Lang Resources & Evaluation (2009) 43:71–85
DOI 10.1007/s10579-008-9075-7

Multilingual collocation extraction with a syntactic parser

Violeta Seretan · Eric Wehrli

Published online: 1 October 2008

Abstract An impressive amount of work was devoted over the past few decades to collocation extraction. The state of the art shows that there is a sustained interest in the morphosyntactic preprocessing of texts in order to better identify candidate expressions; however, the treatment performed is, in most cases, limited (lemmatization, POS-tagging, or shallow parsing). This article presents a collocation extraction system based on the full parsing of source corpora, which supports four languages: English, French, Spanish, and Italian. The performance of the system is compared against that of the standard mobile-window method. The evaluation experiment investigates several levels of the significance lists, uses a fine-grained annotation schema, and covers all the languages supported. Consistent results were obtained for these languages: parsing, even if imperfect, leads to a significant improvement in the quality of results, in terms of collocational precision (between 16.4 and 29.7%, depending on the language; 20.1% overall), MWE precision (between 19.9 and 35.8%; 26.1% overall), and grammatical precision (between 47.3 and 67.4%; 55.6% overall). This positive result bears a high importance, especially in the perspective of the subsequent integration of extraction results in other NLP applications.

Keywords Collocation extraction · Evaluation · Hybrid methods · Multilingual issues · Syntactic parsing

1 Introduction

In a context in which multi-word expressions in general became an increasingly important concern for NLP (Sag et al. 2002), the task of acquiring accurate

V. Seretan (✉) · E. Wehrli
Language Technology Laboratory (LATL), University of Geneva, Geneva, Switzerland
e-mail: violeta.seretan@unige.ch

collocational resources assumes a particular importance. First of all, collocations make up the lion's share of multi-word expressions (Mel'čuk 1998). Typical syntagmatic combinations such as *large majority, great difficulty, grow steadily, meet requirement, reach consensus*, or *pay attention* are prevalent in language, regardless of genre or domain; in fact, according to a recent study, each sentence is likely to contain at least one collocation (cf. Pearce 2001).

Since their meaning is easy to decode from the meaning of the component words, collocations are usually distinguished from idioms, i.e., semantically opaque units such as *pull someone's leg, kick the bucket*, or *be the last straw*. However, unlike regular combinations, collocations are idiosyncratic: the lexical item typically selected with the headword in order to express a given meaning is contingent upon that word (Mel'čuk 2003). Compare, for instance, *large majority* with *great difficulty* or *distinct preference*: the meaning of intensity is typically expressed each time by a different adjective. This idiosyncrasy becomes more apparent across languages: *ask a question* translates into French as *poser une question* (lit., ?*to put a question*) but into Italian and Spanish usually as *fare una domanda* and *hacer una pregunta* (lit., **to make a question*).

Secondly, a critical problem with existing extraction systems is that they generally rely on blind word combinatorics, while completely disregarding linguistic criteria that are essential both for obtaining accurate results and for successfully integrating them in other NLP applications, such as parsing, machine translation, and word sense disambiguation.

Consider a sentence like the following[1]: *The question asked if the grant funding could be used as start-up capital to develop this project*. Most of the existing systems would normally succeed in identifying the pair *question-asked* as a collocation candidate, but fail to recognize that it concerns a subject–verb, and not a verb–object syntactic relation. Not only does the lack of syntactic information for the pairs preclude their proper handling in subsequent applications, but it also negatively affects extraction: whenever candidate pairs are wrongly assimilated to pairs of another syntactic type, their frequency profile, on which the extraction procedure relies, is actually falsified.

In this article we present an approach to collocation extraction that relies on the full syntactic analysis of the source corpus in order to ensure the proper candidate identification and the adequate syntactic description of output pairs. After a language-oriented review of existing extraction work (Sect. 2), the paper discusses several issues that arise when attempting to adapt existing extraction techniques—such as those developed for English—to a new language (Sect. 3), then it describes (in Sect. 4) our multilingual extraction system based on parsing. Section 5 presents several experimental results and an evaluation study that compares the performance of our method with that of a standard, syntactically-uninformed procedure. Finally, Sect. 6 concludes the article by discussing the related work and pointing out future research directions.

[1] All the sample sentences provided in this paper actually occurred in our corpora.

2 Review of extraction work

Collocation is generally seen in NLP as a phenomenon of lexical affinity that can be captured by identifying statistically significant word associations in large corpora by using so-called *association measures* (henceforth AMs), e.g., *t*-score, *z*-score, mutual information (MI), chi-square, log-likelihood ratios (LLR). For their description and discussion of the relative merits see for instance (Barnbrook 1996; Kilgarriff 1996; Manning and Schütze 1999; Pearce 2002; Evert 2004).

