A Tool for Multi-Word Collocation Extraction and Visualization in Multilingual Corpora

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

Abstract
This document describes an implemented system of collocation extraction which is designed as aid to translation and which will be used in a real translation environment. Its main functionalities are: retrieving multi-word collocations from an existing corpus of documents in a given language (only French and English are supported for the time being); visualizing the list of extracted terms and their contexts by using a concordance tool; retrieving the translation equivalent of the sentences containing the collocations in the existing parallel corpora; and enabling the user to create a sublist of validated collocations to be further used as reference in translation. The approach underlying this system is hybrid, as the extraction method combines the syntactic analysis of texts (for selecting the collocation candidates) with a statistical-based measure for the relevance test (i.e., for candidates ranking according to the collocational strength). We present the underlying approach and methodology, the architecture of the systems, we describe the main system components and provide several experimental results.

1. Introduction

Collocations, defined as "arbitrary and recurrent word combinations" in (Benson 1990) or "institutionalized phrases" in (Sag et al. 2002), represent a subclass of multi-word expressions that are prevalent in language and constitute a key problem, not only for natural language processing (NLP), but also for humans - either second language learners or professional translators.

The *collocate,* i.e. the proper word that can be used in combination with a given word (often called *base word*), is unpredictable. It is difficult to choose even from a set of near synonyms. One needs to be already aware of the customary, conventional usage of an expression (e.g. "encounter difficulties") in order to avoid unnatural paraphrases (such as the French-like "feel difficulties"). This is essential for major NLP applications, such as natural language generation and machine translation, but also for humans faced with the task of producing documents in a non-native language that requires a good level of proficiency.

The problem of collocations has to be addressed also in a particular circumstance, that of translating documents in a particular domain. In particular, a multilingual society requires the official documents to be written in all the participating languages. Particular attention must be paid to the expressions that have a conventional usage (i.e., to collocations). Alternative paraphrases generally have to be avoided, either due to the specificity of the

domain (one must use a consistent translation for a specific multi-word term), or because the paraphrases may have unwanted implications.

Our work is situated in a cross-linguistic communication context, namely that of the World Trade Organization (WTO), in which the proper understanding and translation of specific terminology plays an important role. It has been carried out in the framework of a project aimed at extracting multi-word terminology (compound words, idioms, and collocations) from French and English parallel corpora and at visualizing the translation equivalents, by means of an alignment system, in the other languages from the range of WTO official languages (English, French, and Spanish).

We present an implemented system able to accurately identify the collocations in a collection of documents by performing a syntactic text analysis, and then retrieve the sentence that contains the translation of a collocation in the existing translations. The system provides several tools designed for terminologists, which enable them to visualize past translations and to create and maintain a database of collocation terminology including multilingual translations. In addition, the system provides tools to be used by translators, which help them to identify a given collocation in text and to see the proposed translations, together with usage examples in different contexts.

Sections 2 of this paper presents the philosophy underlying the method for collocation extraction. Section 3 outlines the design methodology adopted in building the system. In section 4 we present the architecture of the system, its components and their main functionalities, as well as some technical details about the implementation. Section 5 contains some experimental results obtained using our system. Finally, section 6 concludes the article and settles the possible directions of further development.

2. A Hybrid Approach to Collocation Extraction

The method of collocation extraction we have developed is based on a hybrid approach, which combines both the symbolic and statistical processing.

Generally, a process of collocation extraction comprises two main stages:

- the *candidate selection*, in which the word expressions that could represent a collocation are extracted from text according to some defined collocation patterns (or configurations);
- the *relevance test*, which assigns to each extracted candidate a weight indicating its likelihood to constitute a collocation.

The result of a collocation extraction system is a ranked list (usually called *significance list*) of potential collocations, with the most probable collocation on the top.

