Informatics **2014**, *1*, 11-31; doi:10.3390/informatics1010011

OPEN ACCESS

informatics

ISSN 2227-9709

www.mdpi.com/journal/informatics

Review

On Collocations and Their Interaction with Parsing and Translation

Violeta Seretan

Department of Translation Technology, Faculty of Translation and Interpreting, University of Geneva, 40 Bd. du Pont-d'Arve, Geneva 1211, Switzerland; E-Mail: violeta.seretan@unige.ch; Tel.: +41-22-379-8683

Received: 1 September 2013; in revised form: 3 October 2013 / Accepted: 16 October 2013 / Published: 25 October 2013

Abstract: We address the problem of automatically processing collocations—a subclass of multi-word expressions characterized by a high degree of morphosyntactic flexibility—in the context of two major applications, namely, syntactic parsing and machine translation. We show that parsing and collocation identification are processes that are interrelated and that benefit from each other, inasmuch as syntactic information is crucial for acquiring collocations from corpora and, *vice versa*, collocational information can be used to improve parsing performance. Similarly, we focus on the interrelation between collocations and machine translation, highlighting the use of translation information for multilingual collocation identification, as well as the use of collocational knowledge for improving translation. We give a panorama of the existing relevant work, and we parallel the literature surveys with our own experiments involving a symbolic parser and a rule-based translation system. The results show a significant improvement over approaches in which the corresponding tasks are decoupled.

Keywords: collocations; collocation extraction; long-distance dependency; machine translation; multi-word expressions; syntactic flexibility; syntactic parsing

---

1. Introduction

Multi-word expressions—"idiosyncratic interpretations that cross word boundaries" [1]—are widely acknowledged as a key problem for Natural Language Processing (NLP). Indeed, they are seen as a "pain in the neck" for NLP [1] or a "hard nut to crack" [2]. Multi-word expressions (henceforth, MWEs) cover

a broad range of phenomena,[1] such as named entities, multi-word function words, nominal compounds, verb-particle constructions, verbal expressions, idioms, proverbs, and so on, which all have in common the fact that they have to be treated as a whole rather than on a word-by-word basis and, therefore, require a special, holistic treatment in NLP systems.

A particularly important subclass of MWEs is represented by so-called "institutionalized" phrases, or collocations (e.g., *heavy rain*, *heavy smoker*, *serious injury*, *to meet a need*, *to extend thanks*, *deeply in love*). Collocations are expressions that are relatively regular from a syntactic and semantic point of view, but are statistically idiosyncratic. The component words are, in principle, associated by means of regular grammatical processes, such as the combination of a noun with a modifier yielding a nominal phrase whose meaning can be deduced from the meaning of the parts. Yet, what is peculiar, idiosyncratic or irregular about such expressions is that they are highly preferred over alternative lexicalizations: compare, for instance, *traffic light*, which is a collocation, an institutionalized phrase, with combinations like **traffic director* or **intersection regulator*, which barely occur in language (example from [1]). Such combinations are highly language-specific, and they indicate, to a large extent, the degree of fluency of a language utterance or of the output produced by an NLP system. Even if they are decomposable into parts, they still have to be treated by computational systems in a holistic way, to avoid unnatural or awkward formulations.

According to several researchers (e.g., [4–6]), collocations are the most numerous amongst all types of MWEs. As a matter of fact, "no piece of natural spoken or written English is totally free of collocation" [7]. The importance of collocations "stands in their omnipresence" [5]. It is also important to note that, unlike most other types of multi-word expressions, collocations may occur in a wide range of syntactic patterns. The following is a list of syntactic configurations commonly associated with collocations in English: adjective-noun (*heavy smoker*), noun-(predicate)-adjective (*effort [be] devoted*), noun-noun (*suicide attack*), noun-preposition-noun (*round of negotiations*), noun-preposition (*inquiry into*), adjective-preposition (*crazy about*), subject-verb (*problem occurs*), verb-object (*meet requirement*), verb-preposition-argument (*bring to boil*), verb-preposition (*depend on*), adverb-verb (*fully support*), adverb-adjective (*highly important*), adjective-coordination-adjective (*nice and warm*). In addition, lexicographic evidence shows that this list can be considerably extended [8]. Summing up, collocations are idiosyncratic syntagmatic combinations, which are not restricted to a given word class or to a given set of syntactic patterns [9].

Researchers have long since attempted to characterize the phenomenon of word collocation by addressing it from many different angles. However, there is still no agreed-upon definition, and the collocation concept is generally accompanied by vagueness and confusion. Collocations remain less studied and less well understood than other types of MWEs, notably, idioms (*kick the bucket*), light-verb constructions (*to take a walk*) or verb-particle constructions (*to look up*).

In this article, we put emphasis on a particular aspect that distinguishes collocations from other expressions and that contributes to making them particularly difficult to process by computational systems: the high morphosyntactic flexibility of collocations. The component words in a collocation may, in principle, undergo the full range of morphological

[1] The reader is referred to [3] for a detailed classification of multi-word expressions and an overview of their computational treatment.

and syntactic transformations that are possible for regular combinations in language (see Examples 1 and 2). In contrast, other expressions, like named entities (*New York City*), compounds (*wheel chair*) or idioms (*to be over the moon* 'to be extremely pleased'), are relatively fixed, or frozen, this characteristic acting as a useful discrimination feature and permitting a more local (and, therefore, more computationally inexpensive) automatic treatment.

With respect to the flexible nature of collocations, it is worth noticing that in the fields of NLP and translation, the ISO 12620 data categories standard describe collocations as "[a] recurrent word combination characterized by cohesion in that the components of the collocation must co-occur within an utterance or series of utterances, even though they *do not necessarily have to maintain immediate proximity to one another*" (emphasis added). This definition highlights an essential feature of collocations, namely, the discontiguity of the component words, which is the consequence of the syntactic flexibility of these expressions.

Indeed, this discontiguity is arguably one of the biggest challenges that NLP systems face when processing collocations. As collocations exhibit (almost) full syntactic variability, processing them requires dealing with a wide range of syntactic transformation in which collocations can occur. Their high variability calls for linguistically sophisticated approaches, capable of accurately identifying collocations in many syntactic contexts and of accounting for long-distance dependencies, in order to ultimately enable their proper treatment in applications such as parsing or translation.

Another aspect on which our work is particularly focused is the integration of collocations into actual NLP pipelines, i.e., their use in client natural language applications. Developing accurate collocation identification techniques has been a main concern in the NLP field for a couple of decades already; however, the exploitation of collocations in other NLP applications has received considerably less attention. In this article, we address the problem of connecting the application of collocation extraction (or identification) with two major NLP applications, namely, syntactic parsing and machine translation. We investigate whether a synergetic approach, one in which information is shared between the task of collocation identification and the other two tasks, is more efficient than the standard approach, in which the tasks are performed independently of each other.

