Chapter 2
Bridging Collocational and Syntactic Analysis

Violeta Seretan

Abstract The advent of the computer era, which enabled the development of large text corpora and of sophisticated corpus processing tools, led to unprecedented advances in the area of collocational analysis. These advances were paralleled by significant achievements in the area of syntactic analysis, with parsing technologies becoming available for an increasing number of languages. But more often than not, these developments have taken place independently. The coupling of collocational and syntactic analyses has seldom been considered, despite the fact that one type of analysis could benefit the other. In this chapter, we focus on the integration of syntactic parsing and collocational analysis. First, we review the literature describing syntactically-informed approaches to collocation extraction. Second, we survey the work devoted to exploiting collocational resources for syntactic parsing. Finally, we refer to more recent work that proposes a joint approach to collocational and syntactic analysis, arguing that the two analyses are interdependent to such a degree that only a simultaneous process, one in which structure decoding and pattern identification go hand in hand, can provide a solid bridge between them.

1 Introduction

Quantitative methods have flourished in language-related fields of the humanities, such as linguistics, language learning, or lexicography, ever since the advent of the computer era, which enabled the development of electronic text corpora and of corpus processing technology (Nugues, 2014). These disciplines witnessed the emergence of new subfields, such as corpus linguistics, computational linguistics, computational lexicography and computer-assisted language learning, in which

V. Seretan (✉)
University of Geneva, Geneva, Switzerland

University of Lausanne, Lausanne, Switzerland
e-mail: violeta.seretan@unige.ch

P. Cantos-Gómez, M. Almela-Sánchez (eds.), *Lexical Collocation Analysis*,
Quantitative Methods in the Humanities and Social Sciences,
https://doi.org/10.1007/978-3-319-92582-0_2

collocational analysis – that is, the analysis of patterns of words through techniques like association measures and concordancing – plays an essential role in the study of language. Collocational expressions – e.g. *bright idea*, *heavy smoker*, *break record*, *meet needs* and *deeply sorry* – represent 'the way words combine in a language to produce natural-sounding speech and writing' (Lea and Runcie, 2002, vii); therefore, collocational knowledge has far-reaching implications.

Before computerised tools for corpus processing became available, collocational analysis work has been done manually in different contexts. For instance, in a linguistic context, Maurice Gross compiled very comprehensive information on French nouns, verbs and adverbs (Gross, 1984). In a second language learning context, Harold Palmer and his successor Albert Sydney Hornby carried out pioneering work on compiling lists of frequent collocations. Their work led to the future series of collocation dictionaries known today as the *Oxford Advanced Learner's Dictionary*, one of the major references for the English language (Hornby et al., 1948).

Collocations are important not only for linguistics and lexicographic descriptions, but also for natural language processing and human-computer interaction. As stated by Sag et al. (2002, 2), collocations, along with other types of multi-word expressions or 'idiosyncratic interpretations that cross word boundaries', are 'a pain in the neck for NLP [natural language processing]'. Multi-word expressions are an area of active research in the NLP community, as attested by sustained initiatives, for instances, special interest groups and associations, international projects, book series and scientific events (for an up-to-date review, see Monti et al. 2018). But what makes collocations particularly important is their prevalence in language: 'L'importance des collocations réside dans leur omniprésence' (Mel'čuk, 2003, 26).

The computer-based collocation identification in corpora, known as collocation extraction, has a long tradition. Over the recent decades, a significant body of work has been devoted to the computational analysis of text with the purpose of compiling collocational resources for computerised lexicography, computer-assisted language learning and natural language processing, among others. One of the first large-scale research projects in this area was COBUILD, the Collins Birmingham University International Language Database (Sinclair, 1995). To date, collocation extraction work has been carried out not only for English but for many other languages, including, but not limited to, German, French, Italian, and Korean (as shown in Sects. 2 and 3). Outside an academic setting, commercial software tools such as Sketch Engine (Kilgarriff et al., 2004) and Antidote (Charest et al., 2007) became available that perform collocation extraction from corpora for lexicographic purposes.

In general, the focus of automatic collocation extraction work was on developing appropriate statistical methods, able to pinpoint good collocation candidates in the immense dataset of possible word combinations that quantitative methods consider as their input – a task which has traditionally been described by using the metaphor 'looking for needles in a haystack' (Choueka, 1988). However, purely statistical methods reach their limits as far as low-frequency candidates are concerned. They tend to ignore patterns occurring less than a handful of times, and by doing so they

exclude most of the candidates. Consequently, as Piao et al. (2005, 379) explain, 'the usefulness of pure statistical approaches in practical NLP applications is limited'. It soon became obvious that collocation extraction must have recourse to linguistic information in order to 'obtain an optimal result' (Piao et al., 2005, 379).

