Accurate Collocation Extraction Using a Multilingual Parser

Violeta Seretan
Language Technology Laboratory
University of Geneva
2, rue de Candolle, 1211 Geneva
Violeta.Seretan@latl.unige.ch

Eric Wehrli
Language Technology Laboratory
University of Geneva
2, rue de Candolle, 1211 Geneva
Eric.Wehrli@latl.unige.ch

Abstract

This paper focuses on the use of advanced techniques of text analysis as support for collocation extraction. A hybrid system is presented that combines statistical methods and multilingual parsing for detecting accurate collocational information from English, French, Spanish and Italian corpora. The advantage of relying on full parsing over using a traditional window method (which ignores the syntactic information) is first theoretically motivated, then empirically validated by a comparative evaluation experiment.

1 Introduction

Recent computational linguistics research fully acknowledged the stringent need for a systematic and appropriate treatment of phraseological units in natural language processing applications (Sag et al., 2002). Syntagmatic relations between words — also called *multi-word expressions*, or "idiosyncratic interpretations that cross word boundaries" (Sag et al., 2002, 2) — constitute an important part of the lexicon of a language: according to Jackendoff (1997), they are at least as numerous as the single words, while according to Mel'čuk (1998) they outnumber single words ten to one.

Phraseological units include a wide range of phenomena, among which we mention compound nouns (*dead end*), phrasal verbs (*ask out*), idioms (*lend somebody a hand*), and collocations (*fierce battle*, *daunting task*, *schedule a meeting*). They pose important problems for NLP applications, both text analysis and text production perspectives being concerned.

In particular, collocations[1] are highly problematic, for at least two reasons: first, because their linguistic status and properties are unclear (as pointed out by McKeown and Radev (2000), their definition is rather vague, and the distinction from other types of expressions is not clearly drawn); second, because they are prevalent in language. Mel'čuk (1998, 24) claims that "collocations make up the lions share of the phraseme inventory", and a recent study referred in (Pearce, 2001) showed that each sentence is likely to contain at least one collocation.

Collocational information is not only useful, but also indispensable in many applications. In machine translation, for instance, it is considered "the key to producing more acceptable output" (Orliac and Dillinger, 2003, 292).

This article presents a system that extracts accurate collocational information from corpora by using a syntactic parser that supports several languages. After describing the underlying methodology (section 2), we report several extraction results for English, French, Spanish and Italian (section 3). Then we present in sections 4 and 5 a comparative evaluation experiment proving that a hybrid approach leads to more accurate results than a classical approach in which syntactic information is not taken into account.

2 Hybrid Collocation Extraction

We consider that syntactic analysis of source corpora is an inescapable precondition for collocation extraction, and that the syntactic structure of source text has to be taken into account in order to ensure the quality and interpretability of results.

[1] To put it simply, collocations are non-idiomatical, but restricted, conventional lexical combinations.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 953–960,
Sydney, July 2006. ©2006 Association for Computational Linguistics

As a matter of fact, some of the existing collocation extraction systems already employ (but only to a limited extent) linguistic tools in order to support the collocation identification in text corpora. For instance, lemmatizers are often used for recognizing all the inflected forms of a lexical item, and POS taggers are used for ruling out certain categories of words, e.g., in (Justeson and Katz, 1995).

Syntactic analysis has long since been recognized as a prerequisite for collocation extraction (for instance, by Smadja[2]), but the traditional systems simply ignored it because of the lack, at that time, of efficient and robust parsers required for processing large corpora. Oddly enough, this situation is nowadays perpetuated, in spite of the dramatic advances in parsing technology. Only a few exceptions exists, e.g., (Lin, 1998; Krenn and Evert, 2001).