The article explores four main directions of work that have been pursued to a greater or lesser extent in the NLP field until now. By focusing on the interrelation between the tasks of collocation identification and syntactic parsing, we look at the benefits that can be obtained from relying on syntactic parsing for collocation extraction, and, *vice versa*, from using collocations during parsing. Then, by focusing on the interrelation between collocation identification and translation, we investigate whether translation technologies can contribute to the task of automatically detecting collocations in text corpora and, conversely, whether collocations are useful for machine translation. For each main topic, a literature overview is provided, paralleled by reports on our own experiments, confirming that a synergetic approach is preferable over individual approaches. Backed by findings from other studies on synergetic approaches (for instance, on using parsing for semantic analysis [10]), these results suggest that the NLP work, often fragmented, would benefit from greater interaction between various tasks.

The article is structured as follows. In Section 2, we focus on using syntactic parsing for collocation extraction. We survey related work and outline our own extraction methodology, which relies on full syntactic parsing for acquiring collocations from text corpora in several languages. In Section 3,

we review the work in which translation-related technologies, such as sentence and word alignment, are used for identifying collocations. Furthermore, we outline our own method of detecting translation equivalents for collocations by exploiting translation archives. Sections 4 and 5 focus on exploiting collocational knowledge for parsing and translation. In Section 4, we discuss the extent to which collocations are presently taken into account in parsing systems; then, we present an approach in which collocation identification and syntactic analysis are performed simultaneously, rather than separately, as in previous work. Section 5 addresses the question of integrating collocations and other types of multi-word expressions into machine translation systems. It also presents a study aimed at assessing the impact of collocations on the results of an in-house rule-based translation system. Finally, Section 6 concludes the paper by taking stock of the current treatment of collocations and, where appropriate, indicating more adequate processing alternatives.

2. Using Parsing for Collocation Identification

The development of *collocation extraction* as an area of research has seen the increasing adoption of linguistic analysis as an essential preprocessing step. This step allows for a more accurate identification of candidates, which are then scored by using statistical methods and, in particular, so-called *association measures* (e.g., mutual information, t-score, z-score, χ^2, log-likelihood ratio; see [11–14] for descriptions and comparative evaluations of association measures).

The preprocessing techniques gradually evolved from shallower to deeper forms of analysis, as increasingly advanced technology became available, from tokenization, stemming and lemmatization to chunking, shallow parsing, dependency parsing or full parsing. The need for performing a linguistic analysis of the input text is justified by the necessity to account for the high morphosyntactic variation characterizing collocations. Stubbs [15] analyzed, for instance, the occurrences of the pair *bear-resemblance* in a corpus and found the following distribution of inflected forms for the verbal component, *to bear*: *bears* 18%, *bear* 11%, *bore* 11%, *bearing* 4%. Put together, these forms make up a high proportion (44%) of the total number of collocates of the noun *resemblance* (1,085). Example 1 summarizes this information in Stubbs' notation.

> **Example 1.** Morphological variation in collocations: Stubbs' notation for grouping inflected forms of collocates.
>
> resemblance 1,085 < bears 18%, bear 11%, bore 11%, bearing 4% > 44%

This example illustrates the importance of performing a *lexical analysis* of the input text in order to better pinpoint potential collocations. As a matter of fact, a large body of collocation extraction work [16–19] relies on lexical analysis, combined with part-of-speech (POS)-based filtering of combinations considered in a five-word window called *collocational span*.

In addition to the lexical analysis, the *syntactic analysis* of the input text has often been argued as necessary, particularly for languages that exhibit a freer word order, such as German or Korean. For such languages, the extraction techniques developed for English (e.g., Xtract [20]) are inefficient, since they fail to recover systematic long-distance dependencies and to account for the positional ambiguity of arguments. As reported, for instance, by Breidt [21], even distinguishing subjects from objects in

German is difficult without parsing. The author therefore proposed to shorten the collocational span to three words in order to exclude nouns that are unrelated to verbs. This strategy led to increased precision, but the improvement came at the expenses of recall. Similarly, Kim *et al.* [22] reported that a technique like Xtract [20], which is very popular for English and is based on selecting collocation candidates among the word pairs co-occurring at a stable distance in text, is completely unsuitable for Korean, because of the high syntactic flexibility.

Given the marked flexibility of collocations, some researchers have indicated that collocation extraction should ideally rely on the syntactic analysis of the source corpora [12,13,20,23,24]. However, despite their theoretical arguments, parsing has only been used in a minority of practical works. In such (exceptional) cases, collocation candidates are identified as pairs of words in a syntactic relationship, rather than as pairs of words in a collocational span, as in the prevailing syntax-free approaches. There are reports, for instance, on collocation work making use of *full parsing* for German [25], Chinese [26] and Dutch [27]. Similar work has been performed for English [28], a language in which collocation extraction experiments have also been carried out by exploiting manually annotated syntactic treebanks [29,30]. In addition, *dependency parsing* has been used for a number of languages, including English [31,32], French [33] and Czech [34]. Furthermore, a relatively larger amount of work has been devoted to collocation extraction based on *shallow parsing*, e.g., for English [35], German [36], French [11,37,38] and, notably, in the Sketch Engine multilingual system [39].

An important factor differentiating these syntax-based extraction systems is the performance of the parser involved. In some cases, the authors report a rather high parsing error rate, as well as robustness issues, leading them to exclude longer sentences made up of 20 words or more [27,31,32]. In other cases, the grammatical coverage of the parser is reported as limited, the extraction system being unable to deal with certain types of syntactic transformations, like relativization [28].

Apart from the underlying preprocessing technology and the specific association measures used to rank candidates, the extraction systems also vary greatly in the range of the syntactic configurations they take into account. Some systems identify candidates of a single type or a few specific syntactic types, e.g., verb-preposition [29], preposition-noun-preposition, prepositional phrase-verb [27], verb-object, noun-adjective, verb-adverb [26] or noun phrases [11,37,38]. Yet, other systems aim at a broader coverage, e.g., [25,39].

Although generally considered necessary for obtaining high quality results, syntax-based extraction is not always seen as a viable solution in the NLP community. Sometimes, it is discarded because of the unavailability of syntactic parsers; in other cases, the reason for not preprocessing the source corpora with a syntactic parser is taken on the basis of various arguments, such as time inefficiency, lack of precision or lack of robustness. To augment the skepticism, no comparative evaluation has been performed to clearly prove the superiority of syntax-based extraction over the simpler syntax-free alternative.

Our own work [44] was devoted to devising a fully-fledged extraction methodology based on full syntactic parsing. We relied on the Fips multilingual parser [40] to preprocess the source corpora and to select as candidates the combinations of words found in specific syntactic relations (see Section 1). A range of association measures can be applied to rank the selected candidates; the measure proposed by default is the log-likelihood ratio, which is argued to be efficient for both high-frequency and

low-frequency data [41]. The extraction system, initially developed for English and French, has later been extended to the new languages supported by the Fips parser, i.e., Spanish, Italian, German, Greek and Romanian.

Fips is a robust symbolic parser based on generative grammar concepts. It performs a "deep" syntactic analysis of the input sentence, making use of co-indexation to keep track of extraposed constituents, i.e., constituents that "moved" from the initial (canonical) position to the surface position, due to syntactic transformations, such as the ones shown in Example 2:

Example 2. Syntactic variation in collocations: sample transformations.

1. *relativization*
   various global *challenges* that we inevitably have to *face*
2. *passivisation*
   the *challenges faced* by the pharmaceutical industry today
3. *interrogation*
   Which *challenges* do online media *face* in terms of press freedom?