Syntax-based approaches to collocation extraction put emphasis on the accurate selection of the candidate dataset in the first place. Returning to the 'needles in a haystack' metaphor, syntax-based collocation extraction focuses on optimising the haystack and transforming it into a much smaller pile, containing less hay and more needles.[1] When collocation analysis methods are coupled to syntactic analysis methods, the input dataset is built in a more careful way, which considers the syntactic relationship between the candidate words, rather than blindly associating any co-occurring words.

In this chapter, we review existing work that combines collocational and syntactic analysis and discuss current trends on coupling these two tasks into a synchronous process, one in which structure decoding and collocation identification go hand in hand to offer an efficient solution benefiting both tasks.

2 Using Syntactic Information for Collocation Identification

Generally speaking, the architecture of a collocation extraction system can be described as a sequence of two main processing modules, preceded by an optional preprocessing module.

Linguistic preprocessing The input corpora are first split into sentences; then, for each sentence, linguistically motivated filters are applied in order to discard the items that are considered uninteresting (e.g. conjunctions and determiners). In addition, this module performs text normalisation. During this stage, a lemmatiser is typically used in order to reduce inflected word forms like *goes*, *went* and *going* to base word forms (*go*).

Stage 1: Candidate selection Based on the preprocessed version of the input, a selection procedure takes place in order to build a collocation candidate list. This procedure uses specific filters in order to decide which combinations of co-occurring words will be considered for inclusion in the candidate list. Traditionally, the filters allow for any word combination to be considered as a collocation candidate, as long as there are no more than four intervening words (hence the name 'window method'). When part-of-speech information is available, the filters request that candidate combinations match one of the patterns in a list of allowed collocation patterns (e.g. noun-noun, noun-preposition-noun, noun-verb, verb-adverb, etc.).[2]

[1] The same metaphor is used by Uhrig and Proisl (2012).

[2] Although there is no generally accepted list of collocation patterns (as it is widely accepted that the parameters of a collocation extraction procedure may vary according to the intended use of results),

Stage 2: Candidate ranking Given the list of collocation candidates from Step 1, a statistical procedure is applied in order to rank candidates according to their likelihood to constitute collocations. The simplest ranking procedure is raw frequency, which lists candidates from the most frequent to the least frequent ones. Often, in order to reduce the candidate dataset to a manageable size, a frequency threshold is applied, which discards all candidates that occur less than a given number of times (e.g. five or ten times).[3]

It is worth noting that no extraction system is devoid of error. The output is to be interpreted by professional lexicographers in order to decide on the relevance of a particular candidate or corpus-based usage sample identified. Caution should also be applied to the parameters of the extraction system: No one-size-fits-all solution exists, and the choices pertaining to corpus size, preprocessing method, window size, filters, ranking method, frequency threshold, etc. must be weighted by taking into account the intended purpose of the results (Evert and Krenn, 2005).

2.1 Statistical Processing

As stated in Sect. 1, the focus of most work devoted to collocation extraction has been on advancing the state of the art of the candidate ranking stage, that is, finding ways to pinpoint good collocation candidates in the immense dataset of initial candidates. (As we will discuss later in Sect. 2.2, considerably less attention has been devoted to the preceding stage, namely, that of candidate selection.)

Over the years – and particularly since the adoption of the mutual information measure from the information theory field as a way to model lexical association (Church and Hanks, 1990) – most research efforts have been spent on the statistics of lexical association. Some of the most representative works include Daille (1994), Evert (2004) and Pecina (2008).

In a nutshell, any method aimed at ranking collocation candidates (also called a lexical association measure) is a formula that computes a score for a collocation candidate, given the following information:

- the number of times the first word appears in the candidate dataset (as the first item of a candidate),

most authors agree that typical collocation patterns include the ones enumerated in Hausmann's definition (1989, 1010) – 'We shall call collocation a characteristic combination of two words in a structure like the following: (a) noun + adjective (epithet); (b) noun + verb; (c) verb + noun (object); (d) verb + adverb; (e) adjective + adverb; (f) noun + (prep) + noun'.

[3]This decision is also motivated by statistical considerations, as most statistical methods are unreliable for low-frequency data. However, it is contested by the lexicographic community, because a significant part of lexicographically interesting candidates occurs only once or twice in a corpus (Piao et al., 2005, 379).

Table 2.1 Candidate ranking: contingency table

	Word 2	Any word different from word 2
Word 1	a	b
Any word different from word 1	c	d

- the number of times the second word appears in the candidate dataset (as the second item of a candidate),
- the number of times the two words appear together (as the first and second item, respectively), and
- the total size of the candidate dataset.

