Sentence Analysis and Collocation Identification

Eric Wehrli, Violeta Seretan, Luka Nerima
Language Technology Laboratory
University of Geneva
{Eric.Wehrli, Violeta.Seretan, Luka.Nerima}@unige.ch

Abstract

Identifying collocations in a sentence, in order to ensure their proper processing in subsequent applications, and performing the syntactic analysis of the sentence are interrelated processes. Syntactic information is crucial for detecting collocations, and vice versa, collocational information is useful for parsing. This article describes an original approach in which collocations are identified in a sentence as soon as possible during the analysis of that sentence, rather than at the end of the analysis, as in our previous work. In this way, priority is given to parsing alternatives involving collocations, and collocational information guide the parser through the maze of alternatives. This solution was shown to lead to substantial improvements in the performance of both tasks (collocation identification and parsing), and in that of a subsequent task (machine translation).

1 Introduction

Collocations[1] constitute a central language phenomenon and an impressive amount of work has been devoted over the past decades to the automatic acquisition of collocational resources – as attested, among others, by initiatives like the MWE 2008 shared task aimed at creating a repository of reference data (Grégoire et al., 2008). However, little or no reference exist in the literature about the actual use made of these resources in other NLP applications.

In this paper, we consider the particular application of syntactic parsing. Just as other types of multi-word expressions (henceforth, MWEs), collocations are problematic for parsing because they have to be recognised and treated as a whole, rather than compositionally, i.e., in a word by word fashion (Sag et al., 2002). The standard approach in dealing with MWEs in parsing is to apply a "words-with-spaces" preprocessing step, which marks the MWEs in the input sentence as units which will later be integrated as single blocks in the parse tree built during analysis.

We argue that such an approach, albeit sufficiently appropriate for some subtypes of MWEs[2], is not really adequate for processing collocations. Unlike other expressions that are fixed or semi-fixed[3], collocations do not allow a "words-with-spaces" treatment because they have a high morpho-syntactic flexibility.

There is no systematic restriction, for instance, on the number of forms a lexical item (such as a verb) may have in a collocation, on the order of items in a collocation, or on the number of words that may intervene between these items. Collocations are situated at the intersection of lexicon and grammar; therefore, they cannot be accounted for merely by the lexical component of a parsing system, but have to be integrated to the grammatical component as well, as the parser has to consi-

[1] We adopt the lexicographic understanding for the term collocation (Benson et al., 1986), as opposed to the British contextualist tradition focused on statistical co-occurrence (Firth, 1957; Sinclair, 1991).

[2] Sag et al. (2002) thoroughly discusses the extend to which a "words-with-spaces" approach is appropriate for different kinds of MWEs.

[3] For instance, compound words: *by and large*, *ad hoc*; named entities: *New York City*; and non-decomposable idioms: *shoot the breeze*.

Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), pages 27–35,
Beijing, August 2010

der all the possible syntactic realisations of collocations.

Alternatively, a post-processing approach (such as the one we pursued previously in Wehrli et al. (2009b)) would identify collocations after the syntactic analysis has been performed, and output a parse tree in which collocational relations are highlighted between the composing items, in order to inform the subsequent processing applications (e.g., a machine translation application). Again, this solution is not fully appropriate, and the reason lies with the important observation that prior collocational knowledge is highly relevant for parsing. Collocational restrictions are, along with other types of information like selectional preferences and subcategorization frames, a major means of structural disambiguation. Collocational relations between the words in a sentence proved very helpful in selecting the most plausible among all the possible parse trees for a sentence (Hindle and Rooth, 1993; Alshawi and Carter, 1994; Berthouzoz and Merlo, 1997; Wehrli, 2000). Hence, the question whether collocations should be identified in a sentence before or after parsing is not an easy one. The previous literature on parsing and collocations fails to provide insightful details on how this circular issue is (or can be) solved.

In this paper, we argue that the identification of collocations and the construction of a parse tree are interrelated processes, that must be accounted for simultaneously. We present a processing model in which collocations, if present in a lexicon, are identified in the input sentence during the analysis of that sentence. At the same time, they are used to rank competing parsing hypotheses.