Generally speaking, an extraction procedure comprises two main steps: (1) the identification of candidates, often based on the morphologic and the syntactic preprocessing of source texts, and (2) the candidates ranking according to the collocational strength or association score, computed with a given AM on the basis of the frequency information stored in the contingency table of candidate pairs. The remaining of this section provides a language-oriented overview of the existing extraction work.

English: Earlier methods generally deal with *n*-grams (adjacent words) only, and use the plain co-occurrence frequency as an AM (Choueka 1988; Kjellmer 1994; Justeson and Katz 1995). The last work cited notably applies a POS filter on candidates. Similarly, Church and Hanks (1989, 1990) extract adjacent pairs—more precisely, phrasal verbs—by POS-tagging the source text, except that they further apply MI for ranking. Later, Smadja (1993) detects rigid noun phrases, phrasal templates, and also flexible combinations involving a verb (*predicative collocations*). His system, Xtract, combines the *z*-score with several heuristics, such as the systematic occurrence of two lexical items at the same distance in text. A parser is finally used for validating the results, thanks to which the accuracy of the system is shown to increase considerably (from 40% to 80%).

More recent methods are generally able to extract flexible pairs, as they rely on shallow-, dependency-, or full parsing. Church et al. (1989) already used a shallow parser to detect verb–object pairs, that were further ranked with MI and the *t*-score. In the Sketch Engine (Kilgarriff et al. 2004), collocations candidates are also identified with shallow parsing implemented as regular expression pattern-matching over POS tags. The AM used is an adaptation of MI that gives more weight to the co-occurrence frequency. In Lin (1998, 1999), the candidate identification is based on dependency parsing, while for their ranking are employed LLR and a version of MI. LLR is also used in Goldman et al. (2001), the earlier version of our extractor. This system is based on full parsing and is particularly suited for retrieving long-distance collocation instances, even if subject to complex syntactic transformations (as will be seen in Sect. 4).

German: Breidt (1993) applies MI and *t*-score for German and thoroughly evaluates the performance of these AMs in a variety of settings: different corpus and window size, presence/absence of lemmatization, of POS tagging and (simulated) parsing. This study was focused on V–N pairs[2] and concluded that good accuracy can only be obtained in German with parsing (Breidt 1993, p. 82). Recent work (Krenn

[2] The following abbreviations are used in this paper: N—noun, V—verb, A—adjective, Adv—adverb, C—conjunction, P—preposition, Inter—interjection.

2000; Krenn and Evert 2001; Evert and Krenn 2001; Evert 2004) makes use of chunking for extracting particular types of collocations, mainly P–N–V, and is mostly concerned with the comparative evaluation of AMs. Also, Evert and Kermes (2003) extract A–N pairs using three different methods (adjacent POS tags, window of size 10, and chunking). Unsurprisingly, the highest recall is obtained with chunking, and the highest accuracy with the adjacency method. Zinsmeister and Heid (2003) identify N–V and A–N–V candidates with a stochastic parser and classify them into interesting or trivial combinations by means of machine learning techniques taking into account the LLR score. Finally, Wermter and Hahn (2004) extract PP–V combinations by relying on shallow parsing and on the limited modifiability criterion.

French: Outstanding work carried out on lexicon-grammar before computerized tools even became available makes French one of the most studied languages in terms of the distributional and transformational potential of words (Gross 1984). Automatic extraction was first performed in (Lafon 1984), then, to a certain extent, in the framework of terminology extractors dealing specifically with noun-phrases.

For instance, Bourigault (1992) extracts noun-phrases like N–A and N–P–N with shallow parsing, by first identifying phrase boundaries. Similarly, Daille (1994) relies on POS-tagging and lemmatization in order to extract compound nouns defined by specific patterns, such as N–A, N–N, N-*à*–N, N-*de*–N, N–P–Det–N. The system applies a long series of AMs, whose performance is tested against a domain-specific terminology dictionary and against a gold-standard manually created from the source corpus. Also, Jacquemin et al. (1997) use a 10-words window method coupled with a syntactic filter based on shallow parsing, paying particular attention to the detection of morphosyntactic term variants.

Collocation extraction proper is performed by Tutin (2004) by using the local grammar formalism in the INTEX framework (Silberztein 1993). Also, Goldman et al. (2001) identify collocation candidates with full parsing and rank them with LLR, just as in the case of English.

Other languages: Collocation extraction work has also been performed in a number of other languages, among which *Italian*: Calzolari and Bindi (1990) employ the window method for candidate identification in untagged text coupled with MI for ranking, while Basili et al. (1994) make use of parsing information; *Dutch*: Villada Moirón (2005) extracts P–N–P and PP–V expressions using POS filtering and also, to a limited extent, parsing; *Korean*: Shimohata et al. (1997) use an adjacency *n*-gram model on plain text and an entropy-based AM for ranking, while Kim et al. (1999) rely on POS-tagging; *Japanese*: Ikehara et al. (1995) apply an improved *n*-gram method that allows them to extract interrupted collocations; *Chinese*: Huang et al. (2005) use POS information and patterns borrowed from the Sketch Engine (Kilgarriff et al. 2004), and Lu et al. (2004) employ a method similar to Xtract (Smadja 1993).

