Syntactic concordancing and multi-word expression detection

V. Seretan*

Institute for Language, Cognition and Computation,
Human Communication Research Centre,
University of Edinburgh,
10 Crichton Street, Edinburgh EH8 9AB, UK
Fax: +44-131-651-3435
E-mail: violeta.seretan@gmail.com
*Corresponding author

E. Wehrli

Language Technology Laboratory,
Department of Linguistics,
University of Geneva,
5 rue de Candolle, CH-1211 Geneva, Switzerland
Fax: +41-22-379-7931
E-mail: eric.wehrli@unige.ch

Abstract: Concordancers are tools that display the contexts of a given word in a corpus. Also called key word in context (KWIC), these tools are nowadays indispensable in the work of lexicographers, linguists, and translators. We present an enhanced type of concordancer that integrates syntactic information on sentence structure as well as statistical information on word cooccurrence in order to detect and display those words from the context that are most strongly related to the word under investigation. This tool considerably alleviates the users' task, by highlighting syntactically well-formed word combinations that are likely to form complex lexical units, i.e., multi-word expressions. One of the key distinctive features of the tool is its multilingualism, as syntax-based multi-word expression detection is available for multiple languages and parallel concordancing enables users to consult the version of a source context in another language, when multilingual parallel corpora are available. In this article, we describe the underlying methodology and resources used by the system, its architecture, and its recently developed online version. We also provide relevant performance evaluation results for the main system components, focusing on the comparison between syntax-based and syntax-free approaches.

Keywords: collocation extraction; collocations; concordancers; key word in context; KWIC; lexical acquisition; lexical resources; linguistic analysis; multilingualism; multi-word expression detection; multi-word expressions; MWE; natural language processing; NLP; parallel concordancing tool; syntactic analysis; syntactic concordancing; syntax; translation; word cooccurrence.

Reference to this paper should be made as follows: Seretan, V. and Wehrli, E. (2013) 'Syntactic concordancing and multi-word expression detection', *Int. J. Data Mining, Modelling and Management*, Vol. 5, No. 2, pp.158–181.

Biographical notes: V. Seretan earned her BSc and MSc in Computer Science at the University of Iasi, Romania. She received her PhD in Linguistics from the University of Geneva, Switzerland in 2008. She has been a Lecturer at the University of Geneva until 2010. Presently, she is a Visiting Researcher at the Human Communication Research Centre, University of Edinburgh. Her research interests are in computational linguistics, specifically, natural language understanding, lexical acquisition, and text-to-text generation. She has published a book and over 30 articles in international journals and conference proceedings in these areas. She received the University of Geneva 2010 Latsis Prize for her PhD dissertation 'Collocation extraction based on syntactic parsing'.

E. Wehrli received her PhD in Theoretical Linguistics from McGill University in 1979. He was an Assistant Professor in Computational Linguistics at UCLA from 1984 to 1988. Since then, he is a Professor of Linguistics and Computer Science at the University of Geneva. His research interests include natural language parsing and machine translation.

This paper is a revised and expanded version of a paper entitled 'Tools for syntactic concordancing' presented at International Multiconference on Computer Science and Information Technology, Wis³a, Poland, 18–20 October 2010.

1 Introduction

Language technology plays an ever increasing role in today's information society. In this field, the issue of phraseology – knowing how words combine into larger expressions that form the building blocks of language – is of primordial importance. Indeed, identifying the lexical units that compose a text is one of the first steps that must be performed by virtually all language applications (e.g., machine translation), and is arguably a crucial step on which the final performance of these applications depends.

Importantly, knowledge of a word means knowledge of the relations that this word establishes with other words: "You shall know a word by the company it keeps!" (Firth, 1957). The study of words in context with the help of concordancers – in order to discover how words are actually used, what their typical contexts are, and what expressions they are part of – is a major concern in both theoretically and practically motivated language investigations.

The advent of the computer era and the increasing availability of texts in digital format now allow for virtually unlimited lexical exploration. Yet, this is at the same time one of the biggest issues that users presented with automatically detected word contexts inevitably face. The information comes to them as huge amounts of unstructured data, characterised by a high degree of redundancy.

To help users overcome the problem of information overload, a new generation of concordancers has been developed that is able to pre-process textual data so that the most relevant contextual information comes first (Barnbrook, 1996). This goal is achieved by using so-called lexical association measures, which quantify the degree of interdependence between words. Such measures rely, for instance, on statistical hypothesis tests, on concepts from information theory, on techniques based on data mining, or various other methods (Evert, 2004; Pecina, 2008). The main type of

information taken into account by these measures is information on word cooccurrence, i.e., how many times a given word has been observed in the context of another word.

In addition to cooccurrence information, linguistic information is nowadays increasingly used by concordancers, as a means to achieve lexical disambiguation and, thus, to provide a more structured presentation of lexical data. As such, many concordancers rely on the linguistic pre-processing of the source corpus, often consisting of *part-of-speech (POS) tagging*. This morphological analysis helps distinguishing between different categories of a word (a work like *step*, for instance, can be either a verb or a noun). At the same time, it helps organising multiple morphological variants under the same lemma (thus, *step*, *steps*, *stepped* and *stepping* are recognised as instances of the same verb, *to step*). The concordance tools that make of POS information substantially alleviate the tedious task of corpus exploration. Examples of such concordancers include SARA (Aston and Burnard, 1998), XAIRA (Burnard and Dodd, 2003), BNCweb (Hoffmann et al., 2008), and PIE (Fletcher, 2011), which are designed for accessing the POS-tagged version of the British National Corpus (BNC, 1994), or WordSmith (Scott, 2004), MonoConc (Barlow, 2004), and Wmatrix (Rayson, 2009), which can be used with other corpora as well.

In this paper, we argue that, given the state of the art in language technologies, the linguistic pre-processing of corpora could also be performed at the syntactic level, not only at the lexical level. Thus, syntactic information is another type of linguistic information that can be taken into account when exploring lexical data via concordancing. As syntactic parsers are now available for more and more languages,[1] there is a growing interest in building concordancers which rely on the syntactic pre-processing of the source corpus. The syntactic analysis helps, first of all, to deal with morphosyntactic variation (thus, contexts like *to take appropriate steps*, *the step he took*, and *which steps will be taken* can all be recognised as relevant to the same multi-word expression, *to take a step*). Moreover, syntactic analysis leads to a more precise identification of multi-word expressions, as the system takes into account the syntactic proximity instead of the linear proximity between words. Examples of concordancers which integrate syntactic information are the Sketch Engine (Kilgarriff et al., 2004), based on shallow parsing for a high number of languages, and Antidote RX (Charest et al., 2007), based on deep parsing for French.

In this article, we present a fully-fledged concordancing system, called FipsCo, which is based on a syntactic approach, similarly to Sketch Engine and Antidote RX. Its main distinctive features are that it relies on deep syntactic parsing, as opposed to shallow parsing as in Sketch Engine; it is multilingual, unlike Antidote RX; and it provides functionalities for parallel concordancing and for multi-word expression translation. The system has been designed as a translation aid tool to be integrated in the workbench of translators from an international organisation. Also, it is integrated into a larger language processing framework developed in our laboratory, and is used as a tool which supports the semi-automatic acquisition of lexical resources needed for various NLP applications.

The article is structured as follows. In Section 2, we review the related work devoted to syntax-based corpus exploration, and, in particular, to exploration aimed at detecting complex lexical units. In Section 3, we introduce the notion of *collocation*, which is central in relation to corpus-based language studies and concordancing. In Section 4, we outline the architecture and main components of our concordancing system, FipsCo. We proceed by describing in Section 5 the underlying resources and methods used.

We present details on the collocation extraction and collocation translation methods, while paying particular attention at the manner in which a multilingual deep syntactic parser is used to detect the most flexible types of collocations involving long-range dependencies. We also provide details about the evaluation of these methods. In Section 6, we introduce FipsCoWeb, the recently developed online version of the system. Finally, in Section 7 we show how FipsCo and FipsCoWeb are integrated into the larger language processing environment of LATL – the Language Technology Laboratory of the University of Geneva – and in the last section we provide concluding remarks.

