Extraction of multi-word collocations using syntactic bigram composition

Seretan, Violeta; Nerima, Luka; Wehrli, Eric

Extraction of Multi-Word Collocations Using Syntactic Bigram Composition

Violeta Seretan, Luka Nerima, Eric Wehrli
Language Technology Laboratory (LATL)
University of Geneva
2, rue de Candolle
CH-1211 Geneva 4, Switzerland
{Violeta.Seretan, Luka.Nerima, Eric.Wehrli}@lettres.unige.ch

Abstract

This paper presents a method for extracting multi-word collocations (MWCs) from text corpora, which is based on the previous extraction of syntactically bound collocation bigrams. We describe an iterative word linking procedure which relies on a syntactic criterion and aims at building up arbitrarily long expressions that represent multi-word collocation candidates. We propose several measures to rank candidates according to the collocational strength, and we present the results of a trigram extraction experiment. The methodology used is particularly well-suited for the identification of those collocations whose terms are arbitrarily distant, due to syntactic processes (passivization, relativization, dislocation, topicalization).

1 Introduction

Multi-word expressions are recurrent combinations of two or more words (not necessarily contiguous) that form fixed or semi-fixed lexical or syntactic units. The following examples: *"address book"*, *"to kick the bucket"*, *"to experience a problem"* illustrate particular subclasses of multi-word expressions:

- *compound words* - units of lexical category, i.e. behaving like simple words;
- *idioms* (including phrasal verbs) - units of phrasal level completely uncompositional in meaning and exhibiting a certain degree of freedom in terms of syntactic modifiability;
- *collocations* - conventional associations of words in the sense of (Benson 90), i.e. whose co-occurrence happens more often than by chance, and that sound "less natural" if one term is replaced with a near-synonym, for instance.

Collocations are language dependent and can only be learned by observing their occurrence in language use; they are otherwise not predictable. (Harris 51) observed that word usage in a language obeys the "Likelihood" constraint, stating that "each word has a particular and roughly stable likelihood of occurring as argument, or operator, with a given word, though there are many cases of uncertainty, disagreement among speakers, and change through time". Therefore, the correct identification of collocational expressions in text is necessary in both text understanding and generation, not only for NLP tasks (mainly parsing, machine translation, and natural language generation), but also for humans learning a foreign language.

This phenomenon of conventional word associations has been given particular attention since (Firth 57), both by lexicographers, who tried to collect and integrate them into dictionaries (Benson 85; Mel'cuk et al. 99), and by computational linguists who used statistical methods (Choueka *et al.* 83; Sinclair 91; Smadja 93) to automatically retrieve them from texts. While a large amount of work has been dedicated to the treatment of compounds and idioms, notably to the creation of lexical resources (dictionaries of idioms), and to the extraction of specific compounds (e.g. name entities), relatively few collocation resources have been developed so far. The identification of collocations is more difficult (than that of compounds and idioms) since they cannot be defined more precisely than as expressions "that correspond to a conventional way of saying things" and by stating several properties, such as the limited compositionality, susbstituability and modifiability (Manning & Schütze 99).

Current computational methods of collocation extraction are mainly based on pure statistical approaches (which are discussed in detail in section 2), and have two major limitations. The first one is the combinatorial explosion when considering all possibilities for words combination, the methods being constrained to limit their search space to a fixed, low size window of consecutive words. Consequently, there is a reduction in the coverage of these methods, since those collocations whose

terms appear at longer distances in text are ruled out. The other limitation pertains to the grammaticality of results. Very often the expressions retrieved are made up of syntactically unrelated items.

Complete and accurate extraction results crucially influence the subsequent treatment in other NLP applications, such as machine translation, information retrieval, word sense disambiguation. Thus, collocation extraction systems should try to overcome these limitations, by allowing collocations of unrestricted length (flexibly occurring terms), and by ensuring output's grammaticality. This can only be achieved by taking into account the linguistic dimension of texts and by performing a linguistic analysis (e.g. morphological, syntactic).

Over the last few years, there has been a strong increase in the availability of computational resources and software tools dedicated to large-scale and robust syntactic parsing. Our approach proposes to take advantage of systems able to identify syntactically bound term co-occurrences in order to improve the identification of multi-word collocations in text. We make use of such a system, FipsCo (Goldman *et al.* 01), that extracts collocation bigrams from parsed text, and we combine the results in chains of bigrams sharing common terms, as a way to detect multi-word collocations. We consider several statistical tests aimed at validating the extracted expressions.

