"Collocation Extraction Based on Syntactic Parsing"

Violeta Seretan

University of Geneva

The problem of collocations

Today, unlike in the early days of modern linguistics, the system of language is no longer seen as divided into two distinct and unrelated parts, lexicon and grammar – that is, words and rules governing their combination. Instead, it is nowadays agreed that the two components interact in complex ways, and they are ultimately inseparable.

Collocations represent a wide-spread language phenomenon situated at the intersection of lexicon and grammar. They are grammatical combinations of words, just like any syntagm (phrase, or syntactic construction) obtained through the application of the grammar rules of a language. From this point of view, they can be seen as regular productions of language; for instance, *ask a question* might seem the simple result of applying a rule which combines a verb and a noun into a verbal phrase. But at the same time, a closer examination of such combinations reveals that there are also reasons for considering them as part of the lexicon. Pedagogically-motivated language studies have been among the first to shed light on collocations and to foster the lexicographic interest in collocations. These studies argue that combinations such as *ask a question* should, in fact, be considered as part of the lexicon, since language learners learn them in block and use them as the islands of reliability of their speech, in order to produce more fluent and naturally-sounding utterances. Thus, *ask a question* could also be seen as an item of the lexicon, which is indecomposable into parts to the extent that the substitution of *ask* with another word – for instance, *make* – would lead to formulations that are perceived as less natural, if not completely awkward in English.

Further arguments for the inclusion of collocations into the lexicon are that they are language-specific, and bilingual collocation knowledge is essential for translation. For instance, in Italian, the translation equivalent for *ask a question* is *fare una domanda* (lit., *make a question*), and this is preferred over *chiedere una domanda,* the literal translation of the English collocation. In French, the appropriate equivalent is *poser une question* (lit., *put a question*), rather than *demander une question.* Another collocation example is *narrow majority*, best translated into French as *courte majorité* (lit., *short majority*) rather than literally, as *majorité étroite*; conversely, *short majority*, the literal translation of the French collocation *courte majorité* into English, is perceived as less natural than *narrow majority*.

Collocations can be understood, basically, as "the way words combine in a language to produce natural-sounding speech and writing" (Lea & Runcie, 2002). Despite their practical importance, in theoretical linguistics collocations have been less studied and are consequently less well described than similar language phenomena, such as compounds (e.g., *hot dog, dry-clean, raincoat)* or idioms (e.g., *to be over the moon* "to be very happy", *raining cats and dogs* "raining heavily", *kick the bucket* "to die"). With respect to these phenomena, collocations are much more difficult to describe, both from a syntactic and a semantic point of view. Syntactically, they are much more flexible, as illustrated by the sentences below:

(1) a. I should also like to *ask* the following supplementary *question*.
b. Madam President, ladies and gentlemen, these are the *questions* that you, together with the members of our national parliaments, and our governments, must start to *ask* and answer.

As can be seen from Example (1) and particularly from (1b), the words that make up a collocation may, in some cases, be separated by a very large number of words. However, in other cases (e.g., *pay attention*), the syntactic behaviour of collocations is more restricted. Since their morphological and syntactic properties vary from case to case, collocations cannot really be given a uniform description in morpho-syntactic terms. Such a systematic description seems also impossible from a semantic point of view. In fact, collocations are seen as elements populating the grey area of a continuum ranging from fully transparent combinations, represented by regular combinations, to completely opaque combinations, represented by idioms. Some collocations are more transparent (*ask a question*), while others are more opaque (*pay attention*). It is widely agreed that the syntactic restrictions and the semantic opacity go hand in hand.

Because of their syntactic and semantic idiosyncrasy, collocations pose serious challenges to the automatic processing of natural language (NLP). Knowledge about which word combinations constitute collocations is essential for many NLP tasks, such as machine translation. The problem that arises in this context is that the existing methods designed for automatically acquiring such knowledge, from resources such as text repositories, are not sufficiently developed. Most importantly, these methods did not keep pace with the steady progress made in other NLP

fields, such as the field of *parsing* (i.e., the automatic syntactic analysis of natural language). This is why the main purpose of my work was to cope with this paradoxical situation, by developing better collocation extraction methods that are based on syntactic parsing.

Collocation Extraction and its Applications

Thanks to the advent of the computer era and the development of corpus linguistics, the field of computational linguistics has witnesses a sustained interest in the study of language centred on collocation identification: as put forth by Firth (1957), "You shall know a word by the company it keeps!".

