Collocations in a Rule-Based MT System: A Case Study Evaluation of Their Translation Adequacy

Eric Wehrli, Violeta Seretan, Luka Nerima, Lorenza Russo
Language Technology Laboratory
University of Geneva
Switzerland
{firstname.lastname}@unige.ch

Abstract

Collocations constitute a subclass of multi-word expressions that are particularly problematic for machine translation, due 1) to their omnipresence in texts, and 2) to their morpho-syntactic properties, allowing virtually unlimited variation and leading to long-distance dependencies. Since existing MT systems incorporate mostly local information, these are arguably ill-suited for handling those collocations whose items are not found in close proximity. In this article, we describe an integrated environment in which collocations (and possibly their translation equivalents) are first identified from text corpora and stored in the lexical database of a translation system, then they are employed by this system, which is capable of dealing with syntactic transformations as it is based on a deep linguistic approach. We compare the performance of our system (in terms of collocation translation adequacy) with that of two major MT systems, one statistical, and the other rule-based. Our results confirm that syntactic variation affects translation quality and show that a deep syntactic approach is more robust in this sense, especially for languages with freer word order (e.g., German) and richer morphology (e.g., Italian) than English.

1 Introduction

Collocations, typical word combinations in a given syntactic relation (e.g., *warm greeting, distinct preference, [to] wreak havoc, [to] believe firmly, [to] break a record*) constitute a well-known problem for machine translation. Their identification in the source text and their proper processing by MT systems is the key factor in producing a more acceptable output (Orliac and Dillinger, 2003).

Over the past two decades, intensive efforts have been made to devise accurate techniques for collocation extraction from corpora; see, for instance, Church and Hanks (1990), Smadja (1993), Lin (1998), Evert (2004), among many others. Yet, the existing MT systems generally do not integrate collocational resources, or they are not designed to handle collocations in a specific and appropriate manner, as required by their high morpho-syntactic potential.

Therefore, such systems often achieve an unsatisfactory literal translation, especially when the collocation items are not found in the canonical order or in close proximity.[1] For instance, Example (1b) show the French translation returned by a major MT system, freely available online, for the English sentence in (1a), where the order of the verb and object of the collocation *break - record* is changed due to passivisation, and there are several words occurring between the two.

(1)a. *Records* are made to be *broken*.
b. *Les *dossiers* sont faites pour être *rompu*.
[The files are made for be broken.]

Since the system tested fails to identify that *records* and *broken* are part of a collocation, it is unable to propose a correct translation (in this case the French collocation *battre - record*), as it would normally do for a less problematic sentence, like *I want to break a record* (*Je veux battre un record*).

Another cause of failure appears to be the occurrence of collocations in atypical contexts: for

[1] According to Goldman et al. (2001, 62), as many as 30 words may intervene between the collocation items in a sentence.

instance, *give support* is correctly translated when found in a context like in Example (2a) (*give full support*), but not when it occurs in a less typical context like (2b) (*give massive support*).

(2)a. the people who rely on us to *give* full *support* when it is needed [...] → les gens qui comptent sur nous pour *apporter* leur plein *appui* quand il est nécessaire

b. and it is certainly right to *give* massive *support* to these areas [...] → et il est certainement droit de **donner* un *soutien* massif à ces domaines

Such examples indicate a high sensitivity of MT systems to the syntactic environment of the source collocations, which is clearly an issue given their marked syntactic flexibility. This further suggests that the collocation translation quality may seriously be affected for those source languages in which the word order is particularly free.

This paper describes the way collocations are treated in a large in-house machine translation system. The first condition for achieving an adequate translation for collocations is their accurate identification in the source sentence; in our system, this step is ensured by the detailed syntactic analysis provided by a deep syntactic parser.

Section 2 briefly states what exactly we mean by collocation, and indicates the challenges they pose to MT. Section 3 introduces our MT system, ITS-2, and provides details on its lexical database, as well as on the method used for extracting collocations (and their equivalents) from corpora. Section 4 describes the transfer method used by our system for translating collocations, then Section 5 presents an evaluation of the potential of our system to properly translate collocations.