The classical collocation extraction methods focus mostly on the second stage, by proposing more or less sophisticated association measures for collocation ranking, which are either statistic or information theoretic based (Sinclair 1991; Smadja 1993; Church & Hanks 1990). The first stage of extraction, the pattern definition, is usually ignored. In fact, any combination of two words is considered as a possible collocation candidate[1]. Besides, due to the manner the association measures are conceived, the methods are not appropriate for collocations longer that two words (i.e., multi-word collocations, henceforth MWCs).

In contrast, our method emphasizes the importance of the first step in the process of collocation extraction. The method uses the same association measures for assessing the

strength of the lexical associations (collocational strength), but it focuses on precisely defining what kind of words association may represent a collocation.

In the literature, it is largely agreed that the extraction of collocation should be done ideally from parsed, rather than from raw text. This approach, that commit to the definition of collocation as a syntactically bound expression, contrast with the approach in which the text is seen as un unstructured chain of words, and the collocations are considered, in broader acception, as words co-occurring "within a short space of each other" (Sinclair 1991) more often than by chance.

Some of the existing collocation extraction systems (Smadja 1993; Grishman and Sterling 1994; Lin 1998) already perform different degrees of syntactic processing, such as lemmatization, POS tagging, or syntactic dependency test, for filtering or for validating collocations. Recent work shows a growing interest in performing a deep linguistic analysis in collocation extraction (Goldman et al. 2001; Krenn & Evert 2001), accentuating the role the filtering component has in the performance of extraction systems. Integrating a syntactic component in the process of collocation extraction is nowadays possible thanks to the strong increase, over the last few years, in the availability of computational resources and tools dedicated to the large-scale and robust syntactic parsing.

Our method applies thus a strong filter on the collocation candidates, based on syntactic criteria. First of all, not any word combination can constitute a collocation, but only those combinations that are syntactically well-formed. The syntactic analysis helps to filter out the invalid combinations.

Second, in choosing collocation candidates no restriction should apply on the words form, relative position, and distance. The collocations are fully flexible with respect to morphology and syntax. They can undergo complex operations, due to which the composing words may be inflected, inverted or extraposed at an indeterminate length. Commonly, extraction systems count inflected forms and inverted words as different collocations. They also limit the collocate search space to a window of several words (usually 5), in order to overcome the combinatorial explosion when generating all possible word combinations.

Overcoming these restrictions is only possible by performing a morpho-syntactic analysis. For instance, the sentence normalization that is performed during parsing includes the words lemmatization (the words are considered in their base form) and allows affording words inflection. Also, the words in a sentence are considered in their canonical order, therefore the normalization helps dealing with inversion. Finally, the parser is able to keep traces and create co-indexation, thus the complicated cases of extraposition can be afforded.

Considering the syntactic dimension of collocations contributes to improving the performance of extraction systems, in terms of both precision and recall[2]. But the results of candidates ranking can benefit as well from the syntactic analysis. According to a recent report (Krenn & Evert 2001), the association measures perform differently when ranking different types of collocations (e.g., the mutual information measure is more appropriate for instance for ranking adjective-noun collocations than verb-object collocations). Therefore, by tuning the different measures to the suitable syntactic configurations, the overall performance of extraction systems may be improved.

3. Methodology

The system we implemented relies to a large extent to Fips (Laenzlinger & Wehrli 1991), a syntactic parser for French and English which is robust enough to process large collections of documents, without requiring any preprocessing.

We experimented our system on collections of documents of different types, such as newspapers articles containing text only, and the WTO parallel corpora of documents, which contain HTML text automatically generated by different word processors. These documents may also contain tables, differently formatted across versions, but in general they are not really noisy (no images or OCR output) and pose no problems to Fips parser.

The extraction of collocations from text corpora is based on the parser's results, as the possible collocations are selected on syntactic criteria from the parsed text. Fips identifies all the co-occurrences of words in given syntactic configurations, e.g., adjective-noun, subject-verb, verb-object, noun-preposition-noun, that have been defined in advance. The parser is able to handle the morpho-syntactic transformations like those discussed in the previous section.

