Collocation extraction based on syntactic parsing

Seretan, Violeta

UNIVERSITÉ DE GENÈVE
Département de linguistique

FACULTÉ DES LETTRES
Professeur E. Wehrli

Collocation Extraction Based on Syntactic Parsing

THÈSE

présentée à la Faculté des lettres de l'Université de Genève
pour obtenir le grade de Docteur ès lettres

par

Violeta Seretan

Thèse N° 653

GENÈVE
2008

La Faculté des lettres, sur le préavis d'une commission composée de MM. les professeurs Jacques MOESCHLER, président du jury; Eric WEHRLI, directeur de thèse; Christian BOITET (IMAG, Grenoble); Ulrich HEID (IMS, Stuttgart); Paola MERLO (Genève) autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 9 juin, 2008

Le Doyen: Eric WEHRLI

Thèse N° 653

To my daughter, Nadia Giulia,
for sweet collocation
during thesis writing

Abstract

Pervasive across texts of different genres and domains, collocations (typical lexical associations like *to wreak havoc, to meet a condition, to believe firmly, a deep concern, highly controversial*) constitute a large proportion of the multi-word expressions in a language. Due to their encoding idiomaticity, collocations are of paramount importance to text production tasks. Their recognition and appropriate usage is essential, for instance, in Foreign Language Learning or in Natural Language Processing applications such as machine translation and natural language generation. At the same time, collocations have a wide applicability to tasks concerned with the opposite process of text analysis.

The problem that is tackled in this thesis is the automatic acquisition of accurate collocational information from text corpora. More specifically, the thesis provides a methodological framework for the syntax-based identification of collocation candidates in the source text, prior to the statistical computation step. The development of syntax-based approaches to collocation extraction, which has traditionally been hindered by the absence of appropriate linguistic tools, is nowadays possible thanks to the advances achieved in parsing. Until now, the absence of sufficiently robust parsers was typically circumvented by applying linear proximity constraints in order to detect syntactic links between words. This method is relatively successful for English, but for languages with a richer morphology and a freer word order, parsing is a prerequisite for a good performance.

The thesis proposes (and fully evaluates on data in four different languages, English, French, Spanish and Italian) a core extraction procedure for discovering binary collocations, which is based on imposing syntactic constraints on the component items

instead of linear proximity constraints. This procedure is further employed in several methods of advanced extraction, whose aim is to cover a broader spectrum of collocational phenomena in text. Three distinct but complementary extension directions have been considered in this thesis: extraction of n-ary collocations ($n > 2$), data-driven induction of collocationally relevant syntactic configurations, and collocation mining from an alternative source corpus, the World Wide Web. The possibility to abstract away from the surface text form and to recover, thanks to parsing, the syntactic links between discontinuous elements in text, plays a crucial role in achieving highly efficient results.

The methods proposed in this study were adopted in the development of an integrated system of collocation extraction and visualization in parallel corpora, a system which was intended to enrich the workbench of translators or other users (e.g., terminologists, lexicographers, language learners) wanting to exploit their text archives. Finally, the thesis gives an example of a practical application that builds on this system in order to further process the extracted collocations, by automatically translating them when parallel corpora are available.

Résumé

Largement présentes dans les textes de tout genre et de tout domaine, les collocations se taillent la part du lion dans l'inventaire des expressions à mots multiples d'une langue. En raison du caractère idiomatique de leur encodage, les collocations jouent un rôle primordial dans la production de texte (par exemple, dans des applications telles que la traduction automatique et la génération de texte, ainsi que dans l'apprentissage de langues étrangères). En même temps, elles présentent également un grand intérêt du point de vue de l'analyse du texte.

Le sujet de cette thèse est l'acquisition automatique, basée sur corpus, de ressources lexicales collocationnelles. Dans la thèse, nous proposons des méthodes précises pour l'identification de collocations candidates, qui s'appuient sur une analyse syntaxique détaillée du texte source effectuée préalablement aux calculs statistiques. Le développement des approches syntaxiques pour l'extraction a été jusque maintenant entravé par l'absence d'outils d'analyse appropriés, mais cette situation est en train de changer grâce aux progrès réalisés dans le domaine du parsing. Si pour l'anglais le manque d'analyseurs suffisamment robustes a pu être compensé partiellement par l'application de contraintes de proximité linéaire afin de détecter des liens syntaxiques entre les mots, pour d'autres langues il est impératif de faire appel au parsing afin d'obtenir des résultats d'extraction fiables.

Cette thèse propose (et évalue sur des données en 4 langues, l'anglais, le français, l'espagnol et l'italien) une procédure principale d'extraction de collocations binaires qui se base sur l'application de la contrainte de proximité syntaxique aux éléments d'une collocation candidate, à la place de la contrainte de proximité linéaire qui est la plus répandue dans les travaux existants. Cette procédure est ensuite utilisée dans

la conception de plusieurs méthodes d'extraction avancées, qui visent à couvrir un éventail plus large de phénomènes collocationnels dans le texte. Ces méthodes vont dans trois directions distinctes, mais complémentaires : l'extraction de collocations n-aires ($n > 2$), l'induction dirigée par les données des configurations syntaxiques appropriées aux collocations, et la détection de collocations à partir d'un corpus alternatif, le Web. La possibilité de faire abstraction de la forme superficielle du texte et celle de récupérer, grâce au parsing, les liens syntaxiques entre des éléments lexicaux qui ne sont pas forcement contigus dans le texte jouent un rôle déterminant dans l'obtention de résultats performants.

La méthodologie présentée a été adoptée dans la création d'un système intégré d'extraction et de visualisation de collocations dans des corpus multilingues, destiné à enrichir l'environnement de travail des traducteurs, et, en général, celui des utilisateurs intéressés à exploiter leurs archives textuelles (terminologues, lexicographes, ou apprenants de langues étrangères). La thèse présente aussi une application basée sur ce système, qui traite les résultats d'extraction afin d'effecter leur traduction automatique lorsque des corpus parallèles sont disponibles.

Acknowledgements

I feel very fortunate of having found at LATL, the Language Technology Laboratory of the University of Geneva, a great working environment in which I could conduct my PhD studies during the past years. I am most grateful to my supervisor, Eric Wehrli, who gave me the opportunity to join his team and work on a fascinating topic—first, in the framework of the project "Linguistic Analysis and Collocation Extraction", then as part of my thesis. Eric guided my work with his vast knowledge and infinite patience. I appreciated his good advices, his constant support and quick feedback, as much as I appreciated the freedom he gave me to develop my own ideas, or his permanent good mood and optimism. Thank you very much!

Since part of my thesis work took already shape during the project I mentioned, I wish to gratefully acknowledge financial support from the RUIG-GIAN organisation. I also want to thank our project partner, Olivier Pasteur, for his active involvement in this project and for suggestions on the design of our system. Together with him, with Luka Nerima and Eric Wehrli, we had lengthy regular meetings which set the framework of my thesis. Luka further accompanied my work with his expertise in database technology, and was constantly available for discussions. Thank you for your contribution!

During these years, I had the opportunity to present my work in many local or non-local meetings. The feedback I received from colleagues during research seminars and doctoral summer schools, from invited researchers in our lab, from fellow conference attendees and, particularly, from the anonymous reviewers of my papers, left a definite mark on my work. I wish to thank all the people involved for their invaluable help.

Many of my colleagues in the Department of Linguistics provided me with useful comments and suggestions on various occasions. I received valuable criticism, in particular, from (in office-room order): Paola Merlo, Gabriele Musillo, Eric Joanis, Gerold Schneider, Jacques Mœschler, Christopher Laenzlinger, and Antonio Leoni. My computer-scientist office mates, Jean-Philippe Goldman, Mar Ndiaye, and Yves Scherrer, always gave me good advices and were ready to lend me a hand. JP made many useful suggestions for the extraction tool, especially in the early stage. Mar tested this tool several times, and together with Yves and Antonio, helped me annotate the multilingual data I used in evaluation (my husband Vincenzo did not escape the torture; he had to annotate the Italian data).

Other colleagues have contributed to my work in different ways. Occasionally, Gabriela Soare, Stephanie Durrleman-Tame, Catherine Walther Green and Genoveva Puskas helped me proofread my papers. I am grateful to them, and in general to all the members of our Department for maintaining a stimulating working environment. But more than a team, I found here a family and friends. Since Eva Capitão is our friendship hub, I wish to thank her in particular for her kindness, as well as for her support with various administrative matters.

A different type of support came from the Commission of Equality of our University, which granted me a temporary exemption from teaching so that I could focus on my thesis. During the last stages of thesis writing, Vincenzo's help was essential, as he spend so much time taking care of our new-born daughter, Nadia Giulia. Thank you both for the joy you are bringing into my life!

Special thanks are also due to my Romanian family for love and encouragements; to my first teacher, Vasile Frumos, for making me love simplicity and rigor; and to Dan Cristea from the University of Iaşi, for introducing me to Computational Linguistics.

Extra special thanks go to my friends around the world, who persistently inquired about the progress of my PhD. I am also indebted to Alex Clark, Laura Hasler and Eric Joanis, who proofread a previous version of this manuscript.

Finally, I wish to express my gratitude to Christian Boitet, Ulrich Heid and Paola Merlo, who kindly agreed to be part of the thesis committee, and to Jacques Mœschler, who accepted the role of president of the jury.

Contents

Chapter 1

Introduction

This thesis is about using advanced syntactic parsing technology for enhancing the extraction of collocations from text corpora. The work described here relies on a deep multilingual parser for the design of a versatile extraction engine, whose core part is based on the identification of syntactically-related combinations of words that may (or may not) occur within a short space of each other in text. Contrary to common extraction approaches, it makes use of a syntactic proximity criterion instead of a linear proximity criterion in order to identify collocation candidates in text. This chapter provides a general introduction to the work described in the thesis, by presenting its motivation, stating its main objectives, and giving an overview of the thesis content.

1.1 Motivation

A large part of the vocabulary of a language is made up of *phraseological units* or *multi-word expressions*, i.e., complex lexemes that have "idiosyncratic interpretations that cross word boundaries" (Sag et al., 2002, 2). The importance of these units has been widely recognized both in theoretical linguistics, in which phraseology was recently established as a field of research of its own (Cowie, 1998), and in computational linguistics, where a growing attention is nowadays paid to their recognition and proper treatment by language applications.

Phraseological units cover a wide range of phenomena, including compound nouns (*dead end*), phrasal verbs (*[to] ask out*), idioms (*[to] lend somebody a hand*), and collocations (*sharp contrast, daunting task, widely available, [to] meet a requirement*). According to numerous studies, collocations appear to constitute a large proportion of these units. Virtually, any sentence contains at least one collocation (Howarth and Nesi, 1996; Pearce, 2001a). As put by Mel'čuk (1998, 24), "collocations make up the lion's share of the phraseme [phraseological unit] inventory, and thus deserve our special attention".

While an agreed-upon definition of collocations does not yet exist, they are usually understood as typical combinations of words that differ from regular combinations since the items involved co-occur in a short span of text more often than the chance would predict. Unlike idioms, their meaning is easy to decode; nonetheless, they constitute "idioms of encoding" (Makkai, 1972; Fillmore et al., 1988), since they are unpredictable for the non-native speaker and, in general, do not preserve the meaning of (all of) their components across languages. Compare, for instance, the French collocations displayed in the first column of Table 1.1 with their English counterpart shown in the third column; the literal translation displayed in the second column corresponds to "unnatural readings", called "anti-collocations" (Pearce, 2001a, 43).

French	**Literal translation (anti-collocation)**	**English (correct translation)**
accuser retard	accuse delay	experience delay
établir distinction	establish distinction	draw distinction
gagner argent	win money	make money
relever défi	raise challenge	take up challenge
poser question	put down question	ask question

Table 1.1: Collocations across languages.

The past decades have witnessed significant advances in the work of automatic acquisition of collocations from text corpora, which was aimed, in the first place, at providing lexicographic support. Boosted by the advent of computer era and the development of corpus linguistics, but also by the contextualism current that put emphasis on the study of words in context ("You shall know a word by the company it

keeps!" (Firth, 1957, 179)), this work has led to the development of impressive corpus-based dictionaries which include collocations, like COBUILD, the Collins Birmingham University International Language Database (Sinclair, 1995). Moreover, a couple of dictionaries exist that are entirely devoted to collocations (see Appendix A). In their compilation, lexicographers increasingly rely on automatic extraction methods in order to validate or complement their intuition with corpus evidence.[1]

The collocational information acquired from corpora is crucial for some major Natural Language Processing (NLP) applications, like the ones dealing with text production. For instance, in machine translation and in natural language generation, collocations are claimed to be the key factor in producing more acceptable output (Orliac and Dillinger, 2003, 292). They are considered not only useful, but a real problem in these tasks (Heylen et al., 1994, 1240). The importance of collocations stands in their prevalence in language—"L'importance des collocations réside dans leur omniprésence" (Mel'čuk, 2003, 26)—, while the difficulty in handling them comes, principally, from their ambiguous linguistic status, from their ambivalent position at the intersection of lexicon and grammar, and from the lack of precise criteria for their identification.

In NLP, there is also a marked interest in collocations from the opposite perspective of text analysis, as they proved useful in a variety of tasks. For instance, in parsing they were used for solving attachment ambiguities by giving preference to analyses involving collocations (Hindle and Rooth, 1993; Alshawi and Carter, 1994; Berthouzoz and Merlo, 1997; Wehrli, 2000). In word sense disambiguation, collocations were used for discriminating between senses of polysemous words (Brown et al., 1991b; Yarowsky, 1995) in virtue of the "one sense per collocation" hypothesis, according to which words have a strong tendency to exhibit only one sense in a given collocation (Yarowsky, 1993). Analogously, they were used in information retrieval (Ballestros and Croft, 1996; Hull and Grefenstette, 1998), text classification (Williams, 2002), and topic segmentation (Ferret, 2002). Collocations are also considered helpful for a wide range of other applications, from speech recognition

[1]Some collocation dictionaries also display corpus frequency information in their entries, e.g., the Dictionary of English Collocations (Kjellmer, 1994).

and OCR, where homophonic/homographic words can be decided between by taking into account their collocates (Church and Hanks, 1990), to context-sensitive dictionary look-up, where they can help in identifying the dictionary subentry that best matches the context of the word sought (Michiels, 2000).

In most of the existing extraction approaches, collocations are modeled as sequences of adjacent POS-tagged lemmas (i.e., in the n-gram method), or as interruptible lemma pairs found in a short window of text (i.e., in the window method). Only exceptionally the syntactic structure of the source text is taken into account by extraction procedures, in order to:

- capture the collocation pairs that occur in text in a form that is different from the canonical form,[2] or whose items are not found in the text in the immediate vicinity of each other, like the pair *[to] have – impact* occurring in the corpus excerpt shown in Example (1):

  (1) The communication devotes no attention to the *impact* the newly announced policy measures will *have* on the candidate countries.

- eliminate noisy pairs that co-occur often, but without being related, like the pair *human – organisation* in Example (2):

  (2) *human* rights *organisations*

- provide a syntactic interpretation for results, and tell distinct interpretations apart. Without syntactic information, a pair like *question asked* is ambiguous, as it can be interpreted as either an subject-verb or as a verb-object pair. As the pair instance in Example (3) belongs to a subject-verb pair type, it is wrong to associate it with a verb-object pair type (*[to] ask – question*), because it would count as a false instance.

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

[2] In the canonical form, the head word and the dependent word are next to each other and in the typical order (e.g., the subject before the verb and the object after the verb in a SVO language).

Given that most linguistic phenomena observed in a corpus are infrequent, each single instance of a collocation is important for extraction procedures. Therefore, capturing the pairs whose items are distant from each other due to the various grammatical operations the pairs may undergo (e.g., modification, apposition, passivisation, relativisation; see Section 4.2 for examples) and correctly interpreting these pairs are crucial issues that have to be addressed by extraction procedures, particularly for languages allowing a high degree of word order freedom.

1.2 Objectives

The main objective of the thesis is to take advantage of the recent advances achieved in the syntactic parsing field in order to propose an extraction methodology that is more sensitive to the morphosyntactic context in which collocations occur in the source corpora. In particular, our work aims to exploit the detailed syntactic information provided by a multilingual syntactic parser—namely, the Fips parser developed at the Language Technology Laboratory, University of Geneva (Laenzlinger and Wehrli, 1991; Wehrli, 1997; Wehrli, 2007)—as the major means to select collocation candidates in text.

The central claim of the thesis is that by using the syntactic proximity criterion instead of the linear proximity criterion in choosing candidate pairs, a substantial improvement can be gained in the quality of extraction results, both in terms of recall and precision: more collocations types and collocation instances are retrieved from the source corpus, and the noise of extraction is, at the same time, reduced. This claim was corroborated by the evaluation experiments performed for multiple languages in a variety of settings, which compared the precision obtained by the syntax-method on the one hand, and the window method on the other, as well as by a number of case studies that measured the recall at the instance level. A subclaim, here, is that a syntactic approach to collocation extraction is feasible, i.e., the syntactic analysis of large corpora of (unrestricted) natural text is possible; this is indeed the case, since the parser used is sufficiently robust and fast.

In addition, it is claimed that the use of a syntactic filter on the candidate data has

a positive impact on the statistical measures of association strength which are used for ranking the candidate pairs according to their likelihood to constitute collocations. This is achieved, firstly, by adopting a particular design strategy that consists in applying association measures on syntactically homogeneous material, and, secondly, by providing these measures with frequency information on pairs that is more accurate, thanks to the elimination of false instances. As shown by our case-study evaluations, the presence of false instances which are associated by syntactically-uninformed extraction methods with certain pair types may lead to the artificial promotion of spurious pairs to higher positions in the output list, at the expense of interesting pairs. As a consequence, the quality of extraction results is ultimately affected.

It is also argued that the syntactic information provided for candidate pairs by parsing leads to the easier interpretation of extraction results, both by users (e.g., lexicographers) consulting these results with the help of concordance tools, since they can group the collocations according to their syntactic type, and by NLP applications, since they require syntactic information in order to appropriately process the extracted collocations in one way or another. This thesis describes an application that deals with the corpus-based translation of extracted collocation, and that relies heavily on the syntactic information associated with collocation pairs.

In addition to binary collocations, the thesis focuses on the extraction of collocations made up of more than two lexemes. It is shown that the basic module of binary collocation extraction can be extended in order to efficiently extract collocations of higher arity, like *draw a clear distinction*, *reach a joint resolution*, or *proliferation of weapons of mass destruction*.[3]

The thesis is also concerned with broadening as much as possible the set of syntactic types allowed for the extraction of binary collocations, and, for this purpose, it aims to provide a means for detecting all the collocationally relevant syntactic patterns in a language. In addition to patterns like verb-object or adjective-noun that are the most represented among collocations, a wide range of patterns which

[3]The term *multi-word collocation* is used hereafter for collocations made up of strictly more than two lexemes (note that the most usual sense of *multi* is 'more than two', while in *multi-word expression* it means 'more than one').

include functional categories are also relevant as they may correspond to collocations (e.g., preposition-noun: compare for instance *on page* with *at page*, which is an anti-collocation). Based on generic head-dependency relations identified by the Fips parser, it was possible to devise a semi-automatic method for the data-driven identification of collocationally relevant patters in a given language, which was applied so far on French and English corpora.

Also, in order to deal with the problem of data sparseness that is characteristic of text corpora, the thesis aims at extending the proposed extraction methodology to an alternative resource constituting the largest available corpus, the World Wide Web. Collocates of a specific word are mined for in the matched contexts retrieved by a search engine for that word.

Finally, the research described in this thesis has a practical motivation. Much of the work presented therein is rooted in the research project "Collocation extraction and linguistic analysis" carried out between 2002 and 2004 by the Language Technology Laboratory of the University of Geneva in collaboration with the Division of Linguistic Services and Documentation of the World Trade Organisation, under the auspices of the RUIG-GIAN organisation.[4] The goal of this project was to develop a tool for the syntax-based extraction of collocation and for their visualisation in parallel corpora, allowing translators to collect collocations from existing text archives and to visualise their contexts in both the source document and, simultaneously, in its translations.

1.3 Chapters outline

The thesis content is organised as follows. Chapters 2 and 3 constitute the introductory part, with the first chapter describing the theoretical framework, and the second the practical framework in which the issue of collocations has been addressed in existing work.

Chapter 4 introduces and evaluates the proposed extraction methodology, which

[4]RUIG-GIAN (http://www.ruig-gian.org/) is an international research network whose aim is to foster collaboration between academic institutions and international organizations in Geneva.

is committed to the detailed syntactic analysis of texts prior to the application of association measures.

Chapter 5 builds on this core extraction methodology in order to further extend it in three different, but complementary directions: longer collocations, less arbitrarily-chosen syntactic patterns, and using the Web as an alternative source corpus.

Chapter 6 describes the implemented system for multilingual collocation extraction and visualisation in parallel corpora, as well as the collocation translation application developed in relation to this system.

In the concluding chapter, the main findings of our work are outlined, and a discussion of its shortcomings as well as of the suggested directions for further development is provided (Chapter 7).

1.4 Published work

Part of the work described in this thesis has been discussed in previous publications, including: the extraction methodology (Nerima et al., 2003; Seretan and Wehrli, 2006a), the extraction extensions (Seretan et al., 2003; Seretan et al., 2004a; Seretan et al., 2004c; Seretan, 2005), the evaluation (Seretan and Wehrli, 2006a), earlier versions of the system implemented (Seretan et al., 2004b), and the collocation translation application (Seretan and Wehrli, 2007). The relation to the published work is explicitly stated in the relevant places in the thesis. Apart, naturally, from the methodological framework, the overlap with these publications is insignificant. The thesis provides an updated description and complementary details on the methodology, as well as new extraction results, different evaluation experiments, or more in-depth interpretations for past data. Also, the presentation is made under a different angle, since the thesis aims to offer a global and coherent view on the work carried out.

Chapter 2

On Collocations

This chapter provides an introduction to the existing theoretical descriptions of collocations. Since the term *collocation* is generally accompanied by confusion, an attempt has been made at providing a synthetic review of definitions, while distinguishing between purely statistical vs. linguistically-motivated approaches; following recent proposals in the literature, we reserve the term *collocation* for the linguistically-motivated concept (Section 2.2). After identifying the most salient features of collocations (Section 2.3), we review the theoretical frameworks within which collocations have been accounted for (Section 2.4), and present the characterizations made by various researchers in terms of semantic compositionality and morphosyntactic properties (Section 2.5). Finally, Section 2.6 states the understanding adopted for the collocation concept in this work.

2.1 Introduction

The phenomenon of collocating words was brought to the attention of linguists in the 1930s by the British contextualist John R. Firth (1890–1960), who actually introduced the term *collocation*, derived from the Latins *locare*, 'to locate', and *cum*, 'together'.

But long before contextualism, pedagogical studies on first- and second language acquisition were already concerned with collocations, seen as language chunks that are memorized by speakers as whole units and that constitutes the major means for

achieving language fluency (Pawley and Syder, 1983). Moreover, according to Gitsaki (1996), collocations have even been known and studied by the ancient Greeks.

At the beginning of the XXth century, Harold Palmer (1877–1949), who pioneered the study of EFL (English as a Foreign Language), also noted the presence of so-called *polylogs*, or *known units* in language. He built a list of over 6,000 frequent collocations which he included in his teaching, so that students could learn them in block.

The same concern for phraseological units led his successor, Albert Sydney Hornby (1898–1978), to include collocational information in the dictionaries from the series that he initiated with ISED, the *Idiomatic and Syntactic English Dictionary* (1942), and that continued with *A learner's Dictionary of Current English* (1948), the *Advanced Learner's Dictionary of Current English* (1952), and the *Oxford Advanced Learner's Dictionary* (reprinted several times). This pedagogical trend was continued, most notably, by Anthony P. Cowie, Peter Howarth, and by Michael Lewis, who consider collocations as the "islands of reliability" of speakers' utterances (Lewis, 2000, 173).

Thus, it can be stated that collocations unveiled primarily from the pedagogical observations on language acquisition that associated them with a high level of proficiency, which can only be achieved by speakers through memorization and which is seen as a privilege reserved to native speakers. The pedagogical interest in collocations provided a strong motivation for their study, collection and analysis in the perspective of language teaching.

2.2 A survey of definitions

The most general understanding of the term *collocation*—as introduced in the framework of contextualism or described in previous linguistics studies—is that of a relation of *affinity* which holds between words and which is revealed by the typical co-occurrence of words, i.e., by the frequent appearance of words in the context of each another.[1] Bally (1909) refers to items that, while preserving their autonomy,

[1]Contextualists consider that in characterizing a word, its context—i.e., "the company it keeps"—plays the most important role: "You shall know a word by the company it keeps!" (Firth, 1957,

show an affinity which links them to each other ("conservent leur autonomie, tout en laissant voir une affinité évidente qui les rapproche"). In order to describe the lexical affinity, Coseriu (1967) uses the metaphorical expression "lexical solidarity".

This lexical affinity cannot be accounted for by regulatory language processes, since it is not explainable on the basis of grammar rules applied to word classes.

> "[the phraseme—in particular, the collocation] cannot be constructed [...] from words or simpler phrases according to general rules of [language] L, but has to be stored and used as a whole" (Mel'čuk, 1998).

While the characterisation in terms of affinity provides a good intuition for the concept of collocation, its definition remains quite vague, because nothing is said about its linguistic status and properties. Lacking a precise definition, the term *collocation* was constantly accompanied over the time by confusion, and was used in different places for denoting different linguistic phenomena. This confusion was only augmented by the examples provided by various researchers, which show a persistent disagreement.

As pointed out several times in the literature—inter alia, in Hausmann (1989), Bahns (1993), Lehr (1996)—, the understanding of the term varied with researchers' point of view. In NLP, this understanding was often subject to the desired usage of collocations in an application (Smadja, 1993; Evert, 2004; McKeown and Radev, 2000): "the definition of collocations varied across research projects" (McKeown and Radev, 2000, 523); "the practical relevance is an essential ingredient of their definition" (Evert, 2004, 75).

As Bahns puts it, "collocation is a term which is used and understood in many different ways" (Bahns, 1993, 57). But despite the diversity of understandings and points of view, it is still possible to identify two main perspectives on the concept of collocation: one which is purely statistical, and one which is more linguistically motivated. In what follows, we survey the most representative definitions of each group in chronological order.

---

179).

2.2.1 Statistical approaches

Both the pedagogical and the contextualist definitions of collocation mentioned in Section 2.1 imply a statistical component. In order to be acquired by speakers through memorisation in block, collocations should be identifiable as frequent word co-occurrences. Similarly, in contextualism collocations are described in terms of typical co-occurrence or as words that show a "tendency to occur together" (Sinclair, 1991, 71). The notions of frequency, typicality or tendency refer to features that are usually modeled in statistics.

As a matter of fact, the majority of collocation definitions adopt a statistical view. Moreover, even if the phenomenon described has a linguistic connotation, the linguistic aspects are often ignored; thus, in the purely statistical approaches to collocations, definitions are given exclusively in statistical terms. For instance, Firth (1957) notes:

(1) "Collocations of a given word are statements of the habitual and customary places of that word." (Firth, 1957, 181).

Among the examples provided by Firth, we find word pairs like *night – dark*, *bright – day*, or *milk – cow* (1957, 196). As can be noted, the understanding adopted for the collocation concept in contextualism is a broad one, since, in addition to syntagmatic association that are likely to constitute phraseological units (*dark night, bright day*), it covers non-syntagmatic associations which are semantically motivated (*milk – cow*).

The statistical view is predominant in the work of Firth's disciples, M.A.K. Halliday, Michael Hoey, and John Sinclair. The collocation is again understood in a broad sense, as the frequent occurrence of one word in the context of another (where the context means either the whole sentence, or a window of words called *collocational span*):

(2) "Collocation is the cooccurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening." (Sinclair, 1991, 170).

Even later definitions, like the following which are among the most widely used by NLP practitioners, are given exclusively in statistical terms:

(3) "The term *collocation* will be used to refer to sequences of lexical items which habitually co-occur" (Cruse, 1986, 40)

(4) "A collocation is an arbitrary and recurrent word combination". (Benson, 1990)

(5) "Natural languages are full of collocations, recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages" (Smadja, 1993, 143).

In particular, the definition provided by Smadja (1993) rests on work by Church and Hanks (1990), in which collocations are modeled with the statistical notion of significance that helps distinguish genuine word associations from associations which are due to chance: collocations are those lexical associations whose probability of co-occurrence, estimated on the basis of their co-occurrence frequency, is "much larger than chance" (Church and Hanks, 1990, 23).

A peculiarity of statistical approaches is that they regard collocations as symmetrical relations, and pay no attention to the relative importance of the words involved. Thus, Firth (1957, 196) describe collocations in terms of mutual expectation:

> "One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, its collocation with *night* [...] The collocation of a word or a 'piece' is not to be regarded as mere juxtaposition, it is an order of mutual expectancy" (Firth, 1968, 181).

Cruse (1986, 40) also considers that in a collocation, "the constituent elements are, to varying degrees, mutually selective". Similarly, Sinclair (1991, 173) notes that "collocation is one of the patterns of mutual choice". Still, he distinguishes between *upward* collocations, in which the node word (i.e., the word under examination) collocates with a word that is more frequent, and *downward* collocations, in which it combines with a less frequent word (Sinclair, 1991, 116). Thus, when the word *back* is

inspected, *back from* is an upward collocation since *from* is more frequent than *back*, and *bring back* is a downward collocation, since *bring* is less frequent than *back*.

2.2.2 Linguistic approaches

While in the contextualist (and derived) approaches, the structural relation between items in a collocation is rather ignored—as Sinclair (1991, 170) puts it, the collocation refers to "lexical co-occurrence, more or less independently of grammatical pattern or positional relationship"— in other approaches, the syntactic relationship between these items is a central defining feature.

To the general view that describe collocations in pure combinatorics terms, the linguistically-motivated approaches oppose a more restrictive view, in which collocations are seen first of all as expressions of language. These approaches emphasize the linguistic status of collocations, considering them as well-formed syntactic constructions; consequently, the participating words must be related syntactically. This condition prevails over the contextual condition requiring them to appear in the proximity of each other. The definitions below (which are less frequently used in the NLP practice) clearly state the linguistic status of collocations:

(6) "co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern" (Cowie, 1978, 132)

(7) "a sequence of words that occurs more than once in identical form in a corpus, and which is grammatically well structured" (Kjellmer, 1987, 133)

(8) "On appellera collocation la combinaison caractéristique de deux mots dans une des structures suivantes : a) substantif + adjectif (épithète); b) substantif + verbe; c) verbe + substantif (objet); d) verbe + adverbe; e) adjectif + adverbe; f) substantif + (prép.) + substantif." [We shall call collocation a characteristic combination of two words in a structure like the following: a) noun + adjective (epithet); b) noun + verb; c) verb + noun (object); d) verb + adverb; e) adjective + adverb; f) noun + (prep) + noun.] (Hausmann, 1989, 1010)

(9) "A collocation is a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components." (Choueka, 1988)

(10) "A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things." (Manning and Schütze, 1999, 151)

(11) "lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other" (Bartsch, 2004, 76).

One of the most complete definitions of collocations was provided by Mel'čuk in the framework of the Meaning-Text Theory by means of the lexical functions formalism (Mel'čuk, 1998; Mel'čuk, 2003). This approach, presented later in Section 2.4.3, also regards the collocation as a syntactically-bound word association.

In the linguistically-motivated approaches, the condition for the participating words to occur within a short space of each other is not explicitly stated. Still, the proximity limitation persists, since the syntactic well-formedness criterion implies that the collocational span is the phrase, clause or the sentence containing these words. The recurrence is maintained as a defining feature, and is expressed, for instance, by attributes like "conventional", "characteristic", "recurrent".

Also, the collocation is seen as a directed (asymmetrical) relation, in which the role played by the participating words is uneven, and is mainly determined by their syntactic function in the sentence. Thus, as will be seen in Section 2.5.1, Hausmann and Mel'čuk use distinct terms (such as *base* and *collocate*) in order to account for the distinct role played by the items in a collocation pair. Also, Kjellmer (1991) introduces the notion of *left* and *right predictive collocations* for indicating that a given item in a collocation is predicted by the other.

Several authors attempted to provide a more precise characterisation of collocations from a linguistic point of view; their findings will be discussed in Section 2.5.

2.2.3 Collocation vs. co-occurrence

As indicated in Section 2.2.1, the term *collocation* has been originally used in a broad sense, for describing the general event of typical lexical co-occurrence. However, this purely statistical view has been contrasted by a more restricted view which was later adopted by numerous authors, and which is more linguistically motivated since it assumes that the items in a collocation are syntactically bound (Section 2.2.2).

In order to distinguish between these perspectives, some authors have suggested using a different term for each understanding. More precisely, it has been proposed to use the term *association* or *co-occurrence* for the general statistical understanding, and to reserve the term *collocation* for the restricted understanding corresponding to the linguistically-grounded approach. For example, Manning and Schütze (1999) and Evert (2004) state:

> "It is probably best to restrict the collocations to the narrower sense of grammatically bound elements and use the term *association* and *co-occurrence* for the more general phenomenon of words that are likely to be used in the same context." (Manning and Schütze, 1999, 185)

> "In order to make a clear distinction between the two approaches to collocations,[2] I refer to the distributional notion as *cooccurrences* [...] I reserve the term *collocation* for an intensionally defined concept" (Evert, 2004, 17).

The distinction between co-occurrences and collocations seems to be nowadays unanimously accepted (Bartsch, 2004), and will also be adopted by us.

2.3 Towards a core collocation concept

A multitude of collocation definitions exist in the literature,[3] which are often divergent and may therefore lead to confusion, in spite of the fact that a main distinction can

[2] One statistical, called *distributional*, and the other linguistic, called *intensional* (Evert, 2004).

[3] A selection of the most well-known collocation definitions is provided in Appendix B.

be drawn according to the underlying approach (i.e., a purely statistical one vs. a linguistic one).

This section aims to provide a unified view, by trying to capture what seems to constitute the essential defining features of the collocation concept. Despite the marked divergence of points of view, several defining features can be identified that are more recurrently mentioned and that seem to be accepted by most authors. We consider that these features denote a core collocation concept, and this concept may be further refined by adding more specific elements to the basic definition.

In accordance with Smadja (1993) and Evert (2004), we consider that the variations brought to the basic definition could be motivated by theoretical and practical considerations: "Depending on their interests and points of view, researchers have focused on different aspects of collocation" (Smadja, 1993, 145); "I use collocation thus as a generic term whose specific meaning can be narrowed down according to the requirements of a particular research question or application" (Evert, 2004, 17).

In this section we present the features which we identified as potentially relevant to the definition of the core collocation concept.

Collocations are prefabricated phrases. As mentioned in Section 2.1, collocations emerged from studies of language acquisition showing that children memorize not only words in isolation, but also, to a large extent, groups (or *chunks*) of words. These chunks are viewed as the building blocks of language. They are available to speakers as ready-made, or prefabricated units, and contribute to conferring fluency and naturalness to speakers' utterances.