The parser's ability to account for extraposition is essential for coping with cases of long-distance dependency in collocations, when the component words may not occur in the same clause. In Example 1, for instance, the verb *face* occurs in the subordinate clause, while its object *challenges* is in the main clause. As can be seen from Example 3 showing the (simplified) parsing output, the parser correctly identifies the "deep" object of *face* by creating a co-indexation chain (labeled *i*) that contains the empty constituent in the object position of the verb *face*, e_i, the relative pronoun *that* and the noun phrase headed by *challenges*. Thanks to this mechanism, the verb-object pair *face-challenge* can be successfully identified as a potential collocation.

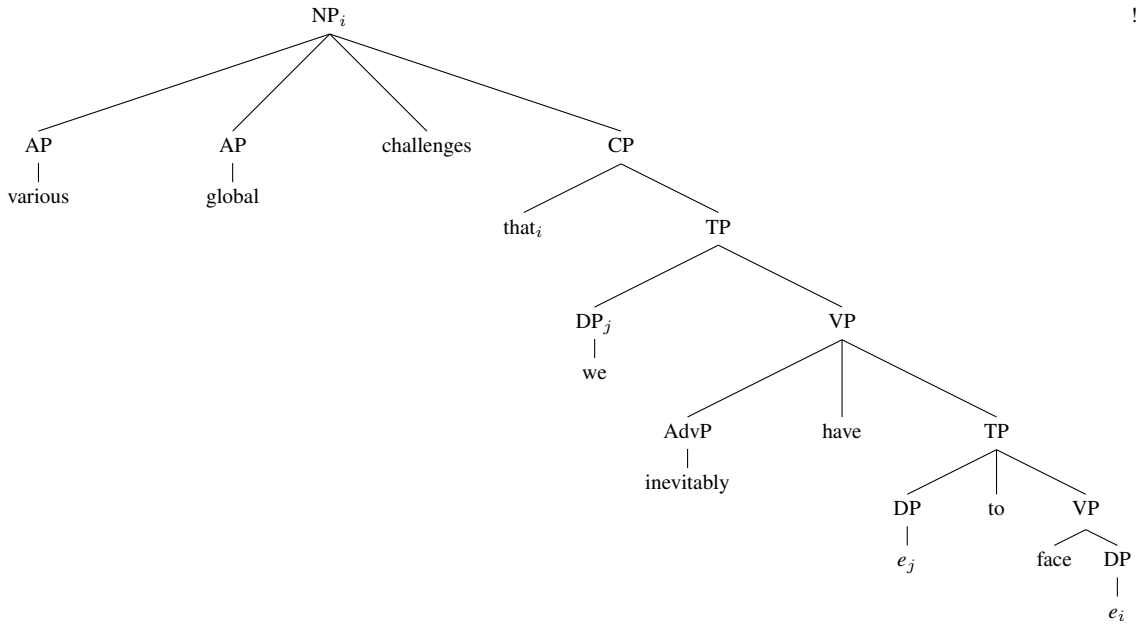
Two large-scale evaluation experiments have been performed to assess the impact of parsing on the quality of collocation extraction results. The results obtained using the syntax-based extraction method have been compared against those of the syntax-free baseline. The baseline consists of applying the so-called sliding window method to lemmatized and POS-filtered data, which means that all possible combinations within a five-word window that comply with the selected POS patterns are considered as candidates. The log-likelihood ratio measure [41] was applied in both cases.

The first experiment was performed in a monolingual setting, on French data from the Hansard corpus of Canadian Parliament proceedings totaling about 1.2 million words. The top 500 pair types have been manually evaluated by three judges, and the results showed a statistically significant improvement in precision with respect to the syntax-free baseline (99% *vs.* 78.3% in terms of grammaticality[2]; 65.9% *vs.* 57% in terms of lexicographic interest of the results.[3])

[2] A two-sample t-test was conducted to compare the number of grammatical pairs in the two methods' output. There was a significant difference in the output: t(982) = 10.78, $p < 0.001$.

[3] A similar two-sample t-test was conducted to compare the number of pairs considered as worth of storing in a lexicon. The difference is statistically significant: two-sample t(982) = 2.90, $p < 0.01$.

Example 3. Sample parsing output (A=Adjective, Adv=Adverb, C=Complementizer, D=Determiner, N=Noun, P=Phrase, T=Tense, V=Verb).



The second experiment was performed cross-linguistically, on French, English, Italian and Spanish parallel data from the Europarl corpus of European Parliament proceedings, totaling about 3.7 million words on average per language. A stratified sampling strategy was used to select the evaluation data, with sequences of 50 pair types taken from various levels in the output list (top—0%, 1%, 3%, 5% and 10%). A total of 2,000 pair types have been manually evaluated by teams of two judges. The results showed, again, statistically significant improvements over the baseline (88.8% *vs.* 33.2% average grammatical precision; 43.2% *vs.* 17.2% average lexicographic precision; and 32.9% *vs.* 12.8% average collocational precision[4]).

It is important to note that the very top results of the baseline are relatively noise-free, since the association measure succeeds in eliminating many erroneous pairs from the top positions. However, the precision degrades rapidly for the items on lower positions: as the frequency of pairs decreases, the measure alone is inefficient in removing the noise. In contrast, the syntax-based approach guarantees a better global quality of the results, meaning that even candidates with lower scores are noise-free. This is particularly important, since lexical data has a skewed, Zipfian distribution, and most of the candidate combinations receive low scores because of their low co-occurrence frequency; yet, they are potentially interesting from a lexicographic point of view. Carrying out lexicographic work on a noise-free list is one of the main benefits of syntax-based approaches to collocation extraction.

The results of these experiments cleared the doubts on the feasibility of a syntax-based approach to collocation extraction and showed the benefits obtained over the syntax-free approach. They confirmed that parsing information contributes to a statistically significant increase in collocation extraction

[4]Two-sample t-tests have been conducted to compare the number of: (1) grammatical pairs; (2) pairs considered as worth storing in a lexicon; and (3) pairs marked as collocations. The differences obtained are statistically significant: (1) two-sample t(1435) = 26.65, $p < 0.001$); (2) two-sample t(1435) = 11.04, $p < 0.001$; 3) two-sample t(1435) = 9.15, $p < 0.001$.

performance and corroborated similar findings obtained by using syntax-based approaches to other tasks, e.g., term extraction [42], semantic role labeling [10] and semantic similarity computation [43].

3. Using Translation for Collocation Identification

In this section, we focus on translation-based approaches to the identification of multi-word expressions (MWEs) in general, and of collocations in particular. In the work reported in the literature on the topic, we distinguish between two distinct trends:

- Approaches that exploit translation archives represented by parallel corpora or source-target pairs of monolingual corpora for identifying translation equivalents for MWEs/collocations;
- Approaches that take into account word alignment information for detecting and ranking monolingual MWE/collocation candidates.

3.1. The First Trend: Exploiting Corpora for Collocation Identification

The first trend is a traditional and more popular trend, represented, for instance, by [26,45–49]. The work in [45–47] was devoted to acquiring translation equivalents for noun phrases, whereas [26,48,53] deal with collocations of several syntactic types and [49] with MWEs in general. In what follows, we provide further details on this work. (Note that evaluation results are systematically reported in this type of work, whereas for collocation extraction, they may be missing or replaced by small output samples.)