A statistic test of independence hypothesis, namely the log-likelihood ratio test (Dunning 1993) is then applied to the sets of word co-occurrences (bigrams) obtained for each configuration. This test assigns each bigram a collocation score used to order bigrams according to their collocational significance, from the best collocation candidate to the candidates that most probably do not constitute a collocation.

Afterwards, the extracted bigrams are presented to the user, who can visualize the sentences and the documents in which they occur. Moreover, an alignment method has been implemented that retrieves the translation equivalents of these sentences in the documents for which versions in other languages exist.

In order to use the extracted terminology for further translations, a manual validation takes place, in which the terminologists compile monolingual or bilingual terminology databases using the visualization tools.

In addition, our system includes a method of syntactic-based composition of extracted bigrams (Seretan et al. 2003) into larger n-grams, in order to identify multi-word collocation in corpora (collocations longer than two words). This is a distinguishing feature of our system, as the large majority of existing collocation extraction systems are only concerned with collocations made up of word pairs (the limitation deriving primarily from the specific design of association measures for pairs of items). Multi-word collocations are instead prevalent in language and recent development in NLP emphasizes the need to integrate their treatment in many applications.

4. The System

The main components of the system are briefly presented below.

- *File Selection module*, used to select a corpus of documents by recursively scanning a folder's structure and applying a files filter (based on file name, file type, and file last modification date).
- *Fips Syntactic Parser*, used to parse the whole collection of documents.

- *Co-occurrences Extraction System*, which retrieves all two-word co-occurrences (word pairs) in pre-defined syntactic patterns, such as: noun-adjective, adjective-noun, noun-noun, noun-preposition-noun, subject-verb, verb-object, verb-preposition, verb-[preposition]-argument. We used the system FipsCo (Goldman et al. 2001), which is based on Fips parser. FipsCo also applies the Log-likelihood test (Dunning 1993) on these co-occurrences and assigns them a collocation score.

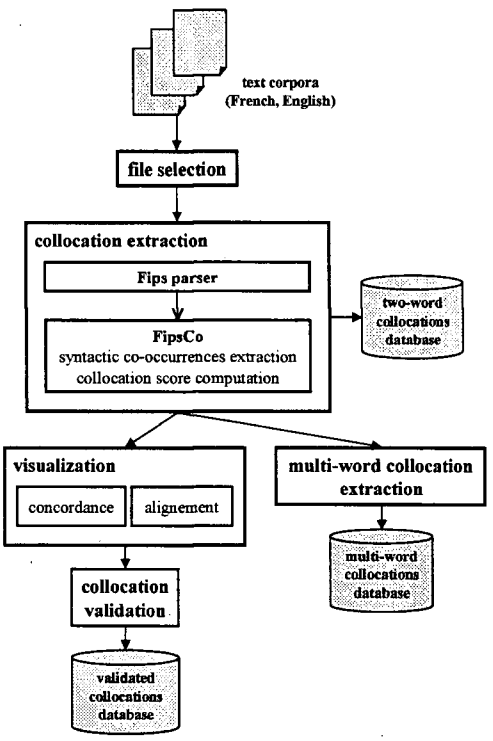

Figure 1: Architecture of the system

- *Multi-Word Collocation Extraction Module*, which uses a syntactic-based method of bigram composition (Seretan et al. 2003) for building up multi-word collocations (collocations longer that two words). We apply this method on the co-occurrences extracted by the previous module.
- *A Concordance Tool*, which displays the extracted list of (multi-word) co-occurrences and their contexts, in the source document as well as in its translations. The list of terms may be ordered/filtered by the collocation score assigned, by frequency, or alphabetically.
- *An Alignment System*, capable of retrieving the target document (i.e., the translated document) in the parallel corpora when available, and of hypothesizing the target sentence that contains the translation equivalent of a collocation.