The paper begins with a brief presentation of some of the existing collocation extraction methods and their main features. Section 3 outlines the method of collocation bigram extraction on which our work relies, pointing out its main distinctive feature, the use of syntactic dependency over simple textual proximity as a criterion for word relatedness. In section 4 we describe in detail the method we propose for extracting multi-word collocations using collocation bigrams. Section 5 presents the experimental results obtained by applying this method on a large collection of English newspaper articles, and the last section draws the conclusion and points out directions for further development.

2 Existing Multi-Word Collocation Discovery Methods

Traditional approaches to automatic collocation extraction from text corpora rely on stochastic measures ranging from simple word co-occurrence frequency to more sophisticated statistical methods, like: the mutual information (Church & Hanks 90), the independence hypothesis test (e.g. likelihood ratios test (Dunning 93), Student's t-test, Pearson's χ^2 test - see chapter 5 of (Manning & Schütze 99) for a rather comprehensive overview).

One feature such methods share is that no syntactic criterion is used to select the candidate collocations. The methods actually consider all possible combinations of words and therefore are forced to limit to a text window of fixed size (usually not more than 5 words). Moreover, they usually take into account two-word collocations (bigrams). Only few methods, e.g. (Choueka *et al.* 83; Smadja 93), are also concerned with n-grams (n>2).

The method proposed by (Choueka *et al.* 83) to find n-word collocations considers the frequency of consecutive word sequences of length n (with n from 2 to 6), with a threshold of 14 for a corpus of 12 million words. The limitation to n=6 is due to the rapid increase of the number of all possible n-grams, for n bigger than 6.

The Xtract system (Smadja 93) is able to retrieve, in its first stage, word bigrams that are not necessarily contiguous in text, but can be separated by a certain number of words. It then studies, in the second stage, the words in the bigrams surrounding positions and identifies n-grams as the repetitive contexts, which can be either "rigid noun phrases", or "phrasal templates" (phrases containing empty slots standing for parts of speech).

Contrary to the former method, the latter is able to extract sequences of words of arbitrary length. It also has the advantage of getting rid of the recursively subsumed n-grams, returning, for each bigram, only the largest n-gram containing it.

Both methods rely only on a superficial text representation, while pointing out that the selection of terms should ideally be done following linguistic criteria.

Since robust large-scale parsers became available in the meantime, such as for instance (Abney 96; Collins 96; Laenzlinger & Wehrli 91)[1], the more recent methods focus on using parsed

[1]For recent advances in robust parsing see (Ballim & Pallotta 02).

rather than raw text for bigram extraction (Alshavi & Carter 94; Grishman & Sterling 94; Lin 98; Goldman *et al.* 01).[2]

In addition to considering syntactic criteria for selecting the candidate data, (Goldman *et al.* 01) also make use of a normalized sentence representation, which allows them to account for long-distance syntactic dependencies due to various linguistic phenomena like passivization, raising, dislocation, topicalization.

Our work relies to a large extent on the features of this method, which we will briefly present in the next section.

3 Automatic Extraction of Collocation Bigrams with FipsCo

FipsCo (Goldman *et al.* 01) is a term extractor system that relies on Fips (Laenzlinger & Wehrli 91), a robust, large-scale parser based on an adaptation of Chomsky's "Principles and Parameters" theory. The system extracts from parsed text all the co-occurrences of words in given syntactic configurations (noun-adjective, adjective-noun, noun-noun, noun-preposition-noun, subject-verb, verb-object, verb-preposition, verb-preposition-argument), thus applying a strong filter on the candidate bigrams. It further applies the likelihood ratio statistical test (Dunning 93) on the sets of obtained bigrams, in order to rank them according to how dependent the bigram's terms are on each other, which gives a measure of collocatedness.

The strength of this approach comes from the combination of a deep syntactic analysis of sentences with statistical tests. The sentence is normalized: the words are considered in their lemmatized form and in their canonical position (e.g. the subject in pre-verbal position, the direct object in post-verbal position); moreover, the system is able to create traces and co-indexation, and can handle complex cases of extraposition, such as relativization, passivization, topicalization, raising, dislocation.

To illustrate this with an example, let's consider the sentence fragment below:

> *"the* **difficulties** *which he might have* **experienced***"*

Extracting the collocation of verb-object type *"experience difficulty"* requires a complex syntactic analysis, made up from several steps: recognizing the presence of a relative clause; identifying the antecedent of the relative pronoun *"which"*; establishing the verb-object link between this pronoun and the verb of the relative clause.