A so-called contingency table is used to synthesise this information (cf. Table 2.1). The letters a, b, c and d represent the frequency 'signature' of the collocation candidate being scored (Evert, 2004).

The correspondence between the letters and the above-stated quantities is established as follows:

- the number of times the first word appears in the candidate dataset (as the first item of a candidate): $a + b$
- the number of times the second word appears in the candidate dataset (as the second item of a candidate): $a + c$
- the number of times the two words appear together (as the first and second item, respectively): a
- the total size of the candidate dataset: $a + b + c + d$.

While the quantities a, b, and c can be computed straightforwardly given the candidate dataset, the number d is to be computed by subtracting the values a, b and c from the total dataset size (usually denoted by N):

$$d = N - (a + b + c) \tag{2.1}$$

Equivalently, since it is easier to compute the quantities $a + b$, $a + c$ (which are called marginal frequencies) and a (which is called joint frequency), we can compute d as follows:

$$d = N - (a + b) - (a + c) + a. \tag{2.2}$$

For the sake of example, we provide below the explicit formula of the log-likelihood ratio association measure, which is one of the most widely used measures for collocation extraction (Dunning, 1993).

$$
\begin{aligned}
LLR = 2(&a\log a + b\log b + c\log c + d\log d - \\
&(a+b)\log(a+b) - (a+c)\log(\mathrm{a}+\mathrm{c}) - \\
&(b+d)\log(b+d) - (c+d)\log(c+d) + \\
&(a+b+c+d)\log(a+b+c+d))
\end{aligned}
\tag{2.3}
$$

An implementation of the computation described above is available, for pedagogical purposes, in the FipsCo Collocation Extraction Toy available in the GitHub software repository.[4] For a comprehensive list of lexical association measures, the interested reader is referred to Pecina (2005, 2008).

From discussing collocation candidate ranking methods, we will now turn to discussing the quality of the information taken into account by such methods.

2.2 *Linguistic Preprocessing and Candidate Selection*

The quality of a collocation extraction system is conditioned by the quality of the candidate dataset. No statistical processing, whatever performant, can improve the quality of the candidate collocational expressions. Given that the extraction output is nothing else than a permutation of the initial candidate list, the importance of linguistic preprocessing and candidate selection becomes evident.

Over the years, there have been repeated calls from researchers working on collocation extraction to use syntactic parsing for collocation extraction. Despite the focus on statistical methods for candidate ranking, there were several early reports acknowledging the fact that successful collocation extraction, particularly for languages other than English, is only possible when performing a careful selection of candidates by using linguistic, as opposed to linear proximity criteria. In the remaining of this section, we review some of the work that stressed the importance of syntax-based collocation extraction.

One of the earliest and better-documented reports in this area is Lafon (1984). The author extracted significant co-occurrences of words from plain French text by considering (oriented, then non-oriented) pairs in a collocational span and by using the *z-score* as an association measure. The preprocessing step consisted in detecting sentence boundaries and ruling out functional words (i.e. non-content words, where a content word is a main verb, a noun, an adjective or an adverb). The author noted that verbs rarely occur among the results, probably as a consequence of the high dispersion among different forms (Lafon, 1984, 193). Indeed, French is a language with a rich morphology,[5] and, in the absence of lemmatisation, the frequency 'signature' values are shrunk, leading to low collocation scores. Apart

[4] https://github.com/seretan/collocation-extraction-toy (accessed 1 February 2018).

[5] A French verb, for instance, may have as many as 48 forms (Tzoukermann and Radev, 1996).

from the lack of lemmatisation, the author also identified the lack of syntactic analysis as one of the main sources of problems faced during extraction. The author pointed out that any interpretation of results should be preceded by the examination of results through concordancing (Lafon, 1984, 201).

A similar report is provided by Breidt (1993) for German. Because syntactic tools for German were not available at that time, Breidt (1993) simulated parsing and used a five-word collocation span to extract verb-noun pairs (such as *[in] Betracht kommen*, 'to be considered', *or [zur] Ruhe kommen*, 'get some peace'). The author used mutual information (MI) and *t-score* as lexical association measures and compared the extraction performance in a variety of settings: different corpus and window size, presence/absence of lemmatisation, part-of-speech (POS) tagging and (simulated) parsing. The author argued that extraction from German text is more difficult than from English text, because of the much richer inflexion for verbs, the variable word order and the positional ambiguity of arguments. She explained that even distinguishing subjects from objects is very difficult in German without parsing. The result analysis showed that in order to exclude unrelated nouns, a smaller window of size 3 is preferable. However, this solution comes at the expense of recall, as valid candidates in long-distance dependencies are missed. Parsing (which was simulated by eliminating the pairs in which the noun is not the object of the co-occurring verb) was shown to lead to a much higher precision of the extraction results. In addition, it was found that lemmatisation alone does not help, because it promotes new spurious candidates. The study concluded that a good level of precision can only be achieved in German with parsing: 'Very high precision rates, which are an indispensable requirement for lexical acquisition, can only realistically be envisaged for German with parsed corpora' (Breidt, 1993, 82).