In [45], Kupiec identifies noun phrase correspondences between English and French by relying on the sentence-aligned Hansard parallel corpus. Both source and target corpora are POS-tagged, then NPs are detected with a finite-state recognizer. The matching is done by using Expectation Maximization (EM), an iterative re-estimation algorithm. The precision reported is 90% for the top 100 translations obtained.

In the same vein, van der Eijk [46] used a similar method for the language pair Dutch-English, except that the matching is done using two main heuristics: the target noun phrase is selected depending on (a) its frequency in the target sentences; and (b) its relative position in the source sentence. The reported performance was lower (68% precision and 64% coverage), which was explained by the fact that the evaluation was performed on a larger test set of 1,100 noun phrases.

In addition, Dagan and Church [47] made use of word alignment to find candidate translations for noun phrases in parallel corpora. Once the source noun phrases have been identified, the text span between the alignments of the first and the last words of the phrase is proposed as a translation candidate. Candidates are sorted in decreasing frequency order. The authors argue that unlike the previous systems, their system, Termight, has the advantage of finding translations even for infrequent terms. The method was tested on 192 English-German correspondences and achieved a precision of 40% (when the first translation alternative was considered only).

All the systems discussed above are limited to a particular type of construction, namely, the nominal compound, which is relatively fixed. In contrast, the Champollion system built by Smadja *et al.* [48] is the first system devoted to collocations proper, and it can handle both rigid and flexible combinations. It relies on the Xtract collocation extractor for English [20]. For each source collocation, it attempts to

detect a translation equivalent in the aligned French sentences from the Hansard corpus. The matching relies on a statistical correlation metric, the Dice coefficient. The method used requires an additional post-processing step in which the order of words in a flexible collocation is decided, given that no syntactic analysis is performed on the target side. The system has been evaluated by three annotators and showed a precision of 77% and 61%, respectively, on two different test sets of 300 collocations each. These collocations have been randomly selected among the medium-frequency results. The difference in the precision obtained is explained by the lower frequency of collocations from the second set.

The method of Lü and Zhou [26] can also deal with flexible collocations; moreover, these are validated syntactic constituents, as they are extracted using a parser. The syntactic types considered are verb-object, adjective-noun and adverb-verb. Collocations are extracted from monolingual corpora in English and Chinese by applying the log-likelihood ratio measure on syntactically related pairs identified by a dependency parser. The matching between a source and target collocation is performed by using a statistical translation model, which estimates word translations with EM. The method (whose reported coverage is 83.98%) was evaluated on a test set of 1,000 randomly selected collocations. It achieves between 50.85% and 68.15% accuracy, depending on the syntactic type. The availability and the quality of bilingual dictionaries are essential for the performance of this method.

Our own method for finding translation equivalents for collocations in parallel corpora [53] consists of source-side and target-side collocation extraction performed with the system presented in Section 2 and a linguistically-motivated matching procedure for pairing a source collocation and its potential translations. The matching procedure takes into account the syntactic type of the collocation, by requiring the translation candidates to be of a compatible syntactic type (for instance, a verb-object collocation in English may have either a verb-object or a verb-preposition-argument equivalent in French: *face challenge, relever défi; meet need, répondre à besoin*). The frequency is also taken into account when trying to pinpoint the translation equivalent. The most frequent target candidate is retained as the potential translation; in the case of tied results, these are ranked according to their association score. Information from bilingual dictionaries is also used, if available for the specific words involved in the source collocation.[5] More precisely, we exploit the translations of the collocation *base* only: according to theoretical stipulations [5], the *base*, i.e., the semantically autonomous component of a collocation, can be translated literally, whereas the *collocate*, the component that is semantically dependent on the base, cannot. For instance, in *meet need*, the base *need* is translated literally as *besoin*; therefore, the combinations with the noun *besoin* are considered as potential translation equivalents. The method has been evaluated on 4,000 pairs extracted from the Europarl corpus [55] in English, French, Spanish and Italian. The results showed a performance of 81.6% according to the F-measure (84.1% precision and 79.2% recall), which means that our method compares favorably against previous methods.

The more recent method of Bai *et al.* [49] detects English equivalents for MWEs extracted from Chinese corpora using a parser. The matching method is very similar to that in [48], also making use of the Dice coefficient, but applying additional frequency filters. The performance has been measured extrinsically by performing a task-based evaluation, in which the extracted translation equivalents have

[5] We experimented with and without dictionary information (in our case, the lexical databases of the Its-2 in-house machine translation system [54]).

been used for statistical machine translation; these have been found to lead to a significant improvement of translation results.

3.2. The Second Trend: Exploiting Word Alignments

Next to the approaches aimed at detecting translation equivalents from (parallel) corpora, we find an emergent class of approaches in which translation information, in particular, word alignment, is used for the monolingual identification of MWEs [50,51]. Such alignment-based approaches rely on tools that use standard models derived from statistical machine translation.

The hypothesis put forward, for instance, by Villada Moirón and Tiedemann [50] is that the idiosyncrasy of MWEs is reflected in their translational entropy: unlike for regular (compositional) constructions, the components of an idiomatic expressions are not translated literally, but are harder to translate. Therefore, there is a larger variety of translation links for them and, consequently, a higher average entropy for such expressions. The authors first extract verb-preposition-noun candidates in Dutch using parsing and association measures, then compute their average translational entropy into English, Spanish and German in order to re-rank these candidates. They evaluate their method on the top 200 candidates by comparing their new ranking against the old ranking using UAP, the uninterpolated average precision [52]. It was found that the alignment significantly improves the ranking of candidate expressions and, thus, the extraction performance.

In the work of Caseli *et al*. [51], the word alignment is used for the actual identification of MWEs, as opposed to their mere ranking as in [50]. The authors consider as MWE candidates the word sequences that are aligned systematically to the same target sequence, regardless of the length of the latter. Candidates are POS-filtered, and a frequency threshold is applied. When evaluated by taking into account human judgments, their method showed an overall precision of 49.28%. The precision was found higher for high-frequency candidates and for specific patterns, such as verb-preposition/particle.

4. Exploiting Collocations for Syntactic Parsing

As stated by Sag *et al*. [1], MWEs are a key problem for Natural Language Processing, because they cannot be treated by compositional methods, as these would lead to overgeneration (e.g., the production of a phrase like **intersection regulator* instead of *traffic light*). Moreover, they cannot be analyzed by means of regular grammatical processes since they are idiosyncratic (for instance, the expression *in line* lacks the determiner: **in the line*).

To account for the MWEs present in language, practical NLP work has generally adopted a solution consisting of listing MWEs in a lexicon and tokenizing the input text by using a *words-with-spaces* approach to recognize MWEs. This approach suffers from two main drawbacks. The first is the lack of *flexibility*, as many expressions allow lexical material to occur between the component words, and the words-with-spaces approach is inadequate for handling such situations. The second is the *lexical proliferation* problem, as expressions generally exhibit variation, and listing all possible forms in a lexicon would be impractical. The words-with-spaces approach therefore fails to generalize and handles variation badly [1].