2 Collocations

An agreed-upon definition of *collocations* does not exist yet; however, they are generally understood as a subtype of multi-word expressions that constitute arbitrary, conventional associations of words within a particular syntactic configuration.[2]

Unlike idioms, which exhibit either an opaque meaning—e.g., *to kick the bucket*, *to pull one's leg*—or very limited syntactic freedom, collocations have a fairly transparent meaning and are not subject to particular syntactic restrictions. Thus, a collocation of the type verb-object, such as *to break - record*, can be found in passive constructions, relatives or wh-interrogative clauses. Both of its components can undergo adverbial and adjectival modification, just like any verb and noun, as illustrated in Example (3):

(3)a. John *broke* the world *record*.

b. The world *record* has been *broken*.

c. The *record* that John *broke* was established in 2003.

d. In 1935, Jesse Owens set a long jump world *record* that was not *broken* until 1960 by Ralph Boston.

What makes collocations important for translation (and, in particular, for MT) is the fact that a large number of them do not translate well literally. It is therefore crucial to properly identify them and to dispose of the necessary bilingual resources to provide an adequate translation. The high frequency of collocations—several authors report a frequency of at least one collocation per sentence on average (Sinclair, 1991; Howarth and Nesi, 1996)—makes them a central issue in translation and motivates our particular interest in that matter.

In the remainder of the discussion we will restrict our attention to collocations of the verb-object type. This is one of the most common types of collocations, along with the adjective-noun type. At the same time, it is arguably the type that is the hardest to identify, due to the high frequency of extraposition of the object (as will be discussed in Section 4).

The non-identification of collocations dramatically affects the quality of the output. Collocations, which are in their vast majority semantically unambiguous (Yarowsky, 1993), are typically made of very common words, which in isolation may be polysemous (e.g., *break* in *break - record*). If the recognition of a collocation fails, the sense disambiguation information it carries is no longer available. This means that (even though a literal translation of collocations could in principle often result in an understandable if not fully adequate translation) the risk of choosing a wrong target word is rather high, making the literal translation option rather risky.

3 Our translation system

3.1 Overview

ITS-2 is a large-scale translation system developed in our laboratory, LATL, in the last couple of years (Wehrli, 1998; Wehrli et al., 2009). The language pairs currently supported are: English, German, Italian and Spanish to French, French-German, and French-English.

[2] See Heid (1994), Fontenelle (2001), Mel'čuk (2003), Grossmann and Tutin (2003) or Seretan (2008) for more detailed descriptions of the concept.

ITS-2 relies on an abstract linguistic level of representation, largely inspired from recent work in generative grammar (Chomsky, 1995; Bresnan, 2001; Culicover and Jackendoff, 2005). This level of representation is both rich enough to express the structural diversity of all the languages taken into account, and abstract enough to capture the generalizations hidden behind obvious surface diversity.

At the software level, an object-oriented design has been used, similar to the design adopted for the Fips multilingual parser on which it relies (Wehrli, 2007). To a large extent, ITS-2 can be viewed as an extension of the parser. It relies heavily on the detailed linguistic analysis provided by the parser for the supported languages, and exploits the lexical information of its monolingual lexicons. Both systems aim to set up a generic module which can be further refined to suit the specific needs of, respectively, a particular language or a particular language-pair.

The translation algorithm follows the traditional pattern of a transfer system. First, the input sentence is parsed by the parser, producing an information-rich phrase-structure representation with associated predicate-argument representations. The parser also identifies multi-word expressions such as idioms and collocations; this point is further detailed in Section 4.

Then, the transfer module maps the source-language abstract representation into the target-language representation. Given the abstract nature of this level of representation, the mapping operation is relatively simple and can be sketched as follows: recursively traverse the source-language phrase structure in the order: head, right subconstituents, left subconstituents. Lexical transfer (the mapping of a source-language lexical item with an equivalent target-language item) occurs at the head-transfer level (provided the head is not empty); it yields a target-language equivalent term, often (but by no means always) of the same category. Following the projection principle used in the parser, the target-language structure is projected on the basis of the lexical item which is its head.