The paper is organised as follows. Section 2 reviews the previous work on the interrelation between parsing and processing of collocations (or, more generally, MWEs). Section 3 introduces our approach, and section 4 evaluates it by comparing it against the standard non-simultaneous approach. Section 5 provides concluding remarks and presents directions for future work.

2 Related Work

Extending the lexical component of a parser with MWEs was proved to contribute to a significant improvement of the coverage and accuracy of parsing results. For instance, Brun (1998) compared the coverage of a French parser with and without terminology recognition in the preprocessing stage. She found that the integration of 210 nominal terms in the preprocessing components of the parser resulted in a significant reduction of the number of alternative parses (from an average of 4.21 to 2.79). The eliminated parses were found to be semantically undesirable. No valid analysis were ruled out. Similarly, Zhang and Kordoni (2006) extended a lexicon with 373 additional MWE lexical entries and obtained a significant increase in the coverage of an English grammar (14.4%, from 4.3% to 18.7%).

In the cases mentioned above, a "words-with-spaces" approach was used. In contrast, Alegria et al. (2004) and Villavicencio et al. (2007) adopted a compositional approach to the encoding of MWEs, able to capture more morpho-syntactically flexible MWEs. Alegria et al. (2004) showed that by using a MWE processor in the preprocessing stage of their parser (in development) for Basque, a significant improvement in the POS-tagging precision is obtained. Villavicencio et al. (2007) found that the addition of 21 new MWEs to the lexicon led to a significant increase in the grammar coverage (from 7.1% to 22.7%), without altering the grammar accuracy.

An area of intensive research in parsing is concerned with the use of lexical preferences, co-occurrence frequencies, collocations, and contextually similar words for PP attachment disambiguation. Thus, an important number of unsupervised (Hindle and Rooth, 1993; Ratnaparkhi, 1998; Pantel and Lin, 2000), supervised (Alshawi and Carter, 1994; Berthouzoz and Merlo, 1997), and combined (Volk, 2002) methods have been developed to this end.

However, as Hindle and Rooth (1993) pointed out, the parsers used by such methods lack precisely the kind of corpus-based information that is required to resolve ambiguity, because many of the existing attachments may be missing or wrong. The current literature provides no indication about the manner in which this circular problem can be circumvented, and on whether flexible MWEs should be processed before, during or after the sentence analysis takes place.

3 Parsing and Collocations

As argued by many researchers – e.g., Heid (1994) – collocation identification is best performed on the basis of parsed material. This is due to the fact that collocations are co-occurrences of lexical items in a specific syntactic configuration. The collocation *break record*, for instance, is obtained only in the configurations where *break* is a verb whose direct object is (semantically) headed by the lexical item *record*. In other words, the collocation is not defined in terms of linear proximity, but in terms of a specific grammatical relation.

As the examples in this section show, the relative order of the two items is not relevant, nor is the distance between the two terms, which is unlimited as long as the grammatical relation holds[4]. In our system, the grammatical relations are computed by a syntactic parser, namely, Fips (Wehrli, 2007; Wehrli and Nerima, 2009). Until now, the collocation identification process took place at the end of the parse in a so-called "interpretation" procedure applied to the complete parse trees. Although quite successful, this way of doing presents a major drawback: it happens too late to help the parser. This section discusses this point and describes the alternative that we are currently developing, which consists in identifying collocations as soon as possible during the parse.

One of the major hurdles for non-deterministic parsers is the huge number of alternatives that must be considered. Given the high frequency of lexical ambiguities, the high level of non-determinism of natural language grammars, grammar-based parsers are faced with a number of alternatives which grows exponentially with the length of the input sentence. Various methods have been proposed to reduce that number, and in most cases heuristics are added to the parsing algorithm to limit the number of alternatives. Without such heuristics, the performance of a parser might not be satisfactory enough for large scale applications such as machine translation or other tasks involving large corpora.