3 Portability issues

This review of collocation extraction work reveals a gradual evolution of the extraction methodology used (from frequency counts to machine learning

techniques), of the phenomenon covered (from rigid sequences of adjacent words to flexible predicative relations without an a priori limitation for the collocational span), and also a general interest in adapting existing techniques to new languages.

A series of issues arise when attempting to apply an extraction procedure—most usually, one that was designed for English—to a new language. These are discussed below.

Richer morphology: In this case, lemmatization is, unlike in English, a true necessity because the form-based frequencies might be too small for the AMs to function properly. It is a well-known fact that AM scores are unreliable when the observed values in the contingency table are very low. Grouping all the inflected variants under the same lemma translates into more significant extraction results (Evert 2004, p. 27).

Freer word-order: As shown in Breidt (1993), Kim et al. (1999) or Villada Moirón (2005, p. 162), extraction is more difficult—i.e., the performance of standard techniques based on a superficial text analysis is low—in languages in which arguments can be scrambled freely. In German, even distinguishing subjects from objects is very difficult without parsing (Breidt 1993). A related issue is the higher syntactic transformation potential, which is responsible for the long-distance extraposition of words. The common practice of using a 5-words span for collocate searching might therefore be too restrictive, as proven for French (Jacquemin et al. 1997; Goldman et al. 2001).[3]

Language-specific syntactic configurations: It has already been proven that the morphosyntactic analysis improves extraction results considerably, e.g., in Church and Hanks (1990), Breidt (1993), Smadja (1993), Lin (1999), Zajac et al. (2003). But in order to take full advantage of it, it is essential to know the collocationally relevant syntactic configurations for the new language. Some configurations are in principle appropriate for many languages (such as N–V, V–N, V–Adv, N–A; that is, the general predicate-argument or head-modifier relations), but others are specific to the syntactic structures of the new language (e.g., P–N–V in German that corresponds to V–P–N in English), or have no straightforward counterpart in the target language (e.g., P–A in French: *à neuf*, might correspond to Conj-A in English: *as new*).

Mapping syntactic configurations—AMs: The performance of AMs appear to be sensitive to the syntactic configuration (Evert and Krenn 2001). But since the lexical distribution varies across languages (for instance, in French there are fewer V–P pairs than in English, where they constitute phrasal verbs and verb-particle constructions), an AM that is suited to a syntactic type in one language might be less suited to that type in another. For successful extraction, it is therefore important to find the best tuning between AMs and syntactic configurations for each language.[4]

[3] Jacquemin et al. (1997, p. 27) argue that a 5-words window is insufficient for French due to the "longer syntactic structures". In fact, Goldman et al. (2001, p. 62) identified some instances of verb–object collocations that had the component items separated by as much as 30 intervening words.

[4] Evert and Krenn (2005) indicate that this choice is also dependent on the specific extraction setting (e.g., domain and size of corpora, frequency threshold applied, type of preprocessing performed).

4 An extraction method based on full parsing

The preceding sections showed that in the multilingual context, the syntactic preprocessing of source corpora represents a more important requirement for collocation extraction than traditionally seen in the English setting. As a matter of fact, only a minority of existing English extractors incorporate syntactic knowledge, despite the recent advances in parsing, and despite the suggestion of researchers like Church and Hanks (1990, p. 25) or Smadja (1993, p. 151) to extract collocations from parsed text, as soon as adequate tools for processing large text corpora will become available.

We present an extraction system for four languages (English, French, Spanish and Italian) that implements a hybrid extraction method combining syntactic and statistical techniques.

4.1 Fips parser

The system relies on Fips, a deep symbolic parser based on generative grammar concepts that was developed over the last decade in our laboratory, LATL (Wehrli 2007). Written in Component Pascal, it adopts an object-oriented implementation design allowing to couple language-specific processing with a generic core module. The parsing algorithm proceeds in a bottom-up fashion, by applying general or language-specific licensing rules, by treating alternatives in parallel, and by using pruning heuristics.

In Fips, each syntactic constituent is represented as a simplified X-bar structure of the form $[_{XP}$ L X R] with no intermediate levels, where X is a variable ranging over the set of lexical categories.[5] L and R stand for (possibly empty) lists of, respectively, left and right subconstituents that bear the same structure in turn. The lexical level contains detailed morphosyntactic and semantic information available from manually-built lexicons.