As pointed out in many places in the literature, the collocational knowledge of speakers does not derive from their awareness of the individual words and of the grammar rules of a language; on the contrary, collocations are acquired as such, through experience: "We acquire collocations, as we acquire other aspects of language, through encountering texts in the course of our lives" (Hoey, 1991, 219).

According to Sinclair (1991), the language is governed by two opposed principles: the *open principle*, which refers to the regular choices in language production, and the *idiom principle*, which refers to the use of prefabricated units already available in

blocks. Collocations correspond to the second principle; they are:

(12) "semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair, 1991, 110).

The idea of collocations as prefabricated units was earlier expressed by Hausmann (1985, 124), who calls them "semi-finished products" of language and associates them with a "déjà-vu" effect in language. The same idea is suggested by Coseriu's metaphor of "lexical solidarity" (Coseriu, 1967).

Collocations are arbitrary. Another collocation feature that is mentioned by most definitions is arbitrariness—also, peculiarity, or idiosyncrasy—, as opposed to regularity (i.e., conformance to rules). Collocations are not regular productions of language, but "arbitrary word usages" (Smadja, 1993), "arbitrary [...] word combinations" (Benson, 1990), or "typical, *specific and characteristic* combination of two words" (Hausmann, 1985). A high number of definitions note this aspect, e.g., Fontenelle (1992), van der Wouden (1997), and Mel'čuk (2003) (see Appendix B).

The availability of collocations as prefabricated phrases in the lexicon of a language implies that they are to be learned and used as such, and prevents their reconstruction by means of grammatical processes. There might be no clear reason for a particular choice of words in a collocation, but once this choice was made and conventionalized—or, in Sag's terms, *institutionalized* (Sag et al., 2002)—, other paraphrases are blocked, as stipulated by Sinclair's *idiom* principle.

Arbitrariness of a collocation may refer not only to the choice of a particular word in conjunction with another in order to express a given meaning (Kahane and Polguère, 2001), but also to its syntactic and semantic properties. As Evert (2004) states,

(13) "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, 17).

Collocations are unpredictable. Closely related to the two properties mentioned so far, unpredictability is another main feature that is often cited in collocation definitions (Choueka, 1988; Evert, 2004).

Since the "institutionalization" of a collocation as a prefabricated unit does not depend upon clear linguistic reasons, it is impossible to predict the morphosyntactic properties of a collocation on the basis of the properties of participating words. Due to their arbitrariness, collocations are not reproducible (i.e., they cannot be predicted) simply by applying the grammatical prescriptions of a language.

Most importantly, the affinity of a word for a particular collocate which is strongly preferred over other words from the same synonymy set is, in the first place, unpredictable.[4] As put by Cruse, "these affinities can not be predicted on the basis of semantic or syntactic rules, but can be observed with some regularity in text" (Cruse, 1986).

Collocations are recurrent. The property of collocations that is most usually remarked in the various definitions is, undoubtedly, their recurrence in language. On the one hand, it is their frequent usage that determines the "institutionalization" of collocations. On the other hand, their frequency enables their recognition and learning based on experience: "we acquire collocations [...] through encountering texts in the course of our lives" (Hoey, 1991, 219). A collocation "springs readily to mind; it is psychologically salient" (Benson et al., 1986b, 252).

As mentioned in Section 2.2.1, the vast majority of collocation definitions indicate the frequent nature of collocations. Collocations are "habitual and customary" (Firth, 1957, 181), "actual words in habitual company" (Firth, 1968, 182), "recurrent word combination" (Benson, 1990), "combinations of words that co-occur more often than expected by chance" (Smadja, 1993, 143), or "groupements usuels" (Bally, 1909), "typical combination" (Hausmann, 1985), "conventional way of saying things" (Manning and Schütze, 1999, 151), "institutionalized phrases" (Sag et al., 2002, 3).

[4] See also Manning and Schütze's (1999) non-substitutability criterion mentioned in Section 2.5.

Collocations are made up of two or more words. Despite the fact that the practical work is concerned almost exclusively with collocations made up of exactly two lexemes, in theory, there is no length limitation for collocations. As Sinclair (1991) points out,

> "In most of the examples, collocation patterns are restricted to pairs of words, but there is no theoretical restriction to the number of words involved" (Sinclair, 1991, 170).

As a matter of fact, the majority of definitions stipulate that collocations may involve *more* than two words: "co-occurrence of two or more lexical items" (Cowie, 1978); "sequence of two or more consecutive words" (Choueka, 1988), "collocation is the cooccurrence of two or more words within a short space of each other" (Sinclair, 1991, 170); "an expression consisting of two or more words" (Manning and Schütze, 1999, 151). Arbitrarily long collocations, like *major turning point, play a central role, conduct a comprehensive study, abolish the death penalty, become an increasingly important concern*, are massively present in language.

A few studies—among which (Hausmann, 1985) and (Mel'čuk et al., 1984 1988 1992 1999)—explicitly mentioned the presence of exactly two lexemes in a collocation. However, these lexemes may in turn be composed of more words. The recursive nature of collocations has been noted, for instance, by Heid (1994, 232):

> "An additional problem for the interaction between syntactic and collocational description is the recursive nature of collocational properties: the components of a collocation can again be collocational themselves: next to the German collocation *Gültigkeit haben* (n + v), we have *allgemeine Gültigkeit haben* [lit., 'general validity have'], with *allgemeine Gültigkeit*, a collocation (n + a), as a component" (Heid, 1994, 232).

2.4 Theoretical perspectives on collocations

This section discusses the main theoretical frameworks within which the collocation phenomenon has been addressed in the linguistic literature.

2.4.1 Contextualism

The concept of word collocation plays a central role in contextualism, the linguistic current that actually brought collocations to the attention of linguists. Contextualists consider that the study of language cannot be done without considering the words' context. In particular, they argue that the meaning of words is defined by their co-occurrence (or collocation) with other words. As Firth states, the words are "separated in meaning at the collocational level" (1968, 180). Firth talks about "meaning by collocation", that he defines as "an abstraction at the syntagmatic level (...) not directly concerned with the conceptual or idea approach to the meaning of the words" (Firth, 1957, 196).

The description of collocations within the framework of contextualism passed through several stages. Initially given in terms of habitual co-occurrence of words within a short space of each other in a text (Firth, 1957; Sinclair, 1991), it was then elaborated by Sinclair, who paid less importance to the distance between collocation items in text. These were no longer required to be in the strict proximity of each other: "On some occasions, words appear to be chosen in pairs or groups and these are not necessarily adjacent." (Sinclair, 1991, 115).

As we have already mentioned in Section 2.3, Sinclair (1991) considers that the language obeys two opposed principles, the *open-choice* and the *idiom principle*. The first refers to the regular choices in language and accounts for the utterances produced by the application of grammatical prescriptions. The second stipulates that these regular choices are further restricted by the presence of prefabricated phrases that are already available to speakers:

> "The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." (Sinclair, 1991, 110)[5]

[5] An analogy can be found between the idiom principle and the *Elsewhere Principle* in linguistics, concerned with the opposition between regularity and specificity. This principle states that whenever two rules can be applied, the more specific overwrites the more general one. An account for idioms based on this principle was proposed by Zeevat (1995).

Collocations illustrate the idiom principle, since "the choice of one word affects the choice of others in its vicinity" (Sinclair, 1991, 173).

2.4.2 Text cohesion

The problem of collocations have also been addressed, though only to a limited extent, from the perspective of text cohesion, which refers to the "relations of meaning that exist within text" (Halliday and Hasan, 1976, 4). Cohesion contributes to the semantic unity of a passage of the language, about which a speaker "can normally decide without difficulty whether it forms a unified whole or is just a collection of unrelated sentences" (Halliday and Hasan, 1976, 1).

Halliday and Hasan distinguish between two types of text cohesion, one of grammatical nature, and one of lexical nature. The collocation, along with the reiteration (a general term encompassing lexical phenomena like repetition, synonyms, and hypernyms) is considered an important element of lexical cohesion. The collocation is basically understood in the same way as in contextualism:

(14) "the association of lexical items that regularly co-occur" (Halliday and Hasan, 1976, 284).

The cohesive effect of collocations derives precisely from words' "tendency to share the same lexical environment" (Halliday and Hasan, 1976, 286). Some examples provided by the authors are: *laugh – joke, blade – shape, ill – doctor, try – succeed*, and *sky – sunshine – cloud – rain*. It is explicitly stated that collocations refer not only to pairs, but also to longer "chains of collocational cohesion" (Halliday and Hasan, 1976, 287). The authors also note "a continuity of lexical meaning" in a collocation (1976, 320) through which the cohesion effect is achieved, but state that the meaning relations are not easy to classify in systematic semantic terms.

2.4.3 Meaning-Text Theory

Collocations received a formal characterization within the Meaning-Text Theory (hereafter, MTT) (Mel'čuk et al., 1984 1988 1992 1999; Mel'čuk, 1998; Mel'čuk, 2003) by

means of *lexical functions*, a language modeling tool that aims to associate a given meaning with the language utterances expressing that meaning.

Lexical functions (henceforth, LFs) are relations that are established between lexical items on the basis of the meaning to express, with the meaning being denoted by the name of the LF. For instance, the *Magn* LF represent the meaning of intensity, and relates words like *rain* and *heavy*.

MTT distinguishes between two main types of LFs:

- paradigmatic LFs, that capture lexical relations generated by morphological derivations (e.g., *surprise – surprising*), as well as basic semantic relations, like synonymy and antonymy (e.g., *timid – shy, like – dislike*);

- syntagmatic LFs, that describe the semantics of lexical combinations, and are used to model collocations (this category comprises a relatively higher number of functions, i.e., more than 60 LFs).

The authors indicate that a LF can be thought of as a function in the mathematical sense, which maps arguments into values. For instance, a syntagmatic function maps a lexical item (called *headword* or *base*) into the typical lexical item (called *collocate*) used for conveying the meaning represented by the LF:

$$Magn(rain) = heavy$$

The argument-value combination corresponds to the language utterance that represents the appropriate form for expressing the required meaning in conjunction with the headword (*heavy rain*).

Unlike mathematical functions, however, LFs may map a headword into more collocates. In order for LFs to be seen as mathematical functions, the authors allows the values to denote sets of lexical items, instead of single lexical items (Kahane and Polguère, 2001):

$$Magn(role) = \{central,\ important,\ major\}.$$

The concept of collocation received the following formal definition, adapted from Mel'čuk (2003).

(15) Let AB be a bipartite language expression, where A and B are lexical items of the language L, and let 'S' be the meaning of AB, 'A' the meaning of A, and 'B' the meaning of B. The expression AB is a *collocation* iff the following three conditions hold:
i)'S' $\supset$ 'A' (the meaning of S contains the meaning of A);
ii) A is selected by the speaker in a regular and non-restricted way;[6]
iii) B is not selected in a regular and non-restricted way, but depending on A and the meaning 'S' to express.

Therefore, in the MTT approach it is considered that the meaning of the collocation preserves the meaning of one of its constituents (i.e., A), which is selected by the speaker in a non restricted way, but not necessarily the meaning of the other (i.e., B), whose selection is contingent on the first.

Mel'čuk (2003) discusses the example of the collocation *café noir* (lit., *coffee black*, 'coffee without milk'). The word *café* is selected without any restriction, while the word *noir* is selected in an irregular way (since *noir* is the only possible choice in expressing the required meaning in combination with *café*, and it has to be specified by a rule), and in a restricted way (*noir* means '*without milk*' only in combination with the specific lexeme *coffee*, and not with other lexemes like *tea* or *chocolate*).[7]

The MTT approach had a big impact on the NLP work devoted to collocations. A number of studies explored the use of lexical functions as a semantic formalism in applications like machine translation and natural language generation (Heylen et al., 1994; Heid and Raab, 1989; Wanner, 1997). A formalization of lexical functions that is computationally tractable was recently proposed by Kahane and Polguère (2001).

[6]A linguistic operation is said to be performed in a *regular way* if it conforms to the lexicon and to the general rules of the grammar (general rules refer to rules involving classes of lexemes, as opposed to individual lexemes). It is performed in a *non-restricted way* if it can use any item of the lexicon and any grammar rule, without mentioning other lexical items (Mel'čuk, 2003).

[7]Note that, differently from many other approaches, in MTT the collocation is understood as an asymmetrical relation, since the collocate is dependent on the base, but not the opposite.

At the same time, in several electronic dictionaries,[8] collocations are encoded using a simplified and more formalised description in terms of LFs, which is based on the description used in ECD dictionary (Mel'čuk et al., 1984 1988 1992 1999). Also, an ambitious lexicographic project led to the creation of a large collocational database by enriching the entries of an entire bilingual dictionary with LF information (Fontenelle, 1997a; Fontenelle, 1997b). The resources built enable the development of LF-based NLP applications at a large scale, although, as pointed out by Fontenelle (1997a, 97) and other researchers, e.g., Alonso Ramos and Tutin (1996), the (syntagmatic) lexical functions alone might not be sufficient for tasks like machine translation, since they describe rather general meaning relations between the lexical items in a collocation.[9]

2.4.4 Other perspectives

In order to capture finer-grained distinctions of collocational meaning, Fontenelle (2001) proposed an account for collocations based on the Frame Semantics theory (Fillmore, 1982; Baker et al., 1998). In this approach, the link between the items in a collocation is expressed as the relation between the central word evoking a frame (i.e., a conceptual structure modelling a situation) and the frame elements. For instance, the link between *teacher*, *fail*, and *student* can be deciphered by recognising the semantic frame evoked by *fail*, namely, the examination frame, and by associating *teacher* and *student* with its elements, *Examiner* and *Examinee*.

The phenomenon of collocation has also been addressed in the literature in connection with studies on *semantic prosody* (Louw, 1993; Stubbs, 1995; Hoey, 1997), which noted the tendency of co-occurring words to share either a positive, or a negative connotation. As Sinclair (1991) remarks:

> "Many uses of words and phrases show a tendency to occur in a certain semantic environment. For example, the verb *happen* is associated with unpleasant things–accidents and the like." (Sinclair, 1991, 112)

[8] E.g., Dico and LAF (Polguère, 2000), Papillon (Boitet et al., 2002; Sérasset, 2004; Mangeot, 2006), DAFLES (Selva et al., 2002).

[9] The paradigmatic lexical functions, on the other hand, have been successfully used in translation since 1965 in Grenoble, as allowed translation paraphrases (C. Boitet, personal communication).

A few researchers indicated that collocations may sometimes constitute metaphors. For instance, Fontanelle (1997a) remarks that associations like *clouds of arrows, storm of applause, wave of protests* are clearly metaphorical. A similar remark is made by Gross (1996), who observes that collocations have a semantic component that is non-compositional, and suggests that this fact might be due to a metaphorical relation: "on pourrait parler de métaphore" (Gross, 1996, 21).

From a syntactic perspective, the collocation phenomenon has been studied in relation to *colligation* or *grammatical patterning*, i.e., the typical syntactic environment in which a word usually occur.[10] According to Hargreaves (2000),

> "knowledge of a collocation, if it is to be used appropriately, necessarily involves knowledge of the patterns or colligations in which that collocation can occur acceptably" (Hargreaves, 2000, 214).

In direct relation to this approach, the collocation was considered by more recent studies on lexis-grammar interface as a typical linguistic phenomenon occurring at the intersection between lexicon and grammar (Sinclair, 1991; Renouf and Sinclair, 1991; Hunston et al., 1997; Hoey, 1997; Hunston and Francis, 1998; Hoey, 2000). These studies consider that grammar and lexicon are interrelated to such a degree that no clear distinction can be drawn between the two. According to Francis (1993),

> "It is impossible to look at one independently of the other [...] The interdependence of syntax and lexis is such that they are ultimately inseparable" (Francis, 1993, 147).[11]

But long before the lexis-grammar approaches, and before the advent of the computer era which boosted the corpus-based linguistics studies, the French linguist Maurice Gross conducted in the 1970s an impressive work of lexicon-grammar resource

[10] Firth (1968, 183) describes the concept of colligation as "the interrelation of grammatical categories in syntactical structure".

[11] Sinclair's position on the syntax-lexicon continuum and the place of collocations is even more radical: "The decoupling of lexis and syntax leads to the creation of a rubbish dump that is called 'idiom', 'phraseology', 'collocation', and the like. [...] The evidence now becoming available casts grave doubts on the wisdom of postulating separate domains of lexis and syntax." (Sinclair, 1991, 104).

compilation. This work was empirically based and aimed at describing the syntax of French nouns, verbs and adverbs in the formalism of transformational grammar (Gross, 1984). The lexicon-grammar model of lexical description is nowadays largely used as the basis of computational work, thanks to the implementation provided by the INTEX system (Silberztein, 1993).

Summing up, the collocation is a language phenomenon that captured the attention of researchers in multiple subfields of linguistics. It has been studied from different perspectives, which were all concerned with the same issue: words associations giving rise to idiosyncratic semantic implications. In contextualism, collocating words *define* each other; from the point of view of text cohesion, they contribute to the *semantic unity* of the text; and in Meaning-Text Theory, the collocate word *expresses* a given meaning in combination with (and depending upon) the base word. Collocations were also taken into account in a number of other semantic theories, while in syntactic theories they are currently considered as elements found at the intersection between lexicon and grammar.

2.5 Linguistic descriptions

Although collocations came a long time ago into the attention of linguists and have a recognized linguistic and lexicographic status (McKeown and Radev, 2000, 507), they still lack a systematic characterization stating precise operational criteria for distinguishing them from other types of phraseological units. Such a characterization falls outside the scope of our present work; nonetheless, this section provides a review of the semantic and morphosyntactic properties of collocations that were mentioned in different theoretical and practical studies.

2.5.1 Semantic compositionality

The characterization of collocations is often made in semantic terms, on the basis of the semantic compositionality criterion, i.e., by checking whether the overall meaning of the collocation is obtained by the composition of the meanings of individual words.

This is because the particular importance of collocations arises from their arbitrariness and unpredictability (see Section 2.3), properties which indicate that the processus of lexical selection is not a semantically regular one. Also, as Moon remarks, "the non-compositionality of a string must be considered when assessing its holism" (Moon, 1998, 8).

In the existing literature, collocations are most often seen as expressions that "fall somewhere along a continuum between free word combinations and idioms" (McKeown and Radev, 2000, 509). Thus, it is commonplace to consider that collocations populate the grey area between one extreme which is represented by entirely compositional combinations, and the other extreme of completely opaque combinations. The boundaries between these three main groups—regular combinations, collocations and idioms—cannot be clearly drawn (Moon, 1998; Wehrli, 2000; McKeown and Radev, 2000).

It is therefore unsurprising that different descriptions of collocations make contradicting observations with respect to compositionality. On the one hand, Cruse (1986, 40) states that collocations are "fully transparent" and that "each lexical constituent is also a semantic constituent". Similarly, Sag et al. (2002, 7) describe them as "syntactically and semantically compositional, but statistically idiosyncratic". On the other hand, Choueka (1988) considers that the meaning of a collocation "cannot be derived directly from the meaning or connotation of its components"[12], while Manning and Schütze (1999) describe collocations using the three criteria below, of which the first two are traditionally associated with idioms.[13]

(16) a. *non-compositionality*: "the meaning of a collocation is not a straightforward composition of the meaning of its parts";
b. *non-modifiability*: "many collocations cannot be freely modified with additional lexical material or through grammatical transformations";
c. *non-substitutability*: "we cannot substitute near-synonyms for the components of a collocation" (Manning and Schütze, 1999, 172–173).

[12]See Appendix B for the complete definitions.
[13]Only the criterion of non-substitutability (16-c) is commonly associated with collocations.

In the same vein, van der Wouden (1997, 5) considers the collocation as a general phenomenon of "idiosyncratic restriction on the combinability of lexical items", which encompasses the class of idioms:

> "I will use the term collocation as the most general term to refer to all types of fixed combinations of lexical items; in this view, idioms are a special subclass of collocations" (van der Wouden, 1997, 9).

The same view on collocations including idioms as a particular case is found in several other places in literature (Moon, 1998; Venkatapathy and Joshi, 2005). But contrary to this perspective, collocations are usually distinguished from idioms, whose meaning is considered much more opaque. Manning and Schütze (1999, 151) themselves observe that not all collocations are completely non-compositional, but there also exists some "milder forms of non-compositionality", in which the meaning of the overall expression is nearly the composition of the parts. However, they indicate that "there is always an element of meaning added to the combination" (1999, 184), arguing, for instance, that the meaning of a collocation like *white wine* contains an added element of connotation with respect to the connotation of *wine* and *white* together.

Most researchers agree that collocations are semantically transparent and that their overall meaning is deducible from the meaning of the parts. While they are easy to decode, they represent idioms of encoding: "Idiomacity applies to encoding for collocations, but not to decoding" (Fillmore et al., 1988). More precisely, collocations are easily interpretable on the basis of the individual words, but are difficult to generate because the collocate is unpredictable.

Unlike in idioms, "the individual words in a collocation can contribute to the overall semantics of the compound" (McKeown and Radev, 2000, 507). However, the contribution of the individual words to the overall meaning of the collocation is uneven. While the meaning of the semantic head is preserved (e.g., the meaning of *wine* in the collocation *white wine*), that of the other word (*white*) does not participate in a straightforward way to the meaning of the collocation.

In fact, Hausmann (1979; 1985) describes collocations as "polar" combinations,

in which the item whose meaning is preserved is called *base*, and the other, which is selected by the base, is called *collocate*. In a collocation, the collocate is lexically selected by the base (Hausmann, 1985). A similar distinction is made in the MTT framework (Section 2.4.3). As Polguère states,

> "the *collocate* is chosen to express a given meaning, in a specific syntactic role, contingent upon the choice of the other component, called the *base* of the collocation" (Polguère, 2000).

Summing up, the semantic accounts of collocations in terms of compositionality generally regard collocations as transparent and easy to decode. Unlike in the case of idioms, whose meaning is rather opaque, the individual words of a collocation contribute to the overall meaning. Their contribution is, however, uneven, as the meaning of the base is completely preserved, while the meaning of the collocate is not directly added to the overall meaning.

2.5.2 Morphosyntactic characterisation

As stated at the beginning of this section, the literature does not provide a precise morphosyntactic characterization of collocations; however, the practical work often relies on particular assumptions about the possible surface realization of collocations.

Thus, many collocation extraction methods consider only certain lexical categories for the items of candidate pairs (i.e., open-class syntactic categories such as N, A, V, and Adv), while excluding other categories (the closed-class categories, e.g., P, D, Conj).[14] In addition, they only retain certain types of combinations as valid; typically, the patterns that are considered collocationally relevant are N-A, N-N, N-V, V-N, V-Adv, A-Adv (Hausmann, 1989; Heid, 1994).[15] In general, there is a marked

[14]The following abbreviations are used thorough this thesis for part-of-speech categories: A – adjective, Adv – adverb, Conj – conjunction, D – determiner, Inter – interjection, N – noun, P – preposition, V – verb.

[15]These combinations correspond the so-called lexical collocations in Benson et al.'s typology (1986a). Benson et al. divide the collocations in two broad classes, one allowing exclusively open-class words (*lexical collocations*) and one involving function words as well (*grammatical collocations*).

divergence among existing methods with respect to the categories and combinations allowed, the choice being made in a rather arbitrary way.

One of the few theoretical definitions concerned with morphosyntactic aspects is (Hausmann, 1989, 1010) (see Section 2.2.2). In addition to clearly specifying the POS for participating words, this definition alludes to the underlying syntactic structure of the collocation (e.g., subject-verb or verb-object). Unlike Hausmann, however, many researchers consider that collocations can be found in virtually any syntactic configuration. Thus, Fontanelle (1992) considers that:

(17) "The term collocation refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure." (Fontenelle, 1992, 222)

Lexicographic evidence confirmed that, in fact, collocations may appear in a large spectrum of configurations. Thus, the BBI collocation dictionary (Benson et al., 1986a) provides a very comprehensive list of grammatical collocations in its preface. Arguing against the a priori exclusion of function words from collocation patterns, van der Wouden states that "lexical elements of almost any class may show collocational effect" (van der Wouden, 2001, 17). On the same line, Kjellmer (1990, 172) notes that functional categories such as articles and prepositions are collocational in nature.

Some attempts have been made to characterize collocations from the point of view of the grammatical operations they can undergo, e.g., adjectival or adverbial modification, extraposition, pronominalization etc. For instance, Cruse (1986, 41) identified a subclass of collocations which are closer to idioms, and which he calls *bound collocations* (an example he provides is *foot the bill*). Unlike idioms, bound collocations are semantically transparent and modifiable (for instance, the noun modification is allowed: *foot the electricity bill*). But, like idioms, they resist certain operations, like separation or pronominalization (**I am expected not only to foot, but also to add up, all the bills; *I hope you don't expect me to foot it*). Generally speaking, syntactic restrictions and semantic opacity go hand in hand (Gross, 1996).

Since most collocations are rather syntactically permissive but at the same time idiosyncratic in the sense that the grammatical operations they allow vary from case

to case, providing a characterization of them in terms of syntactic behaviour is very difficult. Indeed, Heid (1994) points out that:

> "syntactic properties do not seem to have much discriminatory power, as far as collocations and idioms, their borderline and the borderline with 'normal constructions' are concerned" (Heid, 1994, 234).

Analogously, van der Wouden (1997, 19) states that it is impossible to formulate constraints on all collocations in syntactic terms. Collocational restrictions differ from selectional restrictions, as they "typically deal with restrictions between heads and complements at the *individual* level, whereas in cases of syntactic and semantic restriction *any member* of a syntactic or semantic class or category will do to satisfy the restrictions." (1997, 26). Also, whereas heads normally select arguments, in collocation—e.g., of verb-object type—the direction of subcategorisation seems reversed (van der Wouden, 1997, 27).

As van der Wouden (1997, 43–44) points out, collocations have an unclear linguistic status and they do not fit in current linguistic theories. Certain subclasses are better understood, e.g., the phrasal verbs and the light verbs; nonetheless, the collocational effects are ubiquitous and they must be taken into account in NLP applications.

2.6 The understanding adopted in this thesis

As explained in the previous sections, the concept of collocation is differently understood and defined in the literature. Since an agreed-upon definition does not exist to date, we need to state the understanding adopted in this work. We attempt to describe it as precisely as possible, while refraining from providing yet another definition.

First, we make a distinction between contextual and syntactical collocation, the first corresponding to the initial broader understanding adopted in contextualism, and the second to the more restricted, linguistically-motivated view (see Section 2.2). We adopt the restricted view throughout this work. As suggested in the recent literature,

we use the term *collocation* to denote the linguistically-motivated concept, and the term *co-occurrence* for the broad view as purely statistical phenomenon.

Second, we adhere to the five core features of collocations identified in Section 2.3. Collocations denote lexical combinations that are: 1) prefabricated, 2) arbitrary, 3) unpredictable, 4) recurrent, and 5) unrestricted in length. The first three features motivate our work of collocation extraction from corpora, aimed, ultimately, at providing lexicographic support for their inclusion in dictionaries. The fourth feature is not essential in our work. We assume that a single occurrence in the corpus justifies the selection of a pair as a collocation candidate (this decision is mainly motivated by the problem of data sparseness in corpora). The last feature is, on the contrary, of high importance in our work. Binary collocation are only a facet of the phenomenon covered. Our work deals with the identification of longer collocations in which smaller collocations recursively participate, and aims to provide a solution for fragmentary collocations identified by the basic extraction procedure.

Third, we further narrow the collocation acceptation according to the description proposed by Meaning-Text Theory by means of syntagmatic lexical functions (Section 2.4.3). One of the two elements of a pair (potentially involved in longer collocations through recursive embedding) is *autosemantic*, i.e., semantically transparent, and the other *synsemantic*, i.e., dependent of the first. While this choice does not really have an impact on the extraction methodology proposed, it proved useful in the further processing of identified collocations, in particular, for their automatic translation (see Section 6.5). Also, collocations are distinguished in this work from idioms, whose meaning is difficult to decode, as well as from compounds, which are rather limited in terms of syntactic behaviour. Again, this distinction has no bearing on the extraction methodology, but was taken into account in its evaluation.

From a syntactic point of view, the main constraint applied to the items of a candidate pair is to be linked by a syntactic relation. In principle, any generic head-dependent relation and any grammatical operation is allowed. Still, a selection of collocationally relevant syntactic patterns is possible (Section 5.2). The existing implementations of our approach apply a pattern-based filter on the candidate data.

2.7 Summary

This chapter focused on theoretical aspects related to collocation, a concept that occupies a central place in this thesis. From the multitude of collocation definitions that exist in the literature, we surveyed in Section 2.2 the most salient ones in chronological order, and classified them according to the underlying approach, i.e., a purely statistical vs. a linguistically motivated one. From a linguistic point of view, collocations were most often characterized in terms of semantic compositionality and with respect to their morphosyntactic properties, this is why these topics are further developed in Section 2.5.

Despite the heterogeneity of definitions and the pronounced divergence of their points of view, it was still possible to identify some key recurrent features that appear to be unanimously accepted by different authors. These features (discussed in detail in Section 2.3), can be seen as defining a 'core' collocation concept. We adhere to the basic features identified when we need to precise the understanding of the term *collocation* adopted in this thesis, and we further refine this understanding by imposing several restrictions that are consistent with the lexical functions approach (Section 2.6).

In Section 2.4 we presented the different perspectives from which this phenomenon has been studied in the literature. Besides the *lexicographic* and *pedagogical* perspectives already mentioned in Section 2.1, that see collocations as pre-constructed blocks that are learned as whole units, we presented the *contextualist* view, in which the collocability of words refers to their combinatorial profile and is considered as a major means for defining their meaning. We also reviewed several *lexical-semantic* perspectives, in which collocations are accounted for by formalisms such as lexical functions or semantic frames. From a more general perspective, the collocation is seen as one of the most important elements contributing to the text *cohesion*. Also, some studies of collocation are concerned with syntactic issues regarding, for instance, the syntactic context in which the collocations typically occur (*colligation*), or the role of collocations at the intersection of lexis and grammar (*lexis-grammar interface*). Other studies are more concerned with semantic matters such as the possible *metaphoric*

nature of collocations or the shared positive or negative connotation between the collocation's items (*semantic prosody*).

In the next chapter, we will see to what extent these theoretical considerations translate into the practical work dealing with the automatic identification of collocations in text corpora.

Chapter 3

Existing Extraction Methods

This chapter presents the foundations of the practical work on collocation extraction from text corpora. Section 3.2 discusses the extent to which the collocation features stipulated by theoretical studies translate into practice, introduces the basic concepts of statistical modelling of collocations as significant word associations, and describes the typical association measures used in existing work. Section 3.3 discusses the role played by text analysis tools such as lemmatizers, POS taggers, chunkers and deep syntactic parsers in preprocessing the source corpora in order to improve the extraction performance. Finally, section 3.4 contains a review of the state of the art, which provides details about the linguistic preprocessing performed by existing collocation extraction systems.

3.1 Introduction

As it emerged from the survey of definitions provided in the previous chapter, the term *collocation* is associated with a dichotomous acceptation. On the one side, the collocation is seen as a purely statistical phenomenon of word co-occurrence (as in the description put forth by the British contextualist J.R. Firth and then adopted by his disciples, M.A.K. Halliday, M. Hoey, and J. Sinclair). On the other side, a restricted acceptation was later systematically adopted (for instance, by A.P. Cowie, P. Howarth, F.J. Hausmann, and I. Mel'čuk), which is grounded mainly on linguistic

considerations: the collocation is no longer exclusively seen as a habitual association of words, but, most importantly, as a syntactically-bound combination.

In order to set these acceptations apart, the literature tends to use the term *co-occurrence* in the first case, and to reserve the term *collocation* for the linguistically-motivated concept. Such a distinction is, however, harder to identify at a practical level, since the techniques employed for collocation extraction from corpora were subject to the availability of appropriate text analysis tools. Yet, it is still possible to observe a gradual evolution from the statistical acceptation towards the linguistic acceptation, as increasingly sophisticated tools became more and more available.

This chapter aims to provide a picture of the existing collocation extraction methods, with a closer look at the various levels of linguistic preprocessing performed in each case (Section 3.4). Before this survey, we introduce the technical backgrounds of current collocation extraction methodology (Section 3.2) and present the rationale behind performing a linguistic processing of source texts prior to the extraction proper (Section 3.3).

3.2 Extraction techniques

3.2.1 The collocation features modelled

In the early NLP work dealing with their automatic identification in corpora, collocations have been assimilated to frequent sequences of words and have been extracted using n-gram methods (where the term *n-gram* refers to sequences of adjacent words). Among the basic features of collocations mentioned in Section 2.3, *recurrence* is, in fact, the easiest to model in practice.

N-gram methods also model another core feature of collocations, the *unrestricted length*, since they do not impose a specific length constraint on collocation candidates. In fact, the sequences extracted by Choueka (1983) may be up to 7 words long, and the rigid collocations identified by Smadja (1993) have an undetermined length.

Later, collocations were primarily seen as a phenomenon of lexical affinity, or mutual attraction that can be captured by identifying statistically significant word

associations[1] in large corpora (see Section 2.2). A wide range of statistical methods have been used to this end, that were either specifically designed for lexical association, or adapted from related disciplines or applications. Since such methods aim to quantify the degree of dependence or association of words, their are often called *lexical association measures* (hereafter, AMs).

Thus, it could be stated that the collocation feature that AMs aim to model is the unity or holism, i.e., the fact that collocations constitute *prefabricated units* available to speakers in blocks. However, AMs achieve this goal only to a limited extent. Since they are usually limited to pairs of words, the extraction performed on their basis concern almost exclusively binary collocations. Therefore, unlike n-gram methods, AMs are not really suited for modelling the unrestricted length feature, as sometimes they fail to detect complete collocations; on the other hand, they have the advantage of allowing a certain degree of flexibility for the candidates proposed.

To a certain extent, it can also be said that *arbitrariness* is another feature of collocations that is modeled by methods based on AMs, since these try to pinpoint specific, typical, or idiosyncratic word associations, as opposed to merely recurrent combinations that are more likely to be completely regular.

It is less clear, instead, whether AMs succeed to model the last core feature of collocations, the *unpredictability*, which states that the affinity between collocation components is unpredictable from the syntactic and semantic rules of the language, and therefore the synonymous alternatives are blocked. An AM that is considered to model the unpredictability feature is the log-likelihood ratios (Dunning, 1993) (presented in Section 3.2.4), since it is supposed to quantify how "surprising" a collocation candidate is.

A few specific extraction techniques have also been developed that model this feature by testing synonymous alternatives and checking if a base word systematically selects as collocate a specific item from a synonymy set, while excluding the others (Pearce, 2001a; Pearce, 2001b). Such techniques rely on the base-collocate distinction for the items in a collocation pair (Section 2.2), whereas existing AMs do not take

[1] In other words, pairs of words that are dependent on each other and co-occur more often than expected by chance.

into account the asymmetric role played by the components of a collocation.