We would like to argue, along the lines of previous work (section 2), that collocations can contribute to the disambiguation process so crucial for parsing. To put it differently, identifying collocations should not be seen as a burden, as an additional task the parser should perform, but on the contrary as a process which may help the parser through the maze of alternatives. Collocations, in their vast majority, are made of frequently used terms, often highly ambiguous (e.g., *break record*, *loose change*). Identifying them and giving them high priority over alternatives is an efficient way to reduce the ambiguity level. Ambiguity reduction through the identification of collocations is not limited to lexical ambiguities, but also applies to attachment ambiguities, and in particular to the well-known problem of PP attachment. Consider the following French examples in which the prepositions are highlighted:

(1)a. ligne *de* partage *des* eaux ("watershed")
   b. système *de* gestion *de* base *de* données ("database management system")
   c. force *de* maintien *de* la paix ("peacekeeping force")
   d. organisation *de* protection *de* l'environnement ("environmental protection agency")

In such cases, the identification of a noun-preposition-noun collocation will prevent or discourage any other type of prepositional attachment that the parser would otherwise consider.

3.1 The Method

To fulfill the goal of interconnecting the parsing procedure and the identification of collocations, we have incorporated the collocation identification mechanism within the constituent attachment procedure of our parser Fips (Wehrli, 2007). This parser, like many grammar-based parsers, uses left attachment and right attachment rules to build respectively left subconstituents and right subconstituents. Given the fact that Fips' rules always involve exactly two constituents – see Wehrli (2007) for details – it is easy to add to the attachment mechanism the task of collocation identification. To take a very simple example, when the rule attaching a prenominal adjective to a noun applies, the collocation identification procedure is invoked. It first verifies that both terms bear the lexical

[4] Goldman et al. (2001) report examples in which the distance between the two terms of a collocation can exceed 30 words.

feature [+partOfCollocation], which signals that a given word is associated in our lexical database to one or several collocations, and then searches the collocation database for an adjective-noun collocation with those two terms. If successful, the corresponding parse tree will be given a high priority.

With examples such as *loose change*, the identification of the collocation will immediately relegate any (partial) analysis based on the verbal reading of either terms.

To take a somewhat more complex example, consider a verb-object collocation such as *break record*, as in example (2)[5].

(2)a. John broke a record.
b. [$_{TP}$ [$_{DP}$ John] broke [$_{DP}$ a [$_{NP}$ record]]]

Here, it is a right attachment rule which will trigger the identification procedure. To be precise, the right attachment rule in this case concerns the attachment of the noun *record* as complement of the indefinite determiner (head of the DP direct object of the verb). The identification procedure considers, in turn, all the governing nodes dominating the noun *record*, halting at the first node of category Noun[6], Verb or Adjective. In our example, the determiner node and then the verb node will be considered. Notice that the procedure will, quite correctly, identify a collocation in the French example (3a), but not in (3b), although both structures are identical. The reason has to do with the fact that the noun governing *record* in the first example is a [+number] noun, that is a classifier noun which is transparent for the identification procedure[7].

(3)a. Jean a battu un grand nombre de records.
"Jean broke a large number of records"
b. Jean a battu le détenteur du record.
"Jean has beaten the holder of the record"

As in the other examples, an analysis in which a collocation has been found is given high priority over alternatives. In the case of (2), this will relegate potential analyses based on the adjectival reading of *broke* or the verbal reading of *record*.

Notice that exactly the same procedure applies when the trace of an extraposed element is (right) inserted, as in the examples (4), which illustrate the case of *wh*-interrogative (a), relative clause (b), *tough*-movement (c).

(4)a. Which record will Paul try to break ?
b. The record Paul broke was very old.
c. This record is difficult to break.

In all such cases, that is, when the right inserted element is a trace, the identification procedure will consider its antecedent, or to be more precise, the semantic head of its antecedent. Finally, the grammatical processes involved in example (4a,c) can combine as in the more complex example (5), for which we give the slightly simplified structure with the chain of elements with index *i* extending from the fronted *wh*-phrase *which record* to the direct object position of the verb *break*, via the direct object position of the verb *consider* and the subject position of the secondary predicate (FP) headed by the [+tough] adjective *difficult*.