One possible reason for this might be the way that collocations are generally understood, as a purely statistical phenomenon. Some of the best-known definitions are the following: "Collocations of a given word are statements of the habitual and customary places of that word" (Firth, 1957, 181); "arbitrary and recurrent word combination" (Benson, 1990); or "sequences of lexical items that habitually co-occur" (Cruse, 1986, 40). Most of the authors make no claims with respect to the grammatical status of the collocation, although this can indirectly inferred from the examples they provide.

On the contrary, other definitions state explicitly that a collocation is an expression of language: "co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern" (Cowie, 1978); "a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit" (Choueka, 1988). Our approach is committed to these later definitions, hence the importance we lend to using appropriate extraction methodologies, based on syntactic analysis.

The hybrid method we developed relies on the parser Fips (Wehrli, 2004), that implements the Government and Binding formalism and supports several languages (besides the ones mentioned in the abstract, a few other are also partly dealt with). We will not present details about the parser here; what is relevant for this paper is the type of syntactic structures it uses. Each constituent is represented by a simplified X-bar structure (without intermediate level), in which to the lexical head is attached a list of left constituents (its specifiers) and right constituents (its complements), and each of these are in turn represented by the same type of structure, recursively.

Generally speaking, a collocation extraction can be seen as a two-stage process:

I. in stage one, collocation candidates are identified from the text corpora, based on criteria which are specific to each system;

II. in stage two, the candidates are scored and ranked using specific *association measures* (a review can be found in (Manning and Schütze, 1999; Evert, 2004; Pecina, 2005)).

According to this description, in our approach the parser is used in the first stage of extraction, for identifying the collocation candidates. A pair of lexical items is selected as a candidate only if there is a syntactic relation holding between the two items (one being the head of the current parse structure, and the other the lexical head of its specifier/complement). Therefore, the criterion we employ for candidate selection is the syntactic proximity, as opposed to the linear proximity used by traditional, window-based methods.

As the parsing goes on, the syntactic word pairs are extracted from the parse structures created, from each head-specifier or head-complement relation. The pairs obtained are then partitioned according to their syntactic configuration (e.g., noun + adjectival or nominal specifier, noun + argument, noun + adjective in predications, verb + adverbial specifier, verb + argument (subject, object), verb + adjunt, etc). Finally, the log-likelihood ratios test (henceforth LLR) (Dunning, 1993) is applied on each set of pairs. We call this method hybrid, since it combines syntactic and statistical information (about word and co-occurrence frequency).

The following examples — which, like all the examples in this paper, are actual extraction results — demonstrate the potential of our system to detect collocation candidates, even if subject to complex syntactic transformations.

[2] "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. However, in practice, free-style texts contain a great deal of nonstandard features over which automatic parsers would fail. This fact is being seriously challenged by current research (...), and might not be true in the near future" (Smadja, 1993, 151).

1.a) *raise question*: The *question* of political leadership has been *raised* several times by previous speakers.

1.b) *play role*: What *role* can Canada's immigration program *play* in helping developing nations... ?

1.c) *make mistake*: We could look back and probably see a lot of *mistakes* that all parties including Canada perhaps may have *made*.

3 Multilingual Extraction Results

In this section, we present several extraction results obtained with the system presented in section 2. The experiments were performed on data in the four languages, and involved the following corpora: for English and French, a subpart or the Hansard Corpus of proceedings from the Canadian Parliament; for Italian, documents from the Swiss Parliament; and for Spanish, a news corpus distributed by the Linguistic Data Consortium.

Some statistics on these corpora, some processing details and quantitative results are provided in Table 1. The first row lists the corpora size (in tokens); the next three rows show some parsing statistics[3], and the last rows display the number of collocation candidates extracted and of candidates for which the LLR score could be computed[4].

Statistics	English	French	Spanish	Italian
tokens	3509704	1649914	1023249	287804
sentences	197401	70342	67502	12008
compl. parse	139498	50458	13245	4511
avg. length	17.78	23.46	15.16	23.97
pairs	725025	370932	162802	58258
(extracted)	276670	147293	56717	37914
pairs	633345	308410	128679	47771
(scored)	251046	131384	49495	30586

Table 1: Extraction statistics

In Table 2 we list the top collocations (of length two) extracted for each language. We do not specifically discuss here multilingual issues in collocation extraction; these are dealt with in a separate paper (Seretan and Wehrli, 2006).