Much practical work has been devoted over the past few decades to modelling the phenomenon of collocation and designing methods for the automatic identification of lexicographically interesting word combinations in large collections of texts, called *corpora* (the plural of *corpus*, "body"). The very first methods begun by identifying frequent sequences of words, or *n-grams* (Choueka, 1988). Subsequent work made use of statistical methods able to detect word pairs that are not necessarily adjacent, but tend to occur at a stable distance in text (Smadja, 1993). Also very popular – still nowadays – were the "window" methods that allowed any distance between the two words, up to a maximum distance of, typically, 5 words (Church and Hanks, 1990). Because they got too many irrelevant candidates, they filtered them by considering part-of-speech patterns such as verb-noun or adjective-noun, as part-of-speech taggers became available. In addition, they made use of more advanced methods to rank candidate pairs so that the pairs formed by words that are more correlated receive a higher score. Such methods are based, most usually, on mutual information (an information-theoretic measure), and on statistical hypothesis testing, which verifies the statistical significance of the event consisting of observing the two words together within the same window of text. The absence of significance means that the two words only occurred together by chance; significant candidate pairs are, instead, considered as collocations.

As soon as researchers begun investigating languages other than English, the limits of these methods became clear. For languages that have a richer inflection and allow for a higher degree of word order

freedom, such as French, German, or Korean, these methods no longer produced satisfactory results. Statistical significance computation is affected by the lower word frequencies observed in the corpus, since the occurrences of a word are spread through a large number or word forms, particularly in the case of verbs. Moreover, verbs often have their arguments – in particular, the objects – occurring outside the 5-word window allowed in selecting candidate pairs. This leads to unsatisfactory results, researchers pointing out that the only solution is the syntactic analysis of the input text (Breidt, 1993).

As a matter of fact, more recent extraction work tried to deal with these problems by relying more and more on linguistic preprocessing, aimed at syntactic parsing or at least at approximations of parsing:

- chunking, which helps identifying the main chunks or phrases in a sentence; their lexical heads are combined, if found in the same sentence, to form a candidate pair (Krenn and Evert, 2001);
- shallow parsing, which attempts to find the main syntactic relations between words by looking at a limited context (Kilgarriff et al., 2004);
- dependency parsing, which finds the head word on which a word depends (Lin, 1998);
- finally, full parsing, which builds a syntactic tree showing the complete structure of a sentence (Seretan and Wehrli, 2006).

Collocation extraction enables the construction of collocation dictionaries, which have a very large applicability. The following are only a few important NLP tasks for which collocation knowledge is useful, if not essential.

- ***Syntactic parsing.*** Collocational knowledge provides cues for both lexical and structural disambiguation. For example, if the parsing system knows that the words *break* and *record* constitute a collocation, then whenever it finds them in a sentence, it will favour analyses in which the two words have the compatible part-of-speech and meaning, and are in the compatible syntactic configuration (verb-object).
- ***Machine Translation.*** Collocations are a key factor in producing more acceptable output. For instance, *break a record* can only be correctly translated intro French as *battre un record*, and not as *casser un record*, if the system has access to a bilingual collocation lexicon containing an entry which specifies the correct translation.

- ***Word sense disambiguation.*** The majority of words have multiple meanings. However, according to the one sense per collocation hypothesis (Yarowsky, 1993), the meaning of a word can be disambiguated once the system has detected a specific word in its context. Thus, by knowing the collocation *break a record*, the system would avoid interpreting *record* as "something on which sound or visual images have been recorded", or *break* as "to separate into parts with suddenness or violence", and would understand the whole combination as "to surpass in excellence".

The NLP literature consistently reports on significant improvements obtained in the performance of various such tasks, when collocation information is taken into account.

In lexicography, collocation extraction methods are being widely used nowadays to produce the raw material for inclusion in collocation dictionaries, after careful examination by lexicographers. Examples of extraction systems used in lexicography are the Sketch Engine (Kilgarriff et al., 2004) and Antidote (Charest et al., 2007). Both systems are based on syntactic parsing, the first relying on shallow parsing and the second on full parsing.