The parser builds the canonical form for a sentence, in which extraposed elements (relative pronouns, clitics, interrogative phrases etc.) are coindexed with empty constituents in canonical positions (i.e., typical argument or adjunct positions). For instance, the sentence in (1) below is assigned by Fips the syntactic structure in (2), in which the canonical position of object for the verb *address* is taken by the empty constituent *e*. The latter stands for the trace of the noun *issue*, which has been extraposed through relativization. The trace *e*, the relative pronoun ϕ (a zero-pronoun), and the noun *issue* are all linked via the index *i*.

(1) *This too is an issue the Convention must address*.

(2) $[_{TP}$ $[_{DP}$ This] $[_{VP}$ $[_{AdvP}$ too] is $[_{DP}$ an $[_{NP}$ issue$_i$ $[_{CP}[_{DP}$ $\phi_i]$ $[_{TP}$ $[_{DP}$ the $[_{NP}$ Convention]] must $[_{VP}$ address $[_{DP}$ $e_i]]]]]]]]$

[5] The lexical categories are N, A, V, P, Adv, C, Inter, to which we add the two functional categories T (tense) and F (functional).

4.2 Extraction method

Collocation candidates are identified in the parsed text as the analysis goes on. Each (partial or complete) structure returned for a sentence is checked for potential collocational pairs, by recursively examining the pairs consisting of the phrase head X and an element of one of its left or right subconstituents.

For instance, one of the potential collocations identified in the structure shown in Example (2) is the verb–object pair *address-issue*. It is detected in the VP substructure having *address* as a head and e_i as a right constituent ([$_{VP}$ address [$_{DP}$ e_i]]). This pair is retrieved through a sequence of operations, which includes: recognizing the presence of a relative construction; building its normalized form with the empty constituent e in the object position; and finally, linking e to the relative zero-pronoun ϕ and then to the antecedent *issue*. All this computation is done by the parser beforehand. The extraction system recovers the lexical object directly from the argument table of the verb built by Fips.

This first extraction step ensures the existence of a syntactic relationship between the items of a candidate pair. Our approach adopts a syntactic view on collocations, which are seen first of all as "syntagmatic combinations of lexical items" (Fontenelle 1992, p. 222). Therefore, a strong syntactic filter is applied on candidate pairs, based on the syntactic proximity of words (other approaches, instead, simply focus on their linear proximity).

The main strength of our extractor lies in the parser's ability to deal with complex cases of extraposition, such as those highlighted in the constructions below:

passivization: I see that *amendments* to the report by Mr Méndez de Vigo and Mr Leinen have been *tabled* on this subject.
relativization: The communication devotes no attention to the *impact* the newly announced policy measures will *have* on the candidate countries.
interrogation: What *impact* do you expect this to *have* on reducing our deficit and our level of imports?
cleft constructions: It is a very pressing *issue* that Mr Sacrédeus is *addressing*.
enumeration: It is to be welcomed that the Culture 2000 programme has allocated one third of its budget to *cultural*, archaeological, underwater and architectural *heritage* and to museums, libraries and archives, thereby strengthening national action.
coordinated clauses: The *problem* is therefore, clearly a deeply rooted one and cannot be *solved* without concerted action by all parties.
subordinate clauses: The *situation* in the regions where there have been outbreaks of foot-and-mouth disease is *critical*.
parenthesized clauses: Could it be on account of the regulatory *role* which this tax (which applies to international financial transactions) could *play* in relation to currencies, by damping down speculation and reducing the volatility of exchange markets?
apposition: I should like to emphasise that the broad economic policy *guidelines*, the aims of our economic policy, do not *apply* to the euro zone alone but to the entire single European market [...]

Such cases are generally not dealt with by extractors based on shallow-parsing, while window-based approaches simply ignore them.

A more specific morphosyntactic filter is subsequently applied on the selected pairs, so that only the pairs satisfying certain constraints are retained as valid collocation candidates. These constraints may refer both to the lexical items individually, and to the combination as a whole. For instance, proper nouns and auxiliary verbs are ruled out, and combinations are considered valid only if in configurations like the following: N–A: *effort devoted*; A–N: *dramatic event*; N–N: *suicide attack*; N (subject)–V: *river flows*; V–N (object): *face difficulty*; V–P: *point out*; V–P–N (argument or adjunct): *bring to end*; N–P–N: *freedom of expression*; V–A: *steer clear*; V–Adv: *fully support*; Adv–A: *completely different*; A–P: *concerned about*; A&A: *nice and warm*; N&N: *part and parcel*.

The configuration list is actually longer, and it is growing as more and more collocational evidence is considered. It has been used for all the languages mentioned, for which it proved sufficiently appropriate, although, as suggested in Sect. 3, some language-specific amendments might be possible. The full customization of the method for each extraction language also requires finding the best AM for each configuration, an endeavor that falls outside the scope of the present work. Currently, the same AM—LLR (Dunning 1993)—is applied on candidate pairs, after partitioning them into syntactically-homogeneous classes as suggested in Evert and Krenn (2001).