Language	Key1	Key2	LLR score
English	federal	government	7229.69
	reform	party	6530.69
	house	common	6006.84
	minister	finance	5829.05
	acting	speaker	5551.09
	red	book	5292.63
	create	job	4131.55
	right	Hon	4117.52
	official	opposition	3640.00
	deputy	speaker	3549.09
French	premier	ministre	4317.57
	bloc	québécois	3946.08
	discours	trône	3894.04
	vérificateur	général	3796.68
	parti	réformiste	3615.04
	gouvernement	fédéral	3461.88
	missile	croisière	3147.42
	Chambre	commune	3083.02
	livre	rouge	2536.94
	secrétaire	parlementaire	2524.68
Spanish	banco	central	4210.48
	millón	dólar	3312.68
	millón	peso	2335.00
	libre	comercio	2169.02
	nuevo	peso	1322.06
	tasa	interés	1179.62
	deuda	externo	1119.91
	cámara	representante	1015.07
	asamblea	ordinario	992.85
	papel	comercial	963.95
Italian	consiglio	federale	3513.19
	scrivere	consiglio	594.54
	unione	europeo	479.73
	servizio	pubblico	452.92
	milione	franco	447.63
	formazione	continuo	388.80
	iniziativa	popolare	383.68
	testo	interpellanza	377.46
	punto	vista	373.24
	scrivere	risposta	348.77

Table 2: Top ten collocations extracted for each language

The collocation pairs obtained were further processed with a procedure of long collocations extraction described elsewhere (Seretan et al., 2003). Some examples of collocations of length 3, 4 and 5 obtained are: *minister of Canadian heritage*, *house proceed to statement by*, *secretary to leader of gouvernment in house of common* (En), *question adresser à ministre*, *programme de aide à rénovation résidentielle*, *agent employer force susceptible causer* (Fr), *bolsa de comercio local*, *peso en cuota de fondo de inversión*, *permitir uso de papel de deuda esterno* (Sp), *consiglio federale disporre*, *creazione di nuovo posto di lavoro*, *costituire fattore penalizzante per regione* (It)[5].

[3] The low rate of completely parsed sentences for Spanish and Italian are due to the relatively reduced coverage of the parsers of these two languages (under development). However, even if a sentence is not assigned a complete parse tree, some syntactic pairs can still be collected from the partial parses.

[4] The log-likelihood ratios score is undefined for those pairs having a cell of the contingency table equal to 0.

[5] Note that the output of the procedure contains lemmas rather than inflected forms.

4 Comparative Evaluation Hypotheses

4.1 Does Parsing *Really* Help?

Extracting collocations from raw text, without preprocessing the source corpora, offers some clear advantages over linguistically-informed methods such as ours, which is based on the syntactic analysis: *speed* (in contrast, parsing large corpora of texts is expected to be much more time consuming), *robustness* (symbolic parsers are often not robust enough for processing large quantities of data), *portability* (no need to a priori define syntactic configurations for collocations candidates).

On the other hand, these basic systems suffer from the combinatorial explosion if the candidate pairs are chosen from a large search space. To cope with this problem, a candidate pair is usually chosen so that both words are inside a context ('collocational') window of a small length. A 5-word window is the norm, while longer windows prove impractical (Dias, 2003).

It has been argued that a window size of 5 is actually sufficient for capturing most of the collocational relations from texts in English. But there is no evidence sustaining that the same holds for other languages, like German or the Romance ones that exhibit freer word order. Therefore, as window-based systems miss the 'long-distance' pairs, their recall is presumably lower than that of parse-based systems. However, the parser could also miss relevant pairs due to inherent analysis errors.