- *Collocation Validation module*, which allows the user to add various types of information about the selected collocations into a terminological database.

The architecture of the system is mainly pipelined and is sketched in Figure 1. The data flow and the processing order mainly correspond to the order in which the system's components have been presented above. Still, in the visualization part the modules are not organized linearly but in parallel, i.e., one of the two alternatives (concordance or alignment) can be chosen at a time. Also, the MWCs extraction module is not yet connected to the visualization and validation part.

As for technical details, the system was implemented in Component Pascal using the BlackBox development environment for Windows[3]. The stand-alone executable application includes all the data files needed (e.g. lexicons). In order to run it, a standard hardware configuration is sufficient. The only installation requirement is defining a data source for the working database provided with the application.

In what follows, we describe in more detail the system's components we have implemented. The other components used, Fips parser and FipsCo co-occurrence extraction system, are described in the references provided.

4.1. File Selection module

This module allows the user to specify the documents in the input corpus, from which the collocations will be extracted. There is no limitation on the number of files that can be processed. Some experimental details are presented later in the dedicated section. The types of files that are supported for the time being are ASCII file types (e.g., txt, html) and odc (BlackBox specific files format)[4].

The most important functionalities of this module are:

- retrieving all the files belonging to a folder which contains the corpus files and can be located either on the local computer or on a network place;
- applying an automatic filter on the retrieved files, based on several criteria;
- providing a manual filter feature, that allows the user to further select or deselect the objects (files of subfolders) from the input folder.

The automatic filtering can be done by applying different criteria, e.g., the file location, name, type, and last modification date. More specifically, it allows the user to include into the selection only the files on the first level of the input folder, or the files on all the levels by recursively including the subfolders. Also, subfolders with specific names can be excluded. The module allows the user to select only files of given types which can be specified in a list, or to select only the files containing a given string of characters in their name. The filter on the last modification date of files allows to include only files created or modified in a given time interval (between two given dates) or time distance (a given number of days ago).

The role of this module is not only to choose the files to process, but also to launch the Fips parser on the files selected and to gather the results of each parse process.

4.2. Multi-Word Collocation Extraction Module

The components of the system that perform the collocation extraction, i.e., the parser Fips and the FipsCo related system of bigram extraction and ranking, deal basically with two-

words collocations. From the parse structures returned by Fips, FipsCo extracts all the co-occurrences of words in specific syntactic patterns, on which it further applies the log-likelihood ratio test.
Some of the patterns considered may include additional words, e.g., prepositions. Besides, a bigram constituent can be a multi-word term, such as a lexicalised compound, idiom or even collocation. These features allow the system to extract multi-word collocations, i.e. collocations containing more than two words. Still, there is not a full support for multi-word collocations in general, as the method is basically designed for two-terms only and the multi-word constituent must be listed in the parser's lexicon.
Nevertheless, the collocation bigrams constituting the results of this method can be used for generating arbitrarily long collocations. Based on this idea, we designed a method for extracting multi-word collocations by using the syntactic composition of collocation bigrams (Seretan et al. 2003).
One of its main advantages is that no pre-defined syntactic configurations are used to define the multi-word collocations' structure; on the contrary, the method allows the system to discover the most frequent syntactic patterns for multi-word collocations. This information can be used later in the extraction of MWCs at parsing time, for defining the candidates' configuration.
Another advantage of this method is that it allows using appropriate association measures to rank the candidates. We defined several measures of multi-word collocation ranking according to the collocational strength, based on the initial bigrams' log-likelihood score. We also applied the log-likelihood ratio test on the bigrams composing the tri-grams generated, which showed to be yield good preliminary results.
The visualization and validation of extracted multi-word expression will be soon integrated into the system.