However, the projections (i.e., constituents) which have been analyzed as arguments of a predicate undergo a slightly different transfer process, since their precise target-language properties may be in part determined by the subcategorization features of the target-language predicate. To take a simple example, the direct object of the French verb *regarder* in (4a) will be transferred to English as a prepositional phrase headed by the preposition *at*, as illustrated in (5a). This information comes from the lexical database. More specifically, the French-English bilingual lexicon specifies a correspondence between the French lexeme $[_{\text{VP}}$ regarder NP] and the English lexeme $[_{\text{VP}}$ look $[_{\text{PP}}$ at NP]]. For both sentences, we also illustrate the syntactic structures as built by the parser and/or the generator of ITS-2:

(4)a. Paul regardait la voiture.

b. $[_{\text{TP}}$ $[_{\text{DP}}$ Paul] regardait$_i$ $[_{\text{VP}}$ e$_i$ $[_{\text{DP}}$ la $[_{\text{NP}}$ voiture]]]]

(5)a. Paul was looking at the car.

b. $[_{\text{TP}}$ $[_{\text{DP}}$ Paul] was $[_{\text{VP}}$ looking $[_{\text{PP}}$ at $[_{\text{DP}}$ the $[_{\text{NP}}$ car]]]]]]

3.2 The lexical database

The lexical database of ITS-2 is composed of several monolingual and bilingual lexicons. For each language supported by the underlying parser, the monolingual lexicons contain:

i) a table of lexemes, containing the base form and syntactic (as well as some semantic) information for words;

ii) a table of words, containing all the inflected forms for the entries in the table of lexemes;

iii) a table of collocations, which contains, in fact, multi-word expressions (including compound words and idioms as well).

For compound words, the storage structure used is the same as for simple words. Compounds are categorized, according to a lexical category, with their relevant syntactic features, and are recorded with all their inflected forms.

For collocational and idiomatic expressions, a uniform structure is used, which essentially contains the reference to the component words. Unlike compound words, collocations and idioms are assigned a syntactic category. The information stored in the lexicon of collocations includes:

- the type of syntactic relation that holds between the two components (lexical items[3]) of a collocation (e.g., noun-adjective, noun-noun, noun-preposition-noun, subject-verb, verb-object);

- the reference to the two lexical items composing the collocation;

- the preposition, when applicable;

- the frozenness features (plural collocation, determinerless complement, bare noun complement, etc.).

For instance, for the verb-object collocation *to take office* the lexicon entry contains the following information:

```
type: verb-object
lexeme No. 1: lex111038161 (take,
   transitive verb)
lexeme No. 2: lex111026216 (office,
   common noun)
preposition: ∅
frozenness features:
   bareNounComplement
```

As for the bilingual lexicons used by ITS-2, they contain source-target correspondences and information useful for the lexical transfer. For storage, a relational database management system was chosen. For each language pair, the bilingual lexicon is implemented as a relational table containing the associations between lexical items of the source language (SL) to lexical items of the target language (TL). The bilingual lexicon is bi-directional, i.e., it also associates lexical items of TL to lexical items of SL.

In addition to these links, the table contains transfer information such as translation context, preferences between one to many translations, semantic descriptors, and argument matching for predicates (mostly for verbs). The table structures are identical for all pairs of languages.

3.3 Collocation extraction

The number of collocations included in the monolingual and bilingual lexicons of our translation system varies from language to language, and it is currently on the order of several thousand entries. The French monolingual lexicon is the largest, with almost 13000 entries. We estimate that a number of 15000-20000 entries for each language would ensure an acceptable coverage. To achieve this coverage, we built a tool for collocation extraction from text corpora (Seretan, 2008), which we currently employ for discovering collocation candidates for inclusion in the lexicon.[4]

The tool provides advanced functionalities for visualizing the extracted results in their original context in the source corpora, and for managing a list of validated candidates to be added to the lexicon. The tool also integrates a sentence alignment module; therefore, whenever parallel corpora are available, the lexicographer can also visualize the target sentence and identify a translation equivalent for storing it in the bilingual lexicons.

The extraction of collocations from text corpora is done by using a hybrid extraction method, which combines syntactic information provided by our parser with existing statistical methods for detecting typical lexical associations in corpora. Thus, collocation candidates are first identified from each sentence based on the parse structures returned by the parser, as lexeme combinations in a given syntactic configuration (for instance, verb-object). Then, these candidates are ranked according to their probability to constitute collocations, as computed with the log-likelihood ratio association measure (Dunning, 1993). The tool also implements a wide range of other measures that the user can choose for ranking collocation candidates.

The method implemented is similar, in principle, to other hybrid methods that were lately applied for collocation extraction (Lin, 1998; Krenn and Evert, 2001; Orliac and Dillinger, 2003; Kilgarriff et al., 2004; Charest et al., 2007). By selecting candidate collocations as pair of lexemes in a given syntactic relation (such as head-modifier or predicate-argument), these methods are much more appropriate for handling flexible collocations than the standard syntactically-uninformed methods, which rely on the linear proximity of words.