As for precision, the window systems are susceptible to return more noise, produced by the grammatically unrelated pairs inside the collocational window. By dividing the number of grammatical pairs by the total number of candidates considered, we obtain the overall precision with respect to grammaticality; this result is expected to be considerably worse in the case of basic method than for the parse-based methods, just by virtue of the parsing task. As for the overall precision with respect to collocability, we expect the proportional figures to be preserved. This is because the parser-based methods return less, but better pairs (i.e., only the pairs identified as grammatical), and because collocations are a subset of the grammatical pairs.

Summing up, the evaluation hypothesis that can be stated here is the following: parse-based methods outperform basic methods thanks to a drastic reduction of noise. While unquestionable under the assumption of perfect parsing, this hypothesis has to be empirically validated in an actual setting.

4.2 Is More Data Better Than Better Data?

The hypothesis above refers to the overall precision and recall, that is, relative to the entire list of selected candidates. One might argue that these numbers are less relevant for practice than they are from a theoretical (evaluation) perspective, and that the exact composition of the list of candidates identified is unimportant if only the top results (i.e., those pairs situated above a threshold) are looked at by a lexicographer or an application.

Considering a threshold for the *n*-best candidates works very much in the favor of basic methods. As the amount of data increases, there is a reduction of the noise among the best-scored pairs, which tend to be more grammatical because the likelihood of encountering many similar noisy pairs is lower. However, as the following example shows, noisy pairs may still appear in top, if they occur often in a longer collocation:

2.a) les essais du missile de croisière

2.b) essai - croisière

The pair *essai - croisière* is marked by the basic systems as a collocation because of the recurrent association of the two words in text as part or the longer collocation *essai du missile de croisière*. It is an grammatically unrelated pair, while the correct pairs reflecting the right syntactic attachment are *essai missile* and *missile (de) croisière*.

We mentioned that parsing helps detecting the 'long-distance' pairs that are outside the limits of the collocational window. Retrieving all such complex instances (including all the extraposition cases) certainly augment the recall of extraction systems, but this goal might seem unjustified, because the risk of not having a collocation represented at all diminishes as more and more data is processed. One might think that systematically missing long-distance pairs might be very simply compensated by supplying the system with more data, and thus that larger data is a valid alternative to performing complex processing.

While we agree that the inclusion of more data compensates for the 'difficult' cases, we do consider this truly helpful in deriving collocational information, for the following reasons: (1) more data means more noise for the basic methods; (2) some collocations might systematically appear in a complex grammatical environment (such as passive constructions or with additional material inserted between the two items); (3) more importantly, the complex cases not taken into account alter the frequency profile of the pairs concerned.

These observations entitle us to believe that, even when more data is added, the *n*-best precision might remain lower for the basic methods with respect to the parse-based ones.

4.3 How Real the Counts Are?

Syntactic analysis (including shallower levels of linguistic analysis traditionally used in collocation extraction, such as lemmatization, POS tagging, or chunking) has two main functions.

On the one hand, it guides the extraction system in the candidate selection process, in order to better pinpoint the pairs that might form collocations and to exclude the ones considered as inappropriate (e.g., the pairs combining function words, such as a preposition followed by a determiner).

On the other, parsing supports the association measures that will be applied on the selected candidates, by providing more exact frequency information on words — the inflected forms count as instances of the same lexical item — and on their co-occurrence frequency — certain pairs might count as instance of the same pair, others do not.

In the following example, the pair *loi modifier* is an instance of a subject-verb collocation in 3.a), and of a verb-object collocation type in 3.b). Basic methods are unable to distinguish between the two types, and therefore count them as equivalent.

3.a) *Loi modifiant* la Loi sur la responsabilité civile

3.b) la *loi* devrait être *modifiée*

Parsing helps to create a more realistic frequency profile for the candidate pairs, not only because of the grammaticality constraint it applies on the pairs (wrong pairs are excluded), but also because it can detect the long-distance pairs that are outside the collocational window.