2 Related work

As mentioned in the previous section, an example of syntax-based concordancer is the Sketch Engine (Kilgarriff et al., 2004). This tool performs a shallow syntactic analysis of the source corpus and applies an association measure derived from pointwise mutual information (Church and Hanks, 1990) in order to produce, for a given word, its 'sketch' – a one-page summary of its grammatical and collocational behaviour. Figure 1 shows part of the 'sketch' obtained for the verb *reach* by considering BNC as the source corpus. Fixed-size concordance lines can be displayed for each of the collocating words, which, in turn, are ordered by syntactic type (verb-object, subject-verb, verb-adverb, etc.) and by association score.

The shallow syntactic analysis of the corpus relies on automatically assigned POS tags for words. This analysis attempts to identify syntactic relations between words by taking into account a limited POS context and applying pattern matching with regular expressions defined over POS tags. The Sketch Engine is multilingual, and has recently been evaluated by expert lexicographers (Kilgarriff et al., 2010). The evaluation has been performed on English, Dutch, Japanese and Slovene data consisting of the first 20 collocates of 42 headwords (the headwords have been sampled among the nouns, verbs and adverbs with a high, medium and low corpus frequency). It was found that about two thirds of the evaluated results are suitable for inclusion in a dictionary; however, the evaluation disregarded the syntactic label – that is, a pair like *reach – agreement* was considered as a true positive even if the identified grammatical relation was wrong (e.g., subject-verb instead of verb-object).

Another syntactic concordancer is Antidote RX (Charest et al., 2007). The underlying syntactic analysis is performed by a dedicated deep parser, capable of recovering word pairs in a syntactic relation even when the words involved are distant in text due to syntactic transformations. The association measure used is log-likelihood ratio (Dunning, 1993). The pairs with a high association score have been manually validated by expert lexicographers. They are displayed by the system along with usage samples, which have also been manually validated. This system is available for French and has been used to compile a comprehensive dictionary of French cooccurrences. From about 850,000 candidate pairs with a high score, 800,000 pairs have been retained, together with 880,000 corpus examples. Thus, the system precision is very high (around 95%); however, the data evaluated only represents a very small fraction of the total 11 million pairs initially retrieved from the 500 million word corpus used.

Figure 1 Sketch Engine – partial results obtained from the BNC for the verb *reach* (see online version for colours)

object	15134	6.5
agreement	933	9.43
conclusion	494	9.29
peak	253	8.6
stage	365	8.06
final	174	8.0
height	163	7.88
climax	107	7.76
decision	345	7.58
destination	95	7.45
age	271	7.44
top	175	7.43
point	439	7.42
compromise	92	7.41
semi-final	86	7.36
consensus	74	7.07
target	120	7.07
level	335	7.01
summit	77	7.0
maturity	63	6.9
limit	91	6.84
settlement	80	6.66
bottom	67	6.57
end	222	6.52
significance	62	6.47
proportion	78	6.45

subject	5877	4.7
agreement	104	6.61
conclusion	27	5.74
radiation	14	5.67
news	41	5.63
train	24	5.39
convoy	9	5.33
total	16	5.29
temperature	18	5.23
unemployment	17	5.21
staircase	8	5.06
hand	95	5.03
Guy	7	4.9
Lindsey	6	4.88
inflation	11	4.88
flame	8	4.84
ambulance	7	4.76
rain	12	4.75
jury	7	4.61
lordship	6	4.56
rumour	6	4.53
oxygen	6	4.53
compromise	6	4.53
consensus	6	4.49
finger	16	4.48
arm	28	4.46

modifier	2584	0.4
finally	109	8.72
eventually	78	8.39
far	113	8.34
across	41	8.14
easily	57	7.68
yet	59	7.62
inside	23	7.58
almost	85	7.48
nearly	32	7.37
soon	46	7.33
home	52	7.08
now	154	6.76
forward	40	6.75
never	118	6.74
quickly	31	6.69
ever	51	6.69
only	148	6.66
to	16	6.63
already	64	6.55
last	11	6.49
even	71	6.46
reportedly	9	6.36
barely	10	6.28
seldom	7	6.03
actually	35	6.02

According to several reports – e.g., Smadja (1993) – even a precision of 40% would be acceptable for lexicographic purposes. In contrast, for the purpose of using the raw (non-validated) results in other NLP applications, a higher precision is arguably desirable, particularly as far as the identified syntactic labels are concerned.

In addition to work on syntax-based concordancers, there have also been efforts devoted to building general frameworks for syntax-based corpus exploration (though not necessarily from the perspective of detecting complex lexical units): CQP – Corpus Query Processor (Christ, 1994), CQPweb (CQPweb, 2008), and LSE – Linguist's Search Engine (Resnik and Elkiss, 2005).

In the next section, we introduce the concept of *collocation*, which is at the heart of our own approach of syntax-based exploration of (multilingual parallel) corpora for identifying multi-word expressions and, possibly, their translation equivalents.

3 Collocations

Collocation is a very important linguistic concept, and at the same time one which is quite difficult to characterise and for which there is no unique commonly accepted definition in the present literature. For our purpose, we henceforth adopt the following practically-driven definition: "A collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon" (Evert, 2004).

Consequently, the concept of collocation is allowed here to encompass all syntactic word combinations found in a corpus which are relevant to the studied word from a lexicographic point of view. In accordance with Lea and Runcie (2002), we consider that *collocation* refers, broadly, to "the way words combine in a language to produce natural-sounding speech and writing".

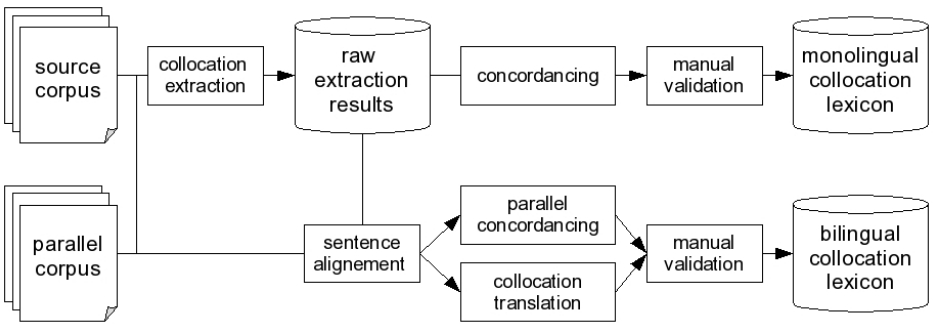
Therefore, this concept may overlap with concepts which correspond to specific subtypes of multi-word expressions usually not referred to as collocations, such as compounds (e.g., *wheel chair*), phrasal verbs (e.g., *to ask* [*somebody*] *out*) or certain types of less figurative idioms (*to open the door* [*for sth*] "to allow [sth] to happen"). Multi-word expressions, in turn, are understood as "idiosyncratic interpretations that cross word boundaries" (Sag et al., 2002).

It is well-known that the boundaries between the different subtypes of multi-word expressions are particularly difficult to draw, and they are rather fuzzy (McKeown and Radev, 2000). From a practical point of view, all multi-word expressions pose similar processing problems, regardless of any finer-grained theoretical classification. Therefore, since our approach is practical, we will refer to the output of our system as *collocations*,[2] without making more elaborate distinctions. For a detailed discussion on the relation between collocations and other types of multi-word expressions in general, and idioms in particular, the interested reader is referred to Seretan (2011).

4 FipsCo: system architecture

The architecture of our syntax-based concordancer, FipsCo, is depicted in Figure 2. In this section we outline each of the processing modules: collocation extraction, concordancing, manual validation, sentence alignment, parallel concordancing and collocation translation.

Figure 2 FipsCo syntax-based concordancer – system architecture



4.1 Collocation extraction

The first processing module performs the task of collocation extraction from the source corpus. The input text is first syntactically analysed with a (full) deep parser, then it is processed using standard methods which measure the strength of association between syntactically-related words.

This module includes a file selection sub-module, which allow users to define the source corpus that will be processed. A corpus is, from a practical point of view, a collection of files in a given folder that satisfies user-defined criteria, based on: file location within the given folder, file name, file type, or file last modification date.