For the English language, one of the earliest and most popular collocation extraction systems was Xtract (Smadja, 1993). The author relied on heuristics such as the systematic occurrence of two words at the same distance in text, in order to detect 'rigid' noun phrases (e.g. *stock market*, *foreign exchange*), phrasal templates (e.g. *common stocks rose *NUMBER* to *NUMBER**) and flexible combinations involving a verb, which the author calls predicative collocations (e.g. *index [. . .] rose*, *stock [. . .] jumped*, *use [. . .] widely*). Syntactic parsing is used in the extraction pipeline in a postprocessing, rather than preprocessing, stage, and ungrammatical results were ruled out. Evaluation by a professional lexicographer showed that parsing led to a substantial increase in the extraction performance, from 40% to 80%. The author noted that 'Ideally, in order to identify lexical relations in a corpus one would need to first parse it to verify that the words are used in a single phrase structure' (Smadja, 1993, 151).

One of the first hybrid approaches to collocation extraction, combining linguistic and statistical information, was Daille's (1994). The author relied on lemmatisation, part-of-speech tagging and shallow parsing in order to extract French compound noun terms defined by specific patterns, such as noun-adjective, noun-noun, noun-*à*-noun, noun-*de*-noun and noun-preposition-determiner-noun (e.g. *réseau national à satellites*, 'national satellite network'). Daille's shallow parsing approach consisted in applying finite state automata over sequences of POS tags. For candidate ranking,

the author implemented a high number of association measures, including MI and LLR. The performance of these measures was tested against a domain-specific terminology dictionary and against a gold standard set which was manually created from the source corpus with help from experts. One of the most important findings of the study was that a high number of terms have a low frequency ($a \leq 2$). LLR was selected as a preferred measure because it was found to perform well on all corpus sizes and to promote less frequent candidates (Daille, 1994, 173). The author argued that by relying on finite state automata for linguistically preprocessing the corpora, it became possible to extract candidates from very heterogeneous environments, without having to impose a limit on the distance between composing words. This shallow parsing method led to a substantial increase in performance over the window method. According to the author, linguistic knowledge helps to drastically improve the quality of statistical systems (Daille, 1994, 192).

After syntactic parsers became available for German, researchers provided additional insights on the need of syntactic information for successful collocation extraction in this language. For instance, Krenn (2000) extracted P-N-V collocations in German (e.g. *zur Verfügung stellen*, lit., *at the availability put*, 'make available'; *am Herzen liegen*, lit., *at the heart lie*, 'have at hearth'). The author relied on POS tagging and partial parsing, i.e. syntactic constituent detection. She compared various association measures, including MI and LLR. Since syntactic information, the set of candidates identified is argued to contain less noise than if retrieved without such information. The author regrets that the window method is still largely used, 'even though the advantage of employing more detailed linguistic information for collocation identification is nowadays largely agreed upon' (Krenn, 2000, 210). On the same lines, Evert (2004), who carried out substantial joint work with Krenn, explained that 'ideally, a full syntactic analysis of the source corpus would allow us to extract the cooccurrence directly from parse trees' (Evert, 2004, 31).

A similar comment is made by Pearce (2002, 1530), who did experimental work for English and argued that 'with recent significant increases in parsing efficiency and accuracy, there is no reason why explicit parse information should not be used'. In a previous study, Pearce (2001) extracted collocations from English treebanks, i.e. corpora manually annotated with syntactic information.[6]

Additional reports on the necessity of performing a syntactic analysis as a preprocessing step in collocation extraction came from authors that attempted to apply methods originally devised for English to new languages, exhibiting richer morphology and freer word order. For instance, Shimohata et al. (1997) attempted to apply to Korean corpora the extraction techniques proposed for English by Smadja (1993). The authors stated that such techniques are unapplicable to Korean because of the freer word order. Villada Moirón (2005) attempted to identify preposition-noun-verb candidates in Dutch by relying on partial parsing (constituent detection). She showed that partial parsing is impractical for Dutch, because of the syntactic flexibility and free word order of this language. In the same vein, Huang et al.

[6]The same approach has been used by Uhrig and Proisl (2012), among others.

(2005) intended to use POS information and regular expression patterns borrowed from the Sketch Engine (Kilgarriff et al., 2004) to extract collocations from Chinese corpora. The authors pointed out that an adaptation of these patterns for Chinese was necessary in order to cope with syntactic differences and the richer POS tagset.