As for the linguistic descriptions of collocations provided by some theoretical studies, these are unfortunately very limited and do not have a sufficient discriminatory power; this might be the reason why these descriptions have not been really taken into account by practical work on collocation identification. In particular, techniques based on semantic criteria that have been successfully used for detecting semantically opaque expressions (Lin, 1999; Fazly, 2007) are not applicable to collocations, which are rather compositional.

The work devoted to collocation extraction was hindered, in particular, by the difficulty to model features like holism, arbitrariness, and unpredictability, which can only be given an intuitive description, but not a formal one which could be translated into practice. Under the present circumstances, in which both a precise characterization of collocations and the appropriate means for implementing most of the identified features could not be found, it seems reasonable that the bulk of existing extraction work concentrates almost exclusively on the more objective, statistical aspects. An impressive number of AMs has been proposed for collocation extraction over the last decades.[2] This work testifies to the continuous efforts put into the design of AMs that are appropriate for collocation extraction, and suggests that the issue of finding a completely satisfactory AM is still open.

3.2.2 General extraction architecture

Generally speaking, a collocation extraction procedure (such as the ones used in the work reviewed in Section 3.4) takes place in two main steps:

- step 1: candidate identification using specific criteria;
- step 2: candidate ranking with a given AM.

The candidate ranking step relies on frequency[3] information about word occurrence and co-occurrence in a corpus (i.e., marginal and joint frequencies), as well as

[2]For instance, the inventory of Pecina (2005), still incomplete, lists more than 80 measures.

[3]Consistently with the related literature, we use the word *frequency* to denote the absolute frequency (number of occurrences) rather than the relative frequency (ratio of the number of occurrences to the corpus size).

about the total number of co-occurrences observed in the source corpus. This information can be organised, for each word pair, in a 2×2 contingency table (as will be explained in Section 3.2.3). Based on the numbers listed in the contingency table, AMs compute a numerical value that represents the association score for a pair.

In the candidate identification step, extraction procedures may either consider word forms and the surface word order found in texts, or may rely on linguistic analysis tools such as lemmatizers, POS taggers, and parsers in order to cope with morphological and syntactic variability. It is often felt that it is necessary to perform linguistic preprocessing for languages with a richer morphology and that exhibit a freer word-order, in order to conflate the instances belonging to the same pair type. The cumulative frequencies obtained for pair types are considered as more reliable for the score computation than the lower frequencies of disparate variants (this issue is further addressed in Section 3.3).

The output of an extraction procedure, often called *significance list*, consists of the list of candidates accompanied by their association score (or, when a score threshold is applied, of a subset of this list). The association score assigned to candidates by the specific AM chosen induces an order on these candidates, which is taken into account in displaying the extraction results. The candidates that received higher scores reflecting a higher association strength are found at the top of the significance list and are considered more likely to constitute collocations. Conversely, the candidates with a lower score found on lower position are, in principle, less likely to be collocational.

Most often, the output of extraction procedures is not a binary, but a fuzzy decision, since no clear-cut distinction is drawn between collocations and non-collocations; a continuous ordering is instead proposed, which is implied by the numerical score. The fuzzy output is compatible with the theoretical views postulating a continuum between collocations and regular combinations (see also Section 2.5). The numerical score has to be interpreted depending on the AM used. Typically, to each AM is associated a critical value against which the collocation score has to be compared, so that one can state, with a given level of confidence, whether or not a word association is statistically significant, i.e., it is likely to constitute a collocation.

Usually, the output list is truncated, so that only the top part of the significance list is eventually retained, all the candidates actually having much higher scores than the critical value associated with the AM. Therefore, all the pairs proposed are statistically significant. Nonetheless, errors are inherent to the statistical tests on which AMs are based. The output list may still contain spurious candidates, or may miss interesting ones.

The extraction output is normally considered as a raw result that must undergo a necessary process of manual validation before its use, e.g., its inclusion in a lexicon. Deciding upon the collocational status of a candidate is a notoriously difficult task; it is, ultimately, the desired usage of the output that determines the validation criteria. For instance, for lexicographic purposes it was indicated that even a precision of 40% would be acceptable (Smadja, 1993, 167).

3.2.3 Contingency tables

In statistics, contingency tables are used for displaying sample values (i.e., values observed in a sample drawn from a population) in relation to two or more random variables that may be contingent (or dependent) on each other.

The process of candidate identification in a corpus which takes place in the first extraction step (see Section 3.2.2) can be seen as a sampling process, in which a subset of pairs are collected from an infinitely large set of pairs in the population, the language. In order to detect an affinity (or dependency) between the component items of a candidate pair, two discrete random variables (X and Y) can be introduced, each associated with one position in the pair: X with the first position, and Y with the second.

For a particular candidate (u, v), where u and v denote lexical items,[4] the values displayed in the contingency table are u and $\neg u$ for variable X, and v and $\neg v$ for variable Y ($\neg u$ means any lexical item except u, and $\neg v$ - any lexical item except v). The contingency table for the pair (u, v) might therefore look similar to Table 3.1, that shows the typical notations for the marginal and joint frequencies.

[4]More precisely, in our work u and v denote lemmas of lexical items. Refer to the Section 3.3.2 for a discussion on using full word forms vs. lemmas.

	$Y = v$	$Y = \neg v$	
$X = u$	a	b	$R_1 = a + b$
$X = \neg u$	c	d	$R_2 = c + d$
	$C_1 = a + c$	$C_2 = b + d$	$N = a + b + c + d$

Table 3.1: Contingency table for the candidate pair (u, v). X, Y - random variables associated with a position in the pair; a – joint frequency; R_1, C_1 – marginal frequencies for u, resp. v; N – sample size.

Thus, a represent the number of items in the sample—i.e., in the candidate data—that have u in the first position and v in the second; b represents the number of items that have u in the first position and $\neg v$ in the second, and so on. In other words, a is the frequency of the candidate pair in the source corpus, and is called *co-occurrence frequency* or *joint frequency*.[5] The sum $R_1 = a + b$ is the frequency of all pairs with u in the first position, also written as $(u, *)$ or $(u, \bullet)$; similarly, $C_1 = a + c$ is the frequency of all pairs with v in the second position, written as $(*, v)$ or $(\bullet, v)$. These sums are referred to as the *marginal frequencies*. The quantity $N = a+b+c+d$ represents the total number of candidate pairs identified in the corpus, or the *sample size*. The tuple (a, R_1, C_1, N) formed by the joint frequency, the marginal frequencies, and the sample size is called the *frequency signature* of a candidate pair (Evert, 2004, 36). Note that the number d of $(\neg u, \neg v)$ pairs is less obvious to compute in practice,[6] but it can be obtained as in Equation 3.1.[7]

$$d = N - (a + b) - (a + c) + a = N - R_1 - C_1 + a \tag{3.1}$$

As explained in Section 3.2.2, the numbers in the contingency table typically refer to pair types, and u and v denote lemmas rather than word forms; in the absence of lemmatization, however, they can denote word forms as well.

[5] We might also refer to this number in the thesis simply as to frequency, or f.

[6] One can normally search for occurrences of a specific string in a corpus, but not for non-occurrences.

[7] That is, the frequency of the pairs $(\neg u, \neg v)$ is equal to the total number of pairs from which we subtract the number of pairs containing u in the first position and that of pairs containing v in the second position, but to which we add up the number of pairs (u, v), because these were subtracted twice.

For certain configurations of the contingency table—for instance, for skewed tables in which a is very small and d very large—, AMs fail to make reliable dependency predictions, especially if they assume that the lexical data is normally distributed[8]; in this case, the mathematical predictions might not be borne out for word cooccurrences (Evert, 2004, 110).

In order to cope with this problem (and, at the same time, to reduce the complexity of the extraction procedure), it is commonplace to apply a frequency threshold on the candidate pairs; for instance, only pairs with $a \geq 30$, $a \geq 10$, or $a \geq 5$ are considered as collocation candidates. This practice leads, however, to the loss of a high proportion of interesting pairs, since the combinations that occur only once or twice in the corpus (*hapax-legomena* and *dis-legomena*) constitute the main body of pairs in a corpus. A more appropriate solution to this problem is to use AMs that do not rely on the assumption of normal distribution (Section 3.2.4). In any case, theoretical studies have shown (Evert, 2004, 133) that a frequency threshold of 5 is sufficient for ensuring a reliable statistical analysis; therefore, it is unnecessary to apply a higher threshold.

3.2.4 Association measures

An association measure (AM) can be defined as "a formula that computes an association score from the frequency information in a pair type's contingency table" (Evert, 2004, 75). This section introduces the AMs standardly used in collocation extraction.

Hypothesis testing

AMs are very often based on statistical hypothesis tests (but not only, as will be seen in the description of AMs provided later in this section). Given a population and a random sample drawn from that population, a *statistical hypothesis test* (also, *statistical test*) is a technique of inferential statistics used for testing if a hypothesis

[8]That is, the frequency curve of words in a corpus is bell-shaped: the frequency of most words is similar to the mean frequency, and there are relatively few words whose frequency is much lower or much higer than the mean.

about the population is supported by evidence data, i.e., by the data observed, or the data in the sample.

Hypothesis testing consists in contrasting the hypothesis that is put forward by the statistical test (called *alternative hypothesis*, H_1) against the default hypothesis (called *null hypothesis*, H_0) that is believed to be true by default and that serves as a basis for the argument. In the case of lexical association, the alternative hypothesis is that the items u and v of a candidate pair are dependent on each other; the null (default) hypothesis is that there is no such dependence between the two items:

- H_0 (null hypothesis): u and v are independent;
- H_1 (alternative hypothesis): u and v are mutually dependent.

The result of a test is given in terms of the null hypothesis, H_0: either H_0 is rejected in favor of H_1 (therefore, it can be concluded that H_1 may be true), or H_0 is not rejected, which means that there was not enough evidence in favor of H_1 (i.e., it is impossible to conclude that H_1 may be true). In our case, if the null hypothesis of independence is rejected, then the two items may be dependent on each other and the candidate pair (u, v) may constitute a collocation. If it is not rejected, it means that there was not enough evidence supporting the alternative hypothesis of mutual dependence; therefore, it cannot be said that (u, v) forms a collocation.

One can never be entirely sure about the outcome of a statistical test, but can only reject the null hypothesis with a certain degree of confidence (which is typically set at 95% or 99%). The value obtained by the test is compared against a threshold value (called *critical value*) in order to decide if the null hypothesis is to be rejected. This threshold depends on the test type[9] and on the desired degree of confidence. The degree of confidence is more usually expressed in terms of *significance level*, or α*-level*, which represents the probability of the test wrongly rejecting the null hypothesis.[10]

The errors that can be made by a hypothesis test are:

[9]One-sided or two-sided (see below). For two-sided tests, a threshold corresponds to twice as lower confidence with respect to one-sided tests.

[10]For instance, a 5% significance level ($\alpha = 0.05$) corresponds to a 95% confidence level, and a 0.1% significance level ($\alpha = 0.001$) corresponds to a 99.9% confidence level.

- *type I errors*: wrongly rejecting the null hypothesis, when it is in fact true;
- *type II errors*: not rejecting the null hypothesis, when it is in fact false.

Type I errors affect the *precision* of the collocation extraction procedure: the candidate tested is wrongly considered as being a collocation, i.e., it is a *false positive*. Type II errors affect its *recall*: the candidate tested is not considered as a collocation as it should, i.e., it is a *false negative*. Using a smaller α-level ensures that the test produces fewer type I errors, but has instead the disadvantage of introducing more type II errors. Therefore, selecting an appropriate α-level means finding a compromise between type I errors and type II errors, i.e., trading off between precision and recall.

A test can be either *one-sided (one-tailed)* or *two-sided (two-tailed)*. For a *one-sided* test, it is known beforehand that the score obtained will be much higher than that expected under the null hypothesis, or it will be much lower. On the contrary, for a *two-sided* test both alternatives are possible and the test does not specify the nature of the difference. As far as lexical association is concerned, this means that one-tailed tests distinguish between positive and negative associations, whereas two-tailed tests do not:[11]

- A *positive association* occurs when the score is sufficiently high to reject the null hypothesis: the items in a candidate pair co-occur more often than expected by chance, if they were independent.
- A *negative association* occurs when the score is sufficiently low to reject this hypothesis: the items co-occur less often than if they were independent.[12]

Finally, statistical tests can be either *parametric* or *non-parametric*:

- *Parametric tests* (e.g., t-score, z-score, LLR) involve numerical data and often make assumptions about the underlying population; most usually, they assume that the data is normally or binomially[13] distributed.

[11] Among the tests described in this section, t-score and z-score are one-sided, while chi-square and LLR are two-sided.

[12] van der Wouden (2001, 31) uses the term *negative collocation* to denote "the relationship of a lexical item has with items that appear with less than random probability in its (textual) context".

[13] In binomially distributed lexical data, it is supposed that the occurrence of a word is comparable to the outcome of a Bernoulli trial, like in a coin-tossing experiment.

- *Non-parametric tests* (e.g., chi-square) involve ordinal data and are more effective than parametric tests when certain assumptions about the population are not satisfied (Oakes, 1998, 11).

Description of standard AMs

This section introduces the typical association measures used in collocation extraction, namely, the t-score, the z-score, the chi-square test, the log-likelihood ratios, and the mutual information. These measures have been presented in more or less detail in a number of other publications (Kilgarriff, 1996; Oakes, 1998; Manning and Schütze, 1999; Evert, 2004; Villada Moirón, 2005). In this introductory survey, we aim to provide a general description in intuitive and explicit terms, which could also be useful for the non-specialist reader.

Below we introduce additional notation used in relation to contingency tables, as shown in Table 3.2 and 3.3, and in Equation 3.2.

	$Y = v$	$Y = \neg v$	
$X = u$	$O_{11} = a$	$O_{12} = b$	$R_1 = a + b$
$X = \neg u$	$O_{21} = c$	$O_{22} = d$	$R_2 = c + d$
	$C_1 = a + c$	$C_2 = b + d$	N

Table 3.2: Contingency table for the candidate pair (u, v): observed values.

	$Y = v$	$Y = \neg v$
$X = u$	$E_{11} = \frac{(a+b)(a+c)}{N}$	$E_{12} = \frac{(a+b)(b+d)}{N}$
$X = \neg u$	$E_{21} = \frac{(c+d)(a+c)}{N}$	$E_{22} = \frac{(c+d)(b+d)}{N}$

Table 3.3: Contingency table for the candidate pair (u, v): expected values under the null hypothesis.

For a cell (i, j) in the contingency table, with $1 \leq i, j \leq 2$, O_{ij} represent the *observed frequency values* in a sample (as introduced in Section 3.2.3). E_{ij} represent the *expected frequencies* under the null hypothesis and are calculated as in Equation 3.2, where R_i are the row marginals ($R_1 = a + b$, $R_2 = c + d$) and C_j are the column marginals ($C_1 = a + c$, $C_2 = b + d$).

$$E_{ij} = \frac{R_i C_j}{N} \tag{3.2}$$

Indeed, if the two items u and v were independent of each other, their expected co-occurrence in a sample of N pairs would be the product of:

1. the probability of seing u as the first item in a pair,
2. the probability of seing v as the second items in a pair,
3. the sample size, N.

Since the individual probabilities are estimated on the basis of the actual frequencies of u and v and the sample size, we obtain the expressions of the expected frequencies from Equation 3.2. Below we illustrate this computation for the cell (1,1):

$$E_{11} = N \times P(u, *) \times P(*, v) = N \times \frac{R_1}{N} \times \frac{C_1}{N} = \frac{R_1 C_1}{N} \tag{3.3}$$

The values in the other cells are computed in a similar way, on the basis of $P(u, *)$, $P(\neg u, *)$, $P(*, v)$, and $P(*, \neg v)$.

The t-score. The t-score AM—used, for instance, in (Church et al., 1991; Breidt, 1993; Krenn, 2000b; Krenn and Evert, 2001)—applies the Student's t test in the task of collocation discovery. The t test is a one-sided parametric test, which assumes that the sample is drawn from a normally-distributed population. It compares the mean of the sample, $\bar{x}$ (i.e., the observed mean), with the mean of the population, μ (i.e., the mean estimated by assuming the null hypothesis).

A high difference indicates that the sample was not drawn from a population in which the null hypothesis holds. Thus, in the case of lexical association, a high t

value suggests that the sample was not drawn from a population in which the two lexical items are independent, and therefore indicates a strong positive association (Pedersen, 1996, 194).

The difference is expressed in standard error of the means units, $\sqrt{\frac{s^2}{N}}$ (s^2 is the sample variance). The t-test uses the following formula:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \tag{3.4}$$

In order to estimate $\bar{x}$, μ and s^2, it is usually assumed that observing the pair (u, v) in the sample is equivalent to a Bernoulli trial (i.e., like in a coin-tossing experiment). Under this assumption, the sample mean, the sample variance and the population mean can be estimated from the sample on the basis of $P(u, v)$, the probability of observing the pair (u, v), as in Equation 3.5 (Manning and Schütze, 1999, 154):[14]

$$\begin{aligned} \bar{x} &= P(u,v) = \frac{O_{11}}{N} \\ s^2 &= P(u,v)(1 - P(u,v)) \approx P(u,v) = \frac{O_{11}}{N} \\ \mu &= P(u,v) = P(u,*)P(*,v) = \frac{R1C1}{N^2} \end{aligned} \tag{3.5}$$

[14]Evert (2004, 77, 83) questions the applicability of the t-score to co-occurrence data, as the normal distribution required by the t test is incompatible with the binomial distribution assumption made when estimating, in particular, the sample variance.

The t-score formula then becomes:

$$t = \frac{\frac{O_{11}}{N} - \frac{R1C1}{N^2}}{\sqrt{\frac{O_{11}}{N^2}}} = \frac{O_{11} - \frac{R1C1}{N}}{\sqrt{O_{11}}} = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} =$$

$$= \frac{a - \frac{(a+b)(a+c)}{N}}{\sqrt{a}} = \frac{aN - (a+b)(a+c)}{N\sqrt{a}} \tag{3.6}$$

The z-score. The z-score is one of the first AMs used for identifying collocations (Berry-Rogghe, 1973; Lafon, 1984; Smadja, 1993). Like the t-score, it is a one-sided parametric test assuming that the data is normally distributed. It computes the difference between an observed value x and the mean of the population μ, expressed in population standard deviations units σ, as in Equation 3.7.

$$z = \frac{x - \mu}{\sigma} \tag{3.7}$$

Smadja (1993) uses this standard z-score formula for detecting collocations that are significantly frequent, by considering x – the frequency of a word, μ – the average frequency of its collocates, and σ – the standard deviation of the frequency of its collocates.

But, in general, the z-score AM is used for measuring the significance of the difference between the observed and expected frequencies for a candidate pair, as in Equation 3.8 below:

$$\textit{z-score} = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}} = \frac{a - \frac{(a+b)(a+c)}{N}}{\sqrt{\frac{(a+b)(a+c)}{N}}} = \frac{aN - (a+b)(a+c)}{\sqrt{N}\sqrt{(a+b)(a+c)}} \tag{3.8}$$

Berry-Rogghe (1973) essentially uses this formula, with a small modification in the denominator (which is multiplied with the probability of the collocate v occurring in a span around u).

The chi-square test. This test is relatively less used for collocation extraction than other AMs; it has been used, for instance, in (Krenn, 2000b; Evert and Krenn, 2001).

It is a two-sided non-parametric hypothesis test, which does not assume a particular distribution for the data. It compares the observed with the expected frequencies (under the null hypothesis) in each cell of the contingency table, as in Equation 3.9. If the overall difference is large, then the null hypothesis of independence is rejected.

$$chi\text{-}square = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{N(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \tag{3.9}$$

The derivation of the explicit formula from the left-hand side expression is shown in Appendix C.

Log-likelihood ratios. LLR (Dunning, 1993) is a two-sided parametric test that has been largely used in relation with collocation extraction, e.g., in (Daille, 1994; Lin, 1999; Orliac and Dillinger, 2003; Lü and Zhou, 2004). As the name suggests, LLR computes the association score for a candidate pair (u, v) by, basically, contrasting two likelihoods and considering the logarithm of the result:

- the likelihood of observing the counts in the contingency table under the null hypothesis of independence;
- the likelihood of observing these counts under the alternative hypothesis of dependence.[15]

Under the null hypothesis H_0, the probability p of observing the second item in the pair (v) is independent of the first item (u) being observed. Therefore, p can be estimated on the basis of the frequency of v in the corpus:

[15] We follow (Manning and Schütze, 1999) in this presentation of LLR. But for the sake of exactness, we must mention that Dunning (1993) does not refer to the second likelihood as the alternative hypothesis, but as the maximum likelihood reached when the independence constraint is relaxed. However, as it can be seen in (Dunning, 1993, 67), this maximum is in fact reached for exactly the same probability values p_1 and p_2 that are computed under the alternative hypothesis.

$$p = P(*, v) = \frac{C_1}{N} = \frac{a+c}{N} \tag{3.10}$$

On the contrary, under the alternative hypothesis H_1, the probability of observing v is dependent on whether or not u is observed; it is denoted by p_1 in the first case and p_2 in the second. When u is observed, p_1 is $P(v|u)$; when u is not observed, p_2 is $P(v|\neg u)$. Using corpus frequencies, p_1 and p_2 are computed as follows:

$$\begin{aligned} p_1 &= \frac{P(u,v)}{P(u,*)} = \frac{\frac{a}{N}}{\frac{a+b}{N}} = \frac{a}{a+b} \\ p_2 &= \frac{P(\neg u,v)}{P(\neg u,*)} = \frac{\frac{c}{N}}{\frac{c+d}{N}} = \frac{c}{c+d} \end{aligned} \tag{3.11}$$

Assuming a binomial distribution $\mathrm{B}(k;n,x) = \binom{n}{k}x^k(1-x)^{n-k}$, the probability of observing the values in the contingency table are computed using the following probabilities:

- under H_0:
  1) probability of observing v when u is present: $P_{H_0}(v|u) = \mathrm{B}(a; a+b, p)$;
  2) probability of observing v when u is not present: $P_{H_0}(v|\neg u) = \mathrm{B}(c; c+d, p)$;
- under H_1:
  1) probability of observing v when u is present: $P_{H_1}(v|u) = \mathrm{B}(a; a+b, p_1)$;
  2) probability of observing v when u is not present: $P_{H_1}(v|\neg u) = \mathrm{B}(c; c+d; p_2)$.

The LLR test considers $L(H_0) = P_{H_0}(v|u)P_{H_0}(v|\neg u)$ the overall likelihood of observing the contingency values under H_0, $L(H_1) = P_{H_1}(v|u)P_{H_1}(v|\neg u)$ the overall likelihood under H_1, and λ their ratio.

Then, the LLR score is defined and computed as in Equation 3.12.[16,17] The computation steps for obtaining the explicit formula are provided in Appendix C.

$$\begin{aligned} LLR &= -2\log\lambda = -2\log\frac{L(H_0)}{L(H_1)} = \\ &= 2(a\log a + b\log b + c\log c + d\log d - \\ &\quad -(a+b)\log(a+b) - (a+c)\log(a+c) - \\ &\quad -(b+d)\log(b+d) - (c+d)\log(c+d) + \\ &\quad +(a+b+c+d)\log(a+b+c+d)) \end{aligned} \tag{3.12}$$

Pointwise Mutual Information. Abbreviated as PMI or more often simply as MI, this measure introduced by Church and Hanks (1990) is probably the most popular AM; it has been used, for instance, in (Calzolari and Bindi, 1990; Breidt, 1993; Daille, 1994; Lin, 1998).

Unlike the AMs presented above, it is not a hypothesis test, but a measure related to Information Theory concepts. For two events in a series of events (like, in our case, the occurrence of two specific words in a corpus), MI quantifies the amount information, in information-theoretic terms, an event (in our case, the occurrence of a word) conveys about the other.

If $I(h;i)$ is the information provided about the event h by the occurrence of event i, defined as $I(h;i) = \log_2 \frac{P(h|i)}{P(h)}$, then this information is equal to the information provided about i by the occurrence of h, i.e., $I(i;h) = \log_2 \frac{P(i|h)}{P(i)}$, hence the name *mutual information*. MI is computed as in the formula in Equation 3.13, where $P(h,i)$ represent the joint probability of h and i (i.e., the probability of observing h and i together).

[16]Given λ, the quantity $-2\log\lambda$ is asymptotically χ^2-distributed (Dunning, 1993, 68).

[17]As Evert (2004, 89) remarks, the LLR formula is equivalent to the following formula for the *average-MI* association measure: $2\sum_{i,j} O_{ij}\log\frac{O_{ij}}{E_{ij}}$.

$$MI = \log_2 \frac{P(h,i)}{P(h)P(i)} \tag{3.13}$$

For lexical co-occurrence, MI is computed as in Equation 3.14, using the following probabilities:

- $P(u,v)$ – the probability of observing u and v together: $P(u,v) = \frac{a}{N}$;
- $P(u,*)$ – the individual probability of item u: $P(u,*) = \frac{R_1}{N} = \frac{a+b}{N}$;
- $P(*,v)$ – the individual probability of item v: $P(*,v) = \frac{C_1}{N} = \frac{a+c}{N}$.

$$MI = \log_2 \frac{P(u,v)}{P(u,*)P(*,v)} = \log_2 \frac{\frac{a}{N}}{\frac{a+b}{N}\frac{a+c}{N}} = \log_2 \frac{aN}{(a+b)(a+c)} \tag{3.14}$$

3.2.5 Criteria for choosing an appropriate AM

The collocation extraction practice has seen various AMs being applied in different settings, without clear-settled criteria for choosing one AM rather than another. However, there are theoretical reasons that make certain AMs to be more appropriate for collocation extraction than other, in a given setting.

Thus, the applicability of certain tests, like the t test and *z-score*, to the discovery of lexical associations has often been contested in the literature,[18] since these tests make the assumption that language data is normally distributed, and this assumption is not justified: whereas in a normal (bell-shaped) distribution most of the data is grouped around the mean and there are only a few extreme values that are either much lower or much higher than the mean,[19] the lexical data has a skewed, Zipfian

[18]For instance, in (Dunning, 1993, 61), (Kilgarriff, 1996, 35), and (Evert, 2004, 83).

[19]In normally distributed data, about two thirds of all values fall within one standard deviation away from the mean.

distribution, with a small proportion of high values (i.e., frequent events) and a majority of low values (i.e., rare events).

The AMs which are based on the assumption of normal distribution work well for the frequent events, but they behave unreliably for the rare events, which actually constitute the main body of textual data (Dunning, 1993). The usage of t-score and z-score is therefore not recommended for low-frequency candidates. In addition, the z-score is not recommended for small samples (small N), for which the t-score is better suited (Oakes, 1998, 12).

Also, *MI* is unreliable for low-frequency pairs, whose scores are considerably overestimated, in particular for the pairs occurring exactly once. In fact, Church and Hanks (1990) suggest a frequency threshold of 5 ($f \geq 5$).

The *chi-square* test overcomes the normal distribution problem, as it makes no assumptions about the data. It is less sensitive to low frequencies, but still overemphasizes rare events. Other disadvantages of chi-square test are that it overemphasizes common events as well (Kilgarriff, 1996, 35), and it is inaccurate when the sample size is very small (Manning and Schütze, 1999, 161).[20]

On the contrary, the *LLR* test works reasonably well with both rare and common phenomena, and with both large and small text samples (Dunning, 1993, 62). As a matter of fact, LLR is generally considered as the most appropriate measure for lexical association (Daille, 1994; Evert, 2004; Orliac, 2006).[21] Still, it has been pointed out that low values of expected frequencies in the contingency table (less than 1) affect the reliability of LLR (Pedersen, 1996, 191). In addition, some researchers, e.g., Stubbs (2002, 73), argue against the assumption of random (e.g., binomial) distribution that some tests, including LLR, make: a text cannot be compared to the outcome of a Bernoulli trial, as in a coin-tossing experiment.[22]

[20] More precisely, when N is smaller than 20, or it is between 20 and 50 and the expected frequencies are 5 or less (Manning and Schütze, 1999, 161).

[21] However, LLR was outperformed by the t-score in a particular extraction setting, namely, when ranking PP-V data in German (Evert and Krenn, 2001; Krenn and Evert, 2001). Yet, when high frequency data was considered only ($f \geq 10$), LLR performed better than t-score on the first 10% of the significance list (Evert and Krenn, 2001). For the low-frequency PP-V data ($f = 3, 4$) LLR performed comparably to other AMs, while for A-N data it was identified as the best measure for all frequency stata (Evert and Krenn, 2001).

[22] The random assumption is considered, however, more plausible than the normality assumption.

To sum up, selecting a specific AM as the best measure for ranking collocation candidates is a difficult process, because one has to take into account the particular setting in which the extraction experiment takes place, like the sample size, the observed and expected frequencies in the contingency table. Moreover, as argued by Evert and Krenn (2005, 452), the type of collocations to extract, the domain of the source corpora, the amount of low-frequency data excluded by the frequency threshold, as well as the preprocessing tools used also play a role in the practical relevance of an AM.

The comparative evaluation of AMs is a topic largely dealt with in the literature, e.g., in (Daille, 1994; Pedersen, 1996; Barnbrook, 1996; Krenn, 2000b; Krenn and Evert, 2001; Schone and Jurafsky, 2001; Pearce, 2002; Thanopoulos et al., 2002; Evert, 2004; Pecina, 2005). But, as a matter of fact, the particular results of the evaluation experiments carried out can hardly be generalized and transferred to a new extraction setting.

Surprisingly, the raw *co-occurrence frequency* was found in some settings to produce results that are as reliable as those obtained with more sophisticated AMs; see, for instance, (Daille, 1994, 154), (Krenn and Evert, 2001), or (Villada Moirón, 2005). The question if AMs correlate better with collocability than the raw frequency still remains open.

The very utility of statistical tests for collocation discovery is actually questioned by Stubbs (2002), who shows that in large corpora (e.g., 200 million words) the expected frequencies are so low that virtually any pair occurs more often than expected by chance; thus, citing a level of significance is pointless. Manning and Schütze (1999, 155–6) also found that most pairs in a corpus of a 14 million words occurred significantly more often than chance. They further pointed out that, while the level of significance is practically never looked at, what matters is the ranking obtained. Stubbs (2002, 75) adopts a rather radical point of view, suggesting that "whatever quantitative findings or statistics are produced by the computer, they must be interpreted by the human analyst". Nonetheless, AMs are nowadays successfully used in extraction systems producing reasonably accurate results, which are used as raw data for lexicography (Kilgarriff et al., 2004; Charest et al., 2007).

3.3 Linguistic preprocessing

According to the general architecture of collocation extraction systems described in Section 3.2.2, the first extraction step consists of the identification of collocation candidates. This step is most usually based on a linguistic preprocessing of the source corpus. Regardless of the subsequent statistical computation dealing with candidate ranking and of the identification method itself (from n-gram methods to more sophisticated methods), the linguistic analysis of the source text is often seen as an inescapable requirement for obtaining better extraction results.

The level of analysis performed by an extraction system depends on i) the kind of targeted results (i.e., semantic co-occurrences or syntactically bound pairs, according to the particular understanding adopted for the concept of collocation); ii) the ranking method applied in the second extraction step (since specific AMs may require larger or smaller candidate set, or higher frequencies that can be obtained by clustering the instances of a pair through lemmatization and syntactic analysis); and iii) the language involved. In most of the cases, this analysis implies at least sentence boundary detection and tokenization. Lemmatization, POS tagging and a syntactic analysis are also often necessary, for reasons that are explained below.

3.3.1 Lemmatization

In general, collocation extraction systems deal with lemmas rather than with full word forms (inflected words). Unless there are practical reasons to extract collocations involving inflected words, like the absence of appropriate tools, extraction systems rely on lemmatization in order to abstract away from morphological variants and to recognize a lexical item in all its inflected forms. The extraction results consist therefore of collocation types encompassing a relatively large number of tokens.

The usefulness of lemmatization was put forth, for instance, by Heid (1994, 250): "it seems useful to have an option in statistical programs to calculate the measures for lemmas rather than word forms". Also, Stubbs (2002), who performed a detailed lexicographic analysis on the collocates of several English words, remarked that lemmatization helps in detecting strong associations more easily. Thus, in (Stubbs, 2002,

82-83) he discusses the example of the word *resemblance*, whose collocates in the corpus Cobuild—among which the verb *bear*—are scattered through different forms, but altogether make up a high proportion of the total number of collocates; see Example (1) which displays, using the author's notation, the total number of collocates and the ratio of each form.

(1) resemblance $1,085$ <bears 18%, bear 11%, bore 11%, bearing 4%> 44%

Also, as indicated by Evert (2004, 35), the grouping of all inflected variants under the same lemma leads to more significant statistical results. The clustering of word forms is particularly important for languages with a rich morphology, since otherwise the low frequencies recorded in the contingency table could compromise the applicability of many AMs (as discussed in Section 3.2.5).

However, as noted, for instance, by Calzolari and Bindi (1990, 57), the lemmatization has the inconvenience that the specific information on inflection is (in principle) lost, and this information might be required for the subsequent treatment of results in other NLP applications.[23] Another remark made by them (Calzolari and Bindi, 1990, 57), and also by Stubbs (2002, 28,69), is that different word forms of the same lemma may have different collocates, as well as different frequencies of occurrence.[24]

3.3.2 POS tagging

The preprocessing of source corpora based on part-of-speech analysis brings considerable improvements to the identification of collocation candidates. Part-of-speech (POS) tagging is therefore performed in most of the extraction work, often in relation with a more detailed analysis at the syntactic level.

Firstly, POS tagging is very useful for better pinpointing candidates in source text, thanks to the lexical disambiguation it provides. Church and Hanks (1990) showed, for instance, that POS tags helps in distinguishing phrasal verbs involving the preposition *to* (e.g., *allude to*) from verbs followed by the infinitive marker (e.g.,

[23] Refer to Table 5.3 in Section 5.1 for an example of collocation in which considering word forms instead of lemmas would have been preferable.

[24] Stubbs (2002, 30) gives the example of *heated argument* vs. **hot argument*.

tend to). The authors note that "we see there is considerable leverage to be gained by preprocessing the corpus and manipulating the inventory of tokens" (Church and Hanks, 1990, 25).

Secondly, POS tagging is largely used for filtering out the combinations that are considered from the start as "noisy" or uninteresting, e.g., the combinations made up exclusively of function words (D-P, Conj-D, etc). As a matter of fact, most extraction systems only retain certain types of POS combinations, like N-V, V-N, N-N, A-N, Adv-V, Adv-A, or V-P.[25] The candidate pairs that do not correspond to these predefined patterns are ruled out.

The literature reports systematically on drastic improvements in the results precision when a POS filter is applied, based on simple syntactic patterns like the ones mentioned above (Church and Hanks, 1990; Breidt, 1993; Smadja, 1993; Daille, 1994; Justeson and Katz, 1995) (see also Section 3.4). However, a strong POS filter works to the detriment of extraction recall. Some authors argue against the a priori exclusion of non-content words (van der Wouden, 2001) and against the relatively arbitrary exclusion of most POS combinations (Dias, 2003), since the collocation phenomenon appears to involve any category of words and any syntactic structure.