4.3. Concordance Tool

The concordance tool allows the user to visualize the terminology extracted and the context of each occurrence of terms in the originating document. The terminology comprises not only the collocation bigrams, but also compound words and idiomatic expressions retrieved by Fips parser in text, whenever these terms are present in the parser's lexicon.
The database that stores the extracted terminology contains various information which enables the retrieval of all the contexts of occurrence in the corpus, for each entry. The context considered for a term is the sentence in which the term occurs. The concordance tool displays these contexts in the whole originating documents, and automatically scrolls the text on the context of the selected occurrence.
The concordance tool shows the whole list of extracted terminology, in which a term is displayed only once. The user can select a term and then see all its occurrences in the corpus, as the tool enables the exhaustive browsing through term instances.
There is support for the advanced visualization of the terms list, based on multiple criteria, such as the collocation data source and language, the corpus frequency, the score etc.
The user interface is similar to the interface of the alignment system, which is shown in Figure 2. Compared to the concordance tool, the alignment system also displays the terms

contexts in the translations of the source document, when available (e.g., in Spanish and French for the English terminology extracted from the WTO corpora).

4.4. Alignment System

This section presents the main features of the alignment method we implemented for retrieving the equivalent of a collocation context in the parallel documents, i.e., in the translated versions of the source document. The method has been first introduced in (Nerima et al. 2003).

It is based on the documents structure, relative size, and content analysis.

When trying to hypothesize which is the translation of a sentence, the system first looks at the paragraphs structure of documents, and then it finds the correspondence between the source and target sentences in the paragraphs found.

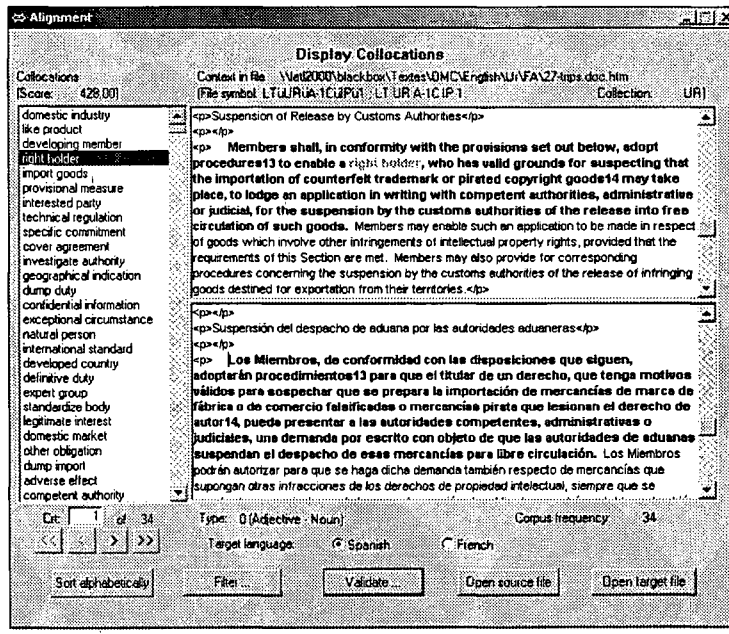

Figure 2: Interface of concordance and alignment systems (screen capture)

The relative size of the two documents is taken into account when looking for an initial candidate (IC) for the target paragraph (TP). Then, the relative proportion of paragraphs sizes around IC is considered when looking for the best candidate. The best candidate is the one whose surrounding in the target document best fits the surrounding of source paragraph (SP), in terms of relative size proportions:

$$\min_{i \in O} \sum_{j=i-c}^{j=i+c-1} \left| \frac{s_j}{s_{j+1}} - \frac{t_j}{t_{j+1}} \right|,$$