3 Syntax-Based Extractors

As shown in the previous section, in early collocation extraction work, integrating syntactic parsing in the extraction pipeline was often seen as an ideal, because robust and fast parsers were unavailable for most languages. The past two decades, however, have witnessed rapid advances in the parsing field, thanks, in particular, to the development of statistical dependency parsers for an increasing number of languages (Nivre, 2006; Rani et al., 2015). But despite these advances, a large body of works in the area of collocation extraction still remained linguistically agnostic. Below we review some of the most notable exceptions, which exploited syntactic parsing for improving the performance of collocation extraction.

One of the most important exceptions is Lin (1998, 1999), which describes a syntax-based collocation extraction approach for English based on dependency parsing. Collocation candidates are identified as word pairs linked by a head-dependent relation. The advantage of this approach is that there is no a priori limitation for the distance between two items in a candidate pair, as in the traditional window-based approach. Since the dependency parser is prone to errors, especially for the longer sentences, the author decided to exclude from the input corpus the sentences longer than 25 words. In addition, the author had attempted to semiautomatically correct some parsing errors before proceeding to the identification of collocation candidates based on the parser output. Evaluation was carried out on a small portion of the top-scored results and showed that 9.7% of the candidates were still affected by parsing errors (Lin, 1999, 320).

A similar work was performed for English and Chinese by Wu, Lü and Zhou (Wu and Zhou, 2003; Lü and Zhou, 2004). In their systems, collocation candidates are identified from syntactically analysed text. A parser is used to identify pairs of words linked by syntactic relations of type verb-object, noun-adjective and verb-adverb. Evaluation was performed on a sample of 2000 pairs that were randomly selected among the top-scored results according the LLR score. The results showed a similar rate of error due to parsing, namely, 7.9%.

In the same vein, Orliac and Dillinger (2003) used a syntactic parser to extract collocations in English for inclusion in the lexicon of a English-French machine translation system. In their approach, collocation candidates are identified by considering pair of words in predicate-argument relations. Their parser is able to handle a variety of syntactic constructions (e.g. active, passive, infinitive and gerundive constructions), but cannot deal with relative constructions. In an experiment that evaluated the extraction coverage, the relative constructions have been

found responsible for nearly half of the candidate pairs missed by the collocation extraction system.

Another substantial work in the same direction was performed by Villada Moirón (2005), who experimented with syntax-based collocation extraction approaches for Dutch. The author used a parser to extract preposition-noun-preposition collocations from corpora. Sentences longer than 20 words were excluded, since they were problematic for the parser. Because of the numerous PP-attachment errors, the parser precision was not high enough to allow for the accurate detection of collocations of the above-mentioned collocation type. Therefore, the author adopted an alternative approach, based on partial parsing.

In the context of a long-standing language analysis project at the University of Geneva, we developed the first broad-coverage syntax-based extractor (Seretan and Wehrli, 2006; Seretan, 2008, 2011).[7] Initially available for English and French, it was later extended to other languages (Spanish, Italian, Greek, Romanian) and used for lexical resource development. As mentioned earlier, we adopted a fully syntactically motivated approach to collocation extraction, considering that the first extraction stage, candidate selection, is the most important one. This was in contrast to mainstream approaches, which paid more attention to candidate ranking than to the quality of the candidate dataset.

In our extractor, collocation candidates are identified as pairs of syntactically related words in predefined syntactic relations, such as the ones listed in Hausmann's definition (see Sect. 2). Our extraction is able to detect collocation candidates even if they occur in very complex syntactic environments. This is illustrated by the example below, in which the candidate *submit proposal* is identified in spite of the intervening relative clause:

(1) A joint *proposal* which addressed such elements as notification, consultations, conciliation and mediation, arbitration, panel procedures, technical assistance, adoption of panel reports and GATTs surveillance of their implementation was *submitted* on behalf of fourteen participants.

We comparatively evaluated the performance of syntax-based extraction and window-based extraction in a series of experiments. For instance, in an experiment involving a stratified sample (i.e. pairs extracted at various levels in the output list, from the top to 10%), the extraction precision was found to rise on average per language from 33.2% to 88.8% in terms of grammaticality and from 17.2% to 43.2% in terms of lexicographic interest of the results. The recall was measured in several case studies, which revealed relative strength and weaknesses of the syntax-based and syntax-free approaches. In one such study, it was found that relative to the number of collocation instances identified in a French corpus by the two methods

[7]The extraction system was named FipsCo, as it relies on the output of the Fips parser (Laenzlinger and Wehrli, 1991; Wehrli, 1997; Wehrli and Nerima, 2015). It is available online at http://latlapps.unige.ch (accessed 1 February 2018).

in total (198 instances), the window method identified 70.2% and the syntax-based method 98%.