As Fontenelle puts it, (1992, 222), "the term *collocation* refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure". Also, according to van der Wouden (1997, 13), "for almost every category it seems, in principle, to be possible to participate in collocations"; "syntactically, (almost) everything is possible in collocation" (van der Wouden, 1997, 14), even if "there are constraints in terms of syntactic structures or on he combination of syntactic categories. For example, I cannot think of many collocations of noun plus adverb, numeral plus adverb, verb plus determiner, etc. Probably, some notion of constituency or projection is needed to establish the necessary relationship between the collocants" (van der Wouden, 1997, 14).

While this criticism is directed against the reckless discarding of certain combinations from the very start, providing an abstraction of collocations at POS level by means of POS tagging remains an incontestably important concern. As Firth states,

[25]The POS labels have been introduced in Section 2.5.2.

"grammatical generalizations of word classes and the setting up of categories for the statement of meaning in terms of syntactical relations is clearly indispensable" (Firth, 1968, 181).

With respect to the distinction drawn in Section 2.2.3 between semantically vs. syntactically motivated combinations (co-occurrences vs. collocations), lemmatisation and POS-tagging preprocessing is sufficient for detecting instances of the first type. On the contrary, the detection of collocations in the restricted linguistic sense requires a finer syntactic analysis that goes beyond the usage of the simple syntactic patterns suggested by some POS combinations.[26] Such an analysis is provided by parsing.

3.3.3 Shallow and full parsing

Shallow parsing (or chunking) is performed more often than the full parsing in the preprocessing stage of extraction procedures, since more widely available. Whereas the full (or deep) parsing aims at building a complete syntactic structure for a sentence and at marking predicate-argument relations, chunking only provides partial and relatively simple structures for contiguous intra-clausal constituents (like AdvP, AP, NP, PP, or verbal complexes).

The identification of collocation candidates from chunked text is considered as more reliable than the identification based on POS-tagged text, thanks to the presence of a (constituent-internal) syntactic link between the items in a pair and to the POS-disambiguation.[27] Chunking also makes the extraction more tractable with respect to the window method, which relies exclusively on POS information and considers all the combinatorial possibilities in a collocation span (usually consisting of 5 content words). Yet, a major drawback of the extraction based on chunking is that the numerous syntactic relations holding between words in two distinct constituents, as well as the predicate-argument relations, are missed.

[26] In particular, as Heid (1994) states, "computational tools which would just look for combinations of adjacent lexemes would not retrieve all combinations which fall under the syntactic definitions".

[27] The percentage of ambiguous tokens in a language depends on its morphology and on the tagset considered. For English, it has been estimated as being roughly 40% (Hajič, 2000).

However, some shallow parsers aim at recovering such relations, or at least part of them. Thus, Basili et al. (1994) present an extraction approach that copes with this problem by relying on a "not so shallow" syntactic analysis. This analysis retrieves long-distance dependencies by means of so-called *skip rules* defined in a "discontinuous grammar" framework. The authors find that "adding more syntactic knowledge to the recipe significantly improves the recall and precision of detected collocations, regardless of any statistical computation" (Basili et al., 1994, 447).

Dependency parsing proper, that captures the relations between a head word and the dependent words (and thus provides a bridge between shallow and full parsing) has also been used for identifying collocation candidates, e.g., by Lin (1998; 1999). The main advantage of this kind of preprocessing is that there is no a priori limitation for the distance between two items in a candidate pair, as in the case of POS-tagging or shallow-parsing preprocessing.

More recently, full parsing—either stochastic or symbolic—has also been used for preprocessing the source corpus, but in a relatively fewer number of works with respect to shallow parsing, undoubtedly because of its much more reduced availability (Blaheta and Johnson, 2001; Pearce, 2001a; Schulte im Walde, 2003; Lü and Zhou, 2004; Villada Moirón, 2005). Full parsing makes it possible to deal with grammatical transformations that certain candidate pairs, notably those involving a verb, like verb-object, often undergo: passivization, relativization, interrogation, cleft sentences, etc. As in the case of dependency parsing, the linear distance between the two candidate items is irrelevant; the criterion used for identifying a candidate pair is the syntactic proximity, instead of the linear proximity of items in text.

There are multiple advantages that are brought to collocation extraction by the syntactic analysis of the source corpora, provided that adequate analysis tools are available that allow the accurate, robust and fast preprocessing of the whole source corpus. Using a syntax-based method of selecting collocation candidates should translate, first, into a higher extraction precision and recall, and second, into an improved tractability of the candidate ranking step, as many spurious candidates are ruled out from the start.

Also, by detecting those pair instances that are subject to complex syntactic operations, syntax-based methods help in computing more accurate frequency information for candidates, which in turn should help AMs to propose a more accurate ranking for candidates. Moreover, AMs become more reliable when different syntactic variants of pairs are clustered thanks to the syntactic analysis, in the same way in which morphological variants are clustered through lemmatization. Another possible advantage of syntactically preprocessing the source corpora is that AMs can also benefit from the tuning to the syntactic configuration, since the performance of AMs appear to be related to the syntactic type of candidate data (Evert and Krenn, 2001).

The possible inconveniences of adopting a syntax-based extraction approach instead of a syntactically-uninformed approach are the dependence on the language or on a specific linguistic theory, as well as the more limited availability of the required tools. In addition, it is necessary to specify a priori the set of relevant syntactic configurations and to ensure that it has a satisfactory coverage, tasks that might reveal quite difficult in practice.

3.3.4 Beyond parsing

It has been pointed out in the literature that in certain cases, even a detailed syntactic analysis might prove insufficient for retrieving some collocation instances. The following example from (van der Wouden, 2001, 33) supports the author's statement that automatic search procedures will never find all instances of some combinations. In fact, virtually no existing NLP technique could succeed in retrieving the verb-object link between *collect* and *stamps*.

(2) As to *collecting*, we're not interested in coins, books, ... Our sole concern is *stamps*.

Similarly, Stone and Doran (1996, 92) discuss the example of the collocation *narrow escape*, in which one of the items (*escape*) does not even occur in the text, but could be identified as an external discourse entity:

(3) Whew! [after burrowing and swimming out of Alcatraz, amid nearby shots and searchlights] That was *narrow*!

The authors suggest that collocations should be seen as holding between "objects in an evolving model of discourse" rather than between words, also because collocations introduce entities that are available for subsequent reference (Stone and Doran, 1996, 92). In fact, it is not unusual that an item of a candidate pair occurs in a different sentence, and is replaced in the current sentence by a referring expression—usually a pronoun, as in Example (4) below (Stone and Doran, 1996, 92). Anaphora resolution techniques are therefore necessary in order to link the pronoun *it* to the collocational antecedent, *escape*.

(4) Their *escape* had been lucky; Bill found *it* uncomfortably *narrow*.

Complementing parsing with discourse-level techniques is expected to benefit collocation extraction, in particular by promoting in the output list those candidate pairs in which the collocate happens to be often pronominalized.

Another situation in which parsing alone does not suffice for retrieving candidate instances is when the items involved undergo a grammatical category change. Pairs like those shown in (5-a) or in (5-b) could, in principle, be considered as equivalent collocations, and therefore should count as instances of the same collocation type. But in order to achieve this more abstract grouping, parsing should be complemented by other techniques such as stemming or syntactic transformation rules.[28]

(5) a. to *make* a *decision*, *decision making*

b. *strong argument*, *strength* of *argument*, to *argue strongly*, to *strengthen* an *argument*

[28] As in (Jacquemin et al., 1997); see also Section 3.4.3.

3.4 Survey of the state of the art

The following survey of existing collocation extraction work is presented in a language-oriented fashion, motivated by the fact that different languages often follow slightly different methodologies, as required by their morphosyntactic characteristics. The review provided is rather exhaustive. Basically, whenever a work concerns the application of techniques for the discovery of significant lexical associations (such as those introduced in Section 3.2), it is included in the review, even if the targeted linguistic phenomenon is more specific, more general, or labeled with a different term than collocation. For instance, we include in this review work on terminology extraction, e.g., (Bourigault, 1992a; Daille, 1994; Dagan and Church, 1994), since closely related to collocation extraction, as well as work dealing with specific syntactic configurations such as compound nouns, verb-particle constructions, or support-verb constructions.

This review does not discuss, however, the efforts of deriving collocations cross-linguistically by exploiting parallel corpora via alignment—e.g., (Kupiec, 1993; Wu, 1994; Smadja et al., 1996; Kitamura and Matsumoto, 1996; Melamed, 1997)—, since it only focuses on monolingual extraction. Also, it does not discuss the related but different methodology used for extracting collocations longer than two items, as this work will be described in more detail in Chapter 5.

3.4.1 English

It goes without saying that the bulk of existing extraction work deals with the English language. The earlier methods developed are generally based on n-gram techniques, and are therefore capable of extracting sequences of adjacent words only; moreover, the AM used is the plain co-occurrence frequency (Choueka, 1988; Kjellmer, 1994; Justeson and Katz, 1995). The last method cited also applies a POS filter on candidates in the preprocessing stage. Similarly, the method of Church and Hanks (1989; 1990) extracts adjacent pairs that are likely to constitute phrasal verbs by POS-tagging the source text, but it further applies MI for candidate ranking.

The Xtract collocation extraction system (Smadja, 1993) detects "rigid" noun phrases (e.g., *stock market*, *foreign exchange*, *New York Stock Exchange*, *The Dow*

Jones average of 30 industrials), phrasal templates (e.g., *common stocks rose *NUMBER* to *NUMBER**),[29] and, notably, flexible combinations involving a verb, called predicative collocations (e.g., *index [...] rose, stock [...] jumped, use [...] widely*). The system combines the z-score AM with several heuristics such as the systematic occurrence of two lexical items at the same distance in text. A parser is eventually used for validating results, leading to a substantial increase in the extraction precision, from 40% to 80%. The validation involved the top 4,000 pairs in the significance list and was performed by a professional lexicographer.

In Termight terminology extraction system (Dagan and Church, 1994), technical terms—more precisely, NPs—are identified by means of regular expressions defining syntactic patterns over POS tags. The candidates are grouped under the same head noun and are sorted according to their frequency; a concordancer then displays the context of terms and the context of their candidate translations, which are found on the basis of word alignments.

Shallow parsing was first used by Church et al. (1989) for detecting verb-object collocations, which were then ranked using MI and t-score AMs. Also, in the Sketch Engine (Kilgarriff et al., 2004), collocations candidates are identified based on shallow parsing implemented as regular expression pattern-matching over POS tags. The AM used is an adaptation of MI that gives more weight to the co-occurrence frequency. As it is the case for these methods, the more recent work is generally able to detect flexible pairs, since it relies on shallow-, dependency-, or full parsing.

Thus, Grefenstette and Teufel (1995) extract V-N pairs by using a robust dependency parser, with the specific goal of identifying nominalisation constructions. Lin (1998; 1999) also uses a dependency parser for identifying candidate pairs or several types. For ranking these pairs, the author uses a version of MI that takes into account the syntactic relation (Lin, 1999), or the LLR measure (Lin, 1999).

Collocation extraction has also been performed from statistically parsed corpora, e.g., by Pearce (2001a), who studies restrictions on synonym substitution, or by Blaheta and Johnson (2001), who extract verb-particle constructions with a log-linear

[29]The term *phrasal template* denotes sequences of words with empty slots standing for POS tags (Smadja, 1993, 149).

model it proposes as an AM. Symbolic parsing is used in the preprocessing stage by Goldman et al. (2001)—the early version of our extractor—and later by Orliac and Dillinger (2003), Wu and Zhou (2003), and Lü and Zhou (2004). All these methods use LLR for ranking, except the method of Wu and Zhou (2003) which uses *weighted MI*, a version of MI that compensates for the tendency of MI to overestimate rare pairs.

As Pearce (2002, 1530) states, "with recent significant increases in parsing efficiency and accuracy, there is no reason why explicit parse information should not be used", at least as far as English language is concerned. Nonetheless, a few extraction works nowadays still limit their preprocessing to POS tagging. For instance, Zaiu Inkpen and Hirst (2002) extract collocations from a POS-tagged corpus, the BNC.[30] They apply a POS filter for removing closed-class words, then use the BSP package[31] to extract adjacent pairs from the remaining tokens and to rank these pairs with several AMs. Also, Dias (2003), who argues against the a priori definition of syntactic patterns, bases his extraction on POS-tag sequences within a text window.

The relative simplicity of English language in terms of both morphology and syntax facilitates the use of more rudimentary techniques, like the application of AMs directly on plain text or on POS-tagged text inside a small collocational span. In other languages, however, a more advanced linguistic preprocessing is required in order to obtain acceptable results.

3.4.2 German

German is the second most investigated language from the point of view of collocability, thanks to the early work of Breidt (1993) and, more recently, to that of Krenn and Evert that was centered on evaluation, e.g., (Krenn, 2000b; Krenn and Evert, 2001; Evert and Krenn, 2001; Evert, 2004; Evert and Krenn, 2005).

Before syntactic tools for German became available, Breidt (1993) extracted V-N

[30] British National Corpus (BNC) is a 100 million word collection of samples of 20th century British English, that has been automatically POS tagged (`http://www.natcorp.ox.ac.uk/corpus/`).

[31] The Bigram Statistic Package (BSP) identifies bigrams in corpora and implements several AMs: MI, Dice coefficient, chi-square, LLR, and Fisher's exact test. Since it was extended to handle n-grams, it is now known as NSP (Banerjee and Pedersen, 2003).

pairs (such as *[in] Betracht kommen*, 'to be considered', or *[zür] Ruhe kommen*, 'get some peace') using the window method. She evaluated the performance of MI and t-score AMs in a variety of settings: different corpus and window size, presence/absence of lemmatization, of POS tagging, and of (simulated) parsing. The author argues that extraction from German text is more difficult than from English text due to the much richer inflection for verbs, the variable word-order, and the positional ambiguity of arguments; she shows that even distinguishing subjects from objects is very difficult without parsing. The study found that in order to exclude unrelated nouns, a smaller window of size 3 is preferable, although at the expense of recall. Increasing the corpus size leads to considerable improvement of recall, but causes a slight decrease in precision. If parsing (which is simulated by eliminating the pairs in which the noun is not the object of the co-occurring verb) leads to a much higher precision of results, the lemmatization alone does not help, for it promotes new spurious candidates. The conclusion of the study was that a good level of precision can only be obtained for German with parsing: "Very high precision rates, which are an indispensable requirement for lexical acquisition, can only realistically be envisaged for German with parsed corpora" (Breidt, 1993, 82).

More recent work, e.g., (Krenn, 2000b; Krenn, 2000a; Krenn and Evert, 2001; Evert and Krenn, 2001; Evert, 2004) made use of chunking for extracting particular types of collocations such as P-N-V, and was mostly concerned with the comparative evaluation of AMs.

Thus, Krenn (2000a; 2000b) describes an extraction experiment concerning adjacent and non-adjacent, full and base form P-N-V (PP-V)[32] candidates, aimed at distinguishing collocational from non-collocational combinations. Since POS information and partial parsing is employed, the set of candidates identified is argued to contain less noise than if retrieved without syntactic information (Krenn, 2000a). The experiment compares several AMs—MI, the Dice coefficient, LLR, the relative entropy—with two newly-proposed identification models: i) the *phrase entropy* for P-N pairs, since the low entropy is considered an an indicator of collocability, and

[32]These patterns describe expressions that are in fact similar to those retrieved by Breidt (1993), e.g., *zur Verfügung stellen* (lit., *at the availability put*, 'make available'), *am Herzen liegen* (lit., *at the heart lie*, 'have at hearth').

ii) the *lexical keys model*, based on lists of typical support verbs. Evaluation based on manually classified data containing more than $30,000$ candidates showed that results depend on the frequency threshold applied, on whether lemmatization is used or not, and on the type of collocation (the author distinguishes between figurative expressions and support-verb constructions). The study found that the P-N entropy model outperforms the AMs on higher frequency, full form data, i.e., data on which a threshold of 5 or 10 is applied. However, its advantage is less pronounced when base forms and lower frequency data is included ($f \geq 3$). Also, while combining the entropy with the lexical keys model proved very helpful in identifying support-verb constructions, it proved less helpful in identifying figurative expressions.

As in a subsequent comparative evaluation experiment (Krenn and Evert, 2001), the PP-V candidates can be grammatically incorrect, since the PP and V items are only required to occur within the same sentence and no parsing is performed to ensure that there is a grammatical relation holding between the two. The reported n-best precision, computed using the same gold standard of manually extracted PP-V collocations, showed that more accurate results are obtained for higher frequency data: thus, the precision is 40% for $f \geq 3$ and higher than 50% for $f \geq 5$. It was also shown that the t-score is the best AM for identifying PP-V collocations, and that no AM performs significantly better than the simple co-occurrence frequency.[33] The experiment involved full form data, but the authors reported that similar differences could be observed among the compared AMs on base form data.

Evert and Krenn (2001) performed a similar evaluation experiment in which, in addition to PP-V collocations, they extracted adjacent A-N pairs using a POS tagger. The authors found that LLR is the best AM for identifying A-N collocations, while chi-square and MI perform worst, as in the case of PP-V data. Evaluation was also performed separately on different frequency strata, i.e., for high and low frequencies. On high frequency PP-V data, LLR outperformed the t-score on the top 10% of the significance list, but the t-score was found better on the remaining part. Another finding of the study was that, contrary to previous claims, MI did not perform better

[33] In particular, LLR was found significantly worse than frequency, although better than other AMs (Krenn and Evert, 2001).

on high frequency data than on low-frequency data: its relative performance with respect to other AMs did not improve with a higher frequency, as one could expect (see Section 3.2.5). As for low-frequency data, all AMs had similar performance on PP-V data, while on A-N data LLR emerged as the best AM. The study also showed that on these data the raw co-coccurrence frequency was significantly worse than all AMs.

Also, Evert and Kermes (2002; 2003) focused on the evaluation of linguistic preprocessing—as opposed to the evaluation of AMs—, since this stage has a crucial influence on the final results obtained. The authors evaluated the grammaticalness of A-N pairs extracted using three different methods: 1) identification of adjacent A-N with a POS-tagger, 2) identification of an A followed by an N within a window of 10 words, and 3) identification based on chunking. Evaluation against A-N pairs extracted from the same corpus (i.e., the Negra treebank containing syntactic annotations) showed that the highest recall is obtained with chunking, but the precision of the chunk-based method is comparable to that of the adjacency method. This result is hardly surprising, given the rigidity of the syntactic type considered.

In addition to the work centered on evaluation, collocation extraction work has also been performed, for instance, by Zinsmeister and Heid (2002; 2004), who compared the collocational behaviour of compound nouns with that of their base noun, and showed that this is correlated with semantic transparency. Also, Zinsmeister and Heid (2003) focused on the detection of A-N-V triples and on classifying them into idiomatic combinations, combinations of collocations, mixed triples,[34] and regular combinations by means of machine learning techniques. In addition, Schulte im Walde (2003) built a collocation database for a variety of syntactic types, which contains subcategorization details for nouns and verbs. All these methods use LLR for candidate ranking and are based on the full stochastic parsing of the source texts. This analysis is argued to be indispensable for identifying verbal collocations,[35] and to lead to a higher precision and recall with respect to partial analyses.

[34] An example of each is: *letzt Wort haben* - 'have the last word' (idiomatic combination), *klar Absage erteilen* - 'clearly reject' (combination of collocations), *rote Zahlen schreiben* - 'be in the red' (mixed triple). Mixed triples associate a collocational with a regular combination.

[35] In particular, the constructions with split particles (Zinsmeister and Heid, 2002).

Yet, chunked text has been used in several other works, e.g., for identifying A-V collocations in predicative constructions (Kermes and Heid, 2003), or for identifying morphosyntactic preferences in collocations: thus, Evert et al. (2004) deal with number and case of nouns in A-N collocations, while Ritz (2006) focuses on number, case, determination, modification for nouns, and on negation and subtype—main, auxiliary, modal—for verbs in V-N collocations.

Finally, Wermter and Hahn (2004) used a POS tagger and a shallow parser in order to extract collocational and idiomatic PP-V combinations by applying the limited modifiability criterion: the pairs which are frequent and in which the adjectival modifiers of the noun in the PP constituent show a lower dispersion are promoted in the significance list.[36] This method was found to perform significantly better than the co-occurrence frequency both in this particular experiment involving high-frequency PP-V pairs ($f > 10$), and in a different extraction experiment concerning high-frequency terms in English ($f > 8$) (Wermter and Hahn, 2006).

3.4.3 French

Outstanding work on lexicon-grammar carried out before computerized tools even became available (Gross, 1984) makes French one of the most studied languages in terms of distributional and transformational potential of words. Automatic extraction of collocations was forsit performed by Lafon (1984), then, to a certain extent, in the framework of terminology extraction systems which deal specifically with NPs, but which apply the same extraction methodology in order to discover significant lexical associations.

Thus, Lafon (1984) extracts significant co-occurrences of words from plain text by considering (oriented, then non-oriented) pairs in a collocational span and by using the z-score as an AM. The preprocessing step merely consists in detecting sentence boundaries and ruling out the functional words. The author noted that verbs rarely occur among the results, probably as a consequence of the high dispersion among different forms (Lafon, 1984, 193). Apart from the lack of lemmatization, the lack of

[36] More precisely, the method promotes the pairs in which the adjectival modifier has a high relative frequency.

a syntactic analysis was identified as one of the main source of problems encountered in the extraction. The author pointed out that any interpretation of results should be preceded by the examination of results through concordancing (Lafon, 1984, 201).[37]

The LEXTER terminology extractor (Bourigault, 1992a; Bourigault, 1992b) detects NPs using surface analysis, by making the assumption that the grammatical form of terminological units is relatively predictable. First, the frontiers of maximal-length NPs are detected, based on POS information; elements like inflected verbs, pronouns, conjunctions, and prepositions other than *de* et *à* are considered as frontier markers. Second, the maximal NPs so identified (e.g., *disque dur de la station de travail*) are further split into smaller units, like N-A (*disque dur*) and N-P-N (*station de travail*), using shallow parsing. Up to 800 different structures are identified by the shallow parser module (Bourigault, 1992b, 979). The author states that given the high quality of results already obtained by relying on a surface analysis, a complete syntactic analysis is unnecessary.

Similarly, the ACABIT system (Daille, 1994) extracts compound nouns defined by specific patterns, e.g., N-A, N-N, N-*à*-N, N-*de*-N, N-P-D-N, by relying on lemmatization, POS tagging, and on shallow parsing with finite state automata over POS tags. For ranking the candidate terms, the system applies a long series of AMs; their performance is tested against a domain-specific terminology dictionary and against a gold-standard manually created from the source corpus with the help of three experts. The evaluation study highlighted LLR as the best performing AM, among other AMs which were retained as appropriate.[38] The raw frequency of pairs was also found as a good indicator of termhood, but it has the disadvantage of not being able to identify rare terms (Daille, 1994, 172–3). In fact, a high number of terms were found between low-frequency pairs, with $f = 2$ (Daille, 1994, 154). LLR was eventually preferred for its good behaviour on all corpus sizes and for promoting less frequent candidates as well (Daille, 1994, 173). The author argues that by using finite state automata for

[37] "[...] c'est moins l'absence de lemmatisation que l'impasse faite sur la dimension syntagmatique au moment de l'émiettement du texte en formes graphiques qui fait problème. Le nécessaire recours aux concordances avant tout commentaire interprétatif d'un dépouillement, atteste suffisamment ce dernier point" (Lafon, 1984, 201).

[38] The retained AMs are FAG – Fager and MacGowan coefficient, cubic MI, LLR, and frequency (Daille, 1994, 137).

extracting candidates in different morphosyntactic contexts without a priori limiting the distance between two words, a better performance is achieved with respect to the mobile window technique. The author's claim is that the linguistic knowledge drastically improves the quality of stochastic systems (Daille, 1994, 192).

Despite the fact that in these extraction systems the targeted phenomenon is quite specific (i.e., NPs), as pointed out in (Jacquemin et al., 1997), the candidates are still subject to numerous linguistic variations, both from a morphological and syntactic point of view. Therefore, Jacquemin et al. (1997) takes into account the derivational morphology and defines transformational rules over syntactic patterns, in order to achieve a broader extraction coverage. The derivational morphology accounts for correspondences like between *modernisateur* and *modernisation*. Syntactic transformations (such as *fruits tropicaux – fruits et agrumes topicaux*, *stabilisation de prix – stabiliser leurs prix*) are inferred from a corpus by first looking at collocations within a window of 10 words,[39] then by applying a syntactic filter based on shallow parsing. The transformation rules are defined as regular expressions over POS tags, and are proposed on the basis of both linguistic and empirical considerations. Nonetheless, certain linguistic phenomena (such as sentential complements for nouns and long distance dependencies) cannot be accounted for by this approach, and the authors suggest that these phenomena might require a deeper syntactic analyzer (Jacquemin et al., 1997, 28).[40]

Collocation extraction proper—as opposed to term extraction—has been performed in (Goldman et al., 2001), the early version of our extractor. Candidate data involving a wide range of syntactic configurations are first identified in French texts using full parsing, then they are ranked using LLR. Long distance dependencies can be retrieved even if subject to complex grammatical transformations; for example, some instances of verb-object collocations were detected in which the constituent items were separated by as many as 30 intervening words (Goldman et al., 2001, 62).

[39]The authors argue that a collocational span of 5 words as commonly used for English is insufficient for French, since French has "longer" syntactic structures (Jacquemin et al., 1997, 27).

[40]The evaluation of extraction results showed that the discovery of term variants lead to a drastic improvement of recall over text simplification techniques based on stemming; in particular, the recall triples and reaches 75.2%.

Also, Tutin (2004) extracts collocations using the local grammar formalism and the INTEX system (Silberztein, 1993), while Archer (2006) detects collocations consisting of a verb and an adverb expressing the meaning of intensity (e.g., *changer radicalement*, 'to change radically') for inclusion in the Papillon lexical database. The candidate pairs, which are identified from syntactically-analysed text, are ranked with the weighted MI measure introduced by Wu and Zhou (2003); then, they are further filtered on the basis of their conceptual similarity to an existing list of intensifying adverbs.

Other extraction work that exists for French, e.g., (Ferret, 2003; Ferret and Zock, 2006),[41] is not actually concerned with collocations, but rather with co-occurrences (or collocations in the broader statistical sense, according to the distinction drawn in Section 2.2.3).

3.4.4 Other languages

Collocation extraction work has also been performed for a number of other languages, notably, Italian, Dutch, Korean, Japanese and Chinese. In most cases, the extraction is based on the reuse of existing techniques or on improved versions. As for the preprocessing stage, this is most usually limited to POS tagging.

- Italian—for instance, Calzolari and Bindi (1990) used the mobile window method on plain (unprocessed) text in order to identify candidates, and used MI for ranking them.[42] Later, Basili et al. (Basili et al., 1994) made use of parsing information, as discussed in Section 3.3.3.

- Dutch—Villada Moirón (2005) retrieves P-N-P and PP-V candidates (such as *in het kader van*, 'in the framework of' and *in pand houden*, 'keep in custody') by using POS filtering and, partly, parsing, then applies several AMs for ranking the candidates. In particular, the author applies the log-linear model defined in

[41] This work (Ferret, 2003; Ferret and Zock, 2006) is focussed on topic segmentation, and relies on the extraction of co-occurrences involving A, N and V lemmas within a window of length 20. The AM used for ranking is MI.

[42] A standard window size of 5 words has been considered.

(Blaheta and Johnson, 2001), but this is found to perform worse than traditional AMs, like LLR or chi-square (Villada Moirón, 2005, 106).[43] Chunking was also previously considered for PP-V identification as an alternative to parsing, but it revealed impractical for Dutch because of the syntactic flexibility and the relatively free word order (Villada Moirón, 2005, 162).

- Korean—for instance, Shimohata et al. (1997) used an adjacency n-gram model applied to plain text, without preprocessing, coupled with entropy for ranking. Kim et al. (1999) relied, instead, on POS-tagging, and proposed an AM that takes into account the relative positions of words. They argued that Xtract-like techniques are inappropriate for Korean due to the freer word order.

- Japanese—Ikehara et al. (1995) applies an improved n-gram method in order to extract interrupted and non-interrupted collocations.

- Chinese—Wu and Zhou (2003) and Lü and Zhou (2004) extracted Chinese collocations by using the same parse-based techniques they used for English collocations (see Section 3.4.1). Also, Lu et al. (2004) employed a method similar to Xtract, while Huang et al. (2005) used POS information and regular expression patterns borrowed from the Sketch Engine (Kilgarriff et al., 2004). The authors pointed out that an adaptation of these patterns for Chinese is necessary in order to cope with syntactic differences and the richer POS tagset.

3.5 Summary

Spanning several decades already, collocation extraction is a very active research area that has witnessed an impressive development both with respect to the techniques used for ranking candidate pairs according to their collocational strength, and to the techniques of text preprocessing used for the identification of candidate pairs in corpora. Collocation extraction experiments have been performed in high number of languages, with English and German being by far the most investigated.

[43] Villada Moirón (2005) also found that LLR performed best on P-N-P data (2005, 84), and second best on PP-V data (2005, 118) after the *salience* AM (2005, 60).

After presenting the state of the art in ranking methodology as well as the most popular association measures (AMs), this chapter provided an inclusive overview of existing extraction work. This review was particularly focussed on the level of linguistic analysis performed in the preprocessing stage by extraction systems. Regardless of the AMs used for ranking candidates and of the various methods used for candidate identification (that may range from the simple consideration of adjacent words to the complex recovery of long-distance links between syntactically-related items), extraction procedures are typically preceded by a preprocessing step in which the linguistic analysis of the text is performed in order to support the identification of collocation candidates in the text (e.g., lemmatization, POS tagging, chunking or, more rarely, parsing). Sometimes, however, candidates are selected directly from plain text, with a combinatorial procedure applied to a limited context (i.e., the mobile-window method).

While this rudimentary procedure performs reasonably well for English, many of the reviewed studies pointed out, however, that for languages which exhibit a richer morphology and a freer word order, a linguistic analysis is an inescapable requirement for obtaining acceptable results. The strategy of widening the collocational window was proven inefficient for such languages (e.g., for German and Korean).

Rudimentary extraction techniques continue to be largely used nowadays even for languages for which syntactic tools became available. We argue that a detailed linguistic analysis (that goes beyond the mere morphological analysis or the superficial techniques based on pattern-matching over POS categories) is necessary for obtaining highly reliable extraction results. The dramatic advances in the parsing field now enable the shift to syntax-based approaches to collocation extraction, as the one we will present in the next chapter.

Chapter 4

Syntax-Based Extraction of Binary Collocations

This is the core chapter of the thesis, which describes and evaluates our methodology of collocation extraction based on deep syntactic parsing. Section 4.1 provides a more detailed account of the existing syntax-based work, and states the specific requirements that our extraction system aims to fulfill. Section 4.2 introduces Fips, the syntactic parser on which our approach is based. The extraction methodology proper is presented in Section 4.3, whereas Section 4.4 deals with its evaluation against a standard baseline, the window method. In the view of the experimental results obtained, Section 4.5 presents a contrastive analysis of these approaches at a more abstract level.

4.1 Introduction

As seen in the previous chapter, the past decades have witnessed a sustained activity in the area of collocation extraction, with a high number of languages being investigated. Most often, the extraction work was based on statistical techniques such as hypothesis testing (see Section 3.2.4). A wide range of word association measures were thus proposed, among which the log-likelihood ratios measure (LLR) (Dunning, 1993) was selected by numerous researchers as one of the most useful, thanks to its

particular suitability to low-frequency data.

Typically, the existing collocation extraction systems rely on the linguistic preprocessing of source corpora in order to better identify the candidates whose association strength is to be quantified by association measures. However, the analysis performed is most of the time limited to a superficial level, e.g., POS-tagging or shallow parsing implemented using pattern matching of regular expressions over POS categories. Yet, many researchers have pointed out that successful collocation extraction requires a more detailed analysis, since collocations are often syntactically flexible. As shown in Example (1) below, they can cross the boundaries of minimal constituents and even those of clauses.[1]

(1) a. *play – role*
It is true, we must combat the menace of alcoholism in young people, and this text successfully highlights the *role* that families, teachers, producers and retailers must *play* in this area.

b. *article – state*
The first *article* of the EC-Vietnam Cooperation Agreement which we signed with the government of Vietnam in 1995 *states* that respect for human rights and democratic principles is the basis for our cooperation.

c. *important – issue*
The *issue* of new technologies and their application in education naturally generates considerable interest and is extremely *important*.

Several researchers—e.g., Smadja (1993), Pearce (2002), Krenn (2000a), and Evert (2004)—pointed out that ideally, collocation extraction would rely on the syntactic analysis of source corpora in order to properly identify candidate pairs since the recent developments in the field of parsing now enable the analysis of large bodies of text:

> "Ideally, in order to identify lexical relations in a corpus one would need to first parse it to verify that the words are used in a single phrase structure" (Smadja, 1993, 151)

[1] All the examples provided in this thesis are drawn from sentences occurring in our corpora.

> "with recent significant increases in parsing efficiency and accuracy, there is no reason why explicit parse information should not be used" (Pearce, 2002, 1530)
>
> "the latter two approaches [window method and POS information] are still state-of-the-art even though the advantage of employing more detailed linguistic information for collocation identification is nowadays largely agreed upon" (Krenn, 2000a)
>
> "Ideally, a full syntactic analysis of the source corpus would allow us to extract the cooccurrence directly from parse trees" (Evert, 2004, 31).

As a matter of fact, the more recent extraction work shows a growing interest in the full parsing of source corpora prior to the statistical computation. As can be seen from the general review from Section 3.4, a number of systems are based on parsers or on syntactically-annotated text (Lin, 1998; Lin, 1999; Pearce, 2001a; Blaheta and Johnson, 2001; Zinsmeister and Heid, 2003; Schulte im Walde, 2003; Orliac and Dillinger, 2003; Wu and Zhou, 2003; Lü and Zhou, 2004; Villada Moirón, 2005). A closer look at this work reveals, however, that there is still a long way to go towards fully-fledged syntax-based extraction.

(Lin, 1998; Lin, 1999) Lin's system based on dependency parsing for English (1998; 1999) has several shortcomings. In order to reduce the number of parsing errors, only the sentences having less than 25 words are permitted, and only the complete analyses are considered for extraction (Lin, 1998, 58). Parsing errors appear to cause such a serious problem, that the author has to proceed to their semi-automatic correction before collecting collocation candidates from the output structures. The author reports that 9.7% of the output pairs checked in a small evaluation experiment involved parsing errors (Lin, 1999, 320). This error rate is rather high, given that the pairs evaluated were taken from the top-scored results.

(Wu and Zhou, 2003; Lü and Zhou, 2004) A similar error rate (7.85%) was obtained for the top collocations extracted with the system of Wu, Lü, and Zhou

(2003; 2004) based on the NLPWin parser. In fact, 157 out of 2000 randomly selected items from the results with a high LLR score were spurious candidates (Lü and Zhou, 2004). Besides, this system only extracts 3 syntactic types, namely V-O, N-A, and V-Adv, which are considered as the most important.