The Need for Syntax-Based Extraction

Computational linguistics research (Smadja, 1993; Heid, 1994; Evert, 2004) has long since recognised the necessity to syntactically analyse the source corpora in order to properly select collocation candidates and to account for the large gamut of syntactic transformations that collocations may undergo (cf. Example 1). However, in the absence of appropriate syntactic tools, the syntax-free methods, and, in particular, the aforementioned window method, have been in use for a long time as the standard alternative for extraction. This situation remained unchanged even after the field of parsing has made significant progress. But at LATL, the Language Technology Laboratory of the University of Geneva, we believed that time was right for a methodological shift, and my research pursued this goal.

Below, I will illustrate by means of a few examples what are the benefits that can be obtained by using syntactic parsing for collocation extraction. Consider again Example (1): the same collocation, *ask a*

question, occurs in sentences (1a) and (1b), but the form of the words and their relative order differ (*ask – question* vs. *questions – ask*). Moreover, the two words occur in (1b) at a distance which is larger than the typical window size of 5 words. The advantage of syntax-based extraction is that it can deal with the morpho-syntactic variation of collocations, thus considering both instances as belonging to the same pair. The frequency of the pair is higher, therefore the statistical significance computation is more reliable and leads to a more accurate ranking for that pair. Another advantage visible in this example is that syntax-based extraction is able to discover more pair instances, whose words fall outside the fixed-size window. This results in more accurate frequency information, and therefore, more accurate statistical significance computation, as above. In addition, it results in the discovery of those pairs which are often realised as long-distance dependencies, and which are otherwise missed by the standard window method.

Syntactic information is essential in correctly interpreting the extraction results and in integrating them into other applications, such as the ones mentioned earlier. Without syntactic information, it is not possible to know, for instance, whether the noun *question* is the subject of the verb *ask*, or it is its object. In Example (1), *question* has the role of object, but in Example (2) below, it is a subject, and the two situations have to be set apart.

(2) My *question* specifically *asked* whether the Commission felt threatened by his words.

The window method incorrectly identifies the pair *question – asked* from this example as an instance of the collocation *ask a question*; with syntax-based extraction, such errors are avoided thanks to the analysis provided by parsing.

Last but not least, the results of syntax-based extraction are validated by a syntactic parser, and, even if the latter inherently makes mistakes, the pairs extracted are more reliable than those obtained without parsing. Errors may still be present among the top extraction results of the window method, despite the strong filter represented by the statistical significance computation. A typical situation leading to errors is, in the case of the window method, the presence in the source text of longer expressions, such as *human rights organisation*. Erroneous pairs like *human – organisation* still appear among the extraction results, because

the component words are found to co-occur often within a 5-word window, without actually being syntactically related. In the absence of a syntactic validation, the pair *human organisation* is misleading, since it looks like a correct adjective-noun pair, when it is in fact incorrect.

Research Contributions

The first step undertaken for achieving the objective of syntax-based collocation extraction was to develop and validate an extraction methodology based on full parsing. The multilingual parser Fips (Wehrli, 2007) developed at LATL has been used in the experiments, and the comparison was made against the standard window method. The "hybrid" extraction system I developed (Seretan and Wehrli, 2006) relies on the syntactic parsing of source corpora to optimize the selection of candidate pairs from text. It combines syntactic information with statistical information in a way that is backed by previous theoretical stipulations (Heid, 1994). The cross-linguistic evaluation performed by human judges showed that the hybrid extraction method that relies on syntactic parsing produces results which are are much superior over those of the standard syntax-free method. This finding is consistent with similar work making use of syntactic parsing for other NLP tasks, and shows the advantage of using parsers for collocation extraction, whenever available. In my case, the wide grammatical coverage of Fips, its robustness and speed in processing large text corpora have been crucial in obtaining high quality results. Fips is available in six major languages – English, French, German, Spanish, Italian, and Greek – and a number of new languages are under development. Worldwide, the parsing field is making steady progress, especially through the development of language-independent frameworks for dependency parsing (Nivre, 2006). There is a visible trend in integrating parsing information into collocation extraction methods, and my work could serve as inspiration to similar syntax-based approaches.

Further steps have been taken to use the newly-developed syntax-based methodology for conducting other practical investigations, in directions less explored in the existing work. The main goal of these investigations was to enlarge the scope of the extraction method to cover a broader spectrum of collocations in text, and thus, to provide a more comprehensive account of this wide-spread language phenomenon.