The collocation extraction module iteratively processes all the files in the selection, sentence by sentence. Collocation candidates are identified from the parse trees and are incrementally added to previous results until an extraction session ends. At the end of the extraction session, several processing statistics are computed for the source corpus, which are derived from parsing information (e.g., the total number of tokens, sentences, and sentences with a complete parse). Then, the candidates identified are ranked according to the association measure chosen by the user. The measure proposed by default is log-likelihood ratio (Dunning, 1993), which is argued to be particularly appropriate to sparse lexical data. The collocation extraction method is further detailed in Section 5.2.

4.2 Concordancing

The raw extraction results obtained are visually presented to the user by the concordancing module, which implements complex display functionalities. The system displays the collocation candidates extracted from text corpora organised into *collocation types* (as opposed to *tokens*). A type conflates all the instances of a candidate pair that have been identified in the source corpus. The collocation types are partitioned into syntactically homogeneous classes, e.g., adjective-noun, adverb-verb, subject-verb, verb-object, etc. They are ranked in the reverse order of their association strength, as given by the particular association measure chosen. Also, collocation types can be filtered according to criteria based on syntactic relation, association score, corpus frequency, composing words, or rank in the output list.

Thanks to the compressed and systematic presentation of syntactically homogeneous lexical data, the users are able to access the most representative contextual information of a word by only consulting a limited amount of text. The data presented to them is manageable, since it is organised and, to a certain extent, free of redundancy – in fact, the possibly very numerous corpus instances (tokens) of a collocation are 'hidden' under a single result (type). Provided that there is lexicographic interest in a type, the users may then display the corresponding tokens to study collocations in context. The concordancer displays various information about the collocation visualised, such as its syntactic type, its association score, its rank in the output list, as well as its status relative to a lexicon of validated collocations.

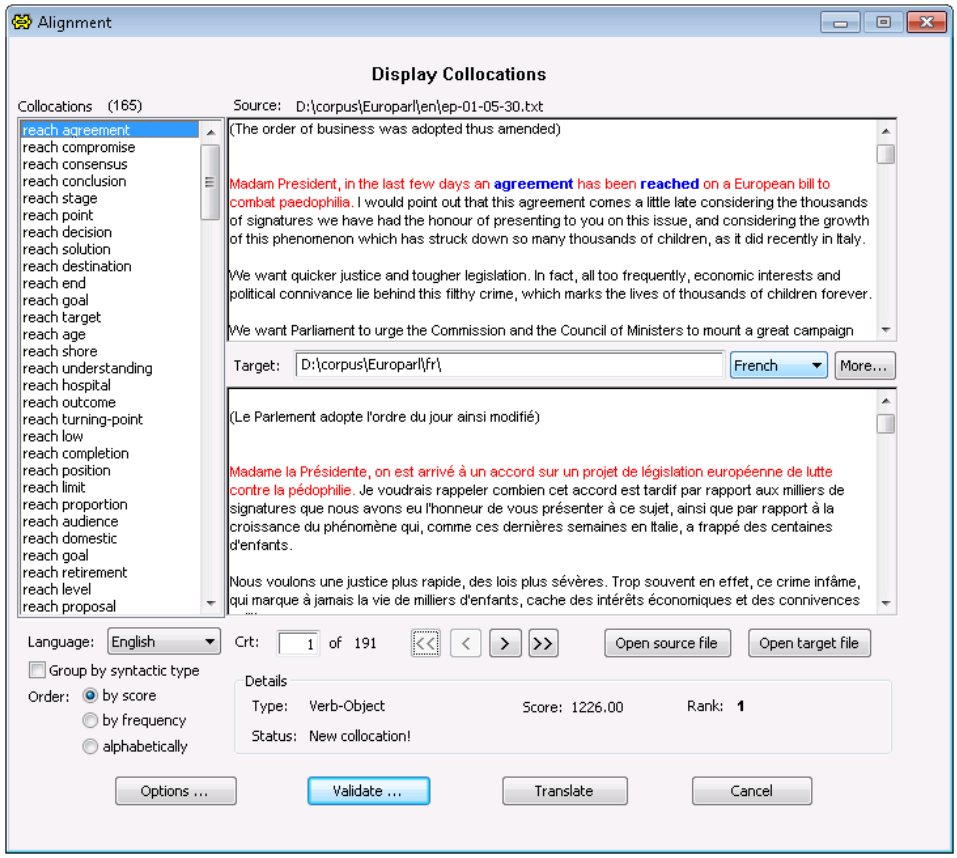
4.3 Manual validation

The manual validation module provides functionalities that allow users to create and maintain a list of manually validated collocations selected among the collocations visualised by the (parallel) concordancing module. An entry contains basic information

about a collocation, such as the collocation keywords, lexeme indexes for the participating items, syntactic type, association score, and corpus frequency. Monolingual entries also contain the source sentence of a particular instance, which provides a naturally-occurring usage sample for the collocation. Bilingual entries store, in addition, the target sentence found via alignment, and the translation proposed for the collocation.

Additional information related to a particular collocation instance is stored, namely, the name of the source and target files, the file position of the collocation's items in these files, as well as the start position of the source and target sentences. This information, which is useful for recovering a larger context, is automatically filled in by the system. The entries in the list of collocations that have been validated can be updated, deleted, or saved by the user in a monolingual and in a bilingual collocation lexicon, which can then be interfaced with other NLP systems.

Figure 3 FipsCo – screen capture of the parallel concordancing interface, displaying filtered results for the verb *reach* (see online version for colours)



4.4 Parallel concordancing

Collocations are a key factor in producing good quality translations (Orliac and Dillinger, 2003). Knowledge of how collocations are translated into another language

is crucial, as collocations usually cannot be translated literally. Parallel concordancing systems access multilingual parallel corpora in order to give users the possibility to discover translation equivalents for collocations. Our concordancing system include functionalities for parallel concordancing, which are implemented by this module.

Figure 3 displays the interface of the parallel concordancer, showing English-French parallel text related to the collocation *reach – agreement*. The collocation extraction results displayed on the left-hand side have been filtered so that they contain the sought word (in this case, the verb *reach*) in a specific syntactic configuration (here, the configuration retained is verb-object). According to the association measure applied, *agreement* is the object that is most strongly related to the verb *reach*, on the basis of observations made in the Europarl corpus (Koehn, 2005). The verb-object combination *reach – agreement* has been detected 191 times in this corpus, and it is, indeed, a lexicographically interesting combination.

The parallel concordancing interface shown in Figure 3 displays the first instance of this collocation type in the corpus. The other instances can be visualised using the browsing arrows. The interface also presents the automatically retrieved French translation of the source sentence, which allow users to find a translation equivalent (here, *arriver à un accord*, lit., *to arrive at an agreement*). In a translation environment, multilingual parallel corpora are typically available from translation archives. Thus, by using a parallel concordancer, translators working on a new document may conveniently identify and reuse existing collocation translations.

4.5 Sentence alignment

When multilingual parallel corpora are available, the target sentence containing the translation of the visualised source sentence is automatically detected and displayed through parallel concordancing, next to the source sentence. The role of the sentence alignment module is to detect this target sentence on-the-fly, i.e., when users actually visualises the source sentence. The parallel corpora do not need to be pre-aligned, although it should be possible to adapt the system so that it can use pre-aligned corpora as well.

First, the target file corresponding to the source file is found by applying file name mapping rules. Once the target file has been found, the sentence that is likely to be the translation of the source sentence is identified using an in-house sentence alignment method (Nerima et al., 2003; Seretan, 2008). This method consists of comparing the relative lengths of paragraphs in the source and target documents in order to find a paragraph alignment, and then assuming a 1:1 match for sentences inside a paragraph. The target paragraph is selected as the paragraph whose surrounding paragraphs best fit, in terms of relative lengths, the surrounding of the source paragraph. Compared to other alignment methods – for an account, see, for instance, Véronis and Langlais (2000) – this method is fast and does not require a macro-structural pre-alignment of the documents (at section or paragraph level). Its precision is around 90% on documents relatively difficult to align, which do not preserve the paragraph structure between versions.[3]

4.6 Collocation translation

This module is used to automatically detect a translation for the extracted collocations, when multilingual parallel corpora are available. The collocation translation method

makes use of the target sentence contexts found via sentence alignment, the monolingual collocation extraction method (described in Section 5.2), and, optionally, of bilingual dictionaries. It applies a series of filters on the collocations extracted from the target contexts until eventually a single target collocation is retained, which is proposed as a translation equivalent. More details on this method are provided in Section 5.5.