Given that the association measures rely heavily on the frequency information, the erroneous counts have a direct influence on the ranking of candidates and, consequently, on the top candidates returned. We believe that in order to achieve a good performance, extraction systems should be as close as possible to the real frequency counts and, of course, to the real syntactic interpretation provided in the source texts[6].

Since parser-based methods rely on more accurate frequency information for words and their co-occurrence than window methods, it follows that the *n*-best list obtained with the first methods will probably show an increase in quality over the second.

To conclude this section, we enumerate the hypotheses that have been formulated so far: (1) Parse methods provide a noise-freer list of collocation candidates, in comparison with the window methods; (2) Local precision (of best-scored results) with respect to grammaticality is higher for parse methods, since in basic methods some noise still persists, even if more data is included; (3) Local precision with respect to collocability is higher for parse methods, because they use a more realistic image of word co-occurrence frequency.

5 Comparative Evaluation

We compare our hybrid method (based on syntactic processing of texts) against the window method classically used in collocation extraction, from the point of view of their precision with respect to grammaticality and collocability.

5.1 The Method

The *n*-best extraction results, for a given *n* (in our experiment, *n* varies from 50 to 500 at intervals of 50) are checked in each case for grammatical well-formedness and for lexicalization. By lexicalization we mean the quality of a pair to constitute (part of) a multi-word expression — be it compound, collocation, idiom or another type of syntagmatic lexical combination. We avoid giving collocability judgments since the classification of multi-word expressions cannot be made precisely and with objective criteria (McKeown and Radev, 2000). We rather distinguish between lexicalizable and trivial combinations (completely regular productions, such as *big house*, *buy bread*, that do not deserve a place in the lexicon). As in (Choueka, 1988) and (Evert, 2004), we consider that a dominant feature of collocations is that they are unpredictable for speakers and therefore have to be stored into a lexicon.

[6] To exemplify this point: the pair *développement humain* (which has been detected as a collocation by the basic method) looks like a valid expression, but the source text consistently offers a different interpretation: *développement des ressources humaines*.

Each collocation from the n-best list at the different levels considered is therefore annotated with one of the three flags: 1. ungrammatical; 2. trivial combination; 3. multi-word expression (MWE).

On the one side, we evaluate the results of our hybrid, parse-based method; on the other, we simulate a window method, by performing the following steps: POS-tag the source texts; filter the lexical items and retain only the open-class POS; consider all their combinations within a collocational window of length 5; and, finally, apply the log-likelihood ratios test on the pairs of each configuration type.

In accordance with (Evert and Kermes, 2003), we consider that the comparative evaluation of collocation extraction systems should not be done at the end of the extraction process, but separately for each stage: after the candidate selection stage, for evaluating the quality (in terms of grammaticality) of candidates proposed; and after the application of collocability measures, for evaluating the measures applied. In each of these cases, different evaluation methodologies and resources are required. In our case, since we used the same measure for the second stage (the log-likelihood ratios test), we could still compare the final output of basic and parse-based methods, as given by the combination of the first stage with the same collocability measure.

Again, similarly to Krenn and Evert (2001), we believe that the homogeneity of data is important for the collocability measures. We therefore applied the LLR test on our data after first partitioning it into separate sets, according to the syntactical relation holding in each candidate pair. As the data used in the basic method contains no syntactic information, the partitioning was done based on POS-combination type.

5.2 The Data

The evaluation experiment was performed on the whole French corpus used in the extraction experiment (section 2), that is, a subpart of the Hansard corpus of Canadian Parliament proceedings. It contains 112 text files totalling 8.43 MB, with an average of 628.1 sentences/file and 23.46 tokens/sentence (as detected by the parser). The total number of tokens is $1,649,914$.