• ***Complex collocation extraction.*** First, the purpose was to go beyond binary collocations and design a tractable method for acquiring complex collocations – i.e., those collocations containing embedded collocations, such as *take a wrong decision*, whose German equivalent is *eine falsche Entscheidung treffen*, lit., "encounter a false decision". Binary collocation is only one facet of the phenomenon of word collocation; they can be embedded within other binary collocations to form complex collocations containing a larger number of words. Indeed, theoretical studies consider collocations as made up of two or more words, but the practical work generally ignores this aspect and focuses almost exclusively on binary collocations. The reason stands in the manner in which collocations are modelled, as significant word pairs that can be discovered by using association measures designed for two variables. Standard window-based methods are affected by combinatorial explosion when considering combinations of more that two words and allowing for a larger window size. On the contrary, by relying on the syntax-based extraction methodology, it was possible to use the mechanism of recursion in order to detect complex collocations. The idea was to check whether binary collocations previously identified by the hybrid system occur significantly in combination with other such collocations – for instance, *take a decision* and *wrong decision*. Thus, I applied the definition of collocation (as significant combination of words) recursively, to collocation themselves (significant combination of collocations). I relied on the concept of "collocation of collocation" to detect more complex combinations typical in a language, whose length is only empirically restricted: *weapons of mass destruction, proliferation of weapons of mass destruction, treaty on the non-proliferation of weapons of mass destruction*, and so on. This method is tractable and efficient, thanks to the advantages brought by the syntactic analysis: only syntactically related word pairs are taken into account; the words in a pair may be far apart in the sentence; the syntactic variation can easily be dealt with, and multiple different instances are conflated as belonging to the same type.

• ***Pattern induction.*** The second investigation was related to the syntactic configurations – or *patterns* – considered as relevant for collocation extraction, e.g., adjective-noun, subject-verb, verb-object. These patterns play a crucial role in the quality of extraction results, but are generally chosen in an arbitrary way. Moreover, a set of patterns

designed for one language cannot be straightforwardly transferred to another language. Very often, extraction systems are designed for a single pattern, or consider only the most representative ones. Some systems take into account only open-class words (nouns, verbs, adjectives), other also consider closed-class words (prepositions, conjunctions etc). As a matter of fact, the patterns considered are highly divergent from one system to another. This situation reflects the syntactic idiosyncrasy of collocations and the lack of a precise description in syntactic terms. I started from the idea that words of any category, and in any syntactic relation can show collocational effect (Fontennelle, 1992; Van der Wouden, 1997). Thus, I sought to find a solution to the problem of pattern selection, again by relying on the extraction methodology designed. The syntactic proximity criterion used for selecting candidate pairs was relaxed so that any pair, in any syntactic relation, was considered as a potential collocation candidate, instead of requiring it be in a specific predefined set of patterns. Subsequently, for each pattern that was found productive in a language, the extraction method continued with the ranking of pairs according to their significance. Finally, human judges inspected the results and decided whether a pattern produces or not collocationally interesting results. This data-driven method of semi-automatic pattern induction lead to the discovery of new patterns for English and French, the majority of which being ignored by existing extractors, in spite of their high relevance for lexicography. A typical example is the preposition-noun pattern, which represent combinations such as *on page* (compare with the French equivalent, *à la page*).

• ***Web extraction.*** The extraction system has been extended so that the words that form collocations with a specific word can be found by harnessing the immense knowledge that is found in world's largest text repository, the World Wide Web. Recent research in computational linguistics considers the Web as a valid source of linguistic knowledge, which is readily available and can be used as an alternative to pre-compiled text corpora (Keller and Lapata, 2003). The motivating scenario is that often a user (e.g., a lexicographer, a language learner) or an application are interested in the collocates of a given word – for instance, *question* in French. The simple solution (adopted in existing related work) is to compare the number of hits returned by a search engine for alternative queries, e.g., *faire une question, demander une*

question, and *poser une question*. The combination with the highest number of hits signals the collocation (*poser une question*). Of course, this solution is efficient only when the words in the collocation are known in advance, which is often an impractical assumption. The method I developed sends a query with the word *question* to the Google search engine, using its APIs (Application Programming Interface). It retrieves the sentences in which *question* occurs, and applies the syntax-based extraction method to find and rank all the verb-object pairs with *question*, ultimately highlighting *poser – question* as the most significant pair. The system can achieve good results even by using a relatively small number of sentences, thanks to the fact that the extraction is based on parsing.