The FipsCo system is freely available for research as a stand-alone tool for the Windows operating system. Its latest developments include the creation of a light-weight online version, FipsCoWeb, which is presented in Section 6.

5 Underlying resources and methodology

In this section, we present the resources used by our system as well as the methods designed for extracting the most relevant expressions with a given word from a corpus and for detecting their translation equivalents in a multilingual parallel corpus.

5.1 Syntactic parsing

The main resource on which the system relies is Fips, a multilingual symbolic parser developed over the past decade in our laboratory (Wehrli, 2007). This parser is based on generative grammar concepts inspired by *The Minimalist Program* (Chomsky, 1995), the simpler syntax model (Culicover and Jackendoff, 2005), and LFG – lexical functional grammar (Bresnan, 2001). Fips can be characterised as a strong lexicalist, bottom-up, left-to-right parser. For each input sentence, it builds a rich structural representation, which combines:

- the constituent structure
- the interpretation of constituents in terms of arguments
- the interpretation of elements like clitics, relative and interrogative pronouns in terms of intra-sentential antecedents
- co-indexation chains linking extraposed elements (e.g., fronted NPs and *wh*-elements) to their canonical positions.

According to the theoretical stipulations on which the parser relies, some constituents of a sentence may move from their canonical 'deep' position to surface positions, due to various grammatical transformations. For instance, in the case of the French sentence shown in example (1), it is considered that the noun *objectif* moved from its original position of direct object into the surface position of subject due to a passivisation transformation. The parser keeps track of this movement by linking the (empty) object position of the verb *atteint* to the extraposed noun, *objectif*. In the normalised sentence representation it builds, the parser identifies this noun as the 'deep' direct object of the verb. Therefore, it succeeds in detecting the combination *atteindre-objectif* as a verb-object pair.

(1) Son *objectif* à long terme ne sera pas *atteint* facilement.

As illustrated by this example, our system is able to cope with syntactic variation, thanks to the syntactic analysis provided by the parser. The sentence normalisation helps to abstract away from the particular surface realisations, and, thus, to detect those

collocation instances in the corpus that are more flexible from a syntactic point of view. The role of the syntactic pre-processing is crucial, as it allows the concordancer to highlight syntactically well-formed combinations and to provide highly accurate extraction results (performance evaluation results are reported in Section 5.3).

The parser is available for English, French, Spanish, Italian, Greek and German. A number of other languages are currently under development, including Romanian, Romansch, Russian and Japanese. Fips is designed as a generic parsing architecture, coupling a language-independent parsing engine with language-specific extensions. The language-independent part implements the parsing algorithm, which is based on three main types of operations:

1 *Project*: assignment of constituent structures to lexical entries

2 *Merge*: combination of adjacent constituents into larger structures

3 *Move*: creation of chains by linking surface positions of extraposed ('moved') constituents to their corresponding canonical positions.

The language-specific part of Fips consists of grammar rules of a given language and a detailed lexicon for that language. In the formalism used by Fips, the role of most grammar rules is to specify the conditions under which two adjacent constituents may be merged into a larger constituent by an attachment operation (*Merge*). The lexica are manually built and contain information such as selectional preferences, subcategorisation, and syntactico-semantic features which are useful for informing the syntactic analysis. Currently, for each of the languages supported there are around a hundred attachment rules defined, while the number of lexemes in the lexicon ranges from about 14,000 (for Greek) to about 57,000 (for English), and is in average almost 36,000. The number of inflected forms varies greatly from one language to another – for instance, there are about 103,000 word forms for English and more than 443,000 for German.

The construction of the lexicon is supported by a morphological generation tool which creates appropriate lexical entries corresponding to a specified inflection paradigm (when applicable). Unlike other parsers, Fips does not require POS-tagged data as input. The POS is assigned to words during analysis, on the basis of the lexical information and the particular parsing hypotheses which are pursued.

It has been shown (Wehrli, 2007) that the precision of the lexical analysis is very high when the parser manages to produce a complete syntactic analysis. For English and French, for instance, this happens for approximately 80% of the sentences from corpora consisting of newspaper articles. The quality of the syntactic analysis has recently been measured in the framework of two parsing evaluation campaigns for French, EASy and PASSAGE.[4] Conclusive results are not yet available, but a separate task-based evaluation performed in a machine translation setting suggested that in case of complete analysis, the identification of lexical items is very good, as well as the identification of the arguments of predicates (i.e., verbs, predicative adjectives, etc.). For other types of attachments, such as prepositional phrase, as adjuncts or as modifiers of nouns, the results are clearly not at the same quality level.

Insofar as the grammatical coverage is concerned, the parser can cope with a very large number of constructions, but some are still problematic, such as complex coordination, enumeration and some cases of parenthetical structures. Ellipsis is not handled at all at the moment.

5.2 Collocation extraction method

As already stated in Section 4, the corpus-based detection of the most important words related to a given word is performed in our system using a collocation extraction method that combines the syntactic analysis provided by the the Fips parser with standard statistical methods able to pinpoint the most strongly associated syntactic combinations.

Thus, in the first extraction step, collocation candidates are identified as combinations of lexical items in predefined syntactic configurations from each sentence of the corpus, by traversing the sentence structure built by the parser. A high number of configurations are taken into account, e.g., for English, adjective-noun, noun-[predicate]-adjective, noun-noun, noun-preposition-noun, noun-preposition, adjective-preposition, subject-verb, verb-object, verb-preposition-argument, verb-preposition, verb-adverb, adverb-adjective, and noun-coordination-noun.[5]

In the second extraction step, the candidates are ranked according to their association score, as computed using association measures such as log-likelihood ratio (Dunning, 1993). Our system implements a wide range of other measures that can be used to rank collocation candidates; this measure is proposed by default since it is a well-established measure for collocation extraction.

The output of the extractor is represented by a so-called *significance list*, which contains at the top the candidates that are most likely to actually constitute collocations. A cutoff point can be applied by the user to the results, in order to retain only the candidates with higher scores. Typically, extraction systems also apply a frequency threshold to eliminate those combinations which occur only few times in the corpus. This is considered necessary because statistical measures are unreliable for low frequency data – more precisely, for combinations occurring less than five times in the corpus (Evert, 2004). However, we opted for keeping all the candidate data without applying a frequency threshold, since relevant collocations may also be found among combinations occurring only very few times in the corpus.

A frequency threshold can be applied after extraction, depending on the specific measure chosen by users and on the intended use of the data. Another reason for our choice is that the syntactic analysis provides a strong filter on the otherwise huge amount of candidate data, making the statistical computation tractable. In the systems which do not rely on syntax, high frequency cutoffs are mainly used as a means to alleviate the computation by drastically reducing the amount of candidate data to process.

5.3 Collocation extraction evaluation

The strength of our extraction method, compared with existing work which is most usually based on POS tagging (Church and Hanks, 1990; Smadja, 1993; Daille, 1994; Justeson and Katz, 1995), chunking (Krenn and Evert, 2001; Evert, 2004) or shallow parsing (Kilgarriff et al., 2004; Tutin, 2004), comes from the detailed sentence analysis provided by deep parsing. The global interpretation provided for the input sentences allows our system to be more sensitive to the morphosyntactic context in which collocations occur, compared to shallow parsers, which risk to make wrong decisions favouring local attachments.

Consider, for instance, the words *question* and *asked* in the corpus sentence below:

(2) The *question asked* if the grant funding could be used as start-up capital to develop this project.

Lacking an analysis for a larger text span, shallow parsers would wrongly infer from the sentence fragment *the question asked* that *question* and *asked* are in a verb-object relation. This is because they typically assign to such sentence fragments a passive interpretation ("the question that was asked by somebody"), instead of the correct active interpretation ("the question asks").

It has been argued (Kilgarriff et al., 2010) that the correctness of grammatical information may be of lesser practical relevance, since the mere presence of the component words in a collocation is sufficient for lexicographers to spot that collocation and consider it for inclusion in a lexicon. Nonetheless, the correctness of this information is essential for the language applications which make use of the raw (non-validated) collocation extraction results.

So far, collocation extraction methods based on syntactic parsing have generally been dismissed by language practitioners, since they are considered as too time-consuming and unreliable on unrestricted data, as opposed to syntax-free approaches which are seen as readily available, robust, and able to produce satisfactory results when applied to large data.