These problems are even more acute in the case of collocations. As stated in Section 1, collocations are the most flexible amongst all kinds of MWEs, which makes the words-with-spaces approach completely inadequate for their representation and treatment. There is no systematic restriction on the number of forms of a collocation component (e.g., a verb), the order of components of a collocation or the number of words that may intervene between them (see Example 2). Collocations are situated at the intersection of lexicon and grammar; therefore, they cannot be accounted for solely by the lexical component of a parsing system. Instead, they have to be integrated into the grammatical component as well, as the parser has to consider all their possible syntactic realisations.

As an alternative to the words-with-spaces preprocessing approach, collocations could be recognized by the parser after the analysis of the input sentence has been performed, following an approach such as the one described in Section 2. Again, this approach is not fully appropriate from a parsing point of view, and the reason lies with the important observation that prior collocational knowledge is highly relevant for parsing. Collocational preferences are, along with other types of information, like selectional restrictions and subcategorization frames, a major means of structural disambiguation. In fact, collocational relations between the words in a sentence proved useful for selecting the most plausible among all the parse trees generated for a sentence [56–59]. As we will show later in this section, a more suitable approach is the inclusion of collocations into the grammatical component of the parser, so that the identification of collocations and the construction of a parse tree become interacting processes, which take place simultaneously and inform each other.

In what follows, we review the extent to which MWE/collocational knowledge has been incorporated into parsing systems in order to improve their performance. A number of studies have provided empirical evidence that, indeed, the recognition of MWE/collocations leads to better parsing results. For instance, Brun [60] 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 led to a significant reduction of the number of alternative parses (from an average of 4.21 to 2.79). The author reports that the eliminated parses were semantically undesirable and that no valid analyses were ruled out. Similarly, Zhang *et al.* [61] extended a lexicon with 373 additional MWE lexical entries and obtained a significant increase in the grammatical coverage of an English parser (of 14.4%, from 4.3% to 18.7%).

In the cases mentioned above, a words-with-spaces approach has been used to represent MWEs. In contrast, Alegria *et al.* [62] and Villavicencio *et al.* [63] adopted a compositional approach to the encoding of MWEs, able to capture more morphosyntactically-flexible MWEs. Alegria *et al.* [62] showed that by using an 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.* [63] found that the simple 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.

In addition, an area of intensive research in parsing is specifically concerned with the use of lexical preferences, co-occurrence frequencies, collocations and contextually similar words for prepositional phrase (PP) attachment disambiguation. Thus, an important number of unsupervised methods [56,64,65], supervised method [57,58] and combined methods [66] have been developed to this end. However, as pointed out in [56], the bottleneck of this strand of work is that the parsers lack precisely the kind

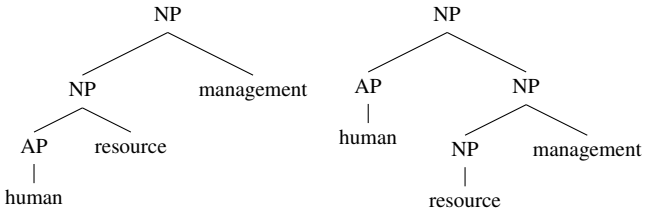
of corpus-based information that is required to resolve ambiguity. Performing parsing and identifying collocations is therefore a circular problem, to which the previous literature provided no solution.

In the remainder of this section, we outline a novel collocation processing approach implemented in relation to the development of the Fips parser, consisting of the simultaneous execution of the two tasks, sentence analysis and collocation identification [67]. Briefly put, the idea is that when the parser processes a lexical item that is marked in the lexicon as part of a collocation, in attempting to attach it to another item, it first checks the collocation lexicon for a syntactically compatible entry and, if found, it gives high priority to the structure in which the two components are attached. Thus, the collocation identification mechanism is incorporated within the constituent attachment procedure of the parser.

It is well known that, given the high frequency of lexical ambiguities and 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. Parsing algorithms use various heuristics to limit the number of alternatives and, thus, to ensure that the parsing performance is satisfactory for processing large corpora. Collocations contribute crucially to the disambiguation process, by helping the parser through the maze of alternatives. The identification of collocation is therefore not a burden—an additional task to solve—but a process that helps the parser. Collocations are generally made of highly ambiguous words and identifying them helps to decide among alternatives. For instance, in *break a record*, both components are ambiguous if taken in isolation, and it is their combination that helps the parser to select the appropriate categories and readings.

Apart from the lexical disambiguation, collocations also contribute to structural disambiguation, as illustrated in Example 4 below for the phrase *human resource management*. To decide between competing analyses, e.g., one in which the word *human* is attached to *resource* and one in which it is attached to *management* , the parser exploits information from its collocation lexicon. Provided that the entry *human resource* is found, listed as an adjective-noun pair, then the parser favors the first analysis over the second, as it accommodates the structure specified in the lexicon.

Example 4. Alternative analyses for the phrase *human resource management*.



To measure the impact of coupling the collocation identification and the parsing tasks, we carried out experiments in which we compared the new version of the parser with the version before, which does not use collocations for attachment decisions. We assessed the impact of the procedure interconnecting parsing and collocation identification on the performance of both tasks, parsing and collocation identification. On a corpus of news articles from *The Economist* [68] totaling slightly more that 0.5 million words, we obtained a sensible increase in the coverage of the parser expressed in terms of the number of completely parsed sentences (83.3% *vs.* 81.7%), as well as an increase in collocation identification precision (93.7% *vs.* 81.6%).

The results are in line with previous reports on the impact of incorporating MWEs into parsing systems; the difference lies in the fact that the syntactic flexibility of MWEs is fully taken into account in our approach. Together, these results show the significant role played by these expressions on the performance of language analysis systems.

5. Exploiting Collocations in Machine Translation

In addition to being useful for NLP applications concerned with language analysis, collocational information derived from corpora is crucial for applications dealing with text production, such as natural language generation and machine translation. Collocations are considered a key factor in producing more acceptable output in these applications [28,69].

In spite of their relatively transparent meaning, collocations pose significant problems from the perspective of language production, since they are "idioms of encoding" [70]. The lexical selection is restricted to the conventionalized form, which is language dependent. Therefore a regular selection and, in the case of machine translation, a literal translation are inappropriate, as they may lead to unnatural, if not awkward, formulations, known as *anti-collocations* [30] (e.g., **accuse delay*, a literal translation of the French *accuser retard*, 'experience delay').

To illustrate the importance of collocations for machine translation, consider the French combinations *grande attention*, *grande diversité* and *grande vitesse*, in which the adjective *grande* 'big' modifies the nouns *attention*, *diversity* and *speed*. A literal translation will lead to inadequate formulations in English: **big attention, *big diversity, *big speed*. The right translations, *great attention, wide range* and *high speed*, show the necessity of using collocations in the target language: the same adjective, *grande*, is translated in three different ways, depending on the noun it modifies.

As discussed in Section 3, a considerable amount of work has been devoted to the extraction of translation equivalents from corpora, e.g., [26,45,47,48], and to the representation of collocational knowledge into computational lexica for machine translation and natural language generation [69,71]. However, there are very few reports on the actual use of collocational knowledge in such systems.