(Villada Moirón, 2005) The problem of parsing precision is also faced by the system of Villada Moirón (2005). The numerous PP-attachment errors made by the Alpino parser forced the author to consider an alternative, chunk-based approach for the detection of P-N-P constructions in Dutch, and an ad-hoc method for the identification of PP-V constructions. This method consists of combining each verb and each PP in a sentence; the parser only contributes with information on phrase boundaries (Villada Moirón, 2005, 97). As in the case of Lin (1998; 1999), sentences longer than 20 words were excluded, since they were problematic for the parser.[2]

(Orliac and Dillinger, 2003) The parser used by Orliac and Dillinger (2003) also suffers from several limitations. Although it succeeds in identifying predicate-argument relations from passive and gerundive constructions, it is unable to handle other constructions, like the relative ones. In an experiment that evaluated the extraction coverage, the relative constructions have been found responsible for nearly half of the candidate pairs missed by their collocation extraction system.

Other systems based on syntax As for the extraction methods based on statistical parsing, like (Pearce, 2001a; Blaheta and Johnson, 2001) for English and (Zinsmeister and Heid, 2003; Schulte im Walde, 2003) for German, their main inconvenience is that they are difficult to apply on new corpora which are different from the ones used in the parsers' training.

Given the limitations discussed above, it is apparent that the existing extraction methodology based on parsing is not yet fully developed, and its major shortcomings are related to the parsing robustness, precision, and coverage. In our opinion, the

[2]The author notes that newer versions of the parser are able to process these sentences as well.

requirements that must be met by an ideal collocation extraction system based on full syntactic parsing are the following:

(R1) being able to process robustly and acceptably fast large bodies of textual data, regardless of the domain or the characteristics of the input text (e.g., sentence length);
(R2) being able to handle the typical syntactic constructions that challenge the candidate identification process by introducing long-distance dependencies (e.g., passive-, relative-, interrogative- and cleft constructions, but also coordinations, subordinations, parenthesized clauses, enumerations, or appositions);
(R3) allowing generality in what concerns the syntactic type of combinations extracted (i.e., instead of being tailored to one or a few specific types, the system should allow the extraction of a wide range of types);
(R4) recovering candidate pairs from partial analyses, whenever a complete parse tree cannot be built for a sentence;
(R5) producing a more accurate output with respect to syntax-free methods;
(R6) if possible, supporting multiple languages, because collocations are more interesting from a cross-lingual perspective than from a monolingual one.

Among the systems reviewed above, only (Wu and Zhou, 2003) and (Lü and Zhou, 2004) could be considered as multilingual (R6), since the parser used supports both English and Chinese. (Lin, 1998; Lin, 1999) and (Villada Moirón, 2005) do not satisfy R1, concerning robustness. Regarding R2, the system of Villada Moirón (2005) uses a parser that licenses many of the syntactic constructions cited, but the extraction does not exploit the complete parser's output. It is not clear to what extent this requirement is satisfied by the statistical parsers, if they can deal with long-distance dependencies. The system of Orliac and Dillinger (2003) only satisfy R2 in part. As for R3, all systems except (Lin, 1998) and (Schulte im Walde, 2003) are limited to one or maximum three syntactic types. (Lin, 1998) does not comply with R4, because it only extracts candidates from complete parses, while ignoring the possible candidate data in the partial structures. Finally, several authors reported precision problems caused by low accuracy of parsers (Lin, 1998; Lin, 1999; Villada Moirón, 2005).

The rest of this chapter presents the underlying methodology of a collocation extraction system that is aimed at fulfilling all of the above-stated requirements. The system relies on the syntactic parser Fips, introduced in the next section.

4.2 The Fips parser

Fips (Laenzlinger and Wehrli, 1991; Wehrli, 1997; Wehrli, 2004; Wehrli, 2007) is a deep symbolic parser developed at the Language Technology Laboratory of the University of Geneva. It currently supports the following languages: English, French, Spanish, Italian, and German, while a number of other languages are under development, including Greek, Romanian, Romansch, and Japanese.

The parser is based on an adaptation of generative grammar concepts inspired by the Minimalist Program (Chomsky, 1995), the Simpler Syntax model (Culicover and Jackendoff, 2005), and by LFG (Bresnan, 2001). Each syntactic constituent is represented as a simplified X-bar structure of the form $[_{XP}$ L X R] with no intermediate level(s), where X denotes a lexical category, like N (noun), A (adjective), D (determiner), V (verb), Adv (adverb), etc. L and R stand for (possibly empty) lists of left and right subconstituents, respectively. The lexical level contains detailed morphosyntactic and semantic information available from the manually-built lexicons, namely selectional properties, subcategorization information, and syntactico-semantic features that are likely to influence the syntactic analysis (the parser thus relies on a strong lexicalist grammar framework).

Written in Component Pascal, Fips adopts an object-oriented implementation design that enables the coupling of language-specific processing modules to a generic module. The generic module is responsible of the parser's main operations, *Project* (assignment of constituent structures to lexical entries), *Merge* (combination of adjacent constituents), and *Move* (creation of chains by linking surface positions of "moved" constituents to their corresponding canonical positions). This module also defines the basic data types and the features applying to all the languages.

The parsing algorithm proceeds in a left-to-right and bottom-up fashion, by applying at each step one of the operations enumerated above. The application of the

Merge operation, in which a left or right subconstituent is attached to the current structure, is constrained by language-specific licensing rules, like the agreement rules. Moreover, the attachment can only be made to a node that is active, i.e., a node that accepts subconstituents. The alternatives are pursued in parallel, and several pruning heuristics are employed for limiting the search space.

For an input sentence, the parser provides both the phrase structure representation and the interpretation of constituents in terms of arguments, stored as a predicate-argument table that is similar to LFG's f-structure. It also provides an interpretation for clitics, wh-elements, and relative pronouns, and creates chains that link extraposed elements to empty constituents in canonical positions. The parser is able to handle a wide range of constructions, like the ones listed in Example (2) below (Fips successfully captured the syntactic relation between the words in italics in each example):

(2)
a. **passivization**: I see that *amendments* to the report by Mr Méndez de Vigo and Mr Leinen have been *tabled* on this subject.
b. **relativization**: The communication devotes no attention to the *impact* the newly announced policy measures will *have* on the candidate countries.
c. **interrogation**: What *impact* do you expect this to *have* on reducing our deficit and our level of imports?
d. **cleft constructions**:[3] It is a very pressing *issue* that Mr Sacrédeus is *addressing*.
e. **enumeration**: It is to be welcomed that the Culture 2000 programme has allocated one third of its budget to *cultural*, archaeological, underwater and architectural *heritage* and to museums, libraries and archives, thereby strengthening national action.
f. **coordinated clauses**: The *problem* is therefore, clearly a deeply rooted one and cannot be *solved* without concerted action by all parties.
g. **interposition of subordinate clauses**:[4] The *situation* in the regions where there have been outbreaks of foot-and-mouth disease is *critical*.

[3]Note that a relative reading is also possible for this example.
[4]The subordinate clause in this example is the relative introduced by *where*.

h. **interposition of parenthesized clauses**: Could it be on account of the regulatory *role* which this tax (which applies to international financial transactions) could *play* in relation to currencies [...]

i. **apposition**: I should like to emphasise that the broad economic policy *guidelines*, the aims of our economic policy, do not *apply* to the euro zone alone but to the entire single European market [...]

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

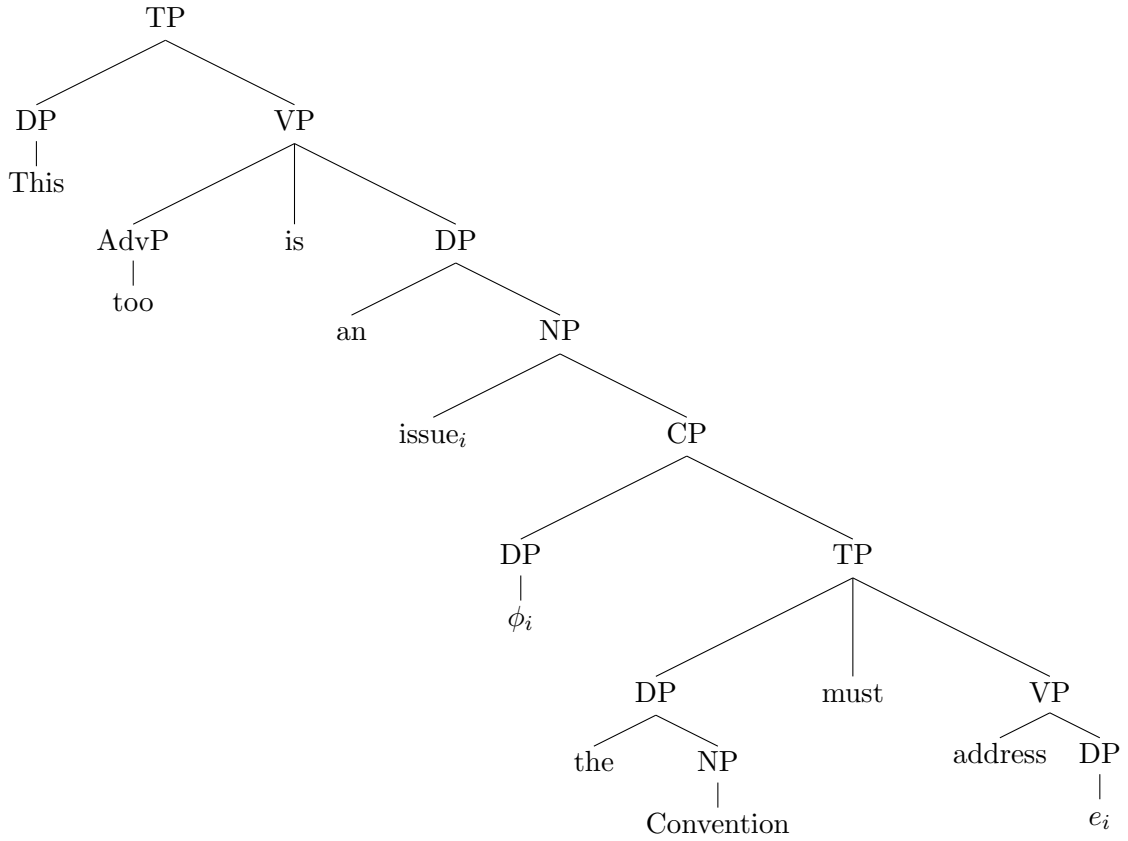


Figure 4.1: Parse tree built by Fips for the sentence *This too is an issue the Convention must address.*

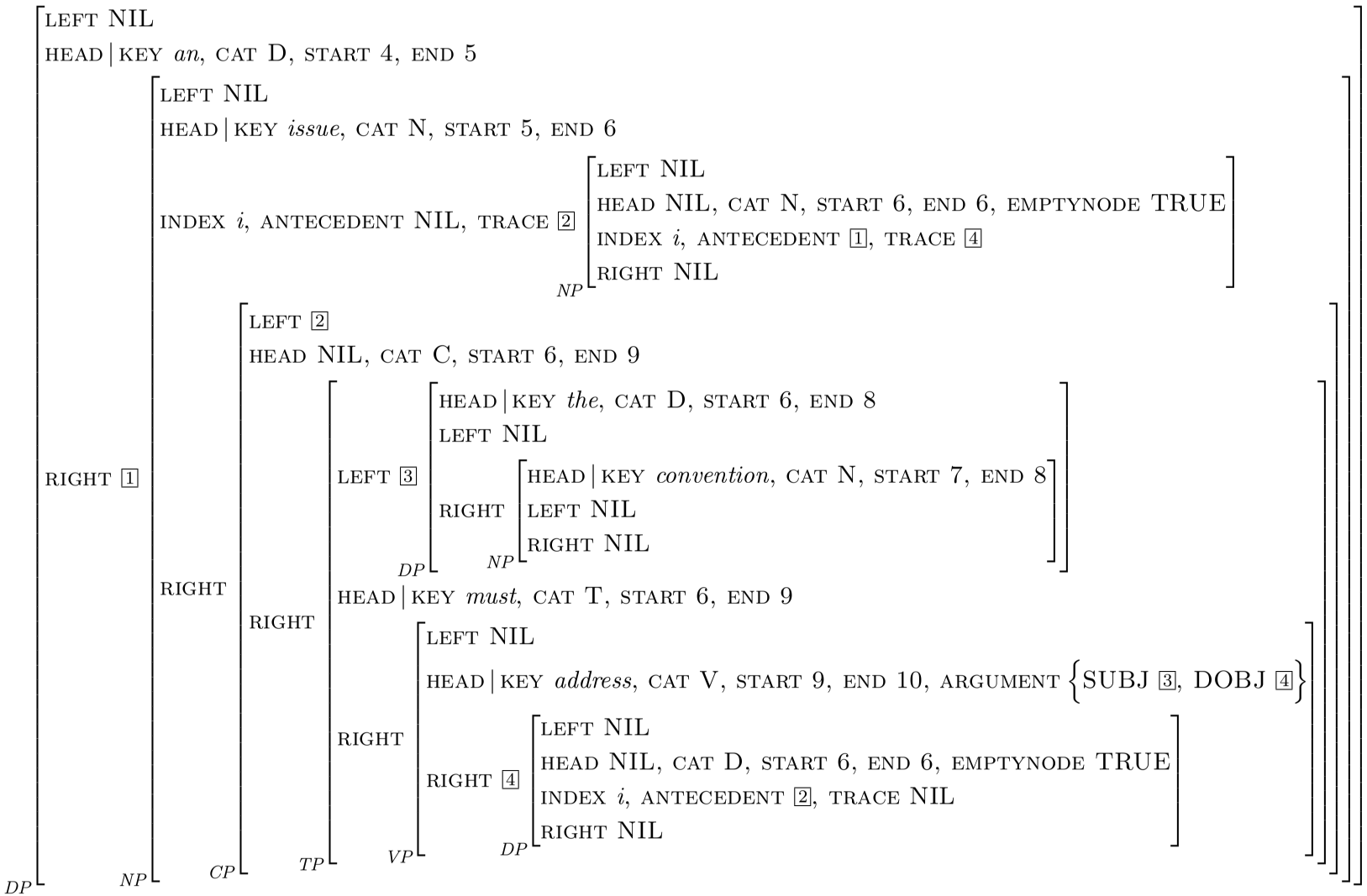


Figure 4.2: Attribute-value representation of the parse tree for the phrase *an issue the Convention must address* (partial information).

Figure 4.1 illustrates the parse tree built by Fips for the sentence in Example (3), which involves a relative construction. The relative pronoun—in this case, the zero-pronoun ϕ—is linked via co-indexation (see index i) to the antecedent *issue* and to its trace denoted by the empty constituent e found in the canonical position of direct object, next to the verb *address*. A simplified representation of the parser output, corresponding to the substructure associated to *an issue the Convention must address*, is shown in Figure 4.2.

Fips can robustly process large text corpora at a reasonable speed (approximately 150 tokens/s). Its precision was measured in the framework of a French parsing evaluation campaign, *EASy – Evaluation des Analyseurs SYntaxiques*,[5] whose definitive results will be made available in the future.

4.3 Extracting collocations with Fips

4.3.1 Candidate identification

Binary collocation candidates are identified from the parse structures built by Fips as the analysis of the text goes on. In most of the similar works based on parsing, the identification is done after parsing was completed, either by re-interpreting the textual representation of the parser's output, as in (Villada Moirón, 2005), or by reading the syntactic annotations of the source corpus, as in (Blaheta and Johnson, 2001; Pearce, 2001a). In our method, the internal representation of syntactic structures built by the parser was used directly. The benefit for the identification process is twofold: the parse information is readily available, and it contains all the rich and complex details provided for the current analysis.

The linguistic preprocessing step is therefore not entirely separated, in our system, from the candidate identification step. The two steps alternate, as candidates are identified each time a sentence was analysed by Fips. However, from an architectural point of view, the parsing and the extraction system are actually separated, the parser Fips permitting the plug-in of customized modules (such as ours that performs the

[5] `http://www.elda.org/easy`

identification of collocation candidates) for post-processing the analyses obtained for a sentence.

Each structure returned by Fips—be it a complete analysis for a sentence, or one of the partial analyses built for subconstituents when a complete analysis is not possible—is checked for potential collocation candidates by recursively examining the head of the current phrase and its left and right subconstituents (details on the typical syntactic structure built by Fips are provided in Section 4.2).

A collocation candidate consists of an oriented pair of lexemes[6] that are syntactically related. Therefore, the main criterion for selecting a lexeme pair as a candidate is the presence of a syntactic link between the two items. In addition, in order to qualify as a valid candidate, the pair must satisfy some more specific constraints. These are discussed later in this section.

Lexeme 1	**Complex 1**	**Lexeme 2**	**Complex 2**	**Type**
ban	–	animal testing	+	V-O
be	–	integral part	+	V-O
budget surplus	+	(of) full employment	+	N-P-N
civil	–	crisis management	+	A-N
draw	–	peace plan	+	V-O
elected representative	+	(of) people	–	N-P-N
electronic equipment	+	(from) mobile phone	+	N-P-N
give	–	green light	+	V-O
integral part	+	(of) social life	+	N-P-N
key point	+	be	–	S-V
local	–	public transport	+	A-N
maintain	–	close contact	+	V-O
promote	–	equal opportunities	+	V-O
protect	–	intellectual property	+	V-O
provide	–	added value	+	V-O
second	–	world war	+	A-N
strengthen	–	rule of law	+	V-O

Table 4.1: Examples of collocation candidates involving complex lexemes.

It is worth noting that each item in a candidate pair can in turn be a complex lexeme (e.g., a compound or a collocation). If present in the lexicon and successfully

[6]Prepositions are also included with noun lexemes for readability reasons.

recognized by the parser, it will be considered as a unit that can participate in other collocation candidates as a single item. Several examples of long candidates identified in an English corpus are displayed in Table 4.1.

Since the collocation identification procedure relies on the normalized form of the sentence as provided by Fips, in which words are assigned their base forms and are considered in the canonical order (e.g., with subjects preceding verbs, and objects following them), it can easily account for variation caused by inflection and inversion.

Morphological constraints on lexemes

The head lexeme X of the currently checked structure must satisfy specific constraints, depending on its category:

- noun: the lexeme must be a common noun not representing a title (e.g., *Mr.* or *General*); proper nouns are excluded.
- verb: the lexeme must be an ordinary verb; auxiliary and modal verbs are excluded (e.g., in English, *has, must*).

Syntactic constraints on candidate pairs

To form a valid candidate, the head lexeme of the current structure can combine with any element of its left or right constituents, provided that the combination involves a specific syntactic relation. For instance, a noun can combine with an adjective that is the head of (one of) its left constituent(s) since they are in a head-modifier relation. Similarly, a noun can combine with an adjective dominated by a right constituent, as in Example (4) below (the right constituent of the noun is an FP, i.e., a functional phrase, or small clause). If the adjective is participial, has a passive sense, and the matrix verb is transitive, then a verb-object relation is hypothesised between the verb (*to prepare*) and the noun (*exams*).

(4) $[_{DP}$ exams$_i$ $[_{FP}$ $[_{DP}$ $e_i]$ $[_{AP}$ prepared]]]
prepare – exam (verb-object)

The candidates of type subject-verb, verb-object, and verb-argument are easily identified from the predicate-argument tables built by Fips, even if they involve long-distance dependencies. Thus, the verb-object pair *address – issue* will be identified in a straightforward way from the sentence shown in Example (3) in which the noun *issue* is extraposed, since the argument table for the verb *address* contains *issue* in the direct object position. All the computation needed for recovering the verb-object link (i.e., recognizing the presence of a relative construction, building its normalized form with the empty constituent e in the object position, then linking e to the relative zero-pronoun ϕ and further to the antecedent *issue*, and finally adding *issue* to the argument table of *address*) is done by the parser beforehand.

Table 4.2 shows some of the most representative syntactic configurations currently used by our extraction system.[7] Several extraction examples are provided at the beginning of this chapter, illustrating the potential of our parse-based system to detect collocation candidates even from highly complex sentences (Example (1) and Example (2)).

Type	**POS combination**	**Syntactic relation**	**Example**
A-N	Adjective-Noun	head-modifier	*wide range*
N-A	Noun-Adjective	head-modifier	*work concerned*
N-N	Noun-Noun	head-modifier	*food chain*
N-P-N	Noun-Preposition-Noun	head-modifier	*fight against terrorism*
S-V	Noun-Verb	subject-verb	*rule apply*
V-O	Verb-Noun	verb-object	*strike balance*
V-P-N	Verb-Preposition-Noun	verb-argument	*bring to justice*
V-Adv	Verb-Adverb	head-modifier	*desperately need*
V-P	Verb-Preposition	verb-particle	*point out*
A-Adv	Adjective-Adverb	head-modifier	*highly controversial*
A-P	Adjective-Preposition		*concerned about*

Table 4.2: Some of the syntactic configurations accepted by our extraction system.

[7]Far from being exhaustive, this list is continuously evolving since many new combinations emerge as collocationally relevant as more data is processed (see also the considerations in Section 3.3.2 on the relative syntactic freedom of collocations).

4.3.2 Candidate ranking

For the candidate pairs identified from the syntactic structures built by Fips, the extraction system stores both syntactic information and information related to their occurrence in the corpus, which makes possible the link back to the source documents. The following information is specified for a candidate pair:

- `key1`, `key2`: the base word form of the two lexical items;
- `lex1`, `lex2`: the lexeme identification number in the parser's lexicon;[8]
- `type`: a numerical code representing the syntactic type of the pair;
- `cat1, cat2`: the grammatical category of the two items;
- `prep_key`: the preposition key, when the syntactic type contains a preposition;
- `poslex1, poslex2`: the linear position of the two items in the input sentence;
- `source`: the name of the source file;
- `charlex1, charlex2`: the position of the two items in the source file;
- `lexicalized`: a boolean value indicating if the identified pair forms a collocation that is already present in the parser's lexicon.

As indicated in Section 3.2.2 which introduced the general architecture of an extraction system, in the second extraction step an association measure is applied to the selected candidates, that assigns to each candidate a score reflecting its collocational strength. But before applying this measure, the candidate data is partitioned into syntactically homogeneous classes, according to the field `type`. This strategy appears to have a positive impact on the ranking proposed by AMs, as their performance is sensitive to the syntactic type (Evert and Krenn, 2001).

A couple of AMs have been implemented in the framework of our extraction system (as will be seen in Section 6.3). Among these, we retained the log-likelihood ratios measure (LLR) (Dunning, 1993) as the default AM for our extraction experiments, this choice being both theoretically and empirically grounded.[9] Thus, the LLR score

[8]Thanks to parsing, the readings of a lexical item are syntactically disambiguated. It might therefore happen that two pairs that are identical in form (the `key` fields are the same), are actually made up of different lexemes.

[9]Section 3.2.5 discusses in detail the issue of choosing the best AM.

is typically used for ranking the items in the candidate sets obtained, although it is equally possible to use a different AM. The method described is not tailored to a specific AM.

Unlike most of the existing extractors, our system does not impose a frequency threshold on candidate pairs, therefore no items are a priori excluded from initial pair sets. This decision is motivated, first, by the high number of infrequent pairs that might constitute collocations; second, by the good performance of LLR on low-frequency data; and, third, by the increased tractability of our method that stems from the elimination of spurious candidates from the very start (many extractors eliminate the candidates occurring less than a given threshold only to prune the initial set in order to cope with the problem of computational complexity).

As in the case of frequency, the system sets no threshold for the LLR score, all the initial pairs being returned regardless of how high a score they obtained.[10] When LLR is not applicable to a candidate pair (this situation occurs when a value in its contingency cell is 0), the pair receives by convention a minimal score (0). Also, the candidate pairs that are already present in the parser's lexicon[11] are marked as `lexicalized` and are not ranked; instead, they receive a conventional maximum score by default. The frequencies of the individual lexemes are, however, taken into account in the LLR computation, because they contribute to the frequency signature of other candidates (i.e., as explained in Section 3.2.3, to the corpus frequencies listed in their contingency table).

4.4 Evaluation

This section presents two evaluation experiments that compare the performance of our method with that of a syntax-free method that is typically used for collocation extraction, namely the mobile-window method. While our method uses syntactic information provided by the parser Fips, this standard procedure only takes into

[10] The selection of higher-scored pairs can be operated a posteriori, according to the desired degree of confidence.

[11] As noted in the previous section, the system can recognise those pairs of lexemes that make up known collocations, i.e., collocations that are stored in the parser's lexicon.

account the linear word proximity and ignores the syntactic relations between words.

From a theoretical point of view, a syntax-based method is expected to perform better (as argued in Section 3.3.3). But this theoretical claim must be empirically proven in an actual extraction setting, because the errors that are inherent to parsing risk to produce more extraction noise (i.e., ungrammatical pairs) than the window method would produce. In fact, the review of syntax-based extractors provided in Section 4.1 showed that the parsing errors may sometimes pose such serious problems that the authors are forced either to correct the parser's mistakes (Lin, 1998; Lin, 1999), or to discard the attachments proposed by the parser and only use information on phrase boundaries (Villada Moirón, 2005).

It is worth mentioning here that the results of the window method tend to be more accurate as they are situated higher in the significance list, thanks to the higher frequency in the source corpus. Ungrammatical pairs tend to be demoted by AMs to lower ranks because their frequency is lower. As more and more data is added to the extraction system, the precision of the top window results is expected to increase. If this precision is comparable to that achieved with syntax-based methods, then there would be no need for parsing (provided that one is interested only in the upper part of the significance list, i.e., in the pairs whose score is higher than a given threshold). Adding more data also compensates for the long-distance pairs that are missed by the window method; thus, again, using a parser to capture these pairs might not be considered a real necessity.[12]

The rest of this section describes the evaluation method applied, provides details on the implementation of the mobile-window method, and presents the two evaluation experiments performed. The difference between these experiments lies in the source corpora used in extraction, the languages dealt with, and, more importantly, in the level of the significance list that was investigated and in the classification used for annotating the extraction output.

[12]See Section 4.5 for a discussion on the effect that ignoring such long-distance pairs has on the extraction results.

4.4.1 Evaluation method

The evaluation experiments compare, on the one hand, the results of our extraction method based on full parsing (described in Section 4.3) with, on the other hand, the results obtained with the standard mobile-window method (that will be described in Section 4.4.2).

Both extraction methods use LLR (Section 3.2.4) as an association measure in the candidate ranking step; therefore, only the candidate identification step is different. In accordance with Evert and Kermes (2003), we consider that the collocation extraction methods should not be evaluated at the end of the extraction process, but separately after each extraction step. More precisely,

- after the candidate identification step, in order to evaluate the quality of proposed candidates with a specific criterion (such as the grammatical wellformedness criterion);
- after the candidate ranking step, in order to evaluate the quality of the extraction output in terms of collocability, and, indirectly, the efficiency of the AM used.

Since the two methods use the same AM in the second extraction step, we can still compare their final results directly. At the same time, we will be able to assess the impact of the candidate identification step on the extraction output.

First, different test samples are extracted from the output of the two methods at various levels in their significance lists. Each item in these test sets is annotated by at least two judges with a label from a given set.[13,14] The items that are identically annotated by the majority of judges are associated with the dominant label (we further refer to this label as *mark*). They are retained in the reference set, while the items that are not agreed upon are discarded.

Finally, the methods are compared by taking into account the precision obtained relative to the reference set. This is computed as the percentage of true positives in

[13]The levels and the annotation labels considered are different in the two experiments performed.
[14]The judges were chosen so that they are native or near-native speakers of the language concerned.

the test sets, where a true positive is a pair from the reference set that was marked as a collocation. The higher the precision, the better the method.[15]

Since it is rather difficult to judge the result pairs isolated from their source context (especially if they do not incorporate syntactic information, as it is the case for the window method), the annotators were assisted in their task by a concordance tool, part of our collocation extraction system (Section 6.3), that displays the context in the source file for the all the instances of an extracted pair. Referring back to the original contexts is a necessary condition for the evaluation of extraction results, since, as Evert and Kermes (2003) point out, a pair may look like a true positive when looked at in isolation, while in reality it is an extraction error with respect to the corpus (i.e., the pair is ungrammatical in the source context).

4.4.2 Implementation of the window method

The window method was implemented as follows:

Step1: Lemmatization. First, the source corpora are POS-tagged and lemmatized. To this end, we used the parser Fips, as in the case of our parse-based method.[16] As a consequence, the POS-ambiguity is practically eliminated, since only the POS tags that are compatible with the sentence context are kept among those possible for a token.[17]

Step2: POS-filter of lexemes. Function words and auxiliary verbs are ruled out, so that only the content words are retained (nouns, adjectives, verbs, and adverbs). Punctuation marks are in this step retained, for the purpose that is explained below (Step 3).

[15] A high recall is also essential for an extraction system in order to achieve a good performance, but recall-based evaluation is currently hampered by the lack of adequate tools and resources.

[16] This choice can be seen as biasing the candidate identification process, since parsing errors are reflected in the POS tags assigned. We argue, however, that the assignment of tags in case of ambiguity is more precise if done with Fips than without parsing information, and that, on the contrary, our choice makes the two methods more comparable: rather than introducing errors with another POS tagger, we end up by having the same errors, and we can more easily highlight the differences between the two extraction approaches.

[17] According to a study by Hajič (2000) cited in Section 3.3.3, about 40% of the tokens in an English text are POS-ambiguous.

Step 3: Generation of candidate pairs. Candidate pairs are subsequently identified as oriented combinations inside a 5 content-word window (i.e., by allowing a maximum of 4 content words in-between). In order to avoid combinations that cross sentence or clause boundaries, no punctuation marks are allowed between the two items of a pair.

Step 4: POS-filter of candidate pairs. In addition, among all of the possible POS combinations, only those that suggest a syntactic relation have been retained, namely, N-A, A-N, N-N, N-V, and V-N, which are typically taken into account in the existing extraction systems.[18] Table 4.3 displays the corresponding syntactic configurations in our extraction system. In order to facilitate the comparison between the two methods, we restricted the output of our method accordingly.

Window method	**Parse-based method**
Adjective-Noun (A-N)	Adjective-Noun (A-N)
Noun-Adjective (N-A)	Noun-Adjective (N-A)
Noun-Noun (N-N)	Noun-Noun (N-N), Noun-Preposition-Noun (N-P-N)
Noun-Verb (N-V)	Subject-Verb (S-V)
Verb-Noun (V-N)	Verb-Object (V-O), Verb-Preposition-Noun (V-P-N)

Table 4.3: POS combinations used by the window method and the corresponding syntactic configurations in the parse-based method.

Step 5: Candidate ranking. Finally, as in our method, the log-likelihood ratios (LLR) score has been computed for all the combinations obtained, after partitioning them into homogeneous sets defined by the POS combination type (no frequency threshold was applied). The aim of this partitioning is to apply LLR on syntactically homogeneous data, but since no syntactic information is available, the POS combination type was taken into account.

It is important to note that the window method implemented as above represents a rather high baseline against which we aim to compare our method based on parsing.

[18]For instance, combinations involving an adverb have not been considered, since ignored in most window-based extraction system. However, adverbs were retained in Step 2, since they are taken into account in determining the distance between the items of a candidate pair.

A number of choices have been made that are likely to alleviate the candidates identification process and, in particular, to increase the precision of the window method: the POS disambiguation based on parsing information, the elimination of pairs with interposed punctuation marks, and the selection of only those pairs that suggest a grammatical relationship.

4.4.3 Comparative evaluation - Experiment 1

The first comparative evaluation experiment was performed on French data from the Hansard corpus of Canadian Parliament debates. It investigated the top 500 pairs returned by each of the two methods described (the parse-based method and the mobile-window method). The pairs were annotated using a rough classification with 3 categories: *ungrammatical pair*, *regular pair*, and *interesting pair*.

Experiment 1 - Data

A number of 112 files were chosen for this experiment, that cover one month of the Canadian Parliament proceedings[19] and total slightly more than 1.2 million words.

Statistic	Value
size (MB)	8.02
files	112
words (approx.)	1209050
tokens	1649914
sentences	70342
average file length (sentences)	628.1
average sentence length (tokens)	23.46
sentences with complete parses	50458
percentage sentences with complete parses	71.7
parsing speed (tokens/s)	172.2

Table 4.4: Statistics on the corpus used in Experiment 1.

Table 4.4 provides more numerical details about the corpus, as well as several statistics available from the syntactic analysis that was performed with Fips in the

[19]Between January 17th, 1994 and February 17th, 1994.

preprocessing stage of our method. Collocation candidates were extracted from this corpus using, on the one hand, our method based on full parsing (Section 4.3), and, on the other, the mobile-window method which, unlike ours, does not rely on syntactic information (Section 4.4.2).

Table 4.5 presents some comparative statistics on the extraction results: the total number of candidate pairs extracted by the two methods, the number of distinct candidate pairs, the total number of candidate pairs by syntactic type (cf. the correspondences presented in Section 4.4.2), the total number of pairs scored (i.e., pairs for which LLR is defined),[20] and the number of distinct pairs scored.

	Window method	**Parse-based method**
pairs extracted	1024887	370932
pairs extracted (distinct)	560073	147293
Adjective - Noun	98461	15847
Noun - Adjective	127498	40558
Noun - Noun	333606	50601
Noun - Verb	174474	38773
Verb - Noun	290848	89966
pairs scored	1018773	308410
pairs scored (distinct)	554202	131384

Table 4.5: Extraction statistics for Experiment 1.

As can be seen, despite the strong filter applied by the window method (see Section 4.4.2), the number of candidate pairs that are generated still outweighs the number of syntactically filtered candidates. This translates into increased computational complexity and higher difficulty in handling the candidate data with respect to the parse-based method, in particular during the score computation.

The number of content words from which the window pairs were generated is 673789; if adverbs are excluded (because they do not participate in combinations), it becomes 587686. The initial POS filter actually retained 822753 tokens, including punctuation marks that play a role in the selection of candidate pairs.

[20]LLR is not defined for those pairs which contain a null value in their contingency table.

Experiment 1 - Test sets

This experiment compares the top 500 output items obtained with each method, where an output item is represented by a pair type (as opposed to a pair instance, or token). A slightly different comparison of the same test data was made in (Seretan and Wehrli, 2006a), where we performed an n-best evaluation, with n ranging from 50 to 500 at intervals of 50. In the present experiment, we no longer measure the precision in a cumulative fashion, but report the values obtained on each segment of 50 pair types separately; thus, we measure the precision on 10 consecutive test sets at the top of the significance lists. This evaluation procedure has the advantage of providing a clearer picture of the performance of one method versus another, since the local precision (as opposed to the top precision) is available for different segments of the output list.

Method	**TS1**	**TS2**	**TS3**	**TS4**	**TS5**	**TS6**	**TS7**	**TS8**	**TS9**	**TS10**	**All**
Parsing	12147	4498	3435	2507	2183	1872	1536	1176	1404	1020	31778
Window	17241	5586	4696	3009	3195	2872	2739	2297	1767	1920	45322

Table 4.6: Number of pair tokens in the test sets from Experiment 1.