Unlike the methods cited above, in our system the syntactic relations identified are "deeper" since the underlying parsing mechanism is more

[3] We call *lexical item* a lexeme or a collocation (more precisely, an entry in the table of lexemes or collocations). Note that through recursive embedding, a collocation may be formed of collocational subparts, as stipulated by theoretical studies, e.g., (Heid, 1994). For instance, *give full support* is made of two collocations, *give - support* and *full - support*.

[4] We believe that the insertion of new collocations in the lexicon cannot be done in a fully automatic way, as it is ultimately a lexicographer who must decide whether a group of words constitutes a collocation or not.

advanced, whereas the former make use of chunking, dependency parsing, or shallow parsing only. Thanks to the syntactic analysis performed by the parser, our extractor is also able to detect instances of a collocation even when they undergo complex grammatical operations, which are typical of constructions involving verbs. Also, with respect to the other extractors mentioned, our extractor has a broader grammatical coverage and supports a larger number of languages.

When parallel corpora are available in two languages which are both supported by the parser, a translation equivalent can automatically be detected for the extracted collocations with a method described elsewhere (Seretan, 2008). Experiments run on multiple corpora of several million words have permitted a substantial increase of collocation coverage in our lexical database.

4 Collocation translation with ITS-2

The translation of collocations in the ITS-2 system takes place in three phases: identification of a source language collocation, lexical transfer, and generation of a target language collocation.

Identification The proper identification of a collocation is arguably the most difficult task in our treatment of collocations. As we have shown, collocations of the verb-object type can occur in sentences in which the two lexemes constituting the collocation can be several words apart and not even in the expected order, due to syntactic processes such as passivization or *wh*-fronting. In extreme cases, the distance between the two lexemes can exceed several dozens of words (Goldman et al., 2001).

In order to adequately handle such sentences, a comprehensive syntactic analysis is necessary, capable of interpreting extraposed (fronted) elements and of resolving intra-sentential pronominal reference, as well as (at least some) extra-sentential pronominal reference. For instance, in order to identify the collocation *break - record* in Example (6a), the parser must be able (i) to recognize the presence of a relative clause, (ii) to determine the role of the relative pronoun with respect to the verb of the relative clause (direct object), and (iii) to identify the antecedent of the relative pronoun.

(6)a. The record that John has broken.

b. [$_{\mathrm{DP}}$ the [$_{\mathrm{NP}}$ record$_i$ [$_{\mathrm{CP}}$ that$_i$ [$_{\mathrm{TP}}$ [$_{\mathrm{DP}}$ John] has [$_{\mathrm{VP}}$ broken [$_{\mathrm{DP}}$ e]$_i$]]]]]

This is what the parser does, returning a syntactic structure such as (6b), in which the index i shows the three-constituent chain connecting *record* with the direct object position of the verb *broken*.

One important function of a ”deep” syntactic parser is to establish a syntactic normalization of the sentence, that is a canonical way of representing the fundamental structure of a sentence, abstracting away from the various surface structure differences due to grammatical (or stylistic) processes. Coindexed empty categories in argument positions or functional structures are examples of normalized structures commonly used in generative grammar.

With respect to the task of collocation identification, normalization is very helpful in the sense that it provides an abstract unified and standardized representation on which the presence (or the absence) of a collocation can be computed. To illustrate this point, consider the following example:

(7)a. The *deadline* that we had *set* could not be *met*.

b. [$_{\mathrm{TP}}$ [$_{\mathrm{DP}}$ the [$_{\mathrm{NP}}$ deadline$_{i,j}$ [$_{\mathrm{CP}}$ [$_{\mathrm{DP}}$ e]$_i$ that [$_{\mathrm{TP}}$ [$_{\mathrm{DP}}$ we] had [$_{\mathrm{VP}}$ set [$_{\mathrm{DP}}$ e]$_i$]]]] could [$_{\mathrm{VP}}$ not be [$_{\mathrm{VP}}$ met [$_{\mathrm{DP}}$ e]$_j$]]]]

As shown in structure (7b), the main subject *deadline* is the head of a double chain represented by the indices i and j, respectively. The first chain, i, expresses the relationship between the head of the relative clause and the direct object position of the embedded verb (as in the previous example), while the second chain, j, represents the fronting of the direct object of the main verb to the subject position, due to the process of passivization. Thanks to the normalization computed by the parser, the task of checking the presence of a verb-object collocation is therefore greatly simplified.