One such report refers to the Logos machine translation system, which uses collocations extracted with the method of Orliac and Dillinger [28]. It is argued that context-dependent selection of target lexical items, enabled by collocations, "achieves significant improvement in readability and perceived quality of the translation produced" [28]. Another report, by Liu *et al.* [72], concerns the integration of collocations into a statistical machine translation (SMT) system. The authors show that their method significantly improves the performance of both word alignment and translation quality. In a subsequent experiment, Liu *et al.* [73] use source language collocations for reordering for SMT, again achieving significant improvements.

More generally, as far as MWEs are concerned, there is additional evidence coming from reports showing that the incorporation of bilingual MWEs into SMT systems leads to an increase in the quality of translation results. For instance, Bai *et al.* [49] added 1,171 Chinese-English MWEs into their SMT systems and obtained a significant improvement in the Bilingual Evaluation Understudy (BLEU) score. Similarly, Tsvetkov and Wintner [74] reported that by adding 2,955 MWE translation pairs into a Hebrew-to-English SMT system, they obtained a statistically significant improvement of BLEU

and Meteor scores. Furthermore, Bouamor *et al.* [75] used different strategies to integrate into their English-French SMT system bilingual phrases extracted from a 100,000 sentence training corpus, finding an increase in the BLEU and Meteor scores.

It is worth noting that phrase-based SMT systems already incorporate MWE/collocational knowledge as an effect of training their language and translation models on large (parallel) corpora. These systems are successful in dealing with local collocations, but are arguably ill-suited for handling collocation whose components are not in close proximity to one another. As Babych *et al.* [76] put it,

> "SMT output is often surprisingly good with respect to short distance collocations, but often (...) correct choices are missed in cases where selectional restrictions take effect on distant words."

In the same vein, Bod [77] points out that discontiguous phrases represent a real challenge for SMT systems, and he provides empirical evidence that such phrases contribute significantly to improving the translation accuracy. Indeed, we also found that SMT systems are very sensitive to the syntactic environment of source collocations, as well as to the lexical environment [78]. As can be seen in Example 5, the same source collocation is correctly translated from English into French when found in a given context and incorrectly translated in another context:

Example 5. Collocation translations from English to French.

1. the people who rely on us to *give* full *support* when it is needed
   les gens qui comptent sur nous pour *apporter* un *soutien* complet quand il est néecessaire

2. 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.

More recently, Carpuat and Diab [79] provided further evidence on the impact of MWEs on SMT performance. By integrating 500 English multi-words from WordNet (corresponding to about 900 tokens) as "words with spaces" into English-Arabic SMT through segmentation of the training and test sentences, they obtained an increase in performance in terms of BLEU and Translation Error Rate (TER). An additional strategy consisted of identifying the 500 most frequent n-grams in the phrase table of the system and biasing the system towards using phrases that do not break these n-grams. This strategy had a less important, but still positive, effect on translation performance in terms of automatic metric scores.

In our own work, we assessed the impact of collocational knowledge on a rule-based translation system, namely, the Its-2 system [54] based on the Fips parser (*cf.* Section 2). Collocations have been integrated into this system in an indirect manner, by adding them into the underlying parsing system. More precisely, the new parsing strategy integrating collocation identification (as described in Section 4) replaced the old parsing strategy. The evaluation was performed on 200 randomly sampled sentences from Europarl, half in English and half in Italian, which contained verb-object collocations. The sentences were translated into French, and the output was manually evaluated by two judges. For both language pairs, the results showed a statistically significant improvement in collocation translation

adequacy when collocational knowledge is integrated in this specific way into the translation system.[6] These findings are in line with the ones previously mentioned in relation to SMT. They confirm the positive impact of collocations in the rule-based machine translation scenario.

Differently from related work, we performed a focused evaluation concerned with the quality of translations proposed for collocations, as opposed to the overall sentence translation quality. We were reluctant to measure the impact on the BLEU score, since this metric is more suited to an overall assessment of the target sentence and the context could easily mask the effect of choosing a wrong translation for collocations. By giving equal weight to the words in a sentence, BLEU underestimates the importance of choosing the right collocate for a base word. Our evaluation strategy corresponds to what has later been coined as evaluation focused on *linguistic checkpoints* [80], i.e., evaluation of machine translation performance for specific linguistic phenomena.

Further investigation is required in order to be able to check whether the positive result of improved collocation translation is accompanied by a similar improvement in the overall sentence quality. However, given the massive presence of collocations in language and their role on language fluency, we hypothesize that improving the translation of collocations is one of the main ingredients in improving the overall quality of translations.

6. Conclusions

Multi-word expressions have long since been the subject of an important body of NLP work. A lot of progress has been achieved in particular in developing technologies for acquiring specific types of MWEs, such as collocations, from text corpora. However, this research remains somewhat endogenous: despite the widely recognized importance of such expressions for parsing and translation, not many efforts are devoted to actually integrating the acquired expressions in these applications.

In this paper, we focused on the interaction between multi-word expressions in general, and collocations in particular, on the one hand, and the applications of syntactic parsing and machine translation, on the other hand. We highlighted the existing work on the use of collocational information for improving parsing and translation performance and, *vice versa*, the work on the use of parsing and translation information for improving corpus-based collocation identification. In addition to giving a panorama of the previous work in these areas, we described our own methods and experiments, which stem into sustained work devoted to collocation processing in a multilingual, syntactically-aware environment. We showed that parsing and translation technologies can contribute significantly to the task of automatic detection of collocations in text corpora. We also focused on the exploitation of collocations in parsing and machine translation and presented experimental results showing the benefits that can be obtained by following a collocation-aware approach in both tasks.

One of the most sensitive issues in relation to MWE/collocation processing is their syntactic flexibility. Our work is specifically focused on this issue and complements existing words-with-spaces approaches, which are easier to implement, but less adequate for modeling MWEs (*cf.* [1]). We expect

[6] A McNemar test was conducted to compare the number of cases in which the translation became better *vs.* worse. For English-French, the difference (14 *vs.* 4) is statistically significant ($p = 0.0339$). For Italian-French, the difference (16 *vs.* 3) is very statistically significant ($p = 0.0014$).

a further integration of parsing and translation technologies in the NLP field in the future, as already witnessed by the increasing interest in syntax-based SMT, and we hope that (flexible) MWEs will take a more prominent place in both the parsing and translation fields. We hope that our present findings will contribute to the understanding of the mutual role that collocational knowledge and parsing/translation information play in better processing natural language.