It is worth noting that each lexical item may in turn be a complex lexeme (e.g., a compound or a collocation), like *death penalty* in *abolish the death penalty*; such a lexeme can be recognized by the parser as a single lexical item as long as it is part of its lexicon.

5 Results and evaluation

Previous extraction experiments performed with our system dealt exclusively with French and English data, e.g., (Goldman et al. 2001; Seretan et al. 2004). Here, we report on extraction from a rather large parallel corpus in 4 languages, including Spanish and Italian which are now supported by our system. The corpus is a subset of Europarl parallel corpus of European Parliament proceedings (Koehn 2005). It contains 62 files per language, corresponding to the complete 2001 proceedings.

The whole source corpus totalling about 15 million words was successfully parsed, thanks to Fips robustness. The processing speed is on average 150–200 tokens/s. More statistics about the corpus and the results obtained with our extractor described in the preceding section are presented in Table 1 (rows 1–5). Table 2 displays the top-scored collocation candidates extracted from the Spanish and Italian corpora.

An evaluation experiment has been carried out that compares our extraction method against the mobile-window method, a standard extraction procedure that is based on linear word proximity and ignores the syntactic structure of text. Although a syntactic approach is in theory better, this must be proven empirically in an actual extraction setting, because the inherent parsing errors could lead to more extraction noise (i.e., ungrammatical results) than the window method.

Table 1 Extraction statistics (corpora size and number of pairs extracted)

Statistics	EN	ES	FR	IT	Unit
Size	21.4	22.9	23.7	22.7	MB
Words	3.7	3.8	3.9	3.5	M
Sentences	161.9	172.1	162.7	160.9	K
Pairs—syntactic method	851.5	901.2	988.9	880.6	K
Distinct pairs	333.4	315.5	327.4	333.8	K
Pairs—window method	3055.3	3204.9	3131.3	3463.8	K
Distinct pairs	1445.7	1359.6	1426.9	1366.0	K

Table 2 Top 10 results obtained for Spanish and Italian, showing the LLR score and the annotation provided by the two human judges

Spanish			Italian		
Key1 + key2	Annot	Score	Key1 + key2	Annot	Score
Medio ambiente	4–4	12250.7	Unione europeo	2–2	29489.5
Parlamento europeo	2–4	12118.1	Parlamento europeo	2–2	10138.5
Derecho humano	4–4	8366.0	Unire stato	2–2	6798.6
Tener en cuenta	3–3	7658.3	Candidare paese	1–1	6444.4
Punto de vista	4–3	6394.8	Diritto umano	4–4	5050.1
Primero lugar	4–1	5481.1	Punto di vista	4–4	4930.6
Millón de euro	1–1	5181.5	Ordine recare	3–1	4890.0
Llevar a cabo	3–3	4480.1	Paese terzo	4–4	4358.5
Votar a favor	3–3	4414.9	Unire nazione	2–2	4190.1
Desempeñar papel	3–3	4138.6	Lavoro svolgere	0–3	4103.1

Another motivation for this comparison is the fact that the accuracy of the window method intuitively increases among the top results, as more and more data is processed. If this accuracy is comparable to that of the syntax-based method, then there is no need for parsing provided that one is only interested in the upper part of the significance list (i.e., in the pairs having the score higher than a given threshold). Moreover, adding more data also compensates for the long-distance pairs missed with the habitual 5-word span; thus, again, parsing might not be really necessary for capturing these pairs.[6]

The window method was implemented as follows. The same source corpora were lemmatized and POS-tagged with the Fips parser. Function words were filtered out, and oriented pairs were extracted inside a 5 content-word window, by taking care not to cross a punctuation mark. These pairs were further filtered according to their POS, so that only combinations suggesting a syntactic link were eventually retained: A–N,

[6] In this case, however, the instances missed for candidate pairs alter the frequency profile of these pairs (the values in the contingency table), on which their ranking in the significance list and, ultimately, the quality of results depend.

N–A, N–N, N–V, and V–N. Finally, LLR was applied on each combination type separately, just as in the case of our method (Sect. 4.2). The number of candidate pairs extracted is reported in the last two rows of Table 1. Note that the window method implemented as above represents a rather high baseline for comparison, since all the design choices made translate into increased precision.

Our evaluation study compared the accuracy of the two methods at different levels of the significance lists: top (0%), 1, 3, 5 and 10%.[7] A test set of 50 contiguous output pairs was extracted at each level considered, for each method and each language; the overall test set comprises 2,000 output pairs. Each pair has been annotated by 2 human judges using the following categories and (briefly-stated) criteria:

0. ungrammatical pair: parsing error or, for the window method, unrelated words (e.g., *gross domestic* extracted from *We have a budget surplus of nearly 5% of our gross domestic product.)*;
1. regular combination: not worth storing it in a dictionary (e.g., *next item*);
2. named entity, or part of it: proper noun (e.g., *European Commission*);
3. collocation, or part of it: meaning of headword is preserved; the headword typically combines with this word (e.g., *play role*);
4. compound, or part of it: acts like a single word, inseparable (e.g., *great deal*);
5. idiom, or part of it: opaque meaning; meaning of headword is not preserved (e.g., *hit nail* extracted from *hit the nail on the head*).

