *Dépouillements et statistiques en lexicométrie*. Geneva/Paris: Slatkine – Champion.

Lewis, M. (2000). *Teaching Collocations. Further Developments in the Lexical Approach*. Hove: Language Teaching Publications.

Lin, D. (1998). “Extracting Collocations from Text Corpora”. In *First Workshop on Computational Terminology*, p. 57–63, Montreal, Canada.

Lü, Y. and M. Zhou (2004). “Collocation Translation Acquisition Using Monolingual Corpora”. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL 2004)*, p. 167–174, Barcelona, Spain.

Mel'čuk, I. (1998). "Collocations and Lexical Functions". In A. P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications*. Oxford: Claredon Press, 23–53.

Orliac, B. and M. Dillinger (2003). "Collocation Extraction for Machine Translation". In *Proceedings of Machine Translation Summit IX*, p. 292–298, New Orleans, Lousiana, U.S.A.

Pearce, D. (2001). "Synonymy in Collocation Extraction". In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, p. 41–46, Pittsburgh, U.S.A.

Pearce, D. (2002). "A Comparative Evaluation of Collocation Extraction Techniques". In *Third International Conference on Language Resources and Evaluation*, p. 1530–1536, Las Palmas, Spain.

Pecina P. (2008). *Lexical Association Measures: Collocation Extraction*. Ph.D. thesis, Charles University in Prague.

Polguère, A. (1998). "La théorie Sens-Texte". *Dialangue,* 8/9, 9–30. Université du Québec à Chicoutimi.

Seretan, V. (2008). *Collocation extraction based on syntactic parsing*. Ph. D. thesis, University of Geneva.

Seretan, V., L. Nerima and E. Wehrli (2004a). "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., L. Nerima and E. Wehrli (2004b). "Multi-word collocation extraction by syntactic composition of collocation bigrams". In N. Nicolov *et al*. (eds) *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003*. Amsterdam & Philadelphia: John Benjamins, 91–100.

Seretan, V., L. Nerima and E. Wehrli (2004c). "Using the Web as a corpus for the syntactic-based collocation identification". In *Proceedings of International Conference on Language Resources and Evaluation (LREC 2004)*, p. 1871–1874, Lisbon, Portugal.

Seretan, V. and E. Wehrli (2006). "Accurate Collocation Extraction Using a Multilingual Parser". In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, p. 953–960, Sydney, Australia.

Seretan, V. and E. Wehrli (2007). "Collocation translation based on sentence alignment and parsing". In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, p. 401–410, Toulouse, France.

Seretan, V. (to appear). "Extraction de collocations et leurs équivalents de traduction à partir de corpus parallèles". *TAL*, 50.

Seretan, V. and E. Wehrli (2009). “Multilingual collocation extraction with a syntactic parser”. *Language Resources and Evaluation*, 43, 71–85.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (1995). *Collins Cobuild English Dictionary*. London: Harper Collins.

Smadja, F. (1993). “Retrieving collocations from text: Xtract”. *Computational Linguistics*, 19, 143–177.

Smadja, F., K. McKeown and V. Hatzivassiloglou (1996). “Translating collocations for bilingual lexicons: a statistical approach”. *Computational Linguistics*, 22, 1–38.

Wanner, L., B. Bohnet and M. Giereth (2006). “Making Sense of Collocations”. *Computer Speech & Language*, 20, 609–624.

Wehrli, E. (1998). “Translating Idioms”. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, p. 1388–1392, Montreal, Quebec, Canada.

Wehrli, E. (2004). “Traduction, traduction de mots, traduction de phrases”. In *Actes du XIe Conférence sur le Traitement Automatique des Langues Naturelles*, 2004, p. 483–491, Fes, Morocco.

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.

Wehrli, E. (2007). “Fips, A ‘Deep’ Linguistic Multilingual Parser”. In *ACL 2007 Workshop on Deep Linguistic Processing*, p. 120–127, Prague, Czech Republic.

Wehrli, E., L. Nerima and Y. Scherrer (2009a). “Deep Linguistic Multilingual Translation and Bilingual Dictionaries”. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, p. 90–94, Athens, Greece.

Wehrli, E., V. Seretan, L. Nerima and L. Russo (2009b). “Collocations in a Rule-Based MT System: A Case Study Evaluation of Their Translation Adequacy”. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, p. 128–135, Barcelona, Spain.

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

Wu, H. and M. Zhou (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.

Yarowsky, D. (1993). “One Sense Per Collocation”. In *Proceedings of ARPA Human Language Technology Workshop*, p. 266–271, Princeton.