The total number of pair types evaluated for the two methods is 1000. We illustrate some of the data tested in Appendix D, which shows the first and the last test set evaluated for each method. Appendix E displays, in addition, the labels assigned by annotators to each item in these test sets. Table 4.6 lists the total number of pair tokens in each test set.

Experiment 1 - Annotation categories

In this first experiment, each pair tested was classified by annotators either as *ungrammatical*, *regular*, or as *interesting* combination. The first two cases correspond to false positives, while the items assigned the third category represent, by convention, true positives. Each category is described below.

ungrammatical pair (label 0) - pair in which an item has a wrong POS category

(e.g., *entreprendre petite*[21]), or in which the items are not syntactically related (e.g., *président de élection*, extracted from a context like *je voudrais donc saisir cette occasion pour féliciter le Président de son élection*);

regular pair (label 2) - pair that is grammatically correct, but that is uninteresting from a lexicographic point of view, since it is completely regular and allows paradigmatic variation (e.g., *économie canadienne*);

interesting pair (label 1) - grammatical pair that is worth storing in the lexicon, since it constitutes (part of) a multi-word expression, be it compound, collocation, idiom, named entity, etc. (e.g., *prendre la parole*, *emploi à long terme*).

As collocations are notoriously difficult to distinguish from other subtypes of multi-word expressions (McKeown and Radev, 2000) and there are no objective criteria that can be applied for this task (Section 2.5), we first used this coarse-grained classification, which does not separate collocations from other MWE pairs. As in (Choueka, 1988) and (Evert, 2004) (see definitions in Appendix B), we consider that the dominant feature of collocations is that they are unpredictable for non-native speakers and therefore have to be stored in a lexicon.

Experiment 1 - Reference set

The performance of the two methods in terms of grammatical and MWE precision is reported by taking into account the annotations produced by the human judges. As an alternative to the expensive manual annotations, the precision can be reported relative to existing gold-standards, i.e., (machine-readable) dictionaries of collocations or multi-word expressions, as in (Pearce, 2001b; Daille, 1994). The only problem with this approach is the low coverage of such resources.[22] As a matter of fact, most

[21] This pair stems from the A-N combination *petite entreprise*, which is syntactically ambiguous and can also be interpreted as an V-N pair (*entreprendre petite*).

[22] For instance, Daille (1994, 145) reported that only 300 out of the 2200 terms she tested were found in a reference list containing about 6000 terms from the same domain as the source corpus, i.e., that of satellite telecommunications. When the domain is not the same, the intersection is virtually insignificant.

extraction work is evaluated on the basis of judgements produced by annotators, who are, in the best case, lexicographers, as in (Smadja, 1993).

In our case, a team of 3 judges evaluated the output of each method (there were two different teams, one per method). The reference sets included those items in each test set that received a consistent annotation from at least 2 annotators. The reference label assigned to a pair (its mark) was the dominant label. The reference set built from the output of the parse-based method contains 496 items overall, and the one corresponding to the window method 488 items.

The first two rows of Table 4.7 display the raw inter-rater agreement for each of the 10 test sets, computed as the percentage of pairs on which at least two annotators agreed. The last two rows report the Fleiss' kappa inter-annotator agreement on each test set.

Fleiss' kappa (Fleiss, 1981) is a measure of chance-corrected inter-annotator agreement that applies to more than two raters. Its computation is, in principle, similar to that of Cohen's kappa statistic (Cohen, 1960) (described in Section 4.4.4) that quantifies the agreement between two raters; the same scale is used for interpreting the results.

Agr		**TS1**	**TS2**	**TS3**	**TS4**	**TS5**	**TS6**	**TS7**	**TS8**	**TS9**	**TS10**	**Avg**
raw	P	98	98	100	100	100	100	98	98	100	100	99.2
raw	W	100	96	100	98	100	96	96	98	96	96	97.6
kappa	P	0.42	0.39	0.27	0.35	0.18	0.37	0.37	0.42	0.40	0.56	0.37
kappa	W	0.34	0.44	0.51	0.50	0.53	0.62	0.45	0.49	0.54	0.33	0.47

Table 4.7: Agreement statistics for the test sets in Experiment 1: raw agreement and Fleiss' kappa score (odd rows correspond to the parse-based method; even rows to the window method).

The overall Fleiss' kappa computed on the entire test data (when the test sets are considered altogether) is 0.39 for the parse-based method and 0.50 for the window method; according to the interpretation scale, the agreement among the three annotators can be considered as *fair* in the first case and *moderate* in the second. As for the percentage of complete raw agreement (i.e., items that were equally annotated by *all* of the three judges), it is 57.0% (285/500) and 55.6% (278/500), respectively.

The agreement results show that despite the rough classification used, there are still numerous cases of ambiguity, in which one category cannot be easily distinguished from another. For instance, the distinction between multi-word expressions and regular combinations is particularly difficult to draw.

In fact, if we leave out the cases in which a pair was annotated by at least one judge as ungrammatical, we end up with 198 cases of ambiguous pairs in the output of the parse-based method. From these, 107 pairs were considered as MWE by the majority of judges, while the other judge considered them as regular; for the other 91, the opposite holds. In the output of the window method, there are 135 such pairs, divided into 86 of the first type (MWEs with tendency towards regularity) and 49 of the second type (regular combinations with MWE tendency).

We provide several examples of such ambiguous pairs below (more examples can be found in Appendix E).

- MWEs with tendency towards regularity: *attention particulier*, *communauté internationale*, *consentement unanime*, *créer emploi*, *développement de ressource*, *dire mot*, *exploitation forestier*, *exprimer point de vue*, *faire travail*, *frai partagé*,[23] *grand nombre*, *grave problème*, *lésion grave*, *membre de famille*, *même chose*, *personne handicapé*, *remettre à travail*, *ressource naturel*, *retirer troupe*, *seuil pauvreté*, *vie humain*, *vote libre*;

- regular combinations with MWE tendency: *an prochain*, *argent de contribuable*, *construction de pont*, *corriger situation*, *déposer rapport*, *dernier année*, *développement régional*, *fin de guerre*, *opposition officiel*, *présidente suppléant*, *processus de consultation*, *programme établi*, *proposer amendement*, *relation de travail*, *représentant élu*, *situation financier*.[24]

[23]This unusual display for *frais* using the singular number (*frais* meaning *cost* does not accept a singular form) was due to a different sense of *frais* assigned by the parser, as plural of *frai*, 'spawn' (the error has now been corrected by modifying the priority of this reading in the parser's lexicon).

[24]Note that the pairs contain lemmas rather than word forms.

Experiment 1 - Results

Table 4.8 below reports the evaluation results obtained in the experiment described above for each of the 10 test sets considered. The same results are graphically presented in Appendix F.

Method	**TS1**	**TS2**	**TS3**	**TS4**	**TS5**	**TS6**	**TS7**	**TS8**	**TS9**	**TS10**	**Avg**
Window	92.0	84.0	82.0	72.0	72.0	78.0	62.0	80.0	74.0	68.0	76.4
Parsing	100.0	95.9	100.0	98.0	100.0	98.0	100.0	100.0	100.0	98.0	99.0
Window	76.0	68.8	70.0	53.1	56.0	43.8	50.0	57.1	50.0	43.8	56.9
Parsing	73.5	73.5	74.0	74.0	54.0	68.0	53.1	63.3	62.0	64.0	65.9

Table 4.8: Evaluation results for Experiment 1: grammatical precision (rows 1, 2) and MWE precision (rows 3, 4).

Grammatical precision and MWE precision are computed by taking into account the mark of annotated pairs (0 – *ungrammatical*, 1 – *interesting*, 2 – *regular*), as follows:[25]

$$Grammatical\ precision\ for\ TS_i = \frac{number\ of\ items\ in\ TS_i\ with\ mark \neq 0}{number\ of\ items\ in\ TS_i\ having\ a\ mark} \quad (4.1)$$

$$MWE\ precision\ for\ TS_i = \frac{number\ of\ items\ in\ TS_i\ with\ mark = 1}{number\ of\ items\ in\ TS_i\ having\ a\ mark} \quad (4.2)$$

As can be observed in Table 4.8, the method based on parsing outperforms the typically-used window method. On the 10 test sets inspected from the top of the significance lists, the parse-based method is on average by 22.6% better in terms of grammaticality. Thus, parsing contributes to a drastic reduction in the error rate, from 23.6% to only 1%.

The very first results of the window method are still acceptable, as hypothesized at the beginning of this section—on average, 86% of the pairs from the first 3 sets are grammatical. But on the remaining test sets, the average drops to 72.3%, and the precision curve shows the tendency to fluctuate; in contrast, that of the parse-based

[25]Recall from Section 4.4.1 that the mark represents the dominant label of an annotated pair.

method remains stable at an optimal level of 99-100%. The average precision of our method on the 10 test sets is 99%.

As for the MWE precision, the parse-based method is on average 9.1% better than the window method on the 10 test sets considered. This means that parsing contributes to the discovery of a higher number of MWEs (i.e., potential collocations). However, the window method performs quite well on the first test sets considered, corresponding to the top 150 items in the significance list. With an average of 71.6% for the first 3 sets, the precision of the window method reaches a level comparable to that of the parse-based method (73.7%), and is even slightly higher on the first set. On the remaining sets, however, it rapidly degrades to an average of 50.5%, while the precision of the method based on parsing remains stable at about 62.6%.

Overall (when the union of the 10 test sets is considered for each method), there are 278 pairs that were marked by annotators as MWE for the window method, and 327 for the parse-based method; similarly, 382 pairs are grammatical in the case of the window method, and 491 in that of the parse-based method. Therefore, parsing helped discover 1.18 times more MWEs than the window method, and retrieved 1.29 times more grammatical results.

The conclusion that can be drawn from Experiment1 is that, according with the theoretical expectations and despite the challenging task of parsing large text corpora, a syntax-based approach to collocation extraction seems very helpful, insofar as it leads to a substantial improvement[26] in the quality of results.

The difference with the window method is particularly big after the very top results (in our experiment performed on a corpus of about 1.2 million words, the difference was noted for the results situated lower than rank 150 in the significance list). It is also worth noting that the window method we compared our method against is an enhanced method that represents a high baseline, since a series of choices were made that would tend to boost the quality of its output (notably, the POS disambiguation and the elimination of candidate pairs that overstep punctuation, as explained in Section 4.4.2).

[26]The overall difference obtained in both grammatical and MWE precision is statistically significant (at α=0.001 in the first case and at α=0.05 in the second). The difference noted for the first test set is *not* statistically significant.

A closer view at the reference sets built with the help of annotators revealed 38 cases of inconsistent annotations between the reference set corresponding to the parse-based method and to the window method (for instance, in test set 1 shown in Appendix E, the following pairs received inconsistent marks: *gouvernement fédéral*, *président suppléant*, *vote libre*).

These cases involve 14 pairs labeled as MWEs in the first set (parsing) and as regular combinations in the second (window); the pairs in the opposite situation are more numerous, namely 24. This means that the difference in MWE precision might be even higher in favour of the parse-based method, if these cases of ambiguity were settled. The inconsistencies are due to the fact that the teams of judges that annotated the output of the two methods were different. At the end of the annotation process, we did not check for inter-team inconsistencies, but only for intra-annotator inconsistencies (the pairs labeled differently by a *same* annotator were solved by that annotator).

We nevertheless tried to solve these disagreements by assigning the overall dominant label to a pair, i.e., the most frequent among the 6 labels produced by the two teams of judges in total. Our attempt only succeeded in a minority of cases (6 out of 38), and therefore their impact on the reported results is very little. In the majority of cases (32 out of 38), there was a perfect balance among the labels of a pair, with 3 judges choosing the label *interesting pair* and the other 3 judges the label *regular pair*. Some examples of pairs in this situation are: *ministre chargé*, *payer impôt*, *réduire déficit*, *réduire dépense*, *région rural*, *répondre à question*, *solution à problème*. This result confirms once again the difficulty of this apparently simple task that consists in separating MWEs from regular combinations.

We believe that, despite the rough classification used to annotate the pairs, the comparison performed in this experiment is nonetheless meaningful, since the extraction results must first be checked for grammaticality and distinguished from regular combinations, before a more detailed evaluation takes place.

4.4.4 Comparative evaluation - Experiment 2

In the second evaluation experiment, we performed a finer classification of the output pairs with respect to the first experiment. The annotation used 6 categories: *ungrammatical pair*, *regular pair*, *named entity*, *collocation*, *compound*, and *idiom*. In addition, multiple levels of the significance lists have been investigated. Also, the experiment involved data in 4 languages (French, English, Spanish and Italian), taken from the Europarl parallel corpus (Koehn, 2005).

Experiment 2 - Data

In this experiment we used 62 files for each language currently supported by the parser. The files correspond to the complete 2001 collection of European Parliament proceedings and total between 3.5 and 3.9 million words per language. More detailed statistics on this corpus are presented in Table 4.9.

Statistic	**English**	**French**	**Italian**	**Spanish**
size (MB)	21.4	23.7	22.9	22.7
files	62	62	62	62
words (approx.)	3698502	3895820	3829767	3531796
tokens	4158622	4770835	4134549	4307360
sentences	161802	162671	160906	172121
average file length (sentences)	2609.7	2623.7	2595.3	2776.1
average sentence length (tokens)	25.7	29.3	25.7	25.0
sentences with complete parses	92778	94100	67276	87859
percentage sentences with complete parses	57.3	57.8	41.8	51.0
parsing speed (tokens/s)	96.1	113.9	138.1	162.0
POS-filtered tokens (incl. punctuation)	2318619	2416287	2457829	2340266
POS-filtered tokens (excl. punctuation)	1911094	2005377	2063329	1915229
POS-filtered tokens (excl. adverbs)	1659092	1736375	1854027	1694642

Table 4.9: Statistics on the corpus used in the second evaluation experiment.

Table 4.10 display some numerical results on the collocations extracted from this corpus by the parse-based method and by the window method. (The correspondences between the syntactic configurations used in the first method and the POS combinations used in the second are listed in Table 4.3.)

Statistic	Method	English	French	Italian	Spanish
pairs extracted	Parsing	851500	988918	880608	901224
	Window	3055289	3131272	3463757	3204916
pairs extracted (distinct)	Parsing	333428	327366	333848	315532
	Window	1445686	1426873	1365957	1359584
Adjective - Noun	Parsing	124176	42082	81438	62275
	Window	459590	341983	399005	362349
Noun - Adjective	Parsing	8067	148978	153051	132628
	Window	294736	463317	489510	445287
Noun - Noun	Parsing	135073	172918	161659	137908
	Window	1031305	1054456	1056203	1024776
Noun - Verb	Parsing	82602	27175	97179	77018
	Window	505983	462460	581554	504364
Verb - Noun	Parsing	223134	233146	194354	207033
	Window	763675	809056	937485	868140
pairs scored	Parsing	823062	962794	855139	883207
	Window	3046384	3123258	3458888	3200576
pairs scored (distinct)	Parsing	314288	310815	327276	306862
	Window	1437070	1419146	1365402	1376028

Table 4.10: Extraction statistics for Experiment 2.

In Table 4.9, the last three rows present the number of POS-filtered tokens in the source corpus that led to the candidate pairs generated by the window method using the POS combination procedure described in Section 4.4.2.

Experiment 2 - Test sets

The performance of the two methods was evaluated on 5 different test sets, for each of the 4 extraction languages. Each test set contains 50 contiguous items situated at different levels in the significance lists: 0% (top), 1%, 3%, 5% and 10%.[27] The number of pair tokens in each test set is displayed in Table 4.11. Overall, a number of 2000 pair types have been evaluated in this experiment.

[27]The levels chosen are not as small as they might seem at the first sight, because they refer to pair types rather than to pair tokens; besides, the corpus processed is fairly large, and no frequency threshold was applied to the candidate pairs.

Language	Method	TS1	TS2	TS3	TS4	TS5	All
English	Parsing	16215	1030	226	185	88	17744
	Window	27960	770	224	270	81	29305
French	Parsing	19912	647	411	233	67	21270
	Window	33232	363	246	212	137	34190
Italian	Parsing	28935	702	362	80	91	30170
	Window	46884	480	371	171	139	48045
Spanish	Parsing	25638	866	348	265	102	27219
	Window	31462	353	430	194	121	32560

Table 4.11: Number of pair tokens in the test sets from Experiment 2.

Experiment 2 - Annotation categories

This evaluation experiment aims at a more precise quantification of the performance of the two methods in terms of their potential to retrieve collocations, since, unlike in Experiment 1, it attempts to distinguish collocations from other multi-word expressions. The general category of MWE has therefore been split into 4 more specific categories. Each item in the test set is associated by annotators with one of the following labels:

ungrammatical pair (label 0) - same interpretation as in Experiment 1: POS error or syntactically unrelated words (e.g., the 'A-N' pair *gross domestic* extracted from a sentence like *We have a budget surplus of nearly 5% of our* gross domestic *product*);

regular pair (label 1) - same interpretation as in Experiment 1: pair that is grammatically correct, but uninteresting from a lexicographic point of view, since it is completely regular and allows paradigmatic variation (e.g., *next item*);

named entity (label 2) - pair that constitutes (part of) a proper noun (e.g., *European Commission*);

collocation (label 3) - pair that constitutes (part of) a collocation, in the acceptation we adopted in Section 2.6: the meaning of headword is preserved; the collocate typically combines with the headword, while paradigmatic variation is usually not allowed (e.g., *play role*);

compound (label 4) - pair that constitutes (part of) a compound word, i.e., a

combination that is inseparable and acts like a single lexeme; (e.g., *great deal*, part of *a great deal*);

idiom (label 5) - pair that constitutes (part of) an expression whose meaning is opaque or figurative; the meaning of the headword is not preserved (e.g., *hit nail*, part of *to hit the nail on the head* which means 'to be right about something').

Experiment 2 - Reference set

The 5 test sets extracted from the output of the two methods for each language were evaluated by 4 teams of two judges (one team per language). Intra-annotator disagreements, i.e., inconsistent annotations for a same annotator, were identified and solved. The items that were identically annotated by both members of a team were included in the reference set. Overall, 1437 out of the 2000 evaluated pairs satisfied this condition: 650 for the parse-based method, and 787 for the window method.

	Method	**TS1**	**TS2**	**TS3**	**TS4**	**TS5**	**TS1**	**TS2**	**TS3**	**TS4**	**TS5**
		EN					ES				
raw	Parsing	86	66	48	70	62	56	64	52	56	50
	Window	88	90	86	80	82	66	84	74	80	72
kappa	Parsing	0.73	0.57	0.20	0.50	0.68	0.43	0.57	0.19	0.52	0.15
	Window	0.86	0.94	0.85	0.87	0.61	0.68	0.73	0.66	0.77	0.65
		FR					IT				
raw	Parsing	74	62	62	54	70	74	78	70	74	72
	Window	70	78	78	84	70	70	82	78	84	78
kappa	Parsing	0.69	0.41	0.45	0.20	0.49	0.61	0.74	0.62	0.63	0.67
	Window	0.73	0.70	0.63	0.90	0.62	0.83	0.77	0.45	0.52	0.80

Table 4.12: Agreement statistics for the test sets in Experiment 2: raw agreement and kappa score.

Table 4.12 displays the agreement statistics for each test set. Rows 1 and 2 show the raw agreement, defined as the percentage of pairs on which both annotators agreed. Rows 3 and 4 display the Cohen's kappa inter-annotator agreement.

The *kappa statistic* (Cohen, 1960) is a measure of agreement that tries to factor out the agreement due to chance. It is computed according to the formula in Equation 4.3, in which *actual* represents the number of observed agreements, *expected* – the

number of agreements due to chance, and *trials* – the total number of items that were annotated.

$$kappa = \frac{actual - expected}{trials - expected} \tag{4.3}$$

Given a label l, the chance-expected agreement on that label, $expected_l$, is computed as follows:

$$expected_l = \frac{total\ of\ labels\ l for\ judge\ A \times\ total\ of\ labels\ l\ for\ judge\ B}{trials} \tag{4.4}$$

The number of chance-expected agreements is then computed as the sum of chance-expected agreements over all labels:

$$expected = \sum_i expected_i \tag{4.5}$$

Kappa values lie in the interval [-1, 1]. A kappa value of 1 indicates a perfect inter-annotator agreement; a kappa value of 0 indicates agreement equivalent to chance. Negative values correspond to less-than-chance agreements, i.e., to disagreements. Intermediate values are usually interpreted according to the following conventional scale (Landis and Koch, 1977):

- $0 \leq$ kappa < 0.2 - *slight* agreement;
- $0.2 \leq$ kappa < 0.4 - *fair* agreement;
- $0.4 \leq$ kappa < 0.6 - *moderate* agreement;
- $0.6 \leq$ kappa < 0.8 - *substantial* agreement;
- $0.8 \leq$ kappa < 1 - *almost perfect* agreement.

The kappa values obtained in our annotation experiments (Table 4.12) indicate that a rather high agreement was achieved, given the difficulty of the classification task. From the 40 test sets, the agreement is *almost perfect* on 6 test sets, *substantial* on 20 sets, *moderate* on 10 sets, *fair* on 2 sets and *slight* on the remaining 2 sets. Lower kappa values were found to originate in repeated disagreements of the same

type (i.e., when two labels are systematically mixed up by the two annotators). The kappa statistic drastically penalizes high values situated outside the main diagonal in the confusion matrix, even if most of the other cases are agreed upon.

On the whole annotations set (when all test sets for both methods are considered together), the raw agreement is 71.9% and the kappa is 0.61, which indicates a *significant* inter-annotator agreement overall.

Experiment 2 - Results

We report in Table 4.13 the comparative evaluation results obtained with the two extraction methods on the 5 test sets (corresponding to 5 different levels in the significance list) that were considered for each language. A graphical representation of these results is provided in Appendix I.

	Method	**TS1**	**TS2**	**TS3**	**TS4**	**TS5**	**TS1**	**TS2**	**TS3**	**TS4**	**TS5**
		EN					ES				
Gram.	Parsing	97.7	97.0	100.0	88.6	71.0	100.0	96.9	92.3	92.9	84.0
	Window	86.4	35.6	32.6	25.0	36.6	72.7	9.5	13.5	15.0	27.8
MWE	Parsing	67.4	75.8	66.7	31.4	25.8	71.4	40.6	46.2	35.7	16.0
	Window	47.7	15.6	7.0	12.5	4.9	54.5	7.1	10.8	12.5	16.7
Colloc.	Parsing	41.9	69.7	58.3	31.4	16.1	39.3	31.3	42.3	32.1	16.0
	Window	31.8	11.1	7.0	10.0	4.9	36.4	7.1	10.8	12.5	16.7
		FR					IT				
Gram.	Parsing	100.0	93.5	83.9	100.0	65.7	94.6	87.2	94.3	67.6	75.0
	Window	74.3	17.9	20.5	33.3	28.6	77.1	17.1	10.3	11.9	28.2
MWE	Parsing	67.6	45.2	38.7	25.9	5.7	78.4	38.5	37.1	29.7	13.9
	Window	54.3	10.3	10.3	11.9	2.9	51.4	4.9	2.6	2.4	15.4
Colloc.	Parsing	45.9	41.9	35.5	22.2	5.7	32.4	28.2	37.1	29.7	5.6
	Window	34.3	10.3	10.3	11.9	2.9	22.9	4.9	2.6	2.4	12.8

Table 4.13: Evaluation results for Experiment 2: grammatical precision (rows 1, 2 for each language); MWE precision (rows 3, 4); collocational precision (rows 5, 6).

As in Experiment 1, the precision is computed relative to the corresponding reference set, i.e., the pairs in the test set that were identically annotated by both annotators in a team. In addition to the grammatical precision defined as in Equation 4.1, we compute the MWE precision and the collocational precision on a test

set TS_i as follows (recall that the annotation categories are 0 - *ungrammatical*, 1 - *regular*, 2 - *named entity*, 3 - *collocation*, 4 - *compound*, and 5 - *idiom*):

$$MWE\ precision\ for\ TS_i = \frac{number\ of\ items\ in\ TS_i\ with\ mark \geq 2}{number\ of\ items\ in\ TS_i\ having\ a\ mark} \quad (4.6)$$

$$Collocational\ precision\ for\ TS_i = \frac{number\ of\ items\ in\ TS_i\ with\ mark = 3}{number\ of\ items\ in\ TS_i\ having\ a\ mark} \quad (4.7)$$

That is, we collapse the last 4 annotation categories (*named entity*, *collocation*, *compound* and *idiom*) into the same category in order to report the MWE precision, as we did in Experiment 1.

The results obtained are in line with those of Experiment 1 conducted on French data. The method based on parsing yields better results than the window method, for all the languages and all the parameters considered: grammatical, collocational and MWE precision. The benefit of using syntactic information for collocation extraction is therefore confirmed on a different, larger corpus[28] and for multiple languages (English, French, Spanish, and Italian).

The parse-based method was found more precise at all the levels in the output list that were investigated (0%, 1%, 3%, 5%, and 10%). In terms of grammaticality, the precision curve remains stable for the parse-based method on the first 4 levels (i.e., up to 5% of the significance list) at a value situated above 90% in most of the cases, and the average value on all languages for these levels is 92.9%. Then on the last level (10%), it shows a visible degradation around the value of 70% (the average on all languages being 73.9%).

As for the window method, the good performance for the top level (0%) observed in Experiment 1 is confirmed in this experiment as well. An average of 77.6% precision is obtained on the first test set for all languages (on which the parse-based method reaches an average of 98.1%).[29] But on the next 4 levels, the grammatical precision drops considerably, to only 22.7% on average. This means that 4 pairs from 5 are

[28]This corpus is on average 3.1 times bigger than the corpus used in Experiment 1.

[29]The numbers in Experiment 1 were quite similar, i.e., 76.4% vs. 99.0% for the top 500 pairs.

actually extraction noise in the output of the window method. In contrast, the average grammatical precision achieved with parsing on these levels is 86.9% (i.e., with 62.4% higher). Thus, we can report on average a difference of 20.5% on the first level investigated, and of 62.4% on the next four levels, in favour of the parse-based method.

A similar pattern is observed for the other two precision parameters, the MWE precision and the collocational precision. On the first level investigated, the performance of the window method approaches that of the parse-based method (the difference being, however, substantial).[30] Then on the next levels, the performance of the parse-based method decreases slightly, while the window method performs very poorly. The average precision values achieved by each method are presented in the first two rows of Table 4.14. The last rows show the difference and the ratio between the precision values for the two methods.

	Gram.			**MWE**			**Colloc.**		
	All	TS1	TS2-5	All	TS1	TS2-5	All	TS1	TS2-5
Parsing	88.8	98.1	86.9	43.2	71.2	35.8	32.9	39.9	31.4
Window	33.2	77.6	22.7	17.2	52.0	9.2	12.8	31.4	8.6
Parsing – Window	55.6	20.5	64.2	26.1	19.2	26.6	20.1	8.5	22.8
Parsing / Window	2.7	1.26	3.82	2.5	1.37	3.88	2.6	1.27	3.64

Table 4.14: Average precision values for the two methods (on all languages).

When the union of all test sets for all languages is considered, a total of 577 grammatical pairs are found among the 650 agreed upon for the parse-based method (that correspond to an overall precision of 88.8%), and only 261 among the 787 agreed upon for the window method (33.2%). Also, there are 281 pairs marked as MWE in the output of the first method (43.2% of the agreed pairs), and 135 in that of the second (17.2% of the agreed pairs). As for collocations, the parse-based methods discovers 214 of them (32.9% of the agreed pairs), and the window method only 101 (12.8%).[31] Figure I.4 in Appendix I provides a proportional view for these values.

[30] The difference in MWE precision is even statistically significant for English and Italian data in TS1.

[31] All the differences reported are statistically significant.

Finally, a look at the disagreements observed in the annotation data (see Table 4.15) revealed that collocations are most often mixed up with regular combinations, followed by compounds and, to a lesser extent, by idioms.[32] In fact, most of the total disagreement cases concern pairs that are labeled as collocation by one annotator and as regular combinations by the other. These make up 39.4% of the 563 total disagreements for both methods (222/563). The confusion between collocations and compounds is responsible for another 11.2% of the total disagreements (63/563).

	regular	**named entity**	**collocation**	**compound**	**idiom**
regular	422	–	–	–	–
named entity	6	26	–	–	–
collocation	222	2	315	–	–
compound	51	11	63	64	–
idiom	7	0	11	5	11

Table 4.15: Confusion matrix for the annotations in Experiment 2.

Several examples of such disagreements are presented below:

- disagreements collocation – regular combination: *achieve turnover, affected area, animal product, definitive solution, frightening statistic, have initiative, honourable member, main priority, member of family, partnership agreement, present system, religion conscience, scientific assessment, support process* (En), *beneficiar de amnistía, explotación agrario, iniciativa concreto, objetivo cuantitativo, mostrar voluntad, practicar tortura, puesto de trabajo, región periférico* (Es), *argent public, associer à remerciement, changement climatique, commercialiser produit, facteur essentiel, flagrant lacune, niveau dans hiérarchie, organiser séminaire* (Fr), *caso particolare, combattere contraffazione, rinnovare appello, sicurezza alimentare* (It);

- disagreements collocation – compound: *civil society, freight container, sea fleet* (En), *agente económico, entendimiento mutuo, estado miembro, interesado principal, potencia extranjero* (Es), *ampoule électrique, mode de vie, pays candidat,*

[32] Collocations are only rarely confused with named entities.

parlementaire européen, salaire minimum (Fr), *diritto di uomo, peggiore ipotesi, processo decisionale* (It)

- disagreements collocation – idiom: *cruel shortage, cut speaker, open door, remove breeding-ground* (En) *avoir lieu, page douloureux, revêtir importance* (Fr), *quadro completare, riprendere filo* (It)

- disagreements collocation – named entity: *parti conservateur* (Fr), *emisfero sud* (It).

4.4.5 Qualitative analysis of results

Error analysis

Experiment 1 Among the top 500 results returned by the parse-based method in Experiment 1 (Section 4.4.3), 5 pairs were annotated as ungrammatical: *entreprendre petite* (rank 51), *faible revenir* (rank 84), *entreprendre moyenne* (rank 200), *ministre de ancien* (rank 257), and *président de élection* (rank 462).

These false positives concern parsing errors caused exclusively by the syntactic ambiguity of source sentences. The analysis of their instances in the corpus showed that in 4 cases, the ambiguity is generated by the POS ambivalence, and more precisely, by the possibility to interpret a noun as the past participle of a verb (e.g., *entreprise*), and an adjective as a noun (e.g., *faible*). Thus, the A-N pairs *petite entreprise*, *moyenne entreprise*, *faible revenu* were analysed as V-O pairs in contexts like those shown in Example (5) below. The nouns (*entreprise, revenu*) were interpreted as past principles of the verbs (*entreprendre, revenir*), and the adjectives as their direct objects (*petite, moyenne*), or subjects (*faible*).

(5) a. *En tant que patron de petite entreprise...*
b. *afin de réduire le déficit et favoriser le développement des petites et moyennes entreprises*
c. *les programmes de logements sociaux permettent d'améliorer les conditions de vie de familles à faible revenu*

The sentence in Example (6) contains an instance of the incorrect pair *ministre de ancien*. Here, *anciens* was interpreted as the head of the noun phrase *Anciens combattants*, instead of adjectival modifier for the actual head, *combattants*.

(6) *ministre de la Défense nationale et ministre des Anciens combattants*

As for the last erroneous pair (*président de élection*), the source of ambiguity was the PP-attachment: in contexts like the one shown in Example (7), the phrase *de son élection* was interpreted by the parser as a right constituent for the noun *Président* instead of the verb (in this case, *féliciter*).

(7) *je voudrais donc saisir cette occasion pour féliciter le Président de son élection*

Both alternatives are possible and in the source contexts for this pair, the attachment to the noun was even more frequent that to the verb. 23 out of the 25 total instances of this pair are found in contexts like in Example (8), and only 2 instances are in fact attachments to the verb. Therefore, we may consider that the pair *président de élection* is actually correct, but it was wrongly annotated by the judges because the first context shown to them was the one in (7).

(8) *Le président d'élection (M. Hopkins)...*

In addition to the pairs that are ungrammatical, the pairs not marked as *interesting* are also considered as extraction errors. The number of pairs in this situation—i.e., pairs that were marked as *regular*—is 164. The most frequent of them are combinations of type N-A (48), N-P-N (48), and V-O (33). We noted that only 65 of these pairs were actually unanimously considered as regular by the three judges, mostly, the N-P-N pairs (21) and V-O pairs (15). The others were divided into regular pairs with MWE tendency and MWE pairs with tendency towards regularity, as shown in Section 4.4.3.

In the output of the window method, as many as 106 of the total 500 pairs are ungrammatical. The distribution among combination types is the following: N-N – 75 pairs, N-A – 14 pairs, A-N – 10 pairs, N-V – 6 pairs, and V-N – 1 pair. The error

rate for each combination type is as follows: N-N – 41.1%, N-V – 26.1%, A-N – 23%, N-A – 9%, V-N – 1%. The main cause of errors is the absence of a syntactic relation between the items of a pair. A detailed analysis of the ungrammatical pairs allowed us to classify them as follows:

1. ungrammatical subparts of frequent longer expressions (32 pairs), e.g.:

   - *arbitrage final*: *arbitrage* des offres *finales*
   - *susceptible mort*: force *susceptible* de causer la *mort* ou des lésions corporelles graves
   - *commune général*, *commune solliciteur*: leader du gouvernement à la Chambre des *communes* et *solliciteur général* du Canada
   - *défense combattant*: ministre de la *Défense* nationale et ministre des Anciens *combattants*
   - *président Canada*: *président* du Conseil privé de la Reine pour le *Canada*
   - *seconde mondial*: *Seconde* Guerre *mondiale*
   - *développement humain*: *Développement* des ressources *humaines*

2. ungrammatical associations with a very frequent domain-specific term, in our case: *Canada*, *gouvernement*, *ministre*, *pays* (46 pairs). For instance, associations with *Canada* include: *Canada Canada, Canada Canadien, Canada Chambre, Canada député, Canada gouvernement, Canada ministre, Canada Monsieur, Canada président, Canada programme, Canada question.*

3. ungrammatical associations, not necessarily with a domain-specific term (19 pairs), e.g., *dollar ministre, député programme, emploi Chambre, question premier.*

4. associations with wrong POS labels originating in tagging errors (6 pairs):

   - *Monsieur présider* (N-V), correct: *Monsieur le Président* (N-N)
   - *petite entreprendre* (N-V), correct: *petite entreprise* (A-N)

- *monsieur présider*(N-V), correct: *monsieur le Président* (N-N)[33]
- *faible revenir* (N-V), correct: *faible revenu* (A-N)
- *nouveau députer* (N-V), correct: *nouveau député* (A-N)
- *moyenne entreprendre* (N-V), correct: *moyenne entreprise* (A-N)

A second class of errors in the output of the window method are the false positives represented by the *regular* pairs: their number is 104 (i.e., 0.63 times less than that of the parse-based method). Among the most frequent combination types, we found N-N combinations (40), N-A combinations (34), and V-N combinations (15). Only 48 regular pairs are common to both methods.