To evaluate the performance of our extractor based on deep parsing, we performed several cross-language evaluation experiments in which we compared the two approaches, i.e., the syntax-based approach vs. the syntax-free approach (Seretan, 2008). For example, in an experiment conducted on English, French, Italian and Spanish data consisting of 2,000 pairs sampled from different levels of the significance list – ranging from top to 10% of the list – we measured the extraction precision by taking into account reference annotations produced by expert linguists. The precision has been computed using three different criteria. First, for reporting the grammatical precision, a pair is considered as a true positive if it is grammatically well-formed. For the multi-word-expression precision, we require it to be judged as interesting from a lexicographic point of view. Finally, for the collocational precision, we require it to comply with the specific collocation definition provided in the framework of the Meaning-Text Theory (Mel'čuk, 1998; Polguère, 2003).

The difference observed in the performance of the two approaches compared is substantial. The precision obtained was, in all three cases, around 2.5 times higher when parsing information was used. The highest difference was observed for the grammatical precision, which is 2.7 times higher (88.8% vs. 33.2%); the multi-word-expression precision is 2.5 times higher (43.2% vs. 17.2%); and the collocational precision is 2.6 times higher (32.9% vs. 12.8%). Note that the precision figures are lower than those reported in related work (Section 2). This is due, first, to the different manner of sampling data by considering pairs from a larger testbed including low-frequency results, and second, to the stricter evaluation criteria requiring all true positives to have correctly assigned grammatical information.

In addition to precision, recall has also been comparatively evaluated but on a smaller dataset, since a larger evaluation would have required specific annotation resources which are not currently available. Recall has been measured at instance level rather than at type level, as this is arguably a more appropriate strategy (Diab and Bhutada, 2009; Fritzinger et al., 2010). The results obtained on a dataset of

602 instances are 99% for the syntax-based method and 90.2% for the syntax-free method. The token-based analysis pointed out relative strengths and weaknesses of the two approaches: using a parser leads to some instances being missed because of the inherent mistakes, but not using it leads to more numerous instances being missed because of syntactic variation. The large difference observed in the performance of the two approaches allows us to conclude that a syntactic approach is worth pursuing.

5.4 Extracting complex collocations

A peculiarity of the FipsCo system is that it is able to identify complex collocations, which can be seen as structures containing embedded collocations: for instance, *reach a turning point* is a complex collocation of verb-object type, which contains an embedded adjective-noun collocation, *turning point*.

The detection of such complex collocations is particularly useful in the case of non-decomposable compounds and that of compositional expressions containing nested compounds. In these cases, it is important to highlight the whole expression rather than its sub-parts (Frantzi and Ananiadou, 1996). For instance, *genetically modified organisms* is a compound, and it is desirable to output it as a whole rather than only the sub-part *modified organisms*. The expression *second world war* is more compositional, as *world war* is also a collocation on its own; however, it is still desirable to eliminate *second war* from the extraction results, if it only occurs as part of the longer expression *second world war*.

The method of detecting complex collocations, described in detail in Seretan et al. (2003) and Nerima et al. (2010), is in principle similar to the method used for extracting binary collocations. It consists of treating already extracted collocations as single lexical items, detecting combinations of these collocations in a parse tree, and applying standard association measures on the resulting combinations. Thus, our approach relies on the recursive nature of collocations, which has, in fact, been remarked by theoretical studies (Heid, 1994) but until now not used in practice.

Previous related work has generally focused on the detection of *n-grams*, i.e., rigid sequences or words (Choueka et al., 1983; Smadja, 1993). There are also a few exceptions of works that detect less rigid combinations including verbs, by making use of parsed data (Blaheta and Johnson, 2001; Zinsmeister and Heid, 2003). However, these are limited to combinations made up of three items and to specific configurations, such as verb-preposition-preposition in English (Blaheta and Johnson, 2001) or adjective-noun-verb in German (Zinsmeister and Heid, 2003).

In contrast, the method we developed is more general, as it allows detecting collocations of virtually unlimited length through iterative embedding of increasingly longer collocations; also, it is not tailored to specific configurations, but allows every configuration actually observed in the parsed (and normalised) sentence representation. Our concordancer includes a dedicated interface for displaying complex collocations, which is similar to the interface shown in Figure 3.

5.5 Collocation translation method

Collocations are semantically transparent expressions, unlike idioms that are more semantically opaque (e.g., *to be over the moon* ‘to be happy’, *to kick the bucket* ‘to die’). Therefore, they do not seem to pose problems for translation. Yet, like idioms, collocations are unpredictable for the non-native speaker, this is why they are called

'idioms of encoding' (Makkai, 1972; Fillmore et al., 1988). In contrast, semantically opaque expressions are also 'idioms of decoding'. Like idioms, collocations have to be known in advance in order to be used as a means of providing text fluency. The particular problem posed by collocations is that they are very numerous (Mel'čuk, 1998).

The literal translation of collocations is in most cases comprehensible, but it is often felt as inappropriate. For instance, the French collocations *grande attention, grande diversité, grande vitesse* cannot be translated literally, as **big attention*, **big diversity*, and **big speed*. These examples show that the choice of the right word to use in the target language is often a subtle process, depending on the collocating word.

Our concordancing system include a component module which attempts to detect translation equivalents for collocations by exploiting existing translation archives and using the strategy presented below.

Step 1 Given a multilingual parallel corpus, collocation extraction is performed for the source language and, for each collocation, a limited number of corpus sentences is retrieved (this number was set to 50 in our experiments).

Step 2 The source sentences are aligned using the method described in Section 4.5, and a list of target sentences is obtained for each source collocation.

Step 3 Collocation extraction is performed for the target language from the list of sentences, and a list of potential translations is obtained for each source collocation.

Step 4 A matching is performed between each source collocation and the corresponding list of potential translations, and the item of the list that is most likely to constitute a valid translation is returned as output.

The matching process used in Step 4 consist of applying a series of filters on the list of potential translations, which gradually reduce their number until a single item is retained, which will be proposed as translation. These filters are described below.

1 The first filter is related to the syntactic configuration of the target collocations, and retains only those pairs that have a compatible configuration. For instance, a verb-object collocation in English (e.g., *achieve consensus*) may be mapped into either a verb-object or a verb-preposition-argument collocation in French (*établir consensus, parvenir à consensus*).

2 The second filter is optional and is based on bilingual dictionary information. It retains only those pairs that contain the translation of the collocation base, where the *base* refers to the semantically autonomous item of a collocation (Hausmann, 1989; Polguère, 2003). Unlike the other item, called *collocate*, the base allows a literal translation. For instance, in the case of the French collocation *grande vitesse*, the base *vitesse* is translated into English literally as *speed*; in constrast, the collocate *grande* does not have a straightforward translation and, in fact, the adjective *high* is used instead of *big*.

3 Finally, the third filter considers the most frequent among the remaining pairs, or, in case of ties, the pair with the highest association score. The pair selected through this sequence of filters is proposed as translation.

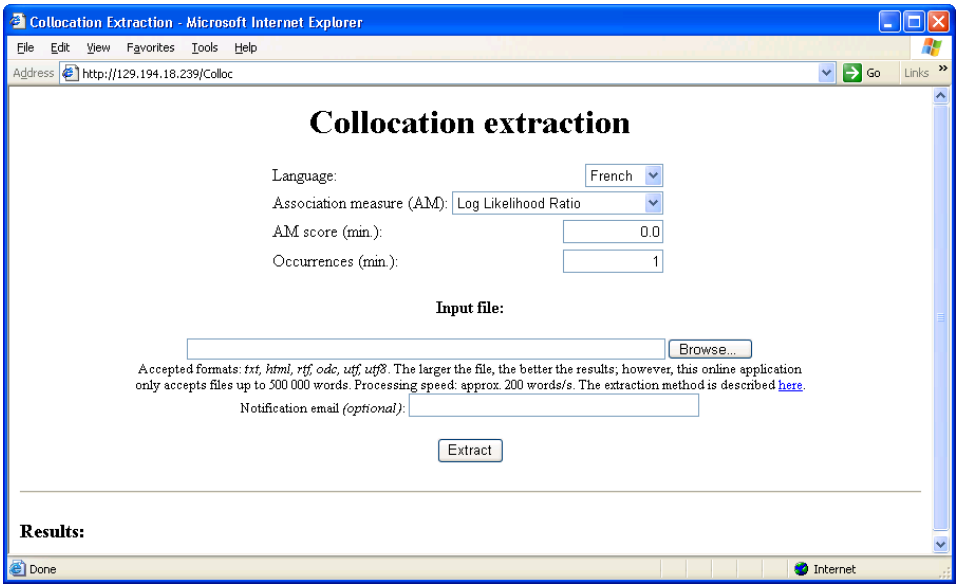
This translation method has been evaluated on data from the Europarl corpus (Koehn, 2005) involving 4,000 collocations and a total of 8 language pairs. It achieved competitive performance (89.9% precision and 70.9% recall), despite its simplicity. We believe that this is due to the availability of syntactic information for both the source and target languages.