Transfer and generation Once a collocation has been identified in a source language sentence, all its members are marked as collocation members in order to prevent their automatic literal translation. Thus, the lexical transfer module will check in the bilingual lexicon whether an entry exists for that collocation. If not, the literal translation will apply. If yes, two different situations can arise:

1. The target language equivalent is a simple lexeme: in this case, the syntactic head of the collocation (in the case of a verb-object collocation, the verb) will be translated by means of that lexeme.

2. The target language equivalent is itself a collocation. This is what would happen in the case of the pairs *meet - deadline* and *set - deadline* in Example (7). For instance, for the first, our English-French bilingual lexicon specifies a correspondence between *meet - deadline* and *respecter - échéance*. Based on this information, *meet* will be translated as *respecter*, and the transfer module will take note that the lexical head of the argument corresponding to the direct object of the source language verb (in that case, also a direct object) will be the French lexeme *échéance*.

The transfer yields a target language abstract representation, to which grammatical transformations (e.g., passivization and other potential extraposition transformations) and morphological generation will apply to create the target sentence. Unless restrictions have been specified in the lexical database, collocations will undergo the exact same grammatical and morphological processes as other lexical items.

5 Evaluation

5.1 The experimental setting

A first evaluation experiment has been conducted to quantify the potential of the ITS-2 system to recognize collocations in the source text and to translate them correctly, and also to compare it against two state-of-the-art translation systems available online: *Google*, a statistical-based MT system,[5] and *Systran*, a rule-based MT system.[6]

The experiment consisted of manually evaluating the adequacy of the translations proposed by the three systems on a small test set of verb-object collocations. The manual evaluation was preferred over established MT evaluation metrics (such as BLEU) since we were interested here in a more focussed evaluation (i.e., the specific subtask of collocation translation evaluation), rather than in a global evaluation of sentence translation quality. Moreover, such metrics based on word-to-word matches are not really appropriate for collocation-oriented evaluation, as they underestimate the impact that the substitution of a single word (the collocate) has on the overall sentence quality.

Two source languages were considered, English and Italian, in order to allow cross-lingual comparison. The target language considered was French. The test set contains 200 collocation instances, half in English, half in Italian, that were attested in the English, and, respectively, Italian version of the Europarl corpus (Koehn, 2005).

The test set was built as follows. First, a number of 10 collocations of type verb-object has been selected in each source language, from among the results of our previous collocation extraction experiments. Their choice was motivated by the non-literal translation into French, the (supposed) high morpho-syntactic modification potential, and the sufficient occurrence in the corpus. The selected types are displayed in Table 1 (first column); the second column shows an adequate translation into French.

Collocation (English, Italian)	**Translation (French)**
bridge gap	combler lacune
draw distinction	établir distinction
foot bill	payer facture
give support	apporter soutien
hold presidency	assurer présidence
meet condition	remplir condition
pose threat	constituer menace
reach compromise	trouver compromis
shoulder responsibility	assumer responsabilité
strike balance	trouver équilibre
assumere atteggiamento	adopter attitude
attuare politica	mener politique
avanzare proposta	présenter proposition
avviare dialogo	entamer dialogue
compiere sforzo	consentir effort
dare contributo	apporter contribution
dedicare attenzione	accorder attention
operare scelta	faire choix
porgere benvenuto	souhaiter bienvenue
raggiungere intesa	conclure accord

Table 1: Collocation types in the test set.

Second, for each collocation type a number of 10 instances was identified in the Europarl corpus,[7] and the corresponding sentences were added to the test set. The method for choosing the instances was the following: the corpus documents were sorted in the reverse order of the document frequency of the noun (i.e., the object in each verb-object pair), then the first collocation occurrence was selected from each document.

The resulting test set was submitted to the 3 systems compared. Each of the 600 total sentences obtained was evaluated by two French native speakers, who performed a binary classifica-

[5] http://www.google.com/language_tools, accessed June 2008.

[6] http://www.systran.co.uk/, accessed June 2008.