Experiment 2 The analysis of the ungrammatical pairs obtained with the parse-based method in Experiment 2 (Section 4.4.4) revealed various causes of errors, among which:

1. wrong assignment of POS labels, caused by the ambiguity of the text: for example, the pair *sous de forme* identified in the context *en tirent un avantage financier et sous d'autres formes* has been considered as a N-P-N pair, the word *sous* being wrongly tagged as a noun (the plural form of *sou*, 'coin') instead of preposition. The two readings are equally possible for this sentence.

2. wrong assignment of syntactic type:[34] for instance, the pair *succéder catastrophe* has been incorrectly interpreted by the parser as a V-O pair instead of S-V in the context *Mais peut-on voir se succéder les catastrophes qui ont frappé mon pays.*

3. absence of a syntactic link between the two pairs, when the parser fails, for instance, to recognize a complex lexical unit (these errors can be get rid of by adding such items in the parser's lexicon): e.g., the pair *fait inapproprié*

[33]The parser distinguishes between *Monsieur* (title) and *monsieur* (common noun) and therefore considers them as two different lexemes.

[34]Note that the window method cannot be subject to such errors as long as no syntactic type is associated with the output pairs, but only POS labels.

extracted from *J'estime qu'il est tout à fait inapproprié que...* is ungrammatical, since *fait* is part of the adverb *tout à fait*.

Among the sources of parsing errors for English data, we identified the nominal compounds (for instance, in *gross domestic product*, the parser analyses *gross domestic* as an A-N pair), and the confusion between gerunds and nouns (e.g., in *a barrage of fire targeting Arafat*, the pair *barrage of targeting* is identified as N-P-N). As far as the Spanish and Italian outputs are concerned, the parsing errors are more numerous because of the more limited lexical and grammatical coverage of the Fips parser for these languages.

As for the window method, ungrammatical pairs are very numerous in the extraction output obtained in Experiment 2. A preliminary analysis of these pairs suggested that the same conclusions drawn in Experiment 1 hold here as well.

Intersection and rank correlation

Experiment 1 A comparison of the output obtained by the two methods has been performed at pair-type level. In Table 4.16, the column *common* reports the number of pair types that were extracted both by the parse-based method and the window method in Experiment 1 (Section 4.4.3).

Top	**All Pairs**			**MWE Pairs**		
	common	perc.	Spearman's ρ	common	perc.	Spearman's ρ
50	30	60.0%	0.775	22	44.0%	0.782
100	56	56.0%	0.832	39	39.0%	0.848
150	84	56.0%	0.861	57	38.0%	0.889
200	103	51.5%	0.871	70	35.0%	0.896
250	120	48.0%	0.866	81	32.4%	0.887
300	145	48.3%	0.871	92	30.7%	0.882
350	168	48.0%	0.882	106	30.3%	0.897
400	192	48.0%	0.873	117	29.3%	0.896
450	210	46.7%	0.863	129	28.7%	0.897
500	231	46.2%	0.854	141	28.2%	0.887

Table 4.16: Output comparison for the two methods in Experiment 1 (Section 4.4.3): intersection and Spearman's rank correlation coefficient for the top 500 results.

The intersection is computed for the top n pairs, with n ranging from 50 to 500 at intervals of 50. Thus, 30 pairs are common to the top 50 results of each method etc., 231 pairs being common overall, in the top 500 results. The percentage of common pairs decreases as we look further in the significance list: from 60% for top 50, it goes down to 56% for top 100 and 150, and reaches only 46.2% overall (third column, *perc*). These results refer to all of the pairs in the test sets, regardless of their annotation mark (*ungrammatical*, *regular*, or *interesting*).

When we consider only the pairs marked as *interesting*, i.e., the pairs that are likely to constitute MWEs, we obtain the figures in the column *MWE pairs*: the number of common MWE pairs identified in the top 50 results is 22 (44%), and so on. Overall, 141 MWE pairs are common to the two methods in the top 500 results, and they represent 28.2% of the whole annotated data.

Table 4.16 also reports the Spearman's ρ rank correlation coefficient on the common pairs returned by the two methods. This value, ranging from -1 to 1, indicates the degree of correlation between the rankings proposed. Given the set of common pairs ordered by each method according to the collocational strength (so that higher-scored pairs are ranked first), the Spearman's ρ is computed as in Equation 4.8:

$$\rho = 1 - \frac{6 \sum_i (RW_i - RP_i)^2}{n(n^2 - 1)} \tag{4.8}$$

where RW_i and RP_i denote the ranks for the pair i proposed by the window method and the parse-based method, respectively, and n is the total number of common pairs. Values of ρ close to 1 indicate a high correlation of rankings, and values below 0 indicate a negative correlation. The rankings correlate significantly if the ρ value is higher than a critical value—in this case, 0.165 (Villada Moirón, 2005, 85).

The ρ values in Table 4.16 indicate a high correlation of the ranks proposed by the two methods for the top 500 pairs returned in Experiment 1. This suggests that, as far as the top results are concerned, the difference between the two methods stands more in the content of the output lists, than in the relative order of the pairs identified.

Experiment 2 In Experiment 2 (Section 4.4.4), the number of pairs in the intersection of the two output lists—one of the window method, the other of the parse-based

method—is very small (see Table 4.17).[35] The common pairs belong almost exclusively to the first test set, TS1 (i.e., to the top 50 results). As in the case of Experiment 1, the Spearman's coefficient indicates a high correlation of the ranks assigned to the common pairs by the two methods.

	English	**French**	**Italian**	**Spanish**
common (TS1–TS5)	24	20	22	31
common (TS1)	23	20	22	31
Spearman's ρ	0.640	0.833	0.586	0.846

Table 4.17: Output comparison for the two methods in Experiment 2 (Section 4.4.4): intersection and Spearman's rank correlation coefficient for the pairs annotated for each language.

In a larger-scale comparison, we considered the complete output lists of the two methods and looked at the overall intersection and rank correlation. These parameters have also been computed on different sets of pairs satisfying the following criteria pertaining to the rank, the log-likelihood score, and the corpus frequency of pairs:

- **rank**: lower than 1000, 2000, and 10000;
- **LLR score**: higher than 1000, 100, and 10;
- **frequency**: higher than 100, 10, and 3.

The results obtained are reported in Appendix J. We found that overall there is a significant correlation between the rankings of the pairs in the intersection (the Sperman's ρ is between 0.477 and 0.521 depending on the language, and 0.487 on average). The correlation is higher for the upper layers of the output lists, as given by the higher frequency or LLR score cutoffs. A relationship between ranks and correlation could not be established (i.e., the correlation function is non-monotone on ranks), a possible explanation being that the layers compared make up very different proportions in the whole output lists (compare column *Perc. P* with *Perc. W*).

[35] This is mainly the consequence of the manner in which the test sets have been constructed, by considering non-adjacent sets at various levels in the output list.

As for the common pairs, they make up a high percentage of the output returned by the parse-based method (75.59% for the English data, 73.48% for French, 79.58% for Italian, 80.81% for Spanish, and 77.36% on average), but only a small fraction of the much larger data retrieved by the window method (17.43% on average).

For all of the languages considered, we can conclude—as in Experiment 1—that the relative order of output pairs is maintained from one method to another as far as the common pairs are concerned, and that the correlation is higher for the upper layers of the output lists as defined by higher frequency and LLR score thresholds.

Instance-level analysis

A more insightful comparison between the two extraction methods can be made by examining the instances of the output pairs in the source corpus. We performed several case studies aimed at answering the following questions: How many instances of a pair type are retrieved from the source corpus by one method with respect to another? How many instances are identical, and, for those that are missed by one method, is there any peculiarity of the source text that prevents their retrieval? How many instances are "true" and how many are "false" (in the sense that the source sentence is incorrectly associated with the pair in question)?[36] And, ultimately, is one method better than another in terms of coverage, in a given setting?

In the first case study, we considered the V-O pair *jouer rôle* from Experiment 1 (Section 4.4.3) and compared the instances identified by the parse-based method, on the one hand, with those identified by the window method, on the other hand. In the output of the window method, this pairs is ranked 39th, with 135 instances; in that of the method based on parsing, it is ranked 14th and has 190 instances.

A number of 131 instances are identical, meaning that almost all the corpus occurrences of this pair that have been detected by the window method have also been detected by the parse-based method; the 4 instances discovered exclusively by the window method are "true" instances that have been missed by the parser. The corpus-based analysis of these 135 instances showed that they are all "true" instances

[36]For instance, Example (7) contains a false instance for the pair *président de élection*, while Example (8) contains a true instance.

for which the average distance between the base *rôle* and the collocate *jouer* is very small (around one word).

The parse-based method discovered 59 more instances. Among these, there are 43 cases of inversion that could not be detected by the window method, since this was designed for extracting oriented pairs. The inversion was caused in the majority of cases by grammatical operations like relativization (37 instances), interrogation (4 instances) and passivization (2 instances). An example of each type is provided below.

- **relativization** En outre, pourrait-il en profiter pour nous préciser le *rôle* que l'État, le gouvernement fédéral plus particulièrement, devrait *jouer* pour aider les familles?
- **interrogation** quel *rôle* de premier plan pouvons-nous *jouer* pour voir à ce que ces résolutions soient respectées?
- **passivization** Le *rôle joué* par le Canada et par M. Pearson dans cette crise

As in the case of the window method, all the instances that were exclusively retrieved with the parse-based method are "true" instances. The conclusion we can draw from this case-study is that both extraction methods are precise (i.e., they both identify "true" instances), but the window method has a lower coverage since it retrieves a lower number of instances. This happens mainly because of a design choice not allowing it the mix N-V and V-N surface pairs, since these POS combinations suggest different syntactic relations: S-V, in the first case, and V-O, in the second. As a consequence, the rank obtained with the window method is lower than that assigned by the parse-based method for some pairs. At a larger scale, this study suggests that the failure of the window method to detect all the instances of some pairs could lead it to artificially demote these pairs to lower positions in the significance list.

Another pair whose instances were studied is the English S-V pair *vote take*, extracted from Experiment 2 (Section 4.4.4).[37] The window method finds 408 instances

[37]This pair is actually part of the longer collocation *vote – take place* that should have been extracted if *take place* was included in the parser's lexicon (Chaper 5 presents a method for obtaining longer collocations from previously extracted pairs).

of it in the source corpus, while with the parse-based method only 325 instances are identified. A number of 325 instances are common to both methods, i.e., all the instances found by the parse-based method are also found by the window method. This suggests that the coverage of the window method is, for this pair, at least as good as that of the parse-based method.

While the precision of the parse-based method is perfect (since all the instances identified are correct), as many as 81 out of the 83 occurrences that were found exclusively by the window method count as "false" instances:

- 10 instances are ungrammatical, for example:

  > I merely wish to ask whether the written explanations of *vote* should not be *taken* first.
  >
  > That means that, as often happens in these cases, the decision we take with our *vote* will not be *taken* into any account by the Council.

- 71 instances belong in reality to the V-O pair *take – vote*, for example:

  > I propose that the sitting be suspended until the *votes* are *taken*.

The remaining 2 instances are "true" S-V instances that have not been identified by the parse-based method.

This second case study suggests that the "false" instances retrieved with the window method for some pairs are very numerous and they are difficult to distinguish from "true" instances as long as the syntactic context is ignored.

In an additional corpus-based investigation of instances of other output pairs, we discovered that there are several situations in which the window method behaves more robustly than our method committed to the syntactic parsing. It is, for instance, the case of pronominalization. Our method is not expected to recover a collocation with a noun when the latter is realized as an anaphoric pronoun (although this would be possible if an anaphora resolution module were coupled to our system). With the window method instead, the noun can be detected as long as it is found in the collocational span considered. For instance, in Example (9) below, the object of the

verb *take* is the anaphoric pronoun *it* whose antecedent is the noun *vote*; therefore, the pair *take – vote* could be inferred as a V-O pair instance. The window method discovers the antecedent accidentally in the proximity of the verb *take* and succeeds in retrieving the pair instance from this sentence.

(9) No, we must take that 'no' *vote* seriously and *take* it on board in our reflections.

Another situation in which the absence of syntactic constraints actually helps extraction is when the text contains a noun phrase with a semantically "transparent" head, e.g., *un sou d'impôt, une partie d'impôt, un minimum d'impôt.* In such a phrase, the semantic head (i.e., the noun bearing the semantic content) does not coincide with the syntactic head (*sou, partie, minimum*), but with its right subconstituent (*impôt*). Collocational links establish between this subconstituent and external items, bypassing the syntactic head. This is why such a link cannot be retrieved in a straightforward manner and with strictly syntactic means. Additional computation and semantic information on nouns is required in order for the extraction procedure to work.[38] In contrast, these collocational links are easily identifiable with the window method, precisely because no syntactic constraint applies. Example (10) below shows several instances of the V-O pair *payer impôt*, in which the object of the verb is an NP with a semantically-transparent head (underlined in text):

(10)
a. sans qu'elles *paient* un cent d'*impôt*
b. ne *payent* toujours pas un sou d'*impôt*
c. les sociétés *paient* leur juste part d'*impôts*
d. qui *paient* déjà la majeure partie des *impôts*
e. *paient* un minimum d'*impôt*
f. devra *payer* 200 $ d'*impôt* de plus

[38] Fontenelle (1999) discusses the problem of transparent nouns, by showing that they may involve a wide range of partitives and quantifiers, as in *shot clouds of arrows, melt a bar of chocolate, suffer from an outbreak of fever, a warm round of applause.* He proposes a lexical-function account for these nouns, in which the transparent nouns are considered as the value of the lexical function *Mult* (e.g., *Mult(arrow)=clowd*).

Although such instances are, strictly speaking, ungrammatical, they could be considered as true instances, since they are semantically correct. Other instances retrieved with the window method are even more difficult to infer syntactically, like the instance in Example (11), but these are at the limit of what can be considered as a true instance or not.

(11) *paye* exactement le même montant en *impôt*

4.5 Discussion

The evaluation experiments carried out showed that by integrating a syntactic filter into a collocation extraction procedure, a considerable improvement is obtained over the standard mobile-window method. This result is far from being obvious, since, first of all, the syntactic analysis of whole source corpora is a difficult endeavor that requires sufficiently fast and robust tools; and, secondly, because this particular implementation of the window method, which favours precision, represents a high baseline for comparison.

The analysis of results provided in Section 4.4.5 pointed out several advantages and disadvantages of one method versus another, related in the first place to the pair instances detected in the corpus. Our case studies revealed that many of the instances identified with the window method are "false" instances; moreover, many instances that are found by the syntax-based method are instead missed by the window method. These (false or missed) instances falsify the frequency signature[39] of the pairs concerned; as a consequence, these pairs are artificially promoted or demoted in the significance list, the quality of results being ultimately affected.

A more directly perceived inconvenience for associating false instances with a pair is that the corresponding source contexts are useless for lexicographic purposes, as well as for the interested NLP applications (i.e., whenever one tries to derive contextual information for an extracted collocation). Parsing helps to overcome this problem, as the high majority of instances identified for a pair in text are correct instances.

[39]This term was introduced in Section 3.2.3.

Another shortcoming of the window method is that the results pairs are highly unreliable. Even if a POS combination suggests a syntactic link (e.g., N-V suggests a S-V relation), the pair may actually be ungrammatical, or may involve a syntactic type that is different from the expected one; in contrast, the pairs returned by the parse-based method contain syntactic information assigned by the parser, which facilitates their interpretation and subsequent processing. The following example shows several pairs extracted by the window method in Experiment 1 (Section 4.4.3). Although they look correct in isolation, these are in reality wrong, as can be seen from the instances provided next to them.

(12) a. *développement humain* - *Développement* des ressources *humaines*
b. *président Canada* - *président* du Conseil privé de la Reine pour le *Canada*
c. *gouvernement Canadien* - le *gouvernement* du Canada invite tous les *Canadiens* à participer...

On the other hand, we found that one advantage of the window method over methods relying on syntax is that it can detect the collocate of a base word in the collocational span, even if this is not directly syntactically related to that word. Such situations include the pronominalization of the base word, as in Example (9), or its embedding in a larger NP whose syntactic head is a semantically-transparent noun, as in Example (10).

The main advantage attributed, in principle, to the window method is that it is fast, robust, and readily available for a new language, whereas syntax-based methods require high-performance, language-specific tools that are difficult to implement. The extraction experiments we performed in the framework of the comparative evaluation studies showed that, actually, the multilingual parser Fips is also fast and robust enough to process source corpora of several million words; that the syntactic filter applied to the candidate pairs leads to a drastic reduction of the candidate data that is later fed into the stochastic procedure of collocation strength computation, and, on the contrary, the much larger data produced by the window method slows down this procedure considerably. Moreover, the window method is not fully language-independent, because its performance relies on the accuracy of POS information that

is in turn dependent on the word disambiguation depending on the syntactical context.

Finally, it is also interesting to note that although Experiment 2 deals with more data than Experiment 1 (i.e., 3 to 4 times as much data), no improvement was obtained in the results of the window method.

As an alternative to the window method, we could have considered comparing our method against methods based on shallow parsing, since these are expected to produce more accurate results than the syntax-free approach.[40] While the precision of chunk-based methods could certainly compete with that of parse-based approaches, we expect, however, a lower recall because of the lower grammatical coverage of shallow parsers, in particular with regard to those constructions responsible for the long distance between the items of candidate pairs.

For the time being, we have not performed a thorough evaluation against methods based on shallow parsing, yet we made a preliminary analysis of the collocations identified with a shallow parser for English. We carried out a case study on the V-O and S-V collocations with the noun *preference* that were extracted from the BNC corpus with the Sketch Engine. The Sketch Engine (Kilgarriff et al., 2004) is a state-of-the-art extractor that identifies collocation candidates of several syntactic types by relying on shallow parsing implemented as regular expression pattern-matching over POS tags. The association measure used for ranking candidates is an adaptation of MI that gives more weight to the co-occurrence frequency.

Among the 201 V-O pair types with *preference* obtained without applying a frequency cutoff, a number of 15 were grammatically incorrect (i.e., 7.5%). Several such pairs are shown in Example (13) below, together with a corpus instance for each. Another 4 pairs used a majority of incorrect instances, and thus got artificially promoted to higher ranks (e.g., the pairs in Example (14)).

(13) a. *vote preference* (rank 20) - we asked about *voting preferences*
b. *focus preference* (rank 24) - *focus preferences* were taken to be strong
c. *elect preference* (rank 39) - most voters will find that a candidate for whom they have expressed some *preference* will have been *elected*

[40] In this case, however, the argument of language independence and ease of implementation does not hold anymore, as chunkers are relatively difficult to develop.

d. *peak preference* (rank 55) - could produce a sharply *peaked preference*

(14) a. *decide preference* (rank 47) - As they walked towards the hotel, she gave him the rates on an ascending scale and spat her gum into the kerbside as if she had already *decided* what his *preference* would be.

b. *count preference* (rank 51) - votes of lower *preference* are *counted* only when transfers have to be made

Similarly, from the 63 S-V pair types returned by the Sketch Engine, a number of 10 (15.9%) were incorrect, among which the pairs in Example (15) below. Another 5 S-V pairs used incorrect instances, like the pair in Example (16).

(15) a. *preference prejudice* (rank 1) - if compliance with the *preference* would *prejudice*

b. *preference support* (rank 6) - The proposal is a clear expression of local *preference supported* by local planning authorities.

(16) *preference lead* - the existence of these *preferences* would clearly *lead* ultimately to a situation

Although the results obtained in this case study are not entirely conclusive due to the small size of data evaluated, they suggest that the extraction based on shallow parsing leaves considerable room for improvement, and we consider that this improvement can be achieved with full parsing.

4.6 Summary

This chapter presented the core of our collocation extraction methodology that relies on the full syntactic parsing of the source corpora in order to optimize the first extraction step, the selection of candidate pairs. With respect to similar extractors based on parsing (described in more detail in Section 4.1), our extractor tries to fill a number of gaps pertaining, among others, to robustness, precision, the syntactic configurations extracted, and, more importantly, to the syntactic constructions covered.

These desiderata (stated at the end of Section 4.1) could successfully be put into practice thanks to the multilingual parser Fips (introduced in Section 4.2), whose robustness and speed permits the syntactic analysis of large text corpora in several languages, and whose accuracy ensures the high grammatical precision of the extraction results, both with respect to the pair types returned and to the instances identified in text.

After describing the extraction procedure proper, we focused on the contrastive evaluation of our method against the mobile-window method, a standard syntax-free extraction method based on the linear proximity of words. This evaluation study was motivated by several issues (discussed at the beginning of Section 4.4) that questioned the efficiency of preprocessing based on full parsing. The latter is often criticised for being error-prone and not really worth applying, while the classical method, which is (at least in principle) easy to implement and robust, is believed to still produce acceptable results when applied to large corpora.

Thus, two evaluation experiments were performed with different settings (i.e., corpus domain, corpus size, source languages, test sets, annotation categories, etc). The results obtained were consistent: in both experiments, for all the languages supported by our extractor, the window method was outperformed by a large extent by the parse-based method. Notably, we found that the major part of window results, with the exception of the pairs situated in the very top of the significance list, are ungrammatical. On the contrary, the parse-based method maintains a high grammatical precision on the lower levels in the significance list as well (Table 4.14). Moreover, the quality and the interpretability of window results suffer from the inescapable mix of "true" and "false" instances identified in the source text for the output pairs. In contrast, the pairs extracted with parsing contain syntactic information and most of their instances are correct, this facilitating their subsequent use for lexicographic or NLP purposes. Finally, the strong syntactic filter applied to candidate pairs alleviates the statistical computation of score in the second extraction stage, whereas the window method is much less tractable because of the huge candidate data it generates.

In line with the grammatical precision, the MWE and collocational precision were also found higher for the parse-based method. In Experiment 1 that evaluated the

MWE precision on the top 500 results, the increase obtained with parsing is of 9% (from 56.9% to 65.9%). Experiment 2 investigated several levels of the output list and reported an increase of 26.1% in MWE precision (from 17.2% to 43.2%), and of 20.1% in collocational precision (from 12.8% to 32.9%), on average for all the levels and languages considered.

Related work that quantified the impact of using parsing for collocation extraction also reported that the syntactic information makes a substantial contribution to the precision of results, e.g., (Smadja, 1993; Lin, 1999; Zajac et al., 2003) (see also Section 3.4). A smaller improvement was instead observed for A-N collocations in German when extracted with a chunker in comparison with a 10-word window method (Evert and Kermes, 2003). This result is unsurprising, given the particularly rigid pattern studied. A different study carried out in German for N-V pairs (Breidt, 1993) came, on the contrary, to the conclusion that good precision can only be achieved for German with parsing.

Differently from the preceding evaluation reports, ours is a more systematic evaluation study, which has been conducted on data in 4 languages totalling a large number of pairs (3000) that, in turn, were taken both among the best-scored and the lower-scored results and were annotated by at least two human judges. Our study also provided a detailed qualitative analysis of results of the methods compared, that revealed possible causes for the difference observed in the pair rankings and in the number of instances retrieved, and highlighted the relative strengths of one method versus another depending on the particularities of the text.

Chapter 5

Advanced Extraction

This chapter presents a series of extensions for the collocation extraction method described in the preceding chapter, that are aimed, firstly, at enlarging the scope of this method so that it covers a broader spectrum of collocational phenomena in text, and, secondly, at providing an alternative solution to the data sparseness problem that is characteristic of text corpora in general. Thus, Section 5.1 presents a method for the extraction of collocations of higher arity, the procedure described so far being exclusively concerned with binary collocations. Section 5.2 deals with the issue of finding an exhaustive set of syntactic configurations that define collocation candidates, and proposes a data-driven solution. The last section describes an approach in which the World Wide Web is used for the discovery of collocations with a given word, as an alternative to static corpora.

5.1 Identification of multi-word collocations

5.1.1 Introduction

The first extension considered for our method was aimed at the discovery of collocations made up of more than two items. Despite the fact that theoretical studies (such as those reviewed in Chapter 2 and, in particular, in Section 2.3) stipulate that collocations may consist of two *or more* words, in practice the collocation extraction

work focuses almost exclusively on binary collocations, mainly as a consequence of the manner in which association measures are designed (Section 3.2.1).

To overcome this limitation pertaining to the length of the collocations retrieved, several solutions could be considered:

1. Extending the existing association measures to more than two items, as in Villada Moirón (2005),[1] or proposing new measures that cope with unrestricted length (Blaheta and Johnson, 2001; Dias, 2003).

2. Using a cyclic extraction strategy, in which the previous results are incorporated into future extraction, so that an already-known collocation is treated as a single item. This solution—which is already implemented in our extraction system, as discussed in Section 4.3.1—best accounts for the recursive nature of collocations noted by researchers like Heid (1994, 232), but is more difficult to put into practice because it requires human intervention in order to validate the output and to update the lexicon of the parser.

3. Post-processing the extraction output in order to automatically infer longer collocations from the binary combinations identified (Smadja, 1993; Kim et al., 1999; Kim et al., 2001).

This section describes a method for extracting collocations of length higher than two (hereafter, *n-grams*)[2] that adopts the third solution. Given the binary collocations (*bigrams*) extracted as shown in Section 4.3, it creates chains of unrestricted length by joining these bigrams whenever they are found together in the same corpus sentence so that they share common items.

For instance, once the bigrams *play role* and *important role* are extracted from a sentence so that *role* is the very same token in that sentence, we can infer the trigram *play important role*. The process can be continued by adding a new item to the current chain, so that we will obtain, for instance, the 4-gram *play important role*

[1] Villada Moirón (2005) extends MI and chi-square in order to cope with candidates of length 3.

[2] For the sake of simplicity, we will use the terms of *bigram*, *trigram*, and in general that of *n-gram* in order to indicate the arity of collocations (be they composed of adjacent words or not).

in, given that the bigram *play in* was extracted from the same sentence and the token *play* is the very same as before.

The method of chain building using bigram composition is introduced in (Seretan et al., 2003; Seretan et al., 2004a). In what follows, we will provide an intuitive description of this method, while focusing more on additional details not discussed in these papers. We will also present new extraction results for French data and an updated review of the related work.

5.1.2 Building collocation chains

The algorithm starts with the set of bigrams extracted by our system, where each bigram contains, inter alia, information regarding the lexemes involved, the syntactic type, and the source context (as described in Section 4.3.2).

In each iteration, n-grams are combined into $(n+1)$-grams; that is, longer chains are obtained from the chains of maximal length built up to that point. The newly-constructed chains are stored, and the procedure is repeated until no new combinations are generated for a given n, the maximum length reached being determined by the length of source sentences (the termination of the procedure is guaranteed, since the length of sentences is limited).

There are multiple ways of combining shorter chains into longer ones, that may lead to the same results. For instance, one could obtain a 4-gram by combining a trigram with a bigram (*play important role*, *play in*), three bigrams (*play role*, *important role*, *play in*), or two trigrams (*play important role*, *play role in*). We opted for the combination of chains of the equal, maximal length—i.e., as in the last option above—, this is why the procedure exclusively uses n-grams for building $(n+1)$-grams. The choice is motivated by the simplicity and the uniformity in the description of the algorithm.

In order to yield an $(n+1)$-gram, two n-grams are combined *iff* they share exactly $n-1$ items; for these items, the source file and the file positions, which are stored in the fields `source` and `charlex1` or `charlex2`, respectively, must be identical.

The order of items in the resulting chain is most often determined by the canonical

order of these items in the participating bigrams, as obtained from the sentence normalization provided by the parser. This allows the convenient grouping of all the instances of the versatile, predicative combinations under the same type. For instance, the following pairs:

(1) *prendre décision*
décision difficile

share the middle lexeme in the resulting trigram, *prendre décision difficile*, and the order of the participating items is unambiguous. This (canonical) order is unique for all the corresponding instances. Thus, both instances shown in Example (2) belong to the trigram type *prendre décision difficile*, even if the textual order for the second is different (*décision difficile prendre*).

(2) a. Le fait est que nous devons *prendre* des *décisions difficiles*.
b. Nous ne voulons pas que les fabriquants de produits du tabac profitent de la *décision difficile* que nous avons *prise* aujourd'hui.

In case of ambiguity, when the order cannot be inferred from the cannonical order of items in bigrams, it is decided by the source contexts (i.e., by the file position of items). For instance, for the bigrams in Example (3), it is not clear what the resulting trigram should be, *tenir référendum sur* or *tenir sur référendum*, and it is the relative position of *référendum* and *sur* in the text that is taken into account.[3]

(3) *tenir référendum*
tenir sur

A drawback of this combinatorial procedure is that for chains of length 4 and more it can produce redundant results, because in certain cases the same chain can be built from distinct pairs of subchains. For instance, the 4-gram *démontrer grand ouverture de esprit* can be built from the following pairs of 3-grams:

[3]If more alternatives are possible, multiple types are generated accordingly.

(4) a. *grand ouverture esprit - démontrer grand ouverture*
b. *démontrer ouverture esprit - grand ouverture esprit*
c. *démontrer ouverture esprit - démontrer grand ouverture*

However, the procedure keeps track of the way in which an n-gram has been built and distinguishes between such redundant cases, in order to avoid reporting a false frequency for that n-gram. The frequency information is used in the subsequent extraction step, which ranks the generated candidates (Section 5.1.3). On the basis of the obtained score, the different variants of a redundant n-gram could be decided between in order to retain only the most successful variant for the next iterations of the algorithm (this filtering is not applied in the current implementation of the method).

5.1.3 Measuring association strength

Longer collocations can be regarded as recursively-embedded binary collocations sharing the base word, as illustrated by the trigram *allgemeine Gültigkeit haben*, lit. *general validity have* (Heid, 1994, 232):

(5) (*allgemeine* <u>*Gültigkeit*</u>) + (<u>*Gültigkeit*</u> *haben*)
((*allgemeine* <u>*Gültigkeit*</u>) *haben*)

It follows then that the association measures that exist for binary collocations can be straightforwardly applied to collocations of arbitrary length in a recursive fashion, by treating complex lexical items like units. In the corresponding contingency table, the frequency values that will be listed will be those of these units, computed from the corpus in the same manner as the frequencies of words (see Section 3.2.3).

	Gültigkeit haben	¬ *Gültigkeit haben*
allgemeine Gültigkeit	a	b
¬ *allgemeine Gültigkeit*	c	d

Table 5.1: Example of contingency table for the trigram *allgemeine Gültigkeit haben*.

Thus, after the creation of n-gram candidates with the procedure described above, contingency values are computed for each n-gram type by taking into account the co-occurrence frequencies of the two component items. The log-likelihood ratios association measure (LLR) is then applied to the output n-grams, for each n, according to the formula shown in Section 3.2.4.

5.1.4 Results

This section reports on the multi-word collocations obtained by applying the extraction method described in the preceding sections on the bigrams previously extracted in Experiment 1 (Section 4.4.3).

From the 370932 total bigram extracted, a number of 173037 trigrams have been generated, that correspond to 143229 trigram types. The combination of trigrams yielded, in turn, 146032 4-grams instances and 140138 types. The distribution of 3-gram and 4-gram types into frequency strata is shown in Table 5.2.

Freq	**3-gram types**	**4-gram types**
$f > 10$	404	63
$6 \leq f \leq 10$	694	96
$3 \leq f \leq 5$	3344	691
$f = 2$	7337	2446
$f = 1$	131450	136842

Table 5.2: Frequency distribution for the 3-grams and 4-grams extracted.

Table 5.3 presents several examples of trigrams and 4-grams selected among the top results according to the LLR score; more (randomly selected) extraction results are presented in Appendix K. The results list the lemmas for the participating items, to which prepositions are added, depending on the syntactic configuration of the initial bigrams (refer to the list of configurations provided in Section 4.3.1).[4]

[4]Note that some items constitute complex units in turn, e.g., *premier plan* in the fifth 4-gram shown. Also, as suggested by the last 4-grams in the list, our strategy that consists of systematically displaying lemmas rather than word forms led to an unusual presentation of some expressions (such as *trouver bon solution possible* instead of *trouver meilleur solution possible*).

Table 5.4 and Table 5.5 display some of the most frequent syntactic configurations for the obtained trigrams and 4-grams, as given by the configuration of the bigrams involved.[5]

3-gram	**4-gram**
accorder attention particulier	avoir effet négatif sur
attirer attention sur	créer grand nombre de emploi
avoir effet néfaste	être à prise avec taux de chômage
compter sur entier collaboration	franchir cap de milliard de dollar
donner bon résultat	jouer rôle de premier plan dans
être sur bon voie	jouer rôle important dans
jouer rôle important	jouer grand rôle dans
prendre décision difficile	poursuivre effort négociation avec
prendre engagement envers	présenter sincère félicitation pour
prendre mesure concret	question faire objet de examen
prêter oreille attentif	ramener partie à table de négociation
revêtir importance particulier	tirer bon profit possible
tirer coup de feu	trouver bon solution possible

Table 5.3: Sample 3-grams and 4-grams among the top-scored results.

5.1.5 Discussion

As can be checked from the random samples of results shown in Appendix K, the bigram composition method produces high-quality results in terms of grammaticality. An analysis of results in terms of collocability (or, in general, of lexicographic interest) has not yet been carried out, but it is apparent that the method is able to retrieve good collocation candidates.

Overall, it can be noted that there are more numerous interesting candidates among the trigrams than among the 4-grams extracted. Indeed, as an effect of the higher fragmentation of data, the collocability strength is expected to decrease as the length of candidates increases. Also, many of the salient long associations discovered in this experiment are composed of collocational subparts, but to which a

[5]Note that, due to the way the extraction procedure works, the words are currently shown in their base form, but in the future the inflected form could be considered for enhanced readability.

Types in 3-grams	Freq (tokens)	Example
S-V, V-O	12703	question revêtir importance
V-O, N-P-N	10926	améliorer qualité de vie
V-O, N-A	8483	avoir effet néfaste
N-P-N, N-P-N	7315	violation de droit de personne
N-P-N, N-A	7184	régime de assistance publique
V-O, V-P	6881	prendre engagement envers
V-O, V-P-N	6486	mettre terme à conflit
S-V, S-V	6433	argent servir financer
S-V, V-P-N	5752	frais élever à milliard
N-P-N, S-V	5533	taux de chômage situer

Table 5.4: The most frequent combinations of syntactic types for 3-grams.

non-collocational item is added because it is characteristic of the domain (e.g., nouns like *ministre*, *député*, *gouvernement*, and so on), as in the trigram *ministre régler problème*. This inconvenience is caused by the choice of this particular source corpus, namely the Hansard corpus of parliamentary debates, but we expect the results on domain-independent corpora to be, for the main part, more interesting from a lexicographic point of view.

Detecting longer collocation chains is particularly important for a special category of binary collocations, i.e., those that are incomplete (e.g., *jouer rôle* is almost never used in isolation, but in combination with modifiers like *essentiel, grand, important, clé, déterminant, de premier plan, de chef de file, actif* etc.), or do not constitute collocations by themselves, because the collocational link involves a longer lexical unit, rather than its shorter subparts. For instance, in the collocation *vote – take place*, the collocate of *vote* is neither *take* nor *place*, but the whole phrase *take place*.[6] The combination *vote take* therefore has the status of a fragment that must be discarded from the list of bigrams extracted as soon as it is used in the construction of trigrams (e.g., *vote take place*).