Related work was mostly focused on the translation of noun phrases (Kupiec, 1993; Dagan and Church, 1994). A method for translating flexible collocations involving verbs has been proposed by Smadja et al. (1996), which relies on a statistical correlation metric in order to perform the matching. The precision achieved varies between 61% and 77%, depending on the corpus frequency of collocations. This method requires further post-processing of the target collocations in order to decide the correct word order. In our case, this is not necessary thanks to the normalisation provided by the parser. Another method has been proposed by Lü and Zhou (2004). It extracts collocations from both source and target corpora using a syntax-based approach like ours, and then relies on a statistical translation model to perform the matching. Unlike our method, this method does not require parallel corpora, but the precision obtained is relatively low (51% to 68.2%, depending on the syntactic configuration).

6 The online version: FipsCoWeb

One of the latest developments of the syntactic concordancer described in the previous sections is related to the creation of an online version, called FipsCoWeb. Figure 4 shows the current interface of the web application built. In this section, we describe the functionalities it provides for server-side collocation extraction and client-side display of the results.

Figure 4 FipsCoWeb – screen capture of the online tool interface (see online version for colours)



FipsCoWeb allows the user to upload a file and to set the most important processing and visualisation parameters, such as the association measure, the cutoff score, and the frequency threshold. After the processing is completed on the server side using the method presented in Section 5.2, the user is presented with the results, as shown in Figure 5. The user then has the possibility to apply different parameters, to select a given syntactic configuration, and to see actual collocation instances retrieved in the input document. As in the stand-alone version of the system, the words of a collocation are displayed in the sentence context, and they are highlighted for readability (see Figure 6).

Figure 5 FipsCoWeb – screen capture showing sample collocation extraction output (see online version for colours)

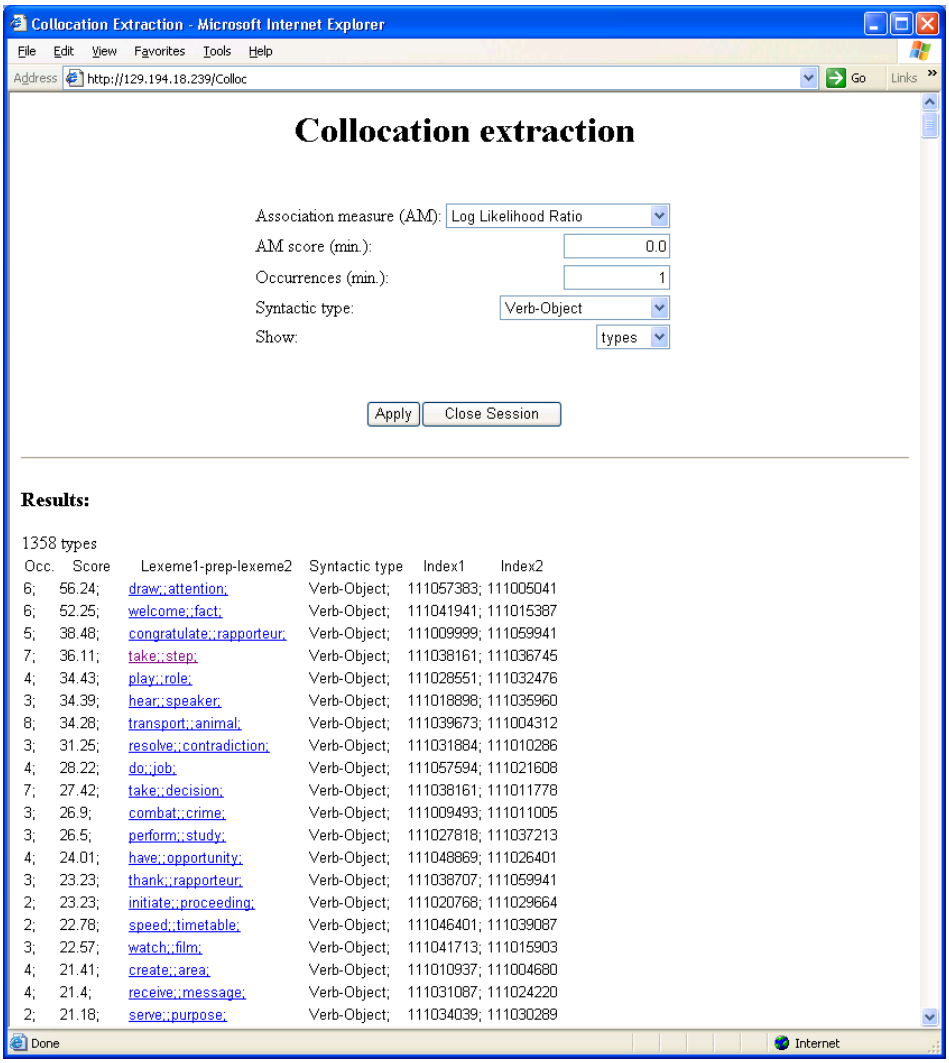
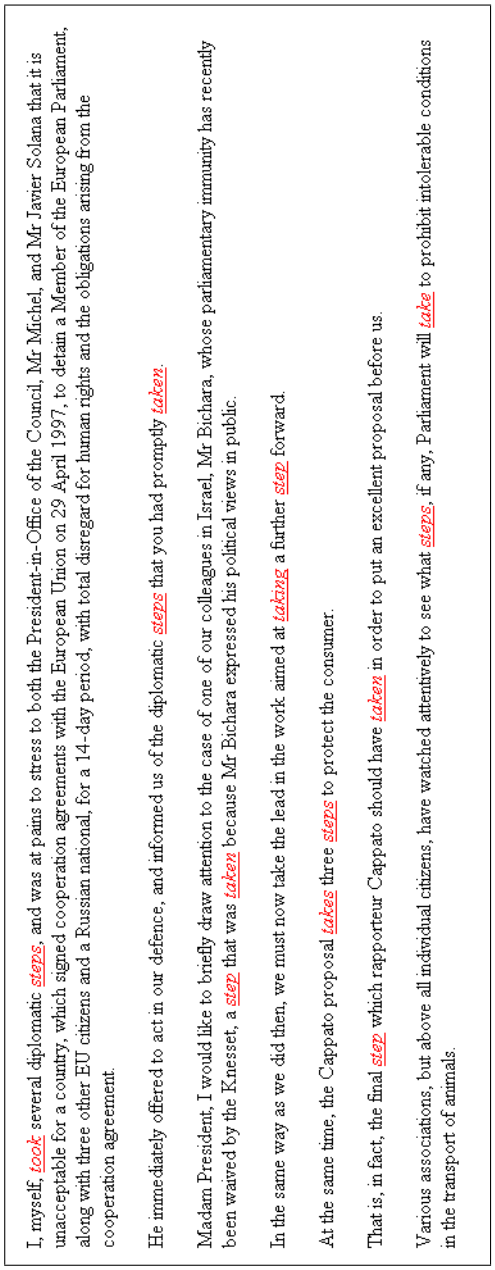


Figure 6 FipsCoWeb – screen capture showing collocation instances in their original context (see online version for colours)


I, myself, *took* several diplomatic *steps*, and was at pains to stress to both the President-in-Office of the Council, Mr Michel, and Mr Javier Solana that it is unacceptable for a country, which signed cooperation agreements with the European Union on 29 April 1997, to detain a Member of the European Parliament, along with three other EU citizens and a Russian national, for a 14-day period, with total disregard for human rights and the obligations arising from the cooperation agreement.