where i is an index for the surrounding paragraphs, s_j and t_j are the paragraphs around the SP and the current target paragraph respectively, and finally c is a constant used to indicate the how many paragraphs are considered in a surrounding ($2c + 1$, i.e., c paragraphs before and c after). The sum expression shows the overall difference in paragraphs sizes proportions, between source and target documents.
The method also performs a shallow content analysis. It uses the paragraphs numbering (if present) to correct the target paragraph choice.
The specificity of this method consists in not presupposing an initial matching between paragraphs, but concentrating on the paragraphs alignment. Sentence level alignment methods (e.g., (Gale & Church 1991)) rely on the preliminary alignment of paragraphs. But this is not a trivial task, especially for differently formatted HTML documents from the WTO corpus. Our method is suited to noisy corpora, in which the document structure, content and layout are not preserved over translations.
Another distinctive feature of this method is the partial alignment. Not the whole document is aligned, but only the paragraph containing the term occurrence currently visualized. The alignment is done on-the-fly and needs no pre-processing.
Using this method, the concordance tool instantly displays, together with the source context, the translations in the other languages for which parallel corpora exist. The user can see how a given collocation has been used in a given context, in the source and in the target languages. The whole system can be seen as a multilingual "dictionary-cum-corpus", which provides for each term examples of actual usage in different contexts and different languages.

4.5. Collocation Validation module

The role of this module is to help the user to compile a database of validated terminology from the automatically extracted collocations. This data can be used later as reference, for instance, in further translations. It can also be used as a resource in further extractions, in which the validated terminology is included in the lexicon used by the parser. This circular procedure would help to retrieve always longer multi-word terminology, by using past extracted terms in newly extracted bigrams.
In a validation session, the user chooses, using the visualization tools, the interesting terms (also terms occurrences) and adds them to a validation list. Most of this information is automatically proposed by the system. The user has the possibility to add and modify the entries and finally, to save the validated list, either part of it or entirely.
The information that is stored for a mono-lingual or bilingual entry contains mainly the term key, the index of lexemes in the lexicon, the syntactic configuration, the source and target languages, the term translation, example of usage in the source and target languages, as well as information concerning the file from which the term has been extracted.

Corpus	Size	Processing Time	Processing Speed	Bigrams Extracted	Tri-grams Extracted
The Economist	6.20 Mb 879'013 words	7'158 s	0.88 Kb/s 121.5 words/s	161'293 total 106'713 distinct	58'398 total 55'351 distinct
Le Monde	8.88 Mb 1'471'270 words	7'936.2 s	1.14 Kb/s 185.4 words/s	276'932 total 182'298 distinct	119'852 total 113'150 distinct

Table 1: Experimental results for bigrams and tri-grams extraction from two corpora

5. Experiments. Discussion and future work

We have done most of the experiments with the developed system on the WTO corpus, which is tri-lingual (French, English, Spanish). The Spanish version is less represented (14.6% from the total of 1.21 Gb, vs. 41.4 % for English and 43.9% for French). The average document size is 32 Kb, or approximately 4'200 words per document. Experimental results on this corpus are reported in (Nerima et al., 2003).

Another corpus we used is monolingual (in English) and includes on-line articles from the newspaper "The Economist". It contains about 1'000 articles, with an average of 920 words per document (6.7 Kb average size). It totals 6.2 Gb, and about 880'000 words. We also used a French monolingual corpus containing articles from "Le Monde". Its size is 8.88 Mb and contains about 1'470'000 words.

Table 1 shows several statistics on the collocation extraction results obtained on these corpora, and on the processing time required[5]. Table 2 lists the top 10 collocation bigrams obtained in each experiment, according to the log-likelihood score assigned, and top 10 tri-grams obtained, in the frequency order (note that the words are always shown in their base form, even if the collocation expressions requires their inflection).