The example below shows an instance that is missed by the syntax-based method (*payer impôt*, 'pay tax'), because of a semantically transparent noun (*partie*, 'part') intervening on the syntactic path between the verb and the object.

(2) qui *paient* déjà la majeure **partie** des *impôts*
'that already *pay* the biggest **part** of the *taxes*'

These recall-related deficiencies are however largely outweighed by the almost perfect precision of the results. Moreover, by drastically reducing the pool of candidates generated, the syntax-based approach makes it possible to extend the extraction in directions that are underexplored because of the combinatorial explosion problem. One of the extensions considered was, for instance, the iterative application of the collocation procedure in order to detect collocations of unrestricted length, such as *take [a] decisive step*, *take [a] bold decisive step* and so on (Seretan et al., 2003).

A limitation of our approach, which we recently overcame, was the identification of verbal collocations in which the nominal argument is pronominalised (cf. Example 3). The syntactic parser was extended to incorporate an anaphora resolution module, which links the pronominal argument of the verb to its antecedent (Wehrli et al., to appear). Thanks to this module, the new version of the extractor is able to retrieve the nominal collocate (*money*) and to link it to the verbal base (*spend*), even if it occurs in a previous sentence.

(3) Lots of EU *money* are owing to Poland and the rest. *It* must be *spent* fast.

This example illustrates the performance achieved by a collocation extraction pipeline that integrates advanced language analysis modules, such as syntactic parsing and anaphora resolution.

4 Using Collocations (and Other Multi-word Expressions) for Parsing

Collocational analysis is performed in order to improve knowledge about words in general and about complex lexical items (phraseology) in particular. Knowledge about lexical items – the units of language – is at the cornerstone of any language application. Phraseological knowledge has been shown to lead to improvements in the performance of a large number of NLP tasks and applications, including POS tagging and parsing, word sense disambiguation, information extraction, information retrieval, paraphrase recognition, question answering and sentiment analysis (Monti et al., 2018).

As far as syntactic parsing is concerned, the literature provides significant evidence for the positive impact of integrating phraseological knowledge, including

collocations, into parsing systems. For instance, Brun (1998) showed that by using a glossary of complex nominal units in the preprocessing component of a parser, the number of parsing alternatives is significantly reduced. Similarly, Nivre and Nilsson (2004) studied the impact that the pre-recognition of phraseological units has on a Swedish parser. They reported a significant improvement in parsing accuracy and coverage when the parser is trained on a treebank in which phraseological units are treated as single tokens. Zhang and Kordoni (2006) used a similar 'words-with-spaces' pre-recognition approach and reported improvements in the coverage of an English parser. A significant increase in coverage was also observed by Villavicencio et al. (2007) when they added phraseological knowledge into the lexicon of their parser. The same 'words-with-spaces' approach was found by Korkontzelos and Manandhar (2010) to increase in the accuracy of shallow parsing of nominal compound and proper nouns. Finally, reports from the PARSEME[8] community also confirmed that the pre-recognition of complex lexical items has a positive impact on both parsing accuracy and efficiency, the parsing search space being substantially reduced when analyses compatible with complex lexical items are promoted (Constant and Sigogne, 2011; Constant et al., 2012).

These reports prove that information on lexical combinatorics is useful in guiding parsing attachments, especially in 'words-with-spaces' pre-recognition approaches, in which complex lexical items are treated as single tokens. But these approaches have two major shortcomings:

- they are not suitable to syntactically flexible items, which are the most numerous of all phraseological units (with the exception of rigid compounds like *by and large*);
- by imposing a predefined structure for the analysis of a complex lexical item, they take an early commitment on the parsing strategy, which may be wrong and compromise the analysis of the context sentence.

An example illustrating the second point is provided below. The first sentence contains an instance of the verb-object collocation *ask question*. In the second sentence, the same combination *question asked* is in a subject-verb syntactic relation. Treating it as a verb-object collocation leads the parser on a wrong path.

(4a) Any *question asked* during the selection and interview process must be related to the job and the performance of that job.
(4b) The *question asked* if the grant funding could be used as start-up capital to develop this project.

When attempting to couple syntactic and collocational analysis, a further complication that arises is the interdependency between the two types of analysis:

[8]PARSEME (2013–2017) was a European COST Action focusing on the link between complex lexical items and a comprehensive linguistic analysis of text. With more than 200 members from 33 countries, the Action fostered research on the integration of complex lexical items in parsing and translation.

we need collocational knowledge for parsing, but we need parsing to acquire collocational knowledge from corpora. To break this deadlock, we proposed a synergetic approach for the two tasks, namely, collocation identification and parsing attachment decision (Wehrli et al., 2010).