(5)a. Which record did Paul consider difficult to break ?

b. [$_{CP}$ [$_{DP}$ which record]$_i$ [$_{TP}$ did [$_{DP}$ Paul] [$_{VP}$ consider] [$_{DP}$ e]$_i$ [$_{FP}$ [$_{DP}$ e]$_i$ [$_{AP}$ difficult [$_{TP}$ to [$_{VP}$ break [$_{DP}$ e]$_i$]]]]]]

3.2 Complex Collocations

As stated, for instance, by (Heid, 1994), collocations can involve more than two (main) terms and it is possible to adopt a recursive definition of collocations, i.e., complex collocations can be viewed as collocations of collocations. The collocation identification procedure has been extended to handle such cases. Consider examples (6) below.

(6)a. La voiture tombera probablement en panne d'essence.
"the car will probably run out of gas"

[5] We use the following labels in our phrase-structure representations: TP-Tense phrase, for simple sentence (the S of standard CFG), CP-Complementizer phrase, for a sentence with a conjunction or a complementizer, DP-Determiner phrase for standard noun phrases (we assume the DP hypothesis, whereby the determiner constitutes the syntactic head of a noun phrase), NP-Noun phrase for nominal projections (nouns with their modifiers/complements), VP-Verb phrase, PP-Prepositional phrase, AP-Adjectival phrase, AdvP-Adverbial phrase, FP-Functional phrase (used for secondary predicates).

[6] Unless the node it marked [+number], as we will see shortly.

[7] See Fontenelle (1999) for a detailed account of transparent nouns.

b. natural language processing

c. He broke a world record.

In the French sentence (6a), *panne d'essence* (literally, "breakdown of gas", "out of gas") is a collocation of type Noun+Prep+Noun, which combines with the verb *tomber* (literally, "to fall") to form a larger collocation of type Verb+PrepObject *tomber en panne d'essence* ("to run out of gas"). Given the strict left to right processing order assumed by the parser, it will first identify the collocation *tomber en panne* ("to break down") when attaching the word *panne*. Then, reading the last word, *essence* ("gas"), the parser will first identify the collocation *panne d'essence*. Since that collocation bears the lexical feature [+partOfCollocation], the identification procedure goes on, through the governors of that item. The search succeeds with the verb *tomber*, and the collocation *tomber en panne d'essence* ("run out of gas") is identified.

4 Evaluation Experiments

In this section, we describe the experiments we performed in order to evaluate the precision and recall of the method introduced in section 3, and to compare it against the previous method (fully described in Wehrli et al. (2009b)). We extend this comparison by performing a task-based evaluation, which investigates the impact that the new method has on the quality of translations produced by a machine translation system relying on our parser (Wehrli et al., 2009a).

4.1 Precision Evaluation

The data considered in this experiment consist of a subpart of a corpus of newspaper articles collected from the on-line version of *The Economist*[8], containing slightly more that 0.5 million words. On these data, we run two versions of our parser:

- V1: a version implementing the previous method of collocation identification,
- V2: a version implementing the new method described in section 3.

[8]URL:http://www.economist.com/ (accessed June, 2010).

The lexicon of the parser was kept constant, which is to say that both versions used the same lexicon (which contains slightly more than 7500 English collocation entries), only the parsing module handling collocations was different. From the output of each parser version, we collected statistics on the number of collocations (present in the lexicon) that were identified in the test corpus. More precisely, we traversed the output trees and counted the items that were marked as collocation heads, each time this was the case (note that an item may participate in several collocations, not only one). Table 1 presents the number of collocations identified, both with respect to collocation instances and collocation types.

	V1	**V2**	**common**	**V1 only**	**V2 only**
Tokens	4716	5412	4347	399	1003
Types	1218	1301	1182	143	368

Table 1. Collocation identification results.

As the results show, the new method (column V2) is more efficient in retrieving collocation instances. It detects 696 more instances, which correspond to an increase of 14.8% relative to the previous method (column V1). As we lack the means to compare on a large scale the corresponding syntactic trees, we can only speculate that the increase is mainly due to the fact that more appropriate analyses are produced by the new method.

A large number of instances are found by both versions of the parser. The difference between the two methods is more visible for some syntactic types than for others. Table 2 details the number of instances of each syntactic type which are retrieved exclusively by one method or by the other.