The annotators were supported in their task by a concordance tool that shows the context of all instances of extracted pairs in the source corpus (Seretan et al. 2004). Inconsistent annotations for a same annotator were identified and solved, and inter-annotator agreement statistics have been computed for each set. The reference sets contain those pairs that were identically annotated by both annotators (1,437 pairs overall).

Table 3 reports the accuracy obtained for the test sets, for each level and each method. Rows 1 and 2 for each language display the collocational accuracy, i.e., the percentage of collocations in the test sets. Rows 3 and 4 show the MWE accuracy, i.e., the percentage of MWEs: since collocations are notoriously difficult to distinguish from other types of multi-word expressions (McKeown and Radev 2000), we collapsed the last four categories into a single one, MWE. Rows 5 and 6 report the grammatical precision, and rows 7–10 display the agreement statistics, namely the raw agreement (the percentage of pairs on which both annotators agree) and the *k*-score (Cohen 1960).[8]

Consistent results are obtained across languages: the method based on parsing outperforms the mobile-window method by a considerable extent, on almost all of the test sets considered. The highest difference can be observed for grammatical precision: on average, when all languages are considered, it varies from 20.5%

[7] These percentages are not as small as they might seem, since the data processed is fairly large and no frequency threshold was applied on the candidate pairs.

[8] The kappa values indicate different degrees of agreement, as follows: 0 to 0.2—*slight*; 0.2 to 0.4—*fair*; 0.4 to 0.6—*moderate*; 0.6 to 0.8—*substantial*; 0.8 to 0.99—*almost perfect*, and 1—*perfect*. The scores we obtained are higher than expected, given the difficulty of the task.

Table 3 Comparative evaluation results at several levels of the significance list

Level	0%	1%	3%	5%	10%	0%	1%	3%	5%	10%
	English					Spanish				
Colloc.	41.9	69.7	58.3	31.4	16.1	39.3	31.3	42.3	32.1	16.0
	31.8	11.1	7.0	10.0	4.9	36.4	7.1	10.8	12.5	16.7
MWE	67.4	75.8	66.7	31.4	25.8	71.4	40.6	46.2	35.7	16.0
	47.7	15.6	7.0	12.5	4.9	54.5	7.1	10.8	12.5	16.7
Gram.	97.7	97.0	100.0	88.6	71.0	100.0	96.9	92.3	92.9	84.0
	86.4	35.6	32.6	25.0	36.6	72.7	9.5	13.5	15.0	27.8
Agr.	86.0	66.0	48.0	70.0	62.0	56.0	64.0	52.0	56.0	50.0
	88.0	90.0	86.0	80.0	82.0	66.0	84.0	74.0	80.0	72.0
K	73.4	57.1	20.0	49.6	67.5	43.0	57.4	18.8	52.3	14.5
	85.5	93.9	85.1	86.6	60.5	67.9	72.7	66.2	77.2	64.8
	French					Italian				
Colloc.	45.9	41.9	35.5	22.2	5.7	32.4	28.2	37.1	29.7	5.6
	34.3	10.3	10.3	11.9	2.9	22.9	4.9	2.6	2.4	12.8
MWE	67.6	45.2	38.7	25.9	5.7	78.4	38.5	37.1	29.7	13.9
	54.3	10.3	10.3	11.9	2.9	51.4	4.9	2.6	2.4	15.4
Gram.	100.0	93.5	83.9	100.0	65.7	94.6	87.2	94.3	67.6	75.0
	74.3	17.9	20.5	33.3	28.6	77.1	17.1	10.3	11.9	28.2
Agr.	74.0	62.0	62.0	54.0	70.0	74.0	78.0	70.0	74.0	72.0
	70.0	78.0	78.0	84.0	70.0	70.0	82.0	78.0	84.0	78.0
K	68.7	41.3	45.3	20.2	49.2	60.7	74.4	62.2	63.1	67.1
	73.4	70.2	62.5	90.0	62.1	82.7	77.3	45.2	52.2	79.6

Colloc.—collocational precision, *MWE*—MWE precision, *Gram.*—grammatical precision, *Agr.*—raw inter-annotator agreement, *K*—*k*-score

Odd rows correspond to the syntax-based method, and even rows to the window method

(for the first level) to 73.6% (for the second). The difference in MWE precision varies between 19.2 and 40.6% on the first 4 levels, and it is only 5.4% on the last one; that in collocational precision—between 8.5 and 35.6% on the first 4 levels, and is only 1.5% on the last.