This collocation will simply be overlooked by the statistical methods, where generally the size of the collocational window is 5. Such situations are quite frequent for example in Romance languages, where words can undergo complex syntactic transformations. (Goldman *et al.* 01) report an average of 29,26% cases of long-distance dependency (i.e. more than 5 words) between the top 100 collocations extracted (in all syntactic configurations) from a French corpus.

4 Multi-Word Collocation Discovery Using Collocation Bigrams

The system presented above is able to extract syntactically related collocation bigrams, that can occur practically unrestrictedly both with respect to the distance between collocates[3], and to the superficial textual realization (thanks to the deep syntactic analysis able to handle the cases of extraposition, where the collocates have undergone different syntactic operations). We will take advantage of these features for identifying multi-word collocations, since they would guarantee the grammaticality of results, as well as the unrestricted distance and realization form[4].

Since this system actually returns not only the best scored collocations, but all the candidate bigrams[5], we will in fact generate all the possible multi-word associations from text. Our goal is to build up, using the set of extracted bigrams, the sequences of bigrams sharing common words. The obtained collocate chains represent well-formed

[2]Note that (Smadja 93) already used a chunker in the third stage of Xtract to identify syntactic relations between collocates, but only to validate the bigrams extracted (verifying, for example, if the relation verb-object holds between the words "make" and "decision" in the collocation "make decision").

[3]At this point, the use of term "collocate" may seem paradoxical, but we would like to underline that words collocation doesn't refer to textual distance, but to the degree of associativity and dependency.

[4]FipsCo is already able to extract a restricted type of multi-word collocations, as bigrams in which a term can be in turn either a compound, an idiom or a collocation. This term must be already present in the lexicon (previusly recognized).

[5]A line separating between bigrams that are collocations from those that are not is in any case difficult to define.

multi-word associations. The configuration of their syntactic structure is defined by the syntactic relations in the bigrams involved.

The shared term must be the same not only lexically (the same word), but also indexically (the very same occurrence of word, i.e. the same position, in the same text). Due to the syntactic relatedness constraint, the shared term will actually appear in the same sentence as the other collocates.

For instance, given two bigrams

$$(w_1\ w_2), \qquad (w_1\text{'}\ w_2\text{'})$$

with w_2 and w_1' identical as index, we can construct the 3-gram:

$$(w_1\ w_2\ w_2\text{'}),$$

as in the case of the following collocations: *"terrorist attack"*, *"attack of September"*; we obtain the 3-gram collocation *"terrorist attack of September"*[6]. Repeating the same procedure we can add further words to the obtained 3-grams, thus obtaining multi-words collocations of arbitrary length. Moving on to n-grams will conserve the inclusion of all terms in the same sentence. We impose no default restrictions on the syntactic configuration of the resulting expression, considering that all the associations are valid.

In subsection 4.1 we present in greater detail the word linkage procedure that allows the construction of longer multi-word collocations using shorter multi-word collocations combinations. In subsection 4.2 we propose several measures for ranking the obtained expressions according to the degree of collocational association. We will also show how we ran the log-likelihood test on the new expressions, a test which provides a finer measure for word association quality, similar to the case of collocation bigrams.

4.1 Iterating on Word Linkage

The procedure of linking new words to a partially constructed collocations in order to discover longer collocations uses the criterion of the existence of a syntactic link between the new words and one of the existing collocation's words. Recursively applied to the set of generated collocations in each step, this procedure allows the incremental composition of longer collocation from shorter subparts. In this manner, it leads to the identification of all collocation candidates in a text, each one virtually limited only by the sentence's boundaries, and possibly by the disconnected substructures in it.

Finding all the 3-grams given a set of bigrams is done, for example, by considering all the pairs of bigrams that share terms. We name "pivot" the term shared by two bigrams. There are three possibilities to construct a 3-gram, that correspond to the position of the pivot in the two bigrams. The most natural pivot position can be seen as the middle (internal) one, as in the example given above in section 3 (*"terrorist **attack** of September"*). The pivot is the last term in the first collocation, and the first in the next one.[7] But the external positions of pivot are also productive. We exemplify with two 3-grams constructed with the pivot in the left and right position respectively: "have impact on", derived from the bigrams "*have* impact" and "*have* on", and "round [of] presidential election, derived from "round [of] *election*" and "presidential *election*".