To measure the precision of the two methods, we randomly selected 20 collocation instances among those identified by each version of the parser, V1 and V2, and manually checked whether these instances are correct. Correctness means that in the given context (i.e., the sentence in which they were identified), the word combination marked as instance of a lexicalized collocation is indeed an instance of that collocation. A counterexample would be, for instance, to mark the pair *decision - make* in the sentence in (7) as

Syntactic type	V1	V2	Difference V2-V1
A-N	72	152	80
N-N	63	270	207
V-O	22	190	168
V-P-N	6	10	4
N-P-N	1	62	61
V-A	25	166	141
P-N	200	142	-58
N&N	6	2	-4
Adv-Adv	4	9	5

Table 2. Differences between the two methods: number of tokens retrieved exclusively by each method.

an instance of the verb-object collocation *to make a decision*, which is an entry in our lexicon.

(7)a. The *decision* to *make* an offer to buy or sell property at price is a management decision that cannot be delegated to staff.

Since judging the correctness of a collocation instance in context is a rather straightforward task, we do not require multiple judges for this evaluation. The precision obtained is 90% for V1, and 100% for V2.

The small size of test set is motivated by the fact that the precision is expected to be very high, since the presence of both collocation components in a sentence in the relevant syntactic relation almost certainly means that the recognition of the corresponding collocation is justified. Exceptions would correspond to a minority of cases in which the parser either wrongly establishes a relation between two items which happen to belong to an entry in the lexicon, or the two items are related but the combination corresponds to a literal usage (examples are provided later in this section).

The errors of V1 correspond, in fact, to cases in which a combination of words used literally was wrongly attributed to a collocation: in example (8a), V1 assigned the words *on* and *business* to the lexical entry *on business*, and in example (8b), it assigned *in* and *country* to the entry *in the country*[9].

(8)a. It is not, by any means, specific to the countryside, but it falls especially heavily *on* small *businesses*.

b. Industrial labour costs in western Germany are higher than *in* any other *country*.

To better pinpoint the difference between V1 and V2, we performed a similar evaluation on an additional set of 20 instances, randomly selected among the collocations identified exclusively by each method. Thus, the precision of V1, when measured on the tokens in "V1 only", was 65%. The precision of V2 on "V2 only" was 90%. The 2 errors of V2 concern the pair *in country*, found in contexts similar to the one shown in example (8b). The errors of V1 also concerned the same pair, with one exception – the identification of the collocation *world trade* from the context *the destruction of the World Trade Centre*. Since *World Trade Centre* is not in the parser lexicon, V1 analysed it and assigned the first two words to the entry *world trade*. *World* was wrongly attached to *Trade*, rather than to *Centre*.

When reported on the totality of the instances tested, the precision of V1 is 77.5% and that of V2 is 95%. Besides the increase in the precision of identified collocations, the new method also contributes to an increase in the parser coverage[10], from 81.7% to 83.3%. The V1 parser version succeeds in building a complete parse tree for 23187 of the total 28375 sentences in the corpus, while V2 does so for 23629 sentences.

4.2 Recall Evaluation

To compare the recall of two methods we performed a similar experiment, in which we run the two versions of the parser, V1 and V2, on a small collection of sentences containing annotated collocation instances. These sentences were randomly selected from the Europarl corpus (Koehn, 2005). The collocations they contain are all verb-object collocations. We limit our present investigation to this syntactic type for two reasons: *a)* annotating a corpus with all instances of collocation entries in the lexicon would be a time-consuming task; and *b)* verb-object collocations are among the most syntactically flexible and therefore difficult to detect in real texts. Thus, this test set provides realistic information on recall.

[9] V1 makes the same error on (8a), but does better on (8b). These expressions are frozen and should not be treated as standard collocations.

[10] Coverage refers more precisely to the ratio of sentences for which a complete parse tree could be built.

The test set is divided in two parts: 100 sentences are in English, and 100 other in Italian, which allows for a cross-linguistic evaluation of the two methods. Each sentence contains one annotated collocation instance, and there are 10 instances for a collocation type. Table 3 lists the collocation types in the test set (the even rows in column 2 display the glosses for the words in the Italian collocations).