Acknowledgments

This work has been performed mainly while the author was affiliated with the Language Technology Laboratory (University of Geneva), under the supervision of Eric Wehrli. The author wishes to acknowledge his support and collaboration and to thank him in particular for providing the parsing and translation infrastructure that made this work possible.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Sag, I.A.; Baldwin, T.; Bond, F.; Copestake, A.; Flickinger, D. 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), Mexico City, Mexico, 17–23 February 2002; pp. 1–15.
2. Villavicencio, A.; Bond, F.; Korhonen, A.; McCarthy, D. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Comput. Speech Lang.* **2005**, *19*, 365–377.
3. Heid, U. Computational Phraseology: An Overview. In *Phraseology: An Interdisciplinary Perspective*; Granger, S., Meunier, F., Eds.; John Benjamins: Amsterdam, The Netherlands, 2008; pp. 337–360.
4. Jackendoff, R. *The Architecture of the Language Faculty*; MIT Press: Cambridge, MA, USA, 1997.
5. Mel'čuk, I. Collocations and Lexical Functions. In *Phraseology. Theory, Analysis, and Applications*; Cowie, A.P., Ed.; Claredon Press: Oxford, UK, 1998; pp. 23–53.
6. Erman, B.; Warren, B. The idiom principle and the open choice principle. *Text* **2000**, *20*, 29–62.
7. *Oxford Collocations Dictionary for Students of English*; Lea, D., Runcie, M., Eds.; Oxford University Press: Oxford, UK, 2002.
8. Benson, M.; Benson, E.; Ilson, R. *The BBI Dictionary of English Word Combinations*; John Benjamins: Amsterdam, The Netherlands, Philadelphia, PA, USA, 1986.
9. Fontenelle, T. Collocation Acquisition from a Corpus or from a Dictionary: A Comparison. In Proceedings of the I-II. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere, Tampere, Finland, 4–9 August 1992; pp. 221–228.
10. Gildea, D.; Palmer, M. The Necessity of Parsing for Predicate Argument Recognition. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 239–246.

11. Daille, B. Approche Mixte Pour l'Extraction Automatique de Terminologie: Statistiques Lexicales et Filtres Linguistiques. Ph.D. Thesis, Université Paris 7, Paris, France, 1994.
12. Pearce, D. A Comparative Evaluation of Collocation Extraction Techniques. In Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, 29–31 May 2002; pp. 1530–1536.
13. Evert, S. The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. Thesis, University of Stuttgart, Stuttgart, Germany, 2004.
14. Pecina, P. Lexical Association Measures: Collocation Extraction. Ph.D. Thesis, Charles University in Prague, Prague, Czech Republic, 2008.
15. Stubbs, M. *Words and Phrases: Corpus Studies of Lexical Semantics*; Blackwell: Oxford, UK, 2002.
16. Church, K.; Hanks, P. Word association norms, mutual information, and lexicography. *Comput. Linguist.* **1990**, *16*, 22–29.
17. Justeson, J.S.; Katz, S.M. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Nat. Lang. Eng.* **1995**, *1*, 9–27.
18. Zaiu Inkpen, D.; Hirst, G. Acquiring Collocations for Lexical Choice Between Near-Synonyms. In Proceedings of the Workshop on Unsupervised Lexical Acquisition (ACL-02), Philadephia, PA, USA, 6–12 July 2002; pp. 67–76.
19. Todiraşcu, A.; Tufiş, D.; Heid, U.; Gledhill, C.; Ştefănescu, D.; Weller, M.; Rousselot, F. A Hybrid Approach to Extracting and Classifying Verb+Noun Constructions. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 28–30 May 2008.
20. Smadja, F. Retrieving collocations from text: Xtract. *Comput. Linguist.* **1993**, *19*, 143–177.
21. Breidt, E. 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, Columbus, Ohio, USA, 22 June 1993; pp. 74–83.
22. Kim, S.; Yoon, J.; Song, M. Automatic extraction of collocations From Korean text. *Comput. Humanit.* **2001**, *35*, 273–297.
23. Heid, U. On Ways Words Work Together—Research Topics in Lexical Combinatorics. In Proceedings of the 6th Euralex International Congress on Lexicography (EURALEX '94), Amsterdam, The Netherlands, 30 August–3 September 1994; pp. 226–257.
24. Krenn, B. Collocation Mining: Exploiting Corpora for Collocation Idenfication and Representation. In Proceedings of the Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2000), Ilmenau, Germany, 9–12 October 2000; pp. 209–214.
25. Schulte im Walde, S. A Collocation Database for German Verbs and Nouns. In Proceedings of the 7th Conference on Computational Lexicography and Corpus Research, Budapest, Hungary, 3 April 2003.
26. Lü, Y.; Zhou, M. Collocation Translation Acquisition Using Monolingual Corpora. In Proceedings of the 42nd Meeting of the Association for Computational Linguistics, ACL'04), Barcelona, Spain, 21–26 July 2004; pp. 167–174.

27. Villada Moirón, M.B.N. Data-Driven Identification of Fixed Expressions and Their Modifiability. Ph.D. Thesis, University of Groningen, Groningen, The Netherlands, 2005.
28. Orliac, B.; Dillinger, M. Collocation Extraction for Machine Translation. In Proceedings of the Machine Translation Summit IX, New Orleans, Lousiana, USA, 23–27 September 2003; pp. 292–298.
29. Blaheta, D.; Johnson, M. Unsupervised Learning of Multi-Word Verbs. In Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation, Toulouse, France, 6–7 July 2001; pp. 54–60.
30. Pearce, D. Synonymy in Collocation Extraction. In Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburgh, PA, USA, 2–7 June 2001; pp. 41–46.
31. Lin, D. Extracting Collocations from Text Corpora. In Proceedings of the First Workshop on Computational Terminology, Montreal, Canada, 15 August 1998; pp. 57–63.
32. Lin, D. Automatic Identification of Non-Compositional Phrases. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, MD, USA, 20–26 June 1999; pp. 317–324.
33. Charest, S.; Brunelle, E.; Fontaine, J.; Pelletier, B. Élaboration Automatique d'un Dictionnaire de Cooccurrences Grand Public. In Proceedings of the Actes de la 14e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007), Toulouse, France, 5–8 June 2007; pp. 283–292.
34. Pecina, P. Lexical association measures and collocation extraction. *Lang. Resour. Eval.* **2010**, *1*, 137–158.
35. Church, K.; Gale, W.; Hanks, P.; Hindle, D. Parsing, Word Associations and Typical Predicate-Argument Relations. In Proceedings of the International Workshop on Parsing Technologies, Pittsburgh, PA, USA, 28–31 August 1989; pp. 103–112.
36. Wermter, J.; Hahn, U. Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05), Vancouver, Canada, 6–8 October 2005; pp. 843–850.
37. Bourigault, D. LEXTER, vers un outil linguistique d'aide à l'acquisition des connaissances. In Actes des 3èmes Journées d'acquisition des Connaissances, Dourdan, France, April 1992.
38. Jacquemin, C.; Klavans, J.L.; Tzoukermann, E. Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. In Proceedings of the 35th Annual Meeting on Association for Computational Linguistics, Madrid, Spain, 7–12 July 1997; pp. 24–31.
39. Kilgarriff, A.; Rychly, P.; Smrz, P.; Tugwell, D. The Sketch Engine. In Proceedings of the Eleventh EURALEX International Congress, Lorient, France, 15–19 July 2004; pp. 105–116.
40. Wehrli, E. Fips, A "Deep" Linguistic Multilingual Parser. In Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing, Prague, Czech Republic, 28 June 2007; pp. 120–127.
41. Dunning, T. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* **1993**, *19*, 61–74.