A similar pattern can be observed for all the precision parameters considered. On the first level, the improvement obtained with parsing is moderate, since the top window results are also sufficiently accurate. On the next three levels, the window method performs very poorly, whereas the performance of the syntax-based method remains relatively stable. Then on the last level, at 10% of the significance list, the precision of the window method tends to rise, sometimes exceeding that of the syntax-based method, except for grammaticality. This might suggest that a bigger ratio or true positives are demoted to lower positions by the window method. On the contrary, an ideal extraction system should promote true positives to the top, while leaving only a few of them on the lower levels.

On the whole test set (when all languages and all significance levels are considered together), the syntax-based method outperforms the window method by 55.6% in terms of grammatical precision (88.8% vs. 33.2%), by 26.1% in terms of MWE precision (43.2% vs. 17.2%) and by 20.1% in terms of collocational precision (32.9% vs. 12.8%).

We believe that this positive result is particularly important from the perspective of further processing of extraction output. Moreover, the high ratio of collocations found among MWEs confirms the magnitude of the phenomenon considered: from the 416 pairs annotated as MWEs by both judges, 75.7% are collocations, 15.4% compounds, 6.3% named entities, and the remaining 2.6% idioms.

6 Conclusion

Collocation is a pervasive language phenomenon of key importance for NLP applications concerned with text production (machine translation, natural language generation), and that has a large applicability to language analysis tasks as well (e.g., parsing, word sense disambiguation).

Our language-oriented review of the considerable amount of work devoted over the last few decades to collocation extraction revealed a growing concern for the morphosyntactic preprocessing of source corpora. The review also showed that in a multilingual context, the syntactic analysis emerges as an inescapable requirement for extraction, without which acceptable results cannot be achieved (Breidt 1993; Kim et al. 1999). A number of the surveyed approaches use, as in our case, the syntactic proximity instead of the linear proximity of words as the main criterion for identifying collocation candidates, e.g., (Church et al. 1989; Basili et al. 1994; Lin 1998; Pearce 2001; Tutin 2004; Kilgarriff et al. 2004). As far as we know, our system (Goldman et al. 2001; Seretan et al. 2004) is the first to rely on full parsing; other similar approaches are based on chunking or on dependency parsing.

As we expect future collocation extraction (and lexical acquisition in general) to increasingly take advantage of syntactic analysis, we consider multilinguality a true concern for these tasks. We identified in Sect. 3 the major issues to be dealt with in order to successfully implement a collocation extractor for a new language.

Our system (described in Sect. 4.2) was applied on a large collection of texts in 4 languages: English, French, Spanish, and Italian. Its performance in terms of grammatical, collocational, and MWE accuracy was compared, for all these languages, to that of the standard mobile-window method, by performing measurements at different levels of the significance lists. The results obtained are in line with those reported by other evaluation studies: even if imperfect, parsing improves extraction considerably (Smadja 1993; Zajac et al. 2003; Seretan and Wehrli 2006). A smaller improvement was instead observed for German A–N collocations (Evert and Kermes 2003), which might seem reasonable given the particularly rigid pattern studied. As far as flexible configurations involving verbs are also concerned, in a previous evaluation experiment on French data we obtained a drastic reduction of noise, as well as a higher MWE precision w.r.t. the window method for the top part of the significance list (the first 500 pairs). Our present study

is extended to the 4 languages currently supported by our extractor, covers different levels of the significance list, and uses a finer classification granularity. Besides, it deals with 3 or 4 times as much data. The results confirmed that parsing leads to a substantial increase in the accuracy of results, of 55.6% for the grammatical precision, 26.1% for the MWE precision, and 20.1% for the collocational precision.

Future work is oriented towards the evaluation of extraction recall and the comparison with shallow-parsing approaches. We conducted a preliminary study on word sketches produced with shallow parsing by the Sketch Engine (Kilgarriff et al. 2004). Its results, although not entirely conclusive because of the small size of data evaluated, suggest that chunking leaves some room for improvement[9], and we believe that this improvement can be achieved with full parsing.

Acknowledgements This work was supported in part by Swiss National Science Foundation grant no. 101412-103999. We wish to thank Jorge Antonio Leoni de León, Yves Scherrer and Vincenzo Pallotta for participating in the annotation task, as well as Stephanie Durrleman-Tame for proofreading the article. We are very grateful to the anonymous reviewers, whose comments and suggestions helped us to improve this paper.

References

Barnbrook, G. (1996). *Language and computers: A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Basili, R., Pazienza, M. T., & Velardi, P. (1994) A "not-so-shallow" parser for collocational analysis. In *Proceedings of the 15th Conference on Computational Linguistics* (pp. 447–453). Association for Computational Linguistics: Kyoto, Japan.