English	Italian
bridge gap	assumere atteggiamento 'assume' 'attitude'
draw distinction	attuare politica 'carry out' 'policy'
foot bill	avanzare proposta 'advance' 'proposal'
give support	avviare dialogo 'start' 'dialogue'
hold presidency	compiere sforzo 'commit' 'effort'
meet condition	dare contributo 'give' 'contribution'
pose threat	dedicare attenzione 'dedicate' 'attention'
reach compromise	operare scelta 'operate' 'choice'
shoulder responsibility	porgere benvenuto 'give' 'welcome'
strike balance	raggiungere intesa 'reach' 'understanding'

Table 3. Collocation types in the test set.

The evaluation results are presented in table 4. V1 achieves 63% recall performance on the English data, and 44% on the Italian data. V2 shows considerably better results: 76% on English and 66% on Italian data. The poorer performance of both methods on Italian data is explained by the difference in performance between the English and Italian parsers, and more precisely, by the difference in their grammatical coverage. The English parser succeeds in building a complete parse tree for more than 70% of the sentences in the test set, while the Italian parser only for about 60%.

As found in the previous experiment (presented in section 4.1), for both languages considered in this experiment, the new method of processing collocations contributes to improving the parsing coverage. The coverage of the English parser increases from 71% to 76%, and that of the Italian parser from 57% to 61%.

	V1	V2	Common	V1 only	V2 only
English	63	76	61	2	15
Italian	44	66	42	2	24

Table 4. Recall evaluation results: number of correct collocation instances identified.

4.3 Task-based Evaluation

In addition to reporting the performance results by using the standard measures of precision and recall, we performed a task-based performance evaluation, in which we quantified the impact that the newly-proposed method has on the quality of the output of a machine translation system. As the examples in table 3 suggest, a literal translation of collocations is rarely the most appropriate. In fact, as stated by Orliac and Dillinger (2003), knowledge of collocations is crucial for machine translation systems. An important purpose in identifying collocations with our parser is to enable their proper treatment in our translation system, a rule-based system that performs syntactic transfer by relying on the structures produced by the parser.

In this system, the translation of a collocation takes place as follows. When the parser identifies a collocation in the source sentence, its component words are marked as collocation members, in order to prevent their literal translation. When the transfer module processes the collocation head, the system checks in the bilingual lexicon whether an entry exists for that collocation. If not, the literal translation will apply; otherwise, the transfer module projects a target-language structure as specified in the corresponding target lexical entry. More precisely, the transfer yields a target language abstract representation, to which grammatical transformations and morphological generation will apply to create the target sentence. The identification of collocations in the source text is a necessary, yet not a sufficient condition for their successful translation.

In this experiment, we considered the test set described in section 4.2 and we manually evaluated the translation obtained for each collocation instance. Both subsets (100 English sentences and 100 Italian sentences) were translated into French. We compared the translations obtai-

Task	Measure	Test set	Language	Increase
Collocation identification	precision	40 instances	English	17.5%
	recall	200 instances	English, Italian	17.5%
		100 instances	English	13%
		100 instances	Italian	22%
Collocation translation	precision	200 instances	{English, Italian}-French	13%
		100 instances	English-French	10%
		100 instances	Italian-French	16%
Parsing	coverage	28375 sentences	English	1.6%
		200 sentences	English	5%
		200 sentences	Italian	4%

Table 5. Summary of evaluation results.

ned by relying on the versions V1 and V2 of our parser (recall that V2 corresponds to the newly-proposed method and V1 to the previous method). The use of automatic metrics for evaluating the translation output was not considered appropriate in this context, since such *n*-gram based metrics underestimate the effect that the substitution of a single word (like in our case, the verb in a verb-object collocation) has on the fluency, adequacy, and even on the interpretability of the output sentence.

The comparison showed that, for both language pairs considered (English-French and Italian-French), the version of parser which integrates the new method is indeed more useful for the machine translation system than the previous version. When V2 was used, 10 more collocation instances were correctly translated from English to French than when using V1. For the Italian-French pair, V2 helped correctly translating 16 more collocation instances in comparison with V1. This corresponds to an increase in precision of 13% on the whole test set of 200 sentences. The increase in performance obtained in all the experiments described in this section is summarized in table 5.