He immediately offered to act in our defence, and informed us of the diplomatic *steps* that you had promptly *taken*.

Madam President, I would like to briefly draw attention to the case of one of our colleagues in Israel, Mr Bichara, whose parliamentary immunity has recently been waived by the Knesset, a *step* that was *taken* because Mr Bichara expressed his political views in public.

In the same way as we did then, we must now take the lead in the work aimed at *taking* a further *step* forward.

At the same time, the Cappato proposal *takes* three *steps* to protect the consumer.

That is, in fact, the final *step* which rapporteur Cappato should have *taken* in order to put an excellent proposal before us.

Various associations, but above all individual citizens, have watched attentively to see what *steps*, if any, Parliament will *take* to prohibit intolerable conditions in the transport of animals.

FipsCoWeb currently allows users to upload files containing a maximum of 500,000 million words. This represents a sufficient amount of data for allowing corpus exploration, which is still manageable for an online application. However, smaller documents lead to a more rapid server response time. Since in most cases it is not realistic to obtain an instantaneous server response due to the processing time required,

users have the possibility to enter an e-mail address to which a link to the results is sent when the server-side computation is completed. The extraction results are stored on the server and can be consulted later, unless users explicitly clear them.

As for technical details, the web version of the system has been implemented in Component Pascal using BlackBox IDE.[6] The main motivation for this choice comes from the fact that the stand-alone system as well as the syntactic parser themselves have been developed on this platform, which is very stable, produces extremely efficient and compact code, and provides good tools for building graphical interfaces. (Needless to say, any advanced object-oriented platform might have been used for this project.)

The web server itself runs as a BlackBox program,[7] which makes the integration between the various software modules easier. However, this server has certain limitations as far as the parallel processing of large files is concerned. This is why we plan to migrate to a different web server. We also plan to implement the online version of the concordancer as a web service, possibly on a distributed software platform.

7 Integration into a larger language processing framework

The syntax-based concordancer we presented in this article is part of a larger language processing framework developed in our laboratory, which integrates several other language tools and resources related to the two main tasks undertaken, i.e., natural language analysis and machine translation.

The corpus-based study of words and their collocates has, in our case, a particular practical motivation. The collocations detected using our method based on syntactic parsing are manually validated and added to the lexical database of the parser. They are used as information which guides the sentence analysis performed by the parser, as described in Wehrli et al. (2010). Syntactic parsing is again used to detect collocations from new corpora, and so on, in a cyclic process.

The collocation translations acquired either manually or automatically are used for expanding the bilingual lexicon of a rule-based machine translation which is based on the parser. In addition, once these collocations have been added to the lexicon, they are used in two applications of terminology assistance, TWiC and TwicPen (Wehrli et al., 2009). Given a word and the sentence in which it occurs, these applications analyse the sentence, disambiguate the syntactic category of the word, and then look up the lexicon in order to retrieve the entry or sub-entry that is compatible with the context in which the word occurs. If the selected word is part of a multi-word expression, then the systems returns a translation for the whole expression, in addition to the translation of the word in isolation.

Thus, the two applications can be seen as context-sensitive dictionaries. The first application, TWiC, is a web browser plug-in which is automatically activated when the user selects a word on a web page. It retrieves the sentence in which the selected word occurs, then automatically detects its language and performs its syntactic analysis using Fips. A pop-up window then displays the translation of the disambiguated word in the target language chosen by the user.

The second application, TwicPen, is a similar tool which is designed for the readers of printed (off-line) material, instead of online material. The readers can select a text span – e.g., a sentence, sentence fragment, or paragraph – by using a hand-held scanner connected to their PDA or personal computer. The tool interface allows them to advance

word by word in the text and see the translation of each word in context, using the same method which is used by the first tool. In addition, the readers may display the translation for the whole text span selected, and compare it with the translation proposed by Google Translate.[8]

8 Conclusions

Concordancers are tools that provide a concise presentation for large amounts of lexical data that are nowadays available. The study of words in context with the help of concordancing tools is important from both theoretical and practical perspectives. As such, the information on typical word collocates presented by monolingual concordancers is used, for instance, as raw material in the compilation of lexical databases, tailored either for the end user or for client applications. Moreover, parallel concordancers are particularly important as they transform existing translation archives into valuable bilingual resources exploitable for future translations.

Presenting lexical information in a way that is useful for users or machines inevitably means dealing with complex phenomena that characterise language data, such as ambiguity, sparseness, and the dispersion due to morphosyntactic variation. In order to deal with these issues, modern concordancers integrate a linguistic pre-processing component, activated prior to the typical statistical computation component. Most of the times, the linguistic pre-processing component is, however, limited to morphological analysis – i.e., POS tagging – and it does not take into account the syntactic relation between words, since these are relatively difficult to obtain for large amounts of unrestricted text.

Nonetheless, robust syntactic parsers recently became operational for an increasing number of languages. Consequently, there is a growing interest in creating syntax-based concordancing systems. In this article, we described such a system that relies on syntactic information provided by a deep multilingual parser in order to detect and display collocations with a given word. The system, called FipsCo, has been developed over the past several years at LATL, the Language Technology Laboratory of the University of Geneva. It provides complex functionalities, such as search and filtering, identification of collocations made up of more than two words, parallel concordancing for multilingual parallel corpora, automatic translation of collocations, and creation of a lexical database which can be used as a translation memory database.

Resorting to a 'deep' syntactic parser rather than a shallow parser – or, worse, a simple window of a few words – to identify collocations provides more useful and accurate results and a more appropriate means of corpus exploration. In particular, it makes it possible to recognise collocations even when the composing words are separated by several words or phrases, or when they do not occur in the expected order, as in the case of the example we provided earlier.

In this article, we introduced our approach to syntax-based concordancing and presented a stand-alone concordancing system as well as its recently developed online version. These tools are part of a larger processing framework dedicated to the processing of multi-word expressions, and are being used to compile lexical resources necessary for the two main long-term NLP projects pursued in our laboratory, namely, a multilingual symbolic parser and a machine translation system based on parsing.

We compared our syntax-based approach against syntax-free approaches, and found that there are substantial differences in their performance in favour of the syntax-based approach, both in terms of recall and precision. In particular, we found that the increase in precision is more than double for all the criteria used for judging result pairs: grammaticality, lexicographic interest, and compliance with the Meaning-Text-Theory collocation definition. These results are in line with previous results obtained when using syntactic information for tasks such as term extraction (Maynard and Ananiadou, 1999), semantic role labelling (Gildea and Palmer, 2002), and semantic similarity computation (Padó and Lapata, 2007).

Future work directions include, in particular, pursuing the goal of interoperability between our own language processing framework and other similar frameworks, extending the concordancing system for the languages for which a parser module is currently under development, as well as further exploring the complex interrelation between multi-word expressions and language processing modules.

Acknowledgements

This work has been partly supported by the Swiss National Science Foundation (grant no. PA00P1_131512). The authors would like to thank the three anonymous reviewers, whose comments and suggestions helped improve the article.

References

Aston, G. and Burnard, L. (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edinburgh.

Barlow, M. (2004) 'MonoConc', available at http://www.athel.com/mono.html (accessed on March 2011).

Barnbrook, G. (1996) *Language and Computers: A Practical Introduction to the Computer Analysis of Language*, Edinburgh University Press, Edinburgh.

Blaheta, D. and Johnson, M. (2001) 'Unsupervised learning of multi-word verbs', in *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, Toulouse, France, pp. 54–60.

BNC (1994) 'British National Corpus', available at http://www.natcorp.ox.ac.uk/corpus/ (accessed on March 2011).

Bresnan, J. (2001) *Lexical Functional Syntax*, Blackwell, Oxford.

Burnard, L. and Dodd, T. (2003) 'Xara: an XML aware tool for corpus searching', in *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster, UK, pp.142–144.

Charest, S. et al. (2007) 'Élaboration automatique d'un dictionnaire de cooccurrences grand public', in *Actes de la 14e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, Toulouse, France, pp.283–292.

Chomsky, N. (1995) *The Minimalist Program*, MIT Press, Cambridge, Mass.

Choueka, Y. et al. (1983) 'Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus', *Journal of the Association for Literary and Linguistic Computing*, Vol. 4, No. 1, pp.34–38.