[7] More precisely, only a subpart of the corpus was considered, namely the 2001 proceedings totalling 62 files and about 4 million words per language.

tion of the translation proposed for the source collocation:

1. correct - the translation corresponds to an adequate expression of the desired meaning in the target language;

2. incorrect - the opposite holds, i.e., either the meaning is not preserved, or it is preserved but the translation proposed is felt as unnatural/weird.

Table 2 shows the inter-rater agreement statistics for each subset <language pair, system>. The kappa statistic indicate a substantial inter-annotator agreement (0.69 on average). Despite this positive result, our analysis of disagreement cases indicated that the task of judging upon the acceptability of a collocation translation is not a trivial one, and that the context plays a very important role in the judgement.

	Language Pair	**Google**	**Systran**	**ITS-2**
Obs	English-French	87	86	88
	Italian-French	72	92	94
k	English-French	0.60	0.72	0.72
	Italian-French	0.42	0.82	0.85

Table 2: Inter-rater agreement: *Obs* – observed agreement, *k* – kappa statistic (Cohen, 1960).

5.2 Results and discussion

The precision of each system was computed as the ratio of correct translations to the number of consistently-annotated instances; the pairs on which the judges disagreed were discarded (their number is quite low, as can be seen from Table 2).

The precision achieved by each system for each language pair is displayed in the first two rows of Table 3. On the English data, our system is outperformed by Google (which is unsurprising, given that the Europarl corpus is extensively used by statistical MT systems for training),[8] but performs better than Systran (which is penalized by the insufficient coverage of its collocation lexicon). However, on the Italian data, our system outperforms both Google and Systran by a large margin. Whereas the performance of these systems dramatically degrades when switching the source language, that of our system remains stable (it is actually slightly better for Italian than for English).

The next rows of Table 3 display the precision obtained when the test set is split in 3 disjoint subsets, according to the distance between the items of a collocation instance: low (distance=1,2); medium (distance=3,4) and high (distance>4).

	Language Pair	**Google**	**Systran**	**ITS-2**
all	English-French	83.9	52.3	71.6
	Italian-French	66.7	30.4	74.5
low	English-French	83.3	48.2	77.0
med	English-French	91.3	66.7	60.0
high	English-French	50.0	33.3	57.1
low	Italian-French	74.5	32.2	81.0
med	Italian-French	57.9	25.0	55.6
high	Italian-French	33.3	33.3	69.2

Table 3: Evaluation results: precision.

The values obtained show that the precision of all systems varies highly with distance, as well as from one source language to another. The collocation instances from the medium-distance subsets (i.e., those that allow 2 or 3 intervening words, like *meet - condition* in *meet the same conditions*, *conditions need to be met*) are those that are better handled by Google and Systran systems in English. In Italian, the systems appear to deal better with the low-distance subset (e.g., *sforzi compiuti*, *compiuto notevoli sforzi*). However, the three systems perform worse on the high-distance subsets. The decrease in precision is, nonetheless, lower for ITS-2: the maximal difference on subsets is 19.9%, whereas for Google is as high as 41.3%, and for Systran it is 33.4%.

This result indicates that the translation of collocations is indeed sensitive to the number of words intervening between the components items, and that beyond 3 words the precision deteriorates drastically. Our test set was, however, not balanced with respect to distance; rather, the distribution reflects the situation of a random sampling (due to the manner in which we built the test set). In the current configuration, only 9% of instances belong to the high-distance subset (while 25.5% belong to the medium-distance set, and 65.5% to the low-distance set). More investigation is needed on larger, balanced data in order to fully confirm the hypothesis that our deep syntactic approach is less affected by distance.[9]

[8] On the other hand, ITS-2 fails to identify some collocation instances and therefore to propose an appropriate translation. A preliminary error analysis has shown that this happens when the source sentences are particularly complex and the parser cannot build its complete syntactic analysis.

[9] The choice of the test corpus, Europarl, might also have an influence on the reported results, as long as the Google system used the very same corpus for training. Future evaluation on a different corpus should provide more realistic results for this system; nonetheless, the results of the current evaluation will at least serve as upperbound reference for future experiments.

6 Conclusion

In this paper we showed how collocations are treated in ITS-2, a rule-based translation system. We argued that the quality of their translation depends in the first place of their successful identification in the input text, and this benefits, in turn, from the fine-grained syntactic analysis provided by a deep parser. At least as far as verb-object collocations are concerned, their identification is a true challenge for MT systems, since they can undergo a wide range of syntactic transformations.