On the one hand, the texts were parsed and $370,932$ candidate pairs were extracted using the hybrid method we presented. Among the pairs extracted, 11.86% ($44,002$ pairs) were multi-word expressions identified at parse-time, since present in the parser's lexicon. The log-likelihood ratios test was applied on the rest of pairs. A score could be associated to $308,410$ of these pairs (corresponding to $131,384$ types); for the others, the score was undefined.

On the other hand, the texts were POS-tagged using the same parser as in the first case. If in the first case the candidate pairs were extracted during the parsing, in the second they were generated after the open-class filtering. From $673,789$ POS-filtered tokens, a number of $1,024,888$ combinations ($560,073$ types) were created using the 5-length window criterion, while taking care not to cross a punctuation mark. A score could be associated to $1,018,773$ token pairs ($554,202$ types), which means that the candidate list is considerably larger than in the first case. The processing time was more than twice longer than in the first case, because of the large amount of data to handle.

5.3 Results

The 500 best-scored collocations retrieved with the two methods were manually checked by three human judges and annotated, as explained in 5.1, as either ungrammatical, trivial or MWE. The agreement statistics on the annotations for each method are shown in Table 3.

Method	**Agr.**	**1,2,3**	**1,2**	**1,3**	**2,3**
parse	observed	285	365	362	340
	k-score	55.4%	62.6%	69%	64%
window	observed	226	339	327	269
	k-score	43.1%	63.8%	61.1%	48%

Table 3: Inter-annotator agreement

For reporting n-best precision results, we used as reference set the annotated pairs on which at least two of the three annotators agreed. That is, from the 500 initial pairs retrieved with each method, 497 pairs were retained in the first case (parse method), and 483 pairs in the second (window method).

Table 4 shows the comparative evaluation results for precision at different levels in the list of best-scored pairs, both with respect to grammaticality and to collocability (or, more exactly, the potential of a pair to constitute a MWE). The numbers show that a drastic reduction of noise is achieved by parsing the texts. The error rate with

	Precision (gram.)		Precision (MWE)	
n	*window*	*parse*	*window*	*parse*
50	94.0	96.0	80.0	72.0
100	91.0	98.0	75.0	74.0
150	87.3	98.7	72.7	73.3
200	85.5	98.5	70.5	74.0
250	82.8	98.8	67.6	69.6
300	82.3	98.7	65.0	69.3
350	80.3	98.9	63.7	67.4
400	80.0	99.0	62.5	67.0
450	79.6	99.1	61.1	66.0
500	78.3	99.0	60.1	66.0

Table 4: Comparative evaluation results

respect to grammaticality is, on average, 15.9% for the window method; with parsing, it drops to 1.5% (i.e., 10.6 times smaller).

This result confirms our hypothesis regarding the local precision which was stated in section 4.2. Despite the inherent parsing errors, the noise reduction is substantial. It is also worth noting that we compared our method against a rather high baseline, as we made a series of choices susceptible to alleviate the candidates identification with the window-based method: we filtered out function words, we used a parser for POS-tagging (that eliminated POS-ambiguity), and we filtered out cross-punctuation pairs.

As for the MWE precision, the window method performs better for the first 100 pairs[7]); on the remaining part, the parsing-based method is on average 3.7% better. The precision curve for the window method shows a more rapid degradation than it does for the other. Therefore we can conclude that parsing is especially advantageous if one investigates more that the first hundred results (as it seems reasonable for large extraction experiments).

In spite of the rough classification we used in annotation, we believe that the comparison performed is nonetheless meaningful since results should be first checked for grammaticality and 'triviality' before defining more difficult tasks such as collocability.

6 Conclusion

In this paper, we provided both theoretical and empirical arguments in the favor of performing syntactic analysis of texts prior to the extraction of collocations with statistical methods.