• ***Bilingual extraction.*** Collocation extraction is normally seen as a monolingual task, which is useful in language analysis and generation. But bilingual extraction (the extraction of both collocations and their translation equivalents) is an even more useful task, from the perspective of machine translation, foreign language learning, and lexicography. Previous work has mainly exploited parallel text corpora, which contain versions of one document in multiple languages, to automatically find translations for a specific type of collocations, namely, noun phrases (e.g., *wheel chair*); these are relatively easier to identify than the collocations which are more syntactically flexible, such as those containing a verb. In my work, starting from the idea that a collocation in the source language has the same (or a compatible) syntactic relation in the target language, and that one of the words is always translated literally, I defined a collocation translation method based on matching the results of collocation extraction for the two languages involved. For instance, the verb-object collocation *reach a compromise* is translated into French by taking a number of sentences (usually, up to 50) in which it occurs, finding their French counterpart, extracting collocations from the French sentences and looking for French collocations that contain *compromis* (the literal translation of *compromise*), and that have a compatible syntactic type: either verb-object (*trouver un compromis*, lit., "find a compromise") or verb-preposition-argument (*parvenir à un compromis*, lit., "arrive at a compromise"). Again, this method yields accurate results from just a small amount of sentences thanks to the availability of syntactic information.

The ability of all these methods to perform well on small amounts of input data is crucial, because of well-known properties of natural language related to data sparsity: language has a Zipfian distribution, with only a small number of common events and a much larger number of rare events (sparse data). The results obtained show that syntactic parsing is a powerful ingredient in designing computational methods aimed at a comprehensive account of collocations in language.

The research on syntax-based collocation extraction I have conducted at LATL has contributed to advancing the state of the art in the field, to filling the gap between theoretical stipulations and practice, and clarifying the role of syntactic parsing (often dismissed as inefficient) in yet another area of computational linguistics. This research has been, first of all, practically motivated. The extraction system developed is being used in ongoing projects at LATL and other institutions, and the publications arising from this research inspired the work of other teams, both in academy and in industry – for instance, the team that authored the Antidote commercial system (Charest et al., 2007).

Acknowledgement

I am grateful to the International Latsis Foundation for generously awarding me the 2010 Geneva University prize for this research, and I am honoured that the Research Committee considered my work worthy of this prestigious distinction. I also wish to thank my supervisor, Prof. Eric Wehrli, for support and for allowing me to carry out this work in the best conditions.

References

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

Charest S, Brunelle E, Fontaine J, Pelletier B (2007) Elaboration automatique d'un dictionnaire de cooccurrences grand public. In: Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007), Toulouse, France, 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, USA, pp 609–623.

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

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

Firth JR (1957) Papers in Linguistics 1934-1951. Oxford University Press, Oxford.

Fontenelle T (1992) Collocation acquisition from a corpus or from a dictionary: A comparison. Proceedings I-II Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere, Tampere, Finland, pp 221–228.

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

Keller F, Lapata M (2003) Using the Web to obtain frequencies for unseen bigrams. Computational Linguistics 29(3):459–484.

Kilgarriff A, Rychly P, Smrz P, Tugwell D (2004) The Sketch Engine. In: Proceedings of the 11th EURALEX International Congress, Lorient, France, pp 105–116.

Krenn B, Evert S (2001) Can we do better than frequency? A case study on extracting PP-verb collocations. In: Proceedings of the ACLWorkshop on Collocation: Computational Extraction, Analysis and Exploitation, Toulouse, France, pp 39–46.

Lea D, Runcie M (eds) (2002) Oxford Collocations Dictionary for Students of English. Oxford University Press, Oxford.

Lin D (1998) Extracting collocations from text corpora. In: First Workshop on Computational Terminology, Montreal, Canada, pp 57–63.

Nivre J (2006) Inductive Dependency Parsing (Text, Speech and Language Technology). Springer-Verlag New York, Inc., Secaucus, NJ.

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 (COLING/ACL 2006), Sydney, Australia, pp 953–960.
Smadja F (1993) Retrieving collocations from text: Xtract. Computational Linguistics 19(1):143–177.
Wehrli E (2007) Fips, a "deep" linguistic multilingual parser. In: ACL 2007 Workshop on Deep Linguistic Processing, Prague, Czech Republic, pp 120–127.
van der Wouden T (1997) Negative Contexts. Collocation, Polarity, and Multiple Negation. Routledge, London, New York.
Yarowsky D (1993) One sense per collocation. In: Proceedings of ARPA Human Language Technology Workshop, Princeton, NJ, USA, pp 266–271.