A case-study comparative evaluation was performed on English-French and Italian-French data against two major MT systems available online. The results showed that i) all three systems perform worse when 3 or more words occur between the collocation items; ii) ITS-2 reaches the highest precision for the verb-object collocations for which the distance between the verb and the object is high (see Table 3, rows 5 and 8); iii) moreover, ITS-2 achieves the best precision for Italian, while the precision of the other systems decreases dramatically when switching from English to Italian (Table 3, rows 1 and 2). The average precision of ITS-2 on both languages is 73.0%, i.e., slightly less than one competing system (75.3%), but higher than the other (41.4%).

Our present evaluation was specifically focused on the quality of translations obtained for verb-object collocations. In future work, this evaluation should be extended to a larger dataset, to other language pairs, other corpora, and other collocation types, in order to gain better insights on how sensitive MT systems are to the syntactic flexibility of collocation. Another possible avenue for future research is the combination of syntactic and statistical techniques, expected to yield better results than either of the two approaches alone.

Acknowledgements

This work has been supported in part by a Swiss National Science Foundation grant (no 100012-113864). The authors wish to thank Alexis Kauffmann for participating in the annotation task.

References

Bresnan, Joan. 2001. *Lexical Functional Syntax*. Blackwell, Oxford.

Charest, Simon, Éric Brunelle, Jean Fontaine, and Bertrand Pelletier. 2007. Élaboration automatique d'un dictionnaire de cooccurrences grand public. In *Proc. of TALN 2007*, pages 283–292, Toulouse, France.

Chomsky, Noam. 1995. *The Minimalist Program*. MIT Press, Cambridge, Mass.

Church, Kenneth and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Culicover, Peter and Ray Jackendoff. 2005. *Simpler Syntax*. Oxford University Press, Oxford.

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

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

Fontenelle, Thierry. 2001. Collocation modelling: from lexical functions to frame semantics. In *Proc. of the ACL Workshop on Collocation*, pages 1–7, Toulouse, France.

Goldman, Jean-Philippe, Luka Nerima, and Eric Wehrli. 2001. Collocation extraction using a syntactic parser. In *Proc. of the ACL Workshop on Collocation*, pages 61–66, Toulouse, France.

Grossmann, Francis and Agnès Tutin, editors. 2003. *Les Collocations. Analyse et traitement*. Éditions De Werelt, Amsterdam.

Heid, Ulrich. 1994. On ways words work together – research topics in lexical combinatorics. In *Proc. of EURALEX '94*, pages 226–257, Amsterdam, The Netherlands.

Howarth, Peter and Hilary Nesi. 1996. The teaching of collocations in EAP. Technical report, University of Leeds.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proc. of EURALEX 2004*, pages 105–116, Lorient, France.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT Summit X)*, pages 79–86, Phuket, Thailand.

Krenn, Brigitte and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proc. of the ACL Workshop on Collocation*, pages 39–46, Toulouse, France.

Lin, Dekang. 1998. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, pages 57–63, Montreal, Canada.

Mel'čuk, Igor. 2003. Collocations: définition, rôle et utilité. In Grossmann, Francis and Agnès Tutin, editors, *Les collocations: analyse et traitement*, pages 23–32. Editions "De Werelt", Amsterdam.

Orliac, Brigitte and Mike Dillinger. 2003. Collocation extraction for machine translation. In *Proc. of MT Summit IX*, pages 292–298, New Orleans, U.S.A.

Seretan, Violeta. 2008. *Collocation Extraction Based on Syntactic Parsing*. Ph.D. thesis, University of Geneva.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

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

Wehrli, Eric, Luka Nerima, and Yves Scherrer. 2009. Deep linguistic multilingual translation and bilingual dictionaries. In *Proc. of the Fourth Workshop on Statistical Machine Translation*, pages 90–94, Athens, Greece.

Wehrli, Eric. 1998. Translating Idioms. In *Proc. of ACL-COLING*, pages 1388–1392, Montreal, Canada.

Wehrli, Eric. 2007. Fips, a "deep" linguistic multilingual parser. In *Proc. of ACL 2007 Workshop on Deep Linguistic Processing*, pages 120–127, Prague, Czech Republic.

Yarowsky, David. 1993. One sense per collocation. In *Proc. of ARPA Human Language Technology Workshop*, pages 266–271, Princeton.