Part of the extraction work that, like ours, relies on parsing was cited in section 2. Most often, it concerns chunking rather than complete parsing; specific syntactic configurations (such as adjective-noun, preposition-noun-verb); and languages other than the ones we deal with (usually, English and German). Parsing has been also used after extraction (Smadja, 1993) for filtering out invalid results. We believe that this is not enough and that parsing is required prior to the application of statistical tests, for computing a realistic frequency profile for the pairs tested.

As for evaluation, unlike most of the existing work, we are not concerned here with comparing the performance of association measures (cf. (Evert, 2004; Pecina, 2005) for comprehensive references), but with a contrastive evaluation of syntactic-based and standard extraction methods, combined with the same statistical computation.

Our study finally clear the doubts on the usefulness of parsing for collocation extraction. Previous work that quantified the influence of parsing on the quality of results suggested the performance for tagged and parsed texts is similar (Evert and Kermes, 2003). This result applies to a quite rigid syntactic pattern, namely adjective-noun in German. But a preceding study on noun-verb pairs (Breidt, 1993) came to the conclusion that good precision can only be achieved for German with parsing. Its author had to simulate parsing because of the lack, at the time, of parsing tools for German. Our report, that concerns an actual system and a large data set, validates Breidt's finding for a new language (French).

Our experimental results confirm the hypotheses put forth in section 4, and show that parsing (even if imperfect) benefits to extraction, notably by a drastic reduction of the noise in the top of the significance list. In future work, we consider investigating other levels of the significance list, extending the evaluation to other languages, comparing against shallow-parsing methods instead of the window method, and performing recall-based evaluation as well.

Acknowledgements

We would like to thank Jorge Antonio Leoni de Leon, Mar Ndiaye, Vincenzo Pallotta and Yves Scherrer for participating to the annotation task. We are also grateful to Gabrielle Musillo and to the anonymous reviewers of an earlier version of

[7] A closer look at the data revealed that this might be explained by some inconsistencies between annotations.

this paper for useful comments and suggestions.

References

Morton Benson. 1990. Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1):23–35.

Elisabeth Breidt. 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, U.S.A.

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

Anthony P. Cowie. 1978. The place of illustrative material and collocations in the design of a learner's dictionary. In P. Strevens, editor, *In Honour of A.S. Hornby*, pages 127–139. Oxford: Oxford University Press.

D. Alan Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Gaël Dias. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL Workshop on Multiword Expressions*, pages 41–48, Sapporo, Japan.

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

Stefan Evert and Hannah Kermes. 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*, pages 83–86, Budapest, Hungary.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

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

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: Some linguistis properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France.

Dekang Lin. 1998. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, pages 57–63, Montreal.

Christopher Manning and Heinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Mass.

Kathleen R. McKeown and Dragomir R. Radev. 2000. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *A Handbook of Natural Language Processing*, pages 507–523. Marcel Dekker, New York, U.S.A.

Igor Mel'čuk. 1998. Collocations and lexical functions. In Anthony P. Cowie, editor, *Phraseology. Theory, Analysis, and Applications*, pages 23–53. Claredon Press, Oxford.

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

Darren Pearce. 2001. Synonymy in collocation extraction. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)*, pages 41–46, Carnegie Mellon University, Pittsburgh.

Pavel Pecina. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City.

Violeta Seretan and Eric Wehrli. 2006. Multilingual collocation extraction: Issues and solutions solutions. In *Proceedings or COLING/ACL Workshop on Multilingual Language Resources and Interoperability*, Sydney, Australia, July. To appear.

Violeta Seretan, Luka Nerima, and Eric Wehrli. 2003. Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of the Fourth International Conference on Recent Advances in NLP (RANLP-2003)*, pages 424–431, Borovets, Bulgaria.

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

Eric Wehrli. 2004. Un modèle multilingue d'analyse syntaxique. In A. Auchlin et al., editor, *Structures et discours - Mélanges offerts à Eddy Roulet*, pages 311–329. Éditions Nota bene, Québec.