For the general case, we consider the following criterion to combine two multi-word collocations (MWCs) into a larger one: two MWCs can combine if they have at least one term that is different and one that is identical with respect to the index (i.e. the position in the document).

The linking procedure that incrementally constructs all n-grams is described below.

Let $\mathcal{D}$ be the initial MWCs database, composed of all syntactically bound bigrams. The following algorithm derives all the n-grams from these database:

$\mathcal{C} := \mathcal{D}$;
repeat
  $\mathcal{N} := \emptyset$;
  for each MWC$_i$ in $\mathcal{C}$
    for each MWC$_j$ in $\mathcal{C}$, i$\neq$ j
      if combine(i, j) then
        add($\mathcal{N}$, combination(i, j));
        remove($\mathcal{D}$, MWC$_i$);
        remove($\mathcal{D}$, MWC$_j$);
  $\mathcal{C} := \mathcal{N}$;
  $\mathcal{D} := \mathcal{D} \bigcup \mathcal{C}$;
until $\mathcal{C} = \emptyset$;

[6] Note that the condition of indexical identity avoids combinations with different readings in case of polysemy, like the combination of bigrams *"terrorist attack"* with *"attack of coughing"*.

[7] Note that the symmetrical case (with the pivot being the first term in one and the second in the another) is equivalent, since there is no order defined on the set of collocation bigrams. The two cases produce the same set of 3-grams.

where *combine(i, j)* is a predicate that is true iff the expressions MWC_i and MWC_j can be combined following the above stated criterion, and *combination(i, j)* is the resulting MWC (obtained by merging the terms involved).

At each step, the procedure tries all the possible combinations of MWCs generated in the previous step, using the composition criterion stated above. When a new combination is possible, it adds it to the database and eliminates the participating (subsumed) MWCs. The process is repeated as long as new MWC can be constructed from the MWCs generated in the previous step.

The procedure is guaranteed to terminate after a finite number of iterations, since the set of new expressions to form is localized in the sentence and it is finite. The final database will contain all the initial bigrams that did not participate in bigger MWCs, and all the syntactic n-grams existing in the text, with n arbitrarily long and limited only by the sentence's length. No subsumed MWCs are present, since the procedure systematically finds the largest bigram and gets rid of its subparts.

It is easy to verify that the complexity of the algorithm is polynomial in the size of the initial bigram database, $|\mathcal{D}|$. We did not consider the optimization issue, which will be the topic of future work.

4.2 Measures of Interestingness

The MWCs extracted with the algorithm described in the previous subsection are all the syntactically bound co-occurrences of terms in the corpus. We considered 4 methods to distinguish between interesting and uninteresting ones, i.e. to identify the good collocation candidates. The first (and simplest) method computes their frequency. The second uses the collocation score initially assigned to the bigrams (based on the statistical test of log-likelihood ratios). For each MWC, we sum up the scores of the composing bigrams and obtain a global score characterizing the collocational behaviour of the MWC as a whole.

The third method tries to find MWCs whose global score is balanced, and is motivated by the intuition that a MWC is a good collocation iff the composing bigrams have similar collocation scores. Thus, we considered the following measure for evaluating n-grams:

$$\frac{n \prod_{i=1}^{n} score(MWC_i)}{\sum_{i=1}^{n} score(MWC_i)} \quad (1)$$

As a fourth method, we adopted a statistical test as a more appropriate measure of n-grams collocational behaviour, namely the log-likelihood. This test was also used by FipsCo in scoring collocation bigrams. It applies to term pairs to whom it assigns a collocativity score computed according to the contingency table of the pair (which contains the frequency of: i) the co-occurrence of the two terms together in the corpus, ii) the co-occurrence of one of the terms with a different one, and iii) all the other co-occurrences, not involving any of the terms in the given pair). The test increases a collocation's score each time the two collocates are found together, and decreases it when one of them co-occurs with a different term.

We were interested in extending the test's application to MWCs, and in doing this we applied it recursively to the sub-MWCs composing a given MWC. Let MWC_1 and MWC_2 be two MWCs that compose a larger MWC (as described in 4.1). The log-likelihood score is computed using a contingency table for the pair (MWC_1, MWC_2), listing co-occurrence frequencies related to each of the two sub-expressions.

The results in multi-word collocation ranking using the proposed measures are showed in the next section, which presents an experiment of building up 3-grams from the collocation bigrams extracted from a large collection of texts.