In this approach, the existing collocation information is taken into account during parsing in order to give preference to attachments involving collocation items, but without, however, making a definitive (possibly risky) commitment. Parsing and collocational analysis go hand in hand in a combined analysis, with no necessity to wait for the results of each analysis.

We evaluated this approach by comparing two versions of the parser, one with and the other without synergetic processing. The evaluation showed that the synergetic approach leads to an increase in the parser performance in terms of coverage while at the same time producing an increase in the collocation identification performance.

5 Conclusion

In this chapter, we explored the relationship between syntactic parsing and collocation extraction. Both tasks are essential for (computer-based) language understanding; both have been extensively addressed by the corresponding research communities, and significant advances have been made on each side. But, paradoxically, communication between the two was only rarely considered. Despite the development of fast and robust parsers for an increasing number of languages, collocation extraction work remains mostly focused on improving candidate ranking methods, instead of candidate selection methods – a situation which leads to the perpetration of the 'garbage in, garbage out' principle and its effects. And, despite the development of collocational resources, syntactic parsing work still lacks (in general) appropriate ways to exploit these resources for improving parsing decisions. The integration of knowledge about complex lexical items is still confined, in parsing and translation, to 'words-with-spaces' approaches. These are appropriate for rigid items but fully inappropriate for collocations, which are morphosyntactically flexible and therefore cannot be treated as single tokens.

Our chapter focused on the few exceptional works, which did take into account the advances made in one area in order to foster the other area and vice versa. We reviewed the most representative collocation extraction work which relied on syntactic parsing (or at least highlighted the need for parsing in the area of collocation extraction). We also reviewed some of the few works on syntactic parsing that exploited collocational information for parsing. These are bricks laid at the end of the bridge that aims to fill the gap between the two sides. Even though the research community has made particular efforts to unite the two ends, the bridge is not yet complete. We expect future years to bring exciting new developments in this direction and thus to enable better communication between

the two research communities and, ultimately, to improve language understanding, thanks to converging language analysis efforts.

Acknowledgements I am grateful to the anonymous reviewers, whose comments and suggestions allowed me to improve the chapter.

References

Breidt, E. (1993). Extraction of V-N-collocations from text corpora: A feasibility study for German. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus (pp. 74–83).

Brun, C. (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*, Morristown (pp. 196–200).

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

Choueka, Y. (1988). Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases. In *Proceedings of the International Conference on User-Oriented Content-Based Text and Image Handling*, Cambridge, MA (pp. 609–623).

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22–29.

Constant, M., & Sigogne, A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (pp. 49–56). Portland: Association for Computational Linguistics.

Constant, M., Sigogne, A., & Watrin, P. (2012). Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 204–212). Jeju Island: Association for Computational Linguistics.

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

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

Evert, S. (2004). The statistics of word cooccurrences: Word pairs and collocations. Ph.D. thesis, University of Stuttgart.

Evert, S., & Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language, 19*(4), 450–466.

Gross, M. (1984). Lexicon-grammar and the syntactic analysis of French. In *Proceedings of the 10th Annual Computational Linguistics and 22nd Meeting of the Association for Computational Linguistics*, Morristown. (pp. 275–282).

Hausmann, F. J. (1989). Le dictionnaire de collocations. In F. Hausmann, O. Reichmann, H. Wiegand, & L. Zgusta (Eds.), *Wörterbücher: Ein internationales Handbuch zur Lexicographie* (pp. 1010–1019). Berlin: Dictionaries, Dictionnaires, de Gruyter.

Hornby, A. S., Cowie, A. P., & Lewis, J. W. (1948). *Oxford advanced learner'dictionary of current English*. London: Oxford University Press.

Huang, C. R., Kilgarriff, A., Wu, Y., Chiu, C. M., Smith, S., Rychly, P., Bai, M. H., & Chen, K. J. (2005). Chinese sketch engine and the extraction of grammatical collocations. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju Island (pp. 48–55).

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. In *Proceedings of the Eleventh EURALEX International Congress*, Lorient (pp. 105–116).

Korkontzelos, I., & Manandhar, S. (2010). Can recognising multiword expressions improve shallow parsing? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 636–644). Los Angeles: Association for Computational Linguistics.

Krenn, B. (2000). Collocation mining: Exploiting corpora for collocation identification and representation. In *Proceedings of the KONVENS 2000*, Ilmenau (pp. 209–214).

Laenzlinger, C., & Wehrli, E. (1991). Fips, un analyseur interactif pour le français. *TA Informations, 32*(2), 35–49.