Bigrams		Tri-grams	
The Economist	Le Monde	The Economist	Le Monde
prime minister	milliard de franc	weapon of mass destruction	ministre de affaire étranger
last year	million de franc	have impact on	Front du salut national
mass destruction	premier fois	go out of	ministre de éducation national
interest rate	milliard de dollar	pull out of	tribunal en grande instance
next year	premier ministre	make difference to	président de conseil général
chief executive	Assemblée national	rise in to	membre de comité central
bin laden	Union soviétique	move from to	membre de bureau politique
poor country	million de dollar	rise from in	réaliser chiffre de affaire
central bank	affaire étranger	play role in	franc de chiffre de affaire
see as	fonction public	have interest in	chiffre de affaire de milliard

Table 2: Top 10 bigrams ordered by the log-likelihood score, and the 10 most frequent tri-grams extracted

A sound evaluation of system's performance must be done using appropriate techniques, such as the precision and recall measurements (when a collocation reference subset will be

available). Possibly, the precision and recall will be quantified at different results strata, as in (Krenn & Evert 2001), in order to evaluate the results ranking too.
Further developments will mainly focus on adding other languages for parsing, using a more comprehensive or more generic set of syntactic patterns for bigrams, connecting the multi-word collocation extractor to the visualization tools, and creating reference collocation resources for evaluation.

6. Conclusion and related work

We presented an implemented system of multi-word terminology extraction and visualization in parallel corpora, that focus primarily on collocations and will be used in a real translation environment. It integrates a hybrid method for extracting collocation bigrams, i.e., a method which is syntactically-based in the candidate filtering stage, and statistically-based in the second stage of candidate ranking according to the significance tests.
The system includes a syntactic method of bigrams composition into longer collocations, which allows the extraction of arbitrarily long expressions. The system is also composed of an alignment method, a visualization tool, and, finally, a validation module through which a terminologist can create monolingual or bilingual terminology. These reference resources may include information on the usage of the expressions in given contexts, that can be further used in human translations or in NLP applications.
Experiments done with this system showed that it can be robustly applied on large corpora in order to extract French and English multi-word terminology.
There exist many collocation extraction methods, alignment methods and visualization tools used for translation aid. The originality of our tool consist, on the one hand, in integrating this kind of methods and tools into one system, and, on the other hand, in using a linguistically motivated approach for (multi-word) collocation extraction, made possible by the robustness of Fips parser.

Acknowledgement

This work has been carried out in the framework of the research project "Linguistic Analysis and Collocation Extraction" (2002-2003) supported by the Geneva International Academic Network (RUIG-GIAN).

Endnotes

[1] Possibly with the exception of combinations involving function words.
[2] The *precision* is defined as the ratio of correctly identified collocations from the number of returned results, and the *recall* as the ratio of correct collocations from the total number of collocations in text.
[3] Component Pascal programming language is a refined version of Pascal, Modula 2, and Oberon. BlackBox Component Builder is an Integrated Development Environment for Component Pascal from Oberon Microsystems Inc., http://www.oberon.ch.
[4] Rich Text Format files (rft, doc) are not yet supported by the development framework.
[5] The computation has been done on a Pentium IV PC (2.4 GHz, 512 MB RAM).

References

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

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

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

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

Gale, W. and Church, K. 1991. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75--102.

Grishman, R. and Sterling, J. 1994. Generalizing automatically generated selectional patterns. In *Proceedings of COLING-94*, Kyoto, Japan, pp. 742—747.

Krenn, B. and 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*, Toulouse, France, pp. 39--46.

Laenzlinger, C. and Wehrli, E.. 1991. Fips, un analyseur interactif pour le français. *TA informations*, 32(2):35--49.

Dekang, L. 1998. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, Montreal, pp. 5--63.

Nerima, L., Seretan, V. and Wehrli, E. 2003. Creating a Multilingual Collocation Dictionary from Large Text Corpora. In *Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics* (EACL'03). Budapest, Hungary, pp. 424--431.

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

Seretan V., Nerima, L. and Wehrli, E. 2003. Extraction of Multi-Word Collocations Using Syntactic Bigram Composition. In *Proceedings of International Conference on Recent Advances in NLP (RANLP-2003)*, Borovets, Bulgaria, pp. 131-138.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

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