5 The Experiment. Results and Discussion

We applied the method of identifying multi-word collocations as presented above on a corpus of 948 English articles from the magazine "The Economist" (on-line version). The collection totaled about 870,000 words. First, the texts were parsed and about 142,000 syntactically related bigrams were extracted using FipsCo (this counts all the co-occurrences in the syntactic patterns used by FipsCo, with no frequency filter). About 7.00% of these bigrams were already multi-word, since one of their terms was either a compound, idiom or another collocation, already included in the lexicon.

We then extracted 3-grams using the linkage method presented in subsection 4.1. We obtained

3-gram	freq	pivot
weapon of mass destruction	38	3
have impact on	17	1
go out of	15	2
pull out of	14	2
make difference to	11	1
rise in to	10	1
move from to	10	1
rise from in	10	1
play role in	9	1
be to in	8	1
have interest in	8	1
rise from to	8	1
come out of	8	2
get out of	8	2
main reason be	8	2
turn blind eye	7	3
make to in	7	1
rise by in	7	1
second world war	7	3
round of presidential election	6	3

Table 1: The 20 most frequent 3-grams extracted

a number of 54,888 3-grams, divided in 13,990, 27,121, and 13,777 for each pivot position case, i.e. left, middle, and right respectively. Table 1 shows the 20 most frequent 3-grams in the whole set[8].

Tables 2-4 present the top 10 3-grams according to the other measures proposed in subsection 4.2 (sum, mean, and log-likelihood score respectively).

3-gram	sum
be prime minister	1152.56
prime minister be	1134.21
prime minister deny	1133.97
appoint prime minister	1131.91
prime minister have	1128.44
prime minister asleep	1127.77
prime minister embarrassed	1127.77
flashpoint prime minister	1126.73
prime minister explain	1126.06
prime minister promise	1123.92

Table 2: Top 10 results for 3-grams according to the sum score measure (described in 4.2)

We considered as the most informative measures the first one, based on the frequency, and the last one, based on the log-likelihood test. Despite its simplicity, the first measure (based on frequency) provides quite good precision at ranking 3-gram collocations. By contrast, we noticed

[8]Only the prepositions that introduce arguments of verbs are considered as a bigram terms. The others are included for readability. We do not apply any function word filter.

3-gram	mean
weapon of mass destruction	377.36
be poor country	187.49
be rich country	180.69
next year be	116.40
be cold war	105.22
be against cold war	105.22
rest of Arab world	104.06
solve problem be	101.87
main reason be	96.13
give nuclear weapon	85.86

Table 3: Top 10 results for 3-grams according to the measure presented in equation (1) in subsection 4.2

3-gram	log
weapon of mass destruction	579.03
have impact on	214.35
move from to	126.10
turn blind eye	124.01
rise from in	120.57
play role in	110.07
make difference to	109.46
rise in to	105.43
second world war	105.42
rise from to	99.08

Table 4: Top 10 results for 3-grams according to the log-likelihood test (see subsection 4.2)

that simply using the sum of scores of participating bigrams cannot give a good measure for evaluating 3-grams: we get as best scored the expressions which contain a top scored bigram (as *"prime minister"*, in our case), but are not necessarily collocations as a whole. The third measure, that gives preference to the uniformly scored collocations, allows us to find out good multi-word collocations (like "weapons of mass destruction", that received the best score). Still, we judge its results less satisfactory than those obtained using the fourth measure, which lists in the first places 3-grams showing actual collocational behaviour ("conventional ways of saying things").

We were interested in the syntactic configurations of the multi-word collocations we obtained, since they could suggest syntactic patterns to use for the extraction of multi-word collocations directly from parsed text[9]. The most frequent association types are listed in Table 5, together with an example for each.

[9]As mentioned earlier, during the extraction no predefined syntactic patterns were used.

rel1	rel2	frequency	example
Adjective-Noun	Noun-Prep-Noun	5607	other part of world
Verb-Object	Verb-Prep	5364	keep eye on
Subject-Verb	Verb-Prep	4904	share fall by
Subject-Verb	Verb-Object	4659	company became leader
Verb-Object	Adjective-Noun	4622	improve public service
Adjective-Noun	Subject-Verb	3834	main reason be
Verb-Prep	Verb-Prep	3232	move from to
Verb-Object	Compound	2366	declare state of emergency
Verb-Object	Subject-Verb	1693	want thing be
Noun-Noun	Noun-Prep-Noun	1627	world standard of prosperity

Table 5: The 10 most frequent association types for 3-grams

6 Conclusion and Future Work

We have presented a method aimed at extracting multi-word collocations, which relies on the previous extraction of collocation bigrams from text, and is based on iteratively associating already constructed collocations using a syntactic criterion. We have used several measures which quantify the strength of the association. In particular, we applied the log-likelihood ratio statistical test (initially used for word bigrams) to the extracted multi-word collocations, which showed to be the best measure for evaluating the collocational strength.