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

Lea, D., & Runcie, M. (Eds.). (2002). *Oxford collocations dictionary for students of English*. Oxford: Oxford University Press.

Lin, D. (1998). Extracting collocations from text corpora. In *Proceedings of the First Workshop on Computational Terminology*, Montreal (pp. 57–63).

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, Morristown (pp. 317–324).

Lü, Y., & Zhou, M. (2004). Collocation translation acquisition using monolingual corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, Barcelona (pp. 167–174)

Mel'čuk, I. (2003). Collocations: Définition, rôle et utilité. In F. Grossmann, & A. Tutin (Eds.), *Les collocations: Analyse et traitement* (pp. 23–32). Amsterdam: Editions De Werelt

Monti, J., Seretan, V., Pastor, G. C., & Mitkov, R. (2018). Multiword units in machine translation and translation technology. In R. Mitkov, J. Monti, G. C. Pastor, & V. Seretan (Eds.), *Multiword units in machine translation and translation technology* (Current issues in linguistic theory, Vol. 341). Amsterdam/Philadelphia: John Benjamins.

Nivre, J. (2006). *Inductive dependency parsing (Text, speech and language technology)*. Secaucus: Springer.

Nivre, J., & Nilsson, J. (2004). Multiword units in syntactic parsing. In *MEMURA 2004 – Methodologies and Evaluation of Multiword Units in Real-World Applications (LREC Workshop)* (pp. 39–46).

Nugues, P. M. (2014). *Corpus processing tools* (pp. 23–64). Berlin/Heidelberg: Springer.

Orliac, B., & Dillinger, M. (2003). Collocation extraction for machine translation. In *Proceedings of Machine Translation Summit IX*, New Orleans (pp. 292–298).

Pearce, D. (2001). Synonymy in collocation extraction. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh (pp. 41–46).

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In *Third International Conference on Language Resources and Evaluation*, Las Palmas (pp. 1530–1536).

Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, Ann Arbor (pp. 13–18).

Pecina, P. (2008). Lexical association measures: Collocation extraction. Ph.D. thesis, Charles University.

Piao, S. S., Rayson, P., Archera, D., & McEnery, T. (2005). Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language Special Issue on Multiword Expressions, 19*(4), 378–397.

Rani, A., Mehla, K., & Jangra, A. (2015). Parsers and parsing approaches: Classification and state of the art. In *Proceedings of the 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, New Delhi (pp. 34–38).

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico City (pp. 1–15).

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

Seretan, V. (2011). *Syntax-based collocation extraction, text, speech and language technology* (Vol. 44). Dordrecht: Springer.

Seretan, V., & Wehrli, E. (2006). Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney (pp. 953–960).

Seretan, V., Nerima, L., & Wehrli, E. (2003). Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of the Fourth International Conference on Recent Advances in NLP (RANLP-2003)*, Borovets (pp. 424–431).

Shimohata, S., Sugio, T., & Nagata, J. (1997). Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Madrid (pp. 476–481).

Sinclair, J. (1995). *Collins cobuild english dictionary*. London: Harper Collins.

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

Tzoukermann, E., & Radev, D. R. (1996). Using word class for part-of-speech disambiguation. In *Proceedings of the Fourth Workshop on Very Large Corpora*, Copenhagen (pp. 1–13).

Uhrig, P., & Proisl, T. (2012). Less hay, more needles – using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates. *Lexicographica, 28*(1), 141–180.

Villada Moirón, M. B. (2005). Data-driven identification of fixed expressions and their modifiability. Ph.D. thesis, University of Groningen.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., & Ramisch, C. (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)*, Prague (pp. 1034–1043).

Wehrli, E. (1997). *L'analyse syntaxique des langues naturelles: Problèmes et méthodes*. Paris: Masson.

Wehrli, E., & Nerima, L. (2015). The fips multilingual parser. In N. Gala, R. Rapp, & G. Bel-Enguix (Eds.), *Language production, cognition, and the lexicon, text, speech and language technology* (Vol. 48, pp. 473–489). Cham: Springer.

Wehrli, E., Seretan, V., & Nerima, L. (2010). Sentence analysis and collocation identification. In *Proceedings of the Workshop on Multiword Expressions: From Theory to Applications (MWE 2010)*, Beijing (pp. 27–35).

Wehrli, E., Seretan, V., & Nerima, L. (to appear) Verbal collocations and pronominalization. In G. C. Pastor & U. Heid (Eds.), *Currrent trends in computational phraseology, research in linguistics and literature*. Amsterdam/Philadelphia: John Benjamins.

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

Zhang, Y., & Kordoni, V. (2006). Automated deep lexical acquisition for robust open texts processing. In *Proceedings of LREC-2006*, Genoa (pp. 275–280).