5 Conclusion

In this paper, we addressed the issue of the interconnection between collocation identification and syntactic parsing, and we proposed an original solution for identifying collocations in a sentence as soon as possible during the analysis (rather than at the end of the parsing process). The major advantage of this approach is that collocational information may be used to guide the parser through the maze of alternatives.

The experimental results performed showed that the proposed method, which couples parsing and collocation identification, leads to substantial improvements in terms of precision and recall over the standard identification method, while contributing to augment the coverage of the parser. In addition, it was shown that it has a positive impact on the results of a subsequent application, namely, machine translation. Future work will concentrate on improving our method so that it accounts for all the possible syntactic configurations of collocational attachments, and on extending its recall evaluation to other syntactic types.

Acknowledgements

Thanks to Lorenza Russo and Paola Merlo for a thorough reading and comments. Part of the research described in this paper has been supported by a grant from the Swiss National Science Foundation, grant no 100015-117944.

References

Alegria, Iñaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. 2004. Representation and treatment of multiword expressions in basque. In *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 48–55, Barcelona, Spain.

Alshawi, Hiyan and David Carter. 1994. Training and scaling preference functions for disambiguation. *Computational Linguistics*, 20(4):635–648.

Benson, Morton, Evelyn Benson, and Robert Ilson. 1986. *The BBI Dictionary of English Word Combinations*. John Benjamins, Amsterdam/Philadelphia.

Berthouzoz, Cathy and Paola Merlo. 1997. Statistical ambiguity resolution for principle-based parsing. In Nicolov, Nicolas and Ruslan Mitkov, edi-

tors, *Recent Advances in Natural Language Processing: Selected Papers from RANLP'97*, Current Issues in Linguistic Theory, pages 179–186. John Benjamins, Amsterdam/Philadelphia.

Brun, Caroline. 1998. Terminology finite-state preprocessing for computational LFG. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 196–200, Morristown, NJ, USA.

Firth, John R. 1957. *Papers in Linguistics 1934-1951*. Oxford Univ. Press, Oxford.

Fontenelle, Thierry. 1999. Semantic resources for word sense disambiguation: a *sine qua non*? *Linguistica e Filologia*, (9):25–43. Dipartimento di Linguistica e Letterature Comparate, Università degli Studi di Bergamo.

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

Grégoire, Nicole, Stefan Evert, and Brigitte Krenn, editors. 2008. *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*. European Language Resources Association (ELRA), Marrakech, Morocco.

Heid, Ulrich. 1994. On ways words work together – research topics in lexical combinatorics. In *Proceedings of the 6th Euralex International Congress on Lexicography (EURALEX '94)*, pages 226–257, Amsterdam, The Netherlands.

Hindle, Donald and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

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

Orliac, Brigitte and Mike Dillinger. 2003. Collocation extraction for machine translation. In *Proceedings of Machine Translation Summit IX*, pages 292–298, New Orleans, Lousiana, USA.

Pantel, Patrick and Dekang Lin. 2000. An unsupervised approach to prepositional phrase attachment using contextually similar words. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 101–108, Hong Kong, China.

Ratnaparkhi, Adwait. 1998. Statistical models for unsupervised prepositional phrase attachment. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 1079–1085, Montreal, Quebec, Canada.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 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)*, pages 1–15, Mexico City.

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

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043, Prague, Czech Republic, June.

Volk, Martin. 2002. Combining unsupervised and supervised methods for PP attachment disambiguation. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 25–32, Taipei, Taiwan.

Wehrli, Eric and Luka Nerima. 2009. L'analyseur syntaxique Fips. In *Proceedings of the IWPT 2009 ATALA Workshop: What French parsing systems?*, Paris, France.

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

Wehrli, Eric, Violeta Seretan, Luka Nerima, and Lorenza 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*, pages 128–135, Barcelona, Spain.

Wehrli, Eric. 2000. Parsing and collocations. In Christodoulakis, D., editor, *Natural Language Processing*, pages 272–282. Springer Verlag.

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

Zhang, Yi and Valia Kordoni. 2006. Automated deep lexical acquisition for robust open texts processing. In *Proceedings of LREC-2006*, pages 275–280, Genoa, Italy.