The methodology used is based on a hybrid (linguistic and statistical) approach aimed at improving the coverage and the precision of multi-word collocation extraction. Unlike purely statistical approaches, the method presented can handle long-distance occurrence of terms (which can often happen due to several types of syntactic transformations). Also, all the results are grammatical, due to the syntactically-based filter of candidates and to the syntactic nature of the criterion used for the composition of longer multi-word collocations.

Further developments of the method include finding finer linguistic criteria for a more precise delimitation of n-grams within the sentence, thus better accounting for subsumed and subsuming expressions.

As for applications, we plan to integrate the method in a concordance and alignment system (Nerima *et al.* 03) that would allow for the visualization of the contexts of multi-word collocation occurrences in the source text, as well as in the parallel text (in its translation), when available.

We believe that many language processing tasks may considerably benefit from the approach of multi-word collocation extraction using linguistic constraints.

Acknowledgement

This work is supported by the Geneva International Academic Network (GIAN), research project "Linguistic Analysis and Collocation Extraction". We are grateful to Catherine Walther Green, Genoveva Puskas, Vincenzo Pallotta, and to the anonymous reviewers for their valuable comments on different versions of this paper.

References

(Abney 96) Steven Abney. Partial parsing via finite-state cascades. In John Carroll, editor, *Proceedings of the Workshop on Robust Parsing at the 8th Summer School on Logic, Language and Information*, number 435 in CSRP, pages 8 – 15. University of Sussex, Brighton, 1996.

(Alshavi & Carter 94) Hiyan Alshavi and David Carter. Training and scaling preference functions for disambiguation. *Computational Linguistics*, 20(4):635–648, 1994.

(Ballim & Pallotta 02) Afzal Ballim and Vincenzo Pallotta. Robust methods in analysing natural language data. *Natural Language Engineering*, 8(2), 2002.

(Benson 85) Morton Benson. Collocations and idioms. In Robert Ilson, editor, *Dictionaries, lexicography and language learning*, pages 61–68. 1985.

(Benson 90) Morton Benson. Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1):23–35, 1990.

(Choueka *et al.* 83) Y. Choueka, S.T. Klein, and E. Neuwitz. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing*, 4(1):34–8, 1983.

(Church & Hanks 90) Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

(Collins 96) Michael John Collins. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191, Philadelphia, PA, 1996.

(Dunning 93) Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

(Firth 57) John R. Firth. Modes of meaning. In J. R. Firth, editor, *Papers in Linguistics*, pages 190–215. 1957.

(Goldman *et al.* 01) Jean-Philippe Goldman, Luka Nerima, and Eric Wehrli. Collocation extraction using a syntactic parser. In *Proceedings of the ACL '01 Workshop on Collocation*, pages 61–66, Toulouse, 2001.

(Grishman & Sterling 94) Ralph Grishman and John Sterling. Generalizing automatically generated slectional patterns. In *Proceedings of COLING-94*, pages 742–747, 1994.

(Harris 51) Zelig Harris. *Structural Linguistics*. University of Chicago Press, Chicago, IL, 1951.

(Laenzlinger & Wehrli 91) Christopher Laenzlinger and Eric Wehrli. Fips, un analyseur interactif pour le français. 32(2), 1991.

(Lin 98) Dekang Lin. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, Montreal, 1998.

(Manning & Schütze 99) Christopher Manning and Heinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, 1999.

(Mel'cuk et al. 99) Igor Mel'cuk et al. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques*. Les Presses de l'Université de Montréal, Montréal, 1999.

(Nerima *et al.* 03) Luka Nerima, Violeta Seretan, and Eric Wehrli. Creating a multilingual collocation dictionary from large text corpora. In *Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, 2003.

(Sinclair 91) John Sinclair. *Corpus, Concordance, Collocation*. Oxfort University Press, Oxford, 1991.

(Smadja 93) Frank Smadja. Retrieving collocations form text: X-tract. *Computational Linguistics*, 19(1):143–177, 1993.