Christ, O. (1994) 'A modular and flexible architecture for an integrated corpus query system', in *Proceedings of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX'94)*, Budapest, Hungary, pp.23–32.

Church, K. and Hanks, P. (1990) 'Word association norms, mutual information, and lexicography', *Computational Linguistics*, Vol. 16, No. 1, pp.22–29.

CQPweb (2008) 'CQPweb', available at http://cqpweb.lancs.ac.uk/ (accessed on March 2011).

Culicover, P. and Jackendoff, R. (2005) *Simpler Syntax*, Oxford University Press, Oxford.

Dagan, I. and Church, K. (1994) '*Termight*: identifying and translating technical terminology', in *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP)*, Stuttgart, Germany, pp.34–40.

Daille, B. (1994) 'Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques', PhD thesis, Université Paris 7.

Diab, M.T. and Bhutada, P. (2009) 'Verb noun construction MWE token supervised classification', in *2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation, Applications*, Suntec, Singapore, pp.17–22.

Dunning, T. (1993) 'Accurate methods for the statistics of surprise and coincidence', *Computational Linguistics*, Vol. 19, No. 1, pp.61–74.

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

Fillmore, C. et al. (1988) 'Regularity and idiomaticity in grammatical constructions: the case of *let alone*', *Language*, Vol. 64, No. 3, pp.501–538.

Firth, J.R. (1957) *Papers in Linguistics 1934–1951*, Oxford Univ. Press, Oxford.

Fletcher, W.H. (2011) 'Phrases in English: online database for the study of English words and phrases', available at http://phrasesinenglish.org (accessed on March 2011).

Frantzi, K.T. and Ananiadou, S. (1996) 'Extracting nested collocations', in *Proceedings of the 15th International Conference on Computational Linguistics (COLING'96)*, Copenhagen, Denmark, pp.41–46.

Fritzinger, F. et al. (2010) 'A survey of idiomatic preposition-noun-verb triples on token level', in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Gildea, D. and Palmer, M. (2002) 'The necessity of parsing for predicate argument recognition', in *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, pp.239–246.

Hausmann, F.J. (1989) 'Le dictionnaire de collocations', in F. Hausmann, O. Reichmann, H. Wiegand and L. Zgusta (Eds.): *Wörterbücher: Ein internationales Handbuch zur Lexicographie. Dictionaries, Dictionnaires*, pp.1010–1019, de Gruyter, Berlin.

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

Hoffmann, S. et al. (2008) *Corpus Linguistics with BNCweb – A Practical Guide*, Peter Lang, Frankfurt am Main.

Justeson, J.S. and Katz, S.M. (1995) 'Technical terminology: some linguistic properties and an algorithm for identification in text', *Natural Language Engineering*, Vol. 1, No. 1, pp.9–27.

Kilgarriff, A. et al. (2004) 'The Sketch Engine', in *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, pp.105–116.

Kilgarriff, A. et al. (2010) 'A quantitative evaluation of word sketches', in *Proceedings of the 14th EURALEX International Congress*, Leeuwarden, The Netherlands.

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

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 Collocation: Computational Extraction, Analysis and Exploitation*, Toulouse, France, pp.39–46.

Kupiec, J. (1993) 'An algorithm for finding noun phrase correspondences in bilingual corpora', in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, USA, pp.17–22.

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

Lü, Y. and Zhou, M. (2004) 'Collocation translation acquisition using monolingual corpora', in *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, Barcelona, Spain, pp.167–174.

Makkai, A. (1972) *Idiom Structure in English*, Mouton, The Hague.

Maynard, D. and Ananiadou, S. (1999) 'A linguistic approach to terminological context clustering', in *Proceedings of Natural Language Pacific Rim Symposium '99*.

McKeown, K.R. and Radev, D.R. (2000) 'Collocations', in R. Dale, H. Moisl and H. Somers (Eds.): *A Handbook of Natural Language Processing*, pp.507–523, Marcel Dekker, New York, USA.

Mel'čuk, I. (1998) 'Collocations and lexical functions', in A.P. Cowie (Ed.): *Phraseology Theory, Analysis, and Applications*, pp.23–53, Claredon Press, Oxford.

Nerima, L. et al. (2003) 'Creating a multilingual collocation dictionary from large text corpora', in *Companion Volume to the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, Hungary, pp.131–134.

Nerima, L. et al. (2010) 'A recursive treatment of collocations', in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

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

Orliac, B. and Dillinger, M. (2003) 'Collocation extraction for machine translation', in *Proceedings of Machine Translation Summit IX*, pp.292–298, New Orleans, Lousiana, USA.

Padó, S. and Lapata, M. (2007) 'Dependency-based construction of semantic space models', *Computational Linguistics*, Vol. 33, No. 2, pp.161–199.

Pecina, P. (2008) 'Lexical association measures: collocation extraction', PhD thesis, Charles University in Prague.

Polguère, A. (2003) *Lexicologie et sémantique lexicale. Notions fondamentales*, Presses de l'Université de Montréal, Montréal.

Rayson, P. (2009) 'Wmatrix: a web-based corpus processing environment', available at http://ucrel.lancs.ac.uk/wmatrix (accessed on March 2011).

Resnik, P. and Elkiss, A. (2005) 'The Linguist's Search Engine: an overview', in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan, pp.33–36.

Sag, I.A. et al. (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.

Scott, M. (2004) *WordSmith Tools Version 4*, Oxford University Press, Oxford.

Seretan, V. (2008) 'Collocation extraction based on syntactic parsing', PhD thesis, University of Geneva.

Seretan, V. (2011) *Syntax-based collocation extraction (Text, Speech and Language Technology)*, Springer, Dordrecht.

Seretan, V. et al. (2003) 'Extraction of multi-word collocations using syntactic bigram composition', in *Proceedings of the Fourth International Conference on Recent Advances in NLP (RANLP-2003)*, pp.424–431.

Smadja, F. (1993) 'Retrieving collocations from text: Xtract', *Computational Linguistics*, Vol. 19, No. 1, pp.143–177.

Smadja, F. et al. (1996) 'Translating collocations for bilingual lexicons: a statistical approach', *Computational Linguistics*, Vol. 22, No. 1, pp.1–38.

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

Véronis, J. and Langlais, P. (2000) 'Evaluation of parallel text alignment systems: the ARCADE project', in J. Véronis (Ed.): *Parallel text processing: Alignment and Use of Translation Corpora*, Text, Speech and Language Technology Series, pp.369–388, Kluwer Academic Publishers, Dordrecht.

Wehrli, E. (2007) 'Fips, a 'deep' linguistic multilingual parser', in *ACL 2007 Workshop on Deep Linguistic Processing*, Prague, Czech Republic, pp.120–127.

Wehrli, E. et al. (2009) 'On-line and off-line translation aids for non-native readers', in *Proceedings of the International Multiconference on Computer Science and Information Technology*, Mragowo, Poland, pp.299–303.

Wehrli, E. et al. (2010) 'Sentence analysis and collocation identification', in *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, Beijing, China, pp.27–35.

Zinsmeister, H. and Heid, U. (2003) 'Significant triples: adjective + noun + verb combinations', in *Proceedings of the 7th Conference on Computational Lexicography and Text Research (Complex 2003)*, Budapest, Hungary.

Notes

1 The parsing field is making steady progress, particularly through the development of language-independent frameworks for dependency parsing, such as Nivre (2006).

2 The output consists, more precisely, of *collocation candidates*, unless these candidates are validated.

3 On such 'difficult' documents, the precision of alignment methods which are almost perfect on 'normal' text may decrease as low as 65% (Véronis and Langlais, 2000).

4 Available at http://www.elda.org/easy/ and http://atoll.inria.fr/passage/ (accessed on March 2011).

5 Examples for each of these configurations are: *heavy smoker, effort [be] devoted, suicide attack, round of negociations, inquiry into, crazy about, war breaks, meet requirement, bring to boil, point out, fully support, highly important, nice and warm*. The configuration list grows as more and more corpus data is explored.

6 Available at http://www.oberon.ch (accessed on March 2011).

7 O_3-web-application-framework (WAF), available at http://o3-software.de/ (accessed on March 2011).

8 Available at http://translate.google.com/ (accessed on March 2011).