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 15th International Conference on Computational Linguistics* (pp. 977–981). Nantes, France.

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, USA.

Calzolari, N., & Bindi, R. (1990). Acquisition of lexical information from a large textual Italian corpus. In *Proceedings of the 13th International Conference on Computational Linguistics* (pp. 54–59). Helsinki, Finland.

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* (pp. 609–623). Cambridge, USA.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1989). Parsing, word associations and typical predicate-argument relations. In *Proceedings of the International Workshop on Parsing Technologies* (pp. 103–112). Carnegie Mellon University: Pittsburgh.

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* (pp. 76–83). Vancouver, B.C.: Association for Computational Linguistics.

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

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

[9] The study considered the top 100 subject–verb pairs extracted with the Sketch Engine from the BNC for the noun *preference*, without a frequency cutoff. We found that as many as 23.8% of the 63 corresponding pair types were derived from ungrammatical instances, e.g., *preference-result*: "to give effect to the *preference* would *result* in ...", or *preference-lead*: "the existence of these *preferences* would clearly *lead* ...".

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., & Kermes, H. (2003). Experiments on candidate data for collocation extraction. In *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics* (pp. 83–86). Budapest, Hungary.

Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 188–195). Toulouse, France.

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

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*, pp. 221–228.

Goldman, J.-P., Nerima, L., & Wehrli, E. (2001). Collocation extraction using a syntactic parser. In *Proceedings of the ACL Workshop on Collocations* (pp. 61–66). Toulouse, France.

Gross, M. (1984). Lexicon-grammar and the syntactic analysis of French. In *Proceedings of the 22nd conference on Association for Computational Linguistics* (pp. 275–282). Morristown, NJ, USA.

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* (pp. 48–55). Jeju Island, Republic of Korea.

Ikehara, S., Shirai, S., & Kawaoka, T. (1995). Automatic extraction of uninterrupted collocations by n-gram statistics. In *Proceedings of First Annual Meeting of the Association for Natural Language Processing*, pp. 313–316.

Jacquemin, C., Klavans, J. L., & Tzoukermann, E. (1997). 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* (pp. 24–31). Association for Computational Linguistics: Morristown, NJ, USA.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1*(1), 9–27.

Kilgarriff, A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition* (pp. 33–40). Sussex, UK.

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

Kim, S., Yang, Z., Song, M., & Ahn, J.-H. (1999). Retrieving collocations from Korean text. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 71–81). Maryland, USA.

Kjellmer, G. (1994). *A dictionary of English collocations*. Oxford: Claredon Press.

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

Krenn, B. (2000). *The usual suspects: Data-oriented models for identification and representation of lexical collocations*, Vol. 7. Saarbrücken, Germany: German Research Center for Artificial Intelligence and Saarland University Dissertations in Computational Linguistics and Language Technology.

Krenn, B., & Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations* (pp. 39–46). Toulouse, France.

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

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

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* (pp. 317–324). Association for Computational Linguistics: Morristown, NJ, USA.

Lu, Q., Li, Y., & Xu, R. (2004). Improving Xtract for Chinese collocation extraction. In: *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp. 333–338.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

McKeown, K. R., & Radev, D. R. (2000). Collocations. In R. Dale, H. Moisl, & H. Somers (Eds.), *A Handbook of natural language processing* (pp. 507–523). New York, USA: Marcel Dekker.

Mel'čuk, I. (1998). Collocations and lexical functions. In A. P. Cowie (Eds.), *Phraseology. Theory, analysis, and applications* (pp. 23–53). Oxford: Claredon Press.

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".

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

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

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)* (pp. 1–15). Mexico City.

Seretan, V., Nerima, L., & Wehrli, E. (2004). A tool for multi-word collocation extraction and visualization in multilingual corpora. In *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004* (pp. 755–766). Lorient, France.

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* (pp. 953–960). Sydney, Australia.

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* (pp. 476–481). Madrid, Spain.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Paris: Masson.

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

Tutin, A. (2004). Pour une modélisation dynamique des collocations dans les textes. In *Proceedings of the Eleventh EURALEX International Congress* (pp. 207–219). Lorient, France.

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

Wehrli, E. (2007). Fips, A "deep" linguistic multilingual parser. In *ACL 2007 Workshop on Deep Linguistic Processing*. Prague, Czech Republic (pp. 120–127). Association for Computational Linguistics.

Wermter, J., & Hahn, U. (2004). Collocation extraction based on modifiability statistics. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)* (pp. 980–986). Geneva, Switzerland.

Zajac, R., Lange, E., & Yang, J. (2003). Customizing complex lexical entries for high-quality MT. In *Proceedings of the Ninth Machine Translation Summit* (pp. 433–438). New Orleans, USA.

Zinsmeister, H., & Heid, U. (2003). Significant triples: Adjective+Noun+Verb combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research (Complex 2003)*, Budapest.