42. Maynard, D.; Ananiadou, S. A Linguistic Approach to Terminological Context Clustering. In Proceedings of Natural Language Pacific Rim Symposium (NLPRS '99), Beijing, China, 5–7 November 1999; pp. 346–351.
43. Padó, S.; Lapata, M. Dependency-based construction of semantic space models. *Comput. Linguist.* **2007**, *33*, 161–199.
44. Seretan, V. *Syntax-Based Collocation Extraction*, Text, Speech and Language Technology; Springer: Dordrecht, The Netherlands, 2011.
45. Kupiec, J. An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, USA, 22–26 June 1993; pp. 17–22.
46. Van der Eijk, P. Automating the Acquisition of Bilingual Terminology. In Proceedings of the Sixth Conference on European chapter of the Association for Computational Linguistics, Utrecht, The Netherlands, 22–26 June 1993; pp. 113–119.
47. Dagan, I.; Church, K. *Termight*: Identifying and Translating Technical Terminology. In Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP), Stuttgart, Germany, 13–15 October 1994; pp. 34–40.
48. Smadja, F.; McKeown, K.; Hatzivassiloglou, V. Translating collocations for bilingual lexicons: A statistical approach. *Comput. Linguist.* **1996**, *22*, 1–38.
49. Bai, M.H.; You, J.M.; Chen, K.J.; Chang, J.S. Acquiring Translation Equivalences of Multiword Expressions by Normalized Correlation Frequencies. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; pp. 478–486.
50. Villada Moirón, B.N.; Tiedemann, J. Identifying Idiomatic Expressions Using Automatic Word-Alignment. In Proceedings of the Workshop on Multi-Word-Expressions in a Multilingual Context, Trento, Italy, 3 April 2006; pp. 33–40.
51. Caseli, H.D.M.; Ramisch, C.; das Graças Volpe Nunes, M.; Villavicencio, A. Alignment-based extraction of multiword expressions. *Lang. Resour. Eval.* **2010**, *44*, 59–77.
52. Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
53. Seretan, V.; Wehrli, E. Collocation Translation Based on Sentence Alignment and Parsing. In Proceedings of the Actes de la 14e Conférence sur le Traitement Automatique des Langues Naturelles, TALN 2007, Toulouse, France, 5-8 June 2007; pp. 401–410.
54. Wehrli, E.; Nerima, L.; Scherrer, Y. Deep Linguistic Multilingual Translation and Bilingual Dictionaries. In Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece, 30–31 April 2009; pp. 90–94.
55. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of the Tenth Machine Translation Summit (MT Summit X), Phuket, Thailand, 12–16 September 2005; pp. 79–86.
56. Hindle, D.; Rooth, M. Structural ambiguity and lexical relations. *Comput. Linguist.* **1993**, *19*, 103–120.
57. Alshawi, H.; Carter, D. Training and scaling preference functions for disambiguation. *Comput. Linguist.* **1994**, *20*, 635–648.

58. Berthouzoz, C.; Merlo, P. Statistical Ambiguity Resolution for Principle-Based Parsing. In *Recent Advances in Natural Language Processing: Selected Papers from RANLP'97*, Current Issues in Linguistic Theory; Nicolov, N., Mitkov, R., Eds.; John Benjamins: Amsterdam, The Netherlands, Philadelphia, PA, USA, 1997; pp. 179–186.
59. Wehrli, E. Parsing and Collocations. In *Natural Language Processing*; Christodoulakis, D., Ed.; Springer Verlag: Berlin/Heidelberg, Germany, 2000; pp. 272–282.
60. Brun, C. 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, Montreal, Canada, 10–14 August 1998; pp. 196–200.
61. Zhang, Y.; Kordoni, V.; Villavicencio, A.; Idiart, M. Automated Multiword Expression Prediction for Grammar Engineering. In Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney, Australia, 23 July 2006; pp. 36–44.
62. Alegria, I.N.; Ansa, O.; Artola, X.; Ezeiza, N.; Gojenola, K.; Urizar, R. Representation and Treatment of Multiword Expressions in Basque. In Proceedings of the Second ACL Workshop on Multiword Expressions: Integrating Processing, Barcelona, Spain, 26 July 2004; pp. 48–55.
63. Villavicencio, A.; Kordoni, V.; Zhang, Y.; Idiart, M.; Ramisch, C. 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), Prague, Czech Republic, 28–30 June 2007; pp. 1034–1043.
64. Ratnaparkhi, A. 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, Montreal, Canada, 10–14 August 1998; pp. 1079–1085.
65. Pantel, P.; Lin, D. An Unsupervised Approach to Prepositional Phrase Attachment Using Contextually Similar Words. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, 1–8 October 2000; pp. 101–108.
66. Volk, M. Combining Unsupervised and Supervised Methods for PP Attachment Disambiguation. In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02), Taipei, Taiwan, 24 August 24–1 September 2002; pp. 25–32.
67. Wehrli, E.; Seretan, V.; Nerima, L. Sentence Analysis and Collocation Identification. In Proceedings of the Workshop on Multiword Expressions: from Theory to Applications, MWE 2010, Beijing, China, 28 August 2010; pp. 27–35.
68. The Economist. Available online at http://www.economist.com (accessed 2002–2013).
69. Heylen, D.; Maxwell, K.G.; Verhagen, M. Lexical Functions and Machine Translation. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994), Kyoto, Japan, 5–9 August 1994; pp. 1240–1244.
70. Fillmore, C.; Kay, P.; O'Connor, C. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* **1988**, *64*, 501–538.

71. Heid, U.; Raab, S. Collocations in Multilingual Generation. In Proceeding of the Fourth Conference of the European Chapter of the Association for Computational Linguistics (EACL'89),Manchester, UK, 10–12 April 1989; pp. 130–136.
72. Liu, Z.; Wang, H.; Wu, H.; Li, S. Improving Statistical Machine Translation with Monolingual Collocation. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; pp. 825–833.
73. Liu, Z.; Wang, H.; Wu, H.; Liu, T.; Li, S. Reordering with Source Language Collocations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 1036–1044.
74. Tsvetkov, Y.; Wintner, S. Extraction of Multi-word Expressions from Small Parallel Corpora. In Proceedings of the Coling 2010: Posters, Beijing, China, 23–27 August 2010; pp. 1256–1264.
75. Bouamor, D.; Semmar, N.; Zweigenbaum, P. Identifying Bilingual Multi-Word Expressions for Statistical Machine Translation. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 23–25 May 2012.
76. Babych, B.; Eberle, K.; Geiß, J.; Ginestí-Rosell, M.; Hartley, A.; Rapp, R.; Sharoff, S.; Thomas, M. Design of a Hybrid High Quality Machine Translation System. In Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), Avignon, France, 23–24 April 2012; pp. 101–112.
77. Bod, R. Unsupervised Syntax-Based Machine Translation: The Contribution of Discontiguous Phrases. In Proceedings of the MT Summit XI, Copenhagen, Denmark, 10–14 September 2007; pp. 51–56.
78. Wehrli, E.; Seretan, V.; Nerima, L.; Russo, L. 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, Barcelona, Spain, 14–15 May 2009; pp. 128–135.
79. Carpuat, M.; Diab, M. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. In Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, USA, 2–4 June 2010; pp. 242–245.
80. Naskar, S.K.; Toral, A.; Gaspari, F.; Way, A. A Framework for Diagnostic Evaluation of MT Based on Linguistic Checkpoints. In Proceedings of the 13th Machine Translation Summit, Xiamen, China, 19–23 September 2011; pp. 529–536.