Eliminating fragments of collocations is a real concern in extraction work dealing with multi-word collocations or terms, the common approach being that shorter chains

[6]Note that this kind of combination cannot be accommodated in the more compositional account of multi-word collocation proposed by Heid (1994).

Types in 4-grams	Freq (tokens)	Example
S-V, V-O, V-P	3290	mesure stimuler entreprise dans
S-V, S-V, V-O	2732	gouvernement tenter régler problème
N-A, S-V, S-V	2704	argument principal être dire
N-P-N, S-V, S-V	2540	taux de intérêt continuer baisser
S-V, V-O, N-P-N	2519	projet faire objet de étude
S-V, V-O, V-P-N	2465	comité axer travail sur question
V-O, A-N, N-P-N	2428	coûter grand nombre de vie
S-V, V-O, N-A	2160	parlement adopter mesure législatif
V-O, V-P, V-P	2153	avoir débat avec au sujet
N-P-N, S-V, V-O	2033	taux de chômage atteindre niveau

Table 5.5: The most frequent combinations of syntactic types for 4-grams.

have to be discarded if they do not occur frequently by themselves in the corpus. A more sophisticated solution is pursued in (Frantzi and Ananiadou, 1996; Frantzi et al., 2000) with the *C-value* method. The underlying idea is that the more frequently a substring occurs in different longer strings, the more "independent" it is, and is therefore more likely to constitute a term by itself. The nested terms that do not comply with this criterion are eliminated by this method, like *soft contact* from *soft contact lense*. Although this method was designed for sequences of adjacent words, we believe that it could be applied to the flexible chains discovered using our method in order to discard collocation fragments.

5.1.6 Related work

Multi-word collocation extraction methods have generally been implemented so far on plain text or on POS-tagged data (Choueka et al., 1983; Smadja, 1993; Daille, 1994; Dias, 2003). They are characterized by the rigidity and dispersion of the results retrieved, due the impossibility of dealing with syntactic variation. With respect to these methods, our method has the advantages of being able to abstract away from the specific text realization, to detect discontinuous items in the case of more syntactically versatile collocations, and to group various instances of a collocation type under the same canonical form.

The method of Choueka (1983) identifies frequent sequences of at most 6 consecutive words, and is computationally expensive due to the high number of candidates that are generated from a corpus. In the second stage of Xtract, Smadja (1993) generates multi-word collocation candidates by analyzing the surrounding positions of already extracted bigrams. Rigid noun phrases and phrasal templates are thus identified (e.g., *The consumer price index*, *The NYSE's composite index of all its listed common stocks *VERB* *NUMBER* to *NUMBER**), which are argued as being useful for language generation.

An onerous method is proposed by Dias (2003), that generates continuous or discontinuous candidate sequences within a 7-word window. These are still rigid, because they follow the linear order in the text. Multiword units are identified among these candidates by combining the association scores obtained, on the one hand, for the sequence of words and, on the other hand, for the corresponding sequence of POS tags. Those sequences of length n whose score is greater than the score of all immediately subsuming and subsumed sequences (i.e., sequences of length $n+1$ and $n-1$) are retained as valid units. The system implemented was successfully used for extracting English trigrams, but failed for longer sequences (the precision reported is very low, between 20% and 40%). Another serious limitation of this system is that it cannot deal with large bodies of text because of the complexity of the approach.

A more linguistically motivated method is adopted by Daille (1994), who tackles the problem of long compound nouns by enumerating the syntactic patterns defining them, and by listing the operations that can lead to their formation from shorter parts (juxtaposition, substitution, modification or coordination). As she acknowledges, it is hard to tell these operations apart. For instance, the N1-A-P-N2 term *réseau national à satellites* can be obtained either by substitution (in *réseau à satellites*, *réseau* is replaced by *réseau national*), or by modification (*réseau à satellites* is modified by the adjective *national*). Our approach takes the opposite direction: the relevant patterns do not have to be known beforehand, since they do not guide the extraction; multiword collocations are found in an unconstrained way, based on the co-occurrence of their already identified subparts in the text. The patterns can be inferred a posteriori from the data extracted. It is indeed difficult to imagine a different process, given the

multitude of syntactic configurations in which the participating items can be found in a collocation; the long compound nouns that Daille deals with are, in comparison, much more restricted from the point of view of syntax.

The (very few) methods that extract long candidates from syntactically parsed text deal with specific types of trigrams, such as multi-word verbs containing a verb and two particles, e.g., *look forward to* (Blaheta and Johnson, 2001), or A-N-V combinations in German, like *rote Zahlen schreiben* (lit. *red numbers write*, 'be in the red') (Zinsmeister and Heid, 2003). Both configurations correspond to particular sequences of types that were also identified in the experiments run with our method (note that the counterpart of A-N-V in our system is V-O, N-A).

More similar to our approach is the method of Kim et al. (1999; 2001) that retrieves n-grams from Korean text by creating clusters of bigrams having similar values for relative entropy and Dice coefficient. Since no syntactic criterion is used for linking the bigrams, it is not clear whether the combinations obtained are indeed grammatical, as suggested by examples that translate as *white sneakers wear* and *give birth to a child*, or the items in the obtained sets just happen to co-occur in the same sentence as a consequence of the fact that both measures rely heavily on co-occurrence frequency.

5.2 Pattern Induction

5.2.1 Motivation

An issue that collocation extraction systems have to address from the very beginning is the definition of the accepted POS combinations, because the quality of their results—in particular, the coverage—depends crucially on this initial choice. In our case, since the extraction is based on syntactic parsing, the question that we have to answer is what syntactic configurations are adequate for describing collocations.

Quite often, extraction systems are only interested in a single configuration, like V-O, P-N-V, or P-N-P (see the review in Section 3.4), or consider only the several most representative configurations (Lin, 1998; Lin, 1999). As a matter of fact, the

POS combinations or the syntactic configurations (which we call *patterns* in short) are highly divergent from one system to another, as can be seen from the synoptic view provided in Table 5.6.[7] One reason for this high divergence is certainly the lack of consensus in the morphosyntactic descriptions of collocations, as discussed in Section 2.5.2.

	A-N	N-N	N-P-N	S-V	V-O	V-P	V-P-N	Adv-A	V-Adv
(Benson et al., 1986a)	✓		P = *of*	✓	✓		✓	✓	✓
(Hausmann, 1989)	✓	✓	✓	✓	✓			✓	✓
(Smadja, 1993)	✓	✓		✓	✓	✓			✓
(Basili et al., 1994)	✓	✓	✓	✓		✓			
(Lin, 1998)	✓	✓		✓	✓				
(Kilgarriff and Tugwell, 2001)	✓	✓	✓	✓	✓	✓			
(Goldman et al., 2001)	✓	✓	✓	✓	✓	✓	✓		

Table 5.6: Patterns used for collocations in the literature.

Some of the most representative patterns that are currently supported by our extraction system are shown in Table 4.2. As mentioned in Section 4.3.1, the complete list of patterns is actually longer, and is evolving as more and more data is inspected.

In our opinion, the definition of the collocationally relevant patterns cannot rely exclusively on linguists' intuition, but has to be supported by a corpus-based investigation of empirical data. This section presents an experiment aimed at the data-driven acquisition of patterns in a semi-automatic way, that has been conducted on both English and French text corpora. A brief report on this experiment can be found in (Seretan, 2005).

5.2.2 The method

We address the problem of finding a set of collocationally relevant patterns that is as exhaustive as possible by adopting a corpus-based approach. The identification method consists of two main steps. In the first step, all the productive patterns in

[7]This table displays for (Benson et al., 1986a) only the lexical collocations listed in the preface of the BBI dictionary. Also, the system of Smadja (1993) deals with the following additional types: V-V, N-P, N-D. Similarly, Basili et al. (1994) state that about 20 patterns were used, while our table displays only those that were explicitly mentioned in their publication.

a language are detected, then in the second, the patterns judged as most interesting are eventually retained.

More precisely, in the first step the method extracts collocations from the source corpora by relaxing the syntactic constraints that apply to the candidate data. As explained in Section 4.3.1, candidate pairs are identified, in our system, from partial or complete parse trees built for the source sentences by the Fips parser. Several constraints are imposed on these candidates, which have to be, first of all, syntactically bound, then must conform to a specific configuration from a set of predefined configurations. For our purpose, we now only impose the syntactic criterion, and consider any possible POS combination as relevant a priori. Given [$_{XP}$ L X R], a syntactic structure built by Fips (refer to Section 4.2 for the relevant details), the set of candidates is built by retaining all the combinations between the current head X, and the heads of the left and right subconstituents L and R. The procedure continues by collecting such pairs from the substructures L and R, recursively.

All the lexical categories considered by Fips are allowed, therefore the initial set of patterns is very permissive. Each combination of two (not necessarily distinct) categories is, in principle, allowed as collocationally relevant. However, some combinations, like P-V or N-Adv, are obviously impossible due to the grammar rules defined in Fips.

In the second step of the method, the identified candidates are ranked using LLR. The most salient pairs that correspond to the productive combinations undergo a process of manual analysis, that decides the "interestingness" of each pattern and therefore its inclusion in the actual list of patterns used by the extraction system.

5.2.3 Experimental results

Two different pattern induction experiments have been carried out using the method described above on English and French corpora of newspaper articles. The English corpus contains online articles from the journal "The Economist", and the French data belongs to the corpus "Le Monde" distributed by the Linguistic Data Consortium.[8]

[8] `http://www.ldc.upenn.edu/`

Several extraction statistics are displayed in Table 5.7: the size of the source corpora,[9] the total number of pairs extracted, and the number of distinct pairs for each language (rows 1–3).

Experimental data	English	French
words (approx.)	0.5 M	1.6 M
extracted pairs (tokens)	188527	748592
extracted pairs (types)	65853	171584
productive POS combinations	60	57

Table 5.7: Statistics for the pattern induction experiment.

The number of productive patterns found is shown in the last row of Table 5.7. The results indicate that a high number of POS combinations are actually productive. Some of these patterns contain very numerous pair instances (e.g., D-N, P-D, N-P, P-N), others very few (Conj-N, V-Conj, V-V). Table 5.8 and Table 5.9 list some of the new collocationally relevant patterns found for English and French, in addition to those already used in our extraction system and that are shown in Table 4.2.

These new patterns are commonly ignored by extraction systems, since they include function words. In the BBI dictionary (Benson et al., 1986a), however, such patterns are widely represented. They are called *grammatical collocations*, as opposed to *lexical collocations* which are made up of open-class words. The preface of the dictionary provides a classification of the grammatical patterns, identified by a series of codes; the last column of Table 5.8 displays, for each pattern identified with our method, the corresponding code in the BBI classification (if a match is found).

5.2.4 Related work

In the trade-off between precision and recall, extraction systems are confronted with the problem of choosing an appropriate set of patterns that describe the type of pairs sought. In order to increase the precision—and also the tractability—of systems, combinations that include functional categories (e.g., D, P, Conj) are usually excluded

[9]Since the experiments are not comparative and were conducted independently, the two corpora are of different sizes.

Pattern	Example	BBI code	BBI example
N-P	*decision on*	G1: N-P	*apathy towards*
N-Conj	*recognition that*	G3: N-that	*agreement that*
P-N	*under pressure*	G4: P-N	*in advance*
A-P	*essential to*	G5: A-P	*angry at*
A-Conj	*necessary for*	G7: A-that	*afraid that*
V-Conj	*judge whether*	–	–
Adv-Adv	*much more*	–	–
Adv-P	*together with*	–	–
Adv-Conj	*rather than*	–	–
P-P	*from near*	–	–

Table 5.8: Some interesting patterns discovered for English and their equivalents in the BBI dictionary.

from the list of accepted patterns. But this choice leads to the failure of systems to capture a whole range of collocational phenomena which, as advocated particularly in (van der Wouden, 2001), are nonetheless important.

An exception is the work of Smadja (1993), which discusses the collocational relevance of patterns like N-P (*accordance with, advantage of, agreement on, allegations of, anxiety about*) and N-D (*some people*). Lexicographic interest in such patterns is proven by the large amount of grammatical collocations listed in specific dictionaries like BBI (Benson et al., 1986a).

We adhere to the view expressed by Fontenelle (1992) and van der Wouden (2001), according to which lexical items of any category can show collocational effect. This is why, in our attempt to induce patterns from data, we considered any possible POS combination as relevant a priori. Then we further narrowed down the obtained set of productive patterns (about 60 in English and in French) through an analysis of the pair instances found, which led to the discovery of a relatively reduced number of patterns (about 20 per language) that could indeed be useful for future extraction.

Part of the patterns found are combinations made up exclusively of closed-class categories, plus adverbs (e.g., Adv-Adv, Adv-Conj, Conj-Adv, P-P, P-Adv). Employing these patterns in extraction is perhaps not useful, because, although highly frequent, they would yield very few pair types, which can be exhaustively listed and

Pattern	Example
N-P	*précaution quant*
N-Conj	*idée que*
P-N	*sur mesure*
A-P	*déterminé comme*
A-Conj	*probable que*
V-Conj	*faire ainsi que*
Adv-Adv	*bien au contraire*
Conj-Adv	*ou bien*
Adv-Conj	*d'autant plus que*
P-P	*jusque sur*
P-Adv	*depuis longtemps*

Table 5.9: Some interesting patterns discovered for French.

known in advance. On the contrary, the patterns that include an open class category are very useful, as they permit the automatic discovery of the collocational properties of a large number of words.

Pursuing the same goal of solving the problem of pattern definition, (Dias, 2003) also proposes a data-driven approach, but one in which the relevant sequences of POS tags are identified with an association measure called Mutual Expectation.[10] As this method is bound to the surface of the text, the patterns produced reflect the order of words in the text; they are affected by dispersion and rigidity, and may be ungrammatical. On the contrary, the patterns detected with our parse-based method are independent of their textual order and are grammatically valid.

The method we proposed can be applied to any language supported by the parser in order to detect collocationally relevant configurations of length two; these, in turn, can be used in the extraction of longer collocations using the method presented in Section 5.1. However, the full customization of our collocation extraction system for a new language should also take into account factors such as differences in lexical distribution across languages, because these are responsible for the performance of

[10]More precisely, multi-word units are then identified in (Dias, 2003) by (1) applying this measure on both sequences of words and their POS tags, (2) combining their scores, and (3) retaining only the local maxima candidates as valid, following the method briefly explained in Section 5.1.6.

association measures when applied to a specific pattern. For instance, in French there are fewer V-P pairs than in English, where they constitute phrasal verbs and verb-particle constructions, therefore a measure that is suited to this pattern in English might be less suited to it in French.[11]

5.3 Web-based extraction

5.3.1 Motivation

The third extension that enhances our extraction method relates to the replacement of the source corpus with data retrieved from the Web with a search engine. The motivating scenario is that often a user (e.g., a lexicographer, a language learner) or an application is interested in the collocates of a specific word. One possible solution is to search that word in the list of collocations extracted from an existing corpus. This kind of search is possible, for instance, with the Sketch Engine (Kilgarriff et al., 2004) that queries the BNC. But in the absence of corpora or of sufficient data for a word, an alternative solution is to mine the Web for collocations with that word.

Researchers are nowadays making increasing use of the Web as an alternative source of linguistic evidence. In several studies, Web co-occurrence frequencies have already been used as indicators of word collocability. By comparing the number of hits for two alternative expressions such as *faire une question* and *poser une question*, it is possible to predict which variant is more plausible. For instance, the first combination obtained around 23300 hits with Google on a search performed by the time this manuscript was written,[12] while the second, with approximately 2180000 hits, is almost a hundred times more frequent and is therefore clearly preferred over the first.

The approach we propose is however different, because the real problem stands less in comparing two alternative combinations (or in validating a single one), and more in discovering the possible combinations, which are not known beforehand. The only

[11] In (Seretan and Wehrli, 2006b) we discuss in detail the problems a collocation extraction system—not necessarily syntax-based—faces when ported from English to a new language with richer a morphology and more flexible word order.

[12] `http://www.google.com`, search performed December 2007.

element that is known to the user is the base of the collocation (for instance, in our previous example, the noun *question*). What is required is the right collocate, that is, the word that typically combines with the base word, and this is unpredictable for the user (*poser*). Our method scans the contexts (or *snippets*) retrieved by a search engine for the base word, in order to find collocates that are in a specific syntactic relation with that word. This method is introduced in (Seretan et al., 2004c).

5.3.2 Identifying collocations from Web data

The Google search engine provides a programming toolkit (the Google Web APIs) for the parametrizable search of its indexed pages. A query can specify the word(s) sought, the language of the documents to be retrieved, and the desired number of results. The format of query results is illustrated in Example (6). The word sought, the noun *comparison*, is highlighted by the search engine with HTML bold tags () in the context of the matched document (shown in the field `Snippet`).

(6) [
URL = "http://swz.salary.com/costoflivingwizard/layoutscripts/coll_start.asp"
Title = "Salary.com's Cost of Living Wizard Tool"
Snippet = "Salary.com's cost-of-living calculator will compare living-cost indexes and salary
 differentials to help you make an informed **comparison**. ..."
Directory Category = {SE="", FVN=""}
Directory Title = ""
Summary = ""
Cached Size = "116k"
Related information present = true
Host Name = ""
]

The Web-based collocation extraction method that we implemented proceeds as follows. In the first step, the snippets are extracted from the query results and are cleaned, so that unwanted elements (such as HTML tags) are removed. Also, those

snippets that do not contain the query word are eliminated.[13] The URL of the source document is listed next to each clean snippet for reference. The format obtained is shown in Example (7).

(7) "Salary.com's cost-of-living calculator will compare living-cost indexes and salary differentials to help you make an informed comparison. ..."
(http://swz.salary.com/costoflivingwizard/layoutscripts/coll_start.asp)

This step allows a small corpus to be built for the base word chosen, which is likely to contain useful collocational information. For instance, the snippet above contains two potential collocates of the noun *comparison*: the verb *make* and the adjective *informed*.

In the second step, this mini-corpus is parsed with Fips and collocation candidates are identified, then ranked according to the LLR measure by employing the extraction method described in Section 4.3. Finally, the syntactically related pairs of words thus obtained are filtered, so that only combinations with the base word having the requested syntactic types are displayed to the user, in the order given by the LLR score or by their frequency in the mini-corpus created.

The concordance tool (described in Section 6.3) is then used to display, one by one, the snippets containing a selected collocation. If interested in a larger context, the user can access the source document by clicking on the link displayed next to the snippet, as in Example (7).

The number of retrieved collocates depends, obviously, on the desired number of matches for a query. The experiments run showed that interesting results can be obtained even from a relatively small number of snippets (such as 100 or 200), although a larger input would certainly help to increase the performance of the association measure employed.[14]

The method implemented can also iteratively process a list of words, rather than a single word. Unfortunately, it cannot be used on a large scale, because the search

[13] Google may return a matching document even if the word sought is not actually found in the text body, but appears in the HTML title of that document.

[14] LLR is argued to have an acceptable performance on low-frequency data anyway (Section 3.2.4).

engine currently limits the number of queries that can be performed on a daily basis by a user (who is identified with a client key).

5.3.3 Sample results

To illustrate the potential of this method to discover collocations from the Web, we display the V-O and A-N pairs retrieved for the base word *argument* from a mini-corpus created from 1000 Web snippets. Since no frequency filter is applied by the method, many of the collocates proposed shown are actually hapaxes (i.e., single occurrences).

(8) a. verb-object: *analyse, answer, assert, associate, attempt, attend, base, be, call, define, describe, design, develop, devise, dismiss, do, enjoy, examine, find, have, know, make, offer, organize, plot, position, present, put, read, reconstruct, refer, refute, represent, require, revise, seek, set, share, sketch, state, strengthen, study, take, understand, use, win, write*;

b. adjective-noun: *a priori, actual, aesthetic, apparent, Aristotelian, big, Chinese, circular, classical, combinatorial, common, complicated, computational, corresponding, decent, different, economic, emotional, empiric, English, erroneous, existing, explicit, external, false, final, first, flawed, following, formal, French, good, heuristic, huge, implicit, intelligent, interesting, knock-down, last, lexical, linear, linguistic, logical, main, mediate, necessary, nice, old, opening, oral, original, parse, pernicious, persuasive, philosophical, possible, private, public, religious, respective, second, sign, simple, single, slippery, sound, standard, strong, technical, teleological, thought-provoking, tired, unique, universal, unpopular, unresolved, unsigned, valid, weak, weighty, worthwhile.*

Apart from a few exceptions, i.e., *Chinese argument* and *slippery argument* that constitute parsing errors due to wrong attachments (the source expressions being *Searle's Chinese room argument* and *slippery slope argument*, respectively), the pairs

retrieved are valid collocation candidates; moreover, as can be observed, a high proportion of these might be of lexicographic interest.

Table 5.10 provides a comparative view of the collocates discovered from the Web and those listed in the BBI dictionary (Benson et al., 1986a) for a different base word, the noun *approach*; the syntactic types shown here are V-O and A-N. The Web-based extraction involved 1000 snippets. The collocates displayed here represent only a selection of the total results, which count 58 V-O pair types and 91 A-N pair types with this noun.

Type	**BBI**	**Web**
V-O	*make,* ***take***	*adapt, adopt, apply, base, build, design, develop, follow, found, implement, offer, provide, pursue,* ***take****, use*
A-N	*audio-visual, careful, cautious, conservative, creative, direct, down-to-earth, easy-going, forthright,* ***fresh****, hard-nosed, holistic, indirect, inflexible,* ***innovative****, judicious,* ***new****, no-nonsense,* ***novel****, objective, oral-aural, pragmatic, rational, realistic, scholarly,* ***scientific****, simplistic, uncompromising, unrealistic*	*alternative, balanced, base, basic, common, complementary, comprehensive, different, distinctive, eclectic, effective, empirical, experimental,* ***fresh****, general,* ***innovative****, integrate, interdisciplinary, logical, modular, multidisciplinary,* ***new****,* ***novel****, oriented, practical, precautionary, principled, reliable, right, robust,* ***scientific****, simple, standard, statistical, straightforward, successful, systematic, technical, theoretic, traditional, true, uncommon, unique*

Table 5.10: BBI collocates and Web collocates for the noun *approach* (common items are marked in bold).

5.3.4 Related work

The Web is a rich, constantly-available, and ever-evolving resource that is now increasingly used for linguistic inquiries, as can be seen from reviews like (Volk, 2002; Lüdeling et al., 2007; Kilgarriff and Grefenstette, 2003). As far as work related to collocations is concerned, Web co-occurrences hits have already been used by Pearce

(2001a) for discriminating between synonymic combinations, in order to find out which of these is more likely to form a collocation (as in the example discussed in Section 5.3.1). Similarly, Zaiu Inkpen and Hirst (2002) applied the MI association measure on frequencies obtained from the Web in order to validate bigrams that were previously extracted from the BNC. The Web frequency information may be criticized for being unreliable, since the counts obtained for a pair might include numerous cases in which the words are actually unrelated[15] or have unwanted categories. However, a recent study showed that Web frequencies correlate well with human plausibility judgments, at least as far as A-N, N-N, and V-O pairs are concerned (Keller and Lapata, 2003).

In the work cited above (Pearce, 2001a; Zaiu Inkpen and Hirst, 2002), it is assumed that the combinations tested are known in advance. Our method focuses, on the contrary, on mining possible collocates from Web contexts when only the base words are known. The syntactic preprocessing of the retrieved contexts is essential for the success of this task, as it enables our method to retrieve relevant results from a limited amount of text, in which collocation candidates have a low frequency. In addition, the pairs discovered are grammatical and can easily be filtered according to syntactic type, as the user is presumably interested in a particular type of syntactic relation.

A similar approach motivates the Linguist's Search Engine (LSE) (Resnik and Elkiss, 2005), that provides a syntactic-based search on corpora built from the Web. Matched documents are downloaded, parsed and annotated, then they are indexed for future searches. The LSE query system, based on syntactic structures, is very powerful. But the whole process is much more time-consuming than our more specialised method, which only downloads and parses the matched contexts. Another advantage of our method is that it returns all the syntactic variants of a collocation in a single search.

Another Web-based tool that allows the discovery of word collocates, but without relying on syntax, is WebCorp.[16] The Web contexts retrieved with the selected search engine for a given word are analysed and co-occurrence tables are produced,

[15]For instance, the items of the pair *faire question* co-occur accidentally in the following snippet: *L'Ecole obligatoire : Pourquoi faire ? Une question trop souvent éludée.*

[16]`http://www.webcorp.org.uk/` (accessed December, 2007).

which show the most frequent words in the surrounding positions.[17] As expected, the function words dominate the lists of the collocates proposed.

5.4 Summary

This chapter addressed several issues that have been largely ignored in previous collocation extraction work. The first issue is related to the length of collocations. Binary collocations like those whose extraction was discussed in the preceding chapters are only one facet of the phenomenon of word collocability as described by theoretical studies (see Chapter 2). Collocations may consist of an arbitrary number of words. As a matter of fact, some of the previously identified binary collocations were found to occur significantly in text in combination with other collocations. By relying on the concept of 'collocation of collocations', we extended the extraction method presented in Chapter 4 to cope with combinations of arbitrary length. The identification of longer collocations is particulary necessary in cases where the corresponding subparts do not constitute collocations by themselves, but only fragments of longer collocations; the problem of fragments is well-known in related fields such as terminology extraction.

A second issue dealt with in this chapter was the problem of deciding which syntactic patterns must be used for collocation extraction in a given language (this problem is also faced by extraction approaches which are not based on syntax, since here the set of allowed POS combinations has to be defined as well). Generally chosen in an arbitrary way, the predefined patterns used in extraction are nonetheless essential for the quality of results. In the tradeoff between precision and recall, functional categories are usually disregarded in existing extraction work. Combinations with

[17] The following collocates were found by WebCorp for the word *argument* with a query executed December, 2007: Top external collocates of "argument": *the, of, is, a, to, that, an, in, for, and, The, be, as, this, not, it, argument, or, I, with*; Key Phrases: *the argument, an argument, Invalid argument, ontological argument, this argument, of argument, The argument, first argument, taxpayer argument, hole argument, that argument, argument is, argument that, argument supplied, argument to, argument and, argument in, argument for, argument was, argument of, argument or, argument I*. At the time this query was run, the option *internal collocates* intended for wildcard searches, e.g., "to * an argument" or "a * argument", was not working.

closed-class words were found, however, to constitute an important part of the new patterns discovered for English and French by the semi-automatic method proposed, which infers collocationally relevant patterns in a data-driven fashion.

A third enhancement of our method is related to the Web-based search for collocates of a given word, proposed as a solution to the absence of static corpora, as well as to the lack of sufficient information about the word in these corpora. There is a growing awareness in the research community of the potential constituted by the Web as a source of linguistic data, as witnessed by the series of workshops dedicated to this topic.[18] In our Web-based collocation extraction method, a mini-corpus is instantly built for the investigated word from the text snippets retrieved by the Google Web APIs. This corpus is then processed using the extraction method based on parsing. The experiments carried out showed that competitive results can be obtained from resources of limited size, like those built from several hundreds of snippets.

In all the implemented extensions, considerable leverage is gained from the syntactic analysis of the input text. This helps in abstracting away from the surface text realization of collocation candidates, enables the convenient grouping of multiple instances belonging to the same type, and permits the retrieval of discontinuous instances of the more flexible collocation types.

[18]E.g., the Web-as-a-corpus workshops endorsed by the SIGWAC, Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus.

Chapter 6

Collocation Extraction and Visualisation Tool

This chapter presents a system dedicated to the extraction of collocations from corpora, which is based on the methodology described in the preceding chapters. The system's full documentation and user guide are available as stand-alone documents, not included in this manuscript due to their size. A preliminary version of the system has been described in (Seretan et al., 2004b). This chapter provides an updated, synthesised description of the system, as well as additional details where necessary. It also presents an application of this tool in the task of collocation translation.

6.1 Introduction

The work described in this thesis originates in a research project with practical motivations, which brought together the Language Technology Laboratory of the University of Geneva and the Division of Linguistic Services and Documentation of the World Trade Organisation (WTO).

In the multilingual environment specific to international organisations such as the WTO, the recognition of the specific expressions used to refer to key concepts during a negotiation process, as well as their consistent translation in other languages, play a crucial role in the success of that negotiation. Therefore, the automatic identification

of collocations in corpora has an immediate application to the work of terminologists and translators of such institutions.

The project's goal was to help them inventory typical expressions in a collection of documents in a given language, and their equivalents in other languages. In particular, the project aimed at building a tool for the automatic acquisition of collocations from the WTO text archives in French and English. This tool had to enable the visualisation of the extracted collocations in the source context, and to find, via text alignment, the counterpart of this context in the target documents, whenever a translation of the source document exists.[1] It also had to be easy to integrate in the translators' workbench, and was therefore customised, to a certain extent, for the specific needs of these translators.

6.2 General description

The tool I developed is an integrated system for multilingual collocation extraction, visualisation in parallel corpora via sentence alignment, and management of a collocational database functioning like a translation memory. It was designed as an aid for human translators wanting to exploit their translation archives in future translations, but can also be used by lexicographers, second language learners, or—as shown in Section 6.5—by NLP applications.

The system has the following capabilities:

1. selection of the documents to be processed (source corpus) based on the recursive scanning of a file directory structure and on the application of multiple criteria (involving, for instance, the directory and file name, the file type, the file's last modification date), optionally complemented by a manual selection of files.
2. accurate collocation extraction from text corpora in multiple languages, based on the syntactic analysis of input text produced by the multilingual parser Fips (Laenzlinger and Wehrli, 1991; Wehrli, 2007) (described in Section 4.2).

[1] Each WTO document typically exists in 3 versions corresponding to the WTO official languages, English, French, and Spanish.

The languages that are currently supported are French, English, Spanish, and Italian.[2] Collocations of a wide range of syntactic types are identified by using the extraction methodology presented in Chapters 4 and 5. They may consist of two or more lexemes that are not necessarily contiguous in the text.[3] The collocation strength can be assessed by using any of the association measures implemented, which are listed in Appendix L.

3. a concordancing functionality that enables the visualisation of extracted collocations in the source documents, by automatically scrolling the document to the source sentence and highlighting both the sentence and the current collocation within it.
4. an alignment functionality that retrieves the corresponding sentence in the target documents, thus helping the user to see how a given collocation has been previously translated in a certain context.
5. a filtering functionality allowing the user to specify which collocations should be visualized, depending on their syntactic type, corpus frequency, collocation score, etc.
6. functionalities dedicated to the export of manually validated collocations into monolingual and bilingual databases that store relevant information with each entry, including text samples for the source and target contexts.

A complementary functionality of the system is the Web-based extraction of collocates for a given word, using the method described in Section 5.3.

6.3 Detailed description

The system is implemented in Component Pascal under BlackBox Component Builder IDE,[4] just as the syntactic parser Fips, on which it relies. It makes an extensive use

[2]At the time this project was carried out, Spanish was not yet supported by the Fips parser, and therefore was only considered as a target language.

[3]Recall from Section 1.2 that we use the term *multi-word collocation* to refer to collocations made up of strictly more than two lexemes.

[4]BlackBox is developed by Oberon Microsystems (`http://www.oberon.ch`). A peculiarity of this development environment is the ease of editing graphical user interfaces components, which turned into a big advantage for our system, since data visualisation plays a major role.

of the SQL database query language in order to store the extraction results, compute the collocation scores, filter the data that will be displayed, etc.

The architecture of the system is mainly pipelined, the execution flow typically following the order in which the main components—corresponding, basically, to the functionalities mentioned in the previous section—are described below. However, there are no restrictions to the order in which the various components can be used, since the extracted and validated results can be stored between sessions for later visualisation.[5]

1. Selection of source files The source corpus used in an extraction session is specified by selecting the folder which contains the desired files and, optionally, by applying a filter on its content, based on the file location (inclusion or exclusion of the subfolders; exclusion of subfolders having a specific name), file name (it must contain a given string of characters), file type (it must belong to a list of allowed types),[6] last modification date (from `date1` to `date2`; in the last `n` days). In addition, the selection can be further narrowed manually, the user being able to select or deselect items in the first level of the source folder with a mouse click or by using standard selection commands (check all; uncheck all; invert selection).

2. Processing of source files This is the main component of the system, which controls the extraction process. The number of files that can be processed is virtually unlimited. The selected files are processed one by one, and the results—the collocation candidates identified by parsing—are accumulated either in a database or in a text file; as an option, they can also be stored in a destination folder whose structure mirrors the structure of the source folder. At the end of the extraction process, several processing statistics are computed for the source corpus that are derived from parsing information (e.g., the total number of tokens, sentences, sentences with a complete parse), and the identified candidates are ranked according to the chosen association measure (by default, LLR). Multi-word collocations are not extracted at

[5]Most of the user options are also kept between sessions (e.g., the corpus selection parameters).

[6]The system supports all the file formats that can be currently imported by BlackBox (e.g., `odc` - Oberon document, `txt`, `htm`, `html` - text, `rtf` - rich text format, `utf` - Unicode).

this stage, but only later depending on the visualisation options set by the user (see the concordancing component).

3. Collocation filter This component specifies the results that are to be displayed in the visualisation interfaces (concordance and alignment interfaces). The extracted collocations can be filtered according to several criteria: syntactic type (the user can select one or more types from a list that is automatically built from the database containing the extracted collocations), collocation score (from `score1` to `score2`), corpus frequency (from `freq1` to `freq2`),[7] or collocation keywords (in this manner, the user can search for collocations containing a specific word). Moreover, the user can specify the range of results to display (from `rank1` to `rank2`), according to the order given by the collocation score or by the corpus frequency. The range restrictions can be applied both to collocation types and to collocation instances. In the case of multi-word collocations, the implementation of the filtering features is currently limited to the corpus frequency and the syntactic type.[8]

4. Concordancing This component is responsible for the visualisation of extraction results according to the selection made by the user. The (filtered) list of collocations is displayed on the left handside on the concordance interface, and can be ordered by score, by frequency in the corpus, or alphabetically. On the right handside, a text panel displays the context of the currently selected collocation in the source document. The whole content of the document is accessible, and is automatically scrolled to the current collocation; this collocation and the sentence in which it occurs are highlighted with different colors.

Each item in the list represents a collocation type; its corresponding instances are read from the database when the user clicks on it. The right panel automatically displays the first instance, then the user has the possibility to navigate through all the instances using the standard browsing arrows (`<< - first`, `< - previous`, `> -`

[7] The user is not required to know the actual maximal values; the corresponding fields can be left blank and these values will be retrieved by the system.

[8] More precisely, the syntactic filter takes into account the POS categories in a multi-word collocation, rather than the syntactic types of the participating bigrams.

`next`, `>>` - `last`), and to skip to a given instance by entering its order number. The visualisation interface also displays information about the rank of the currently selected collocation, its syntactic type, its score, and its status relative to the parser's lexicon (new collocation, or collocation in lexicon). The user can easily switch to a different source language in order to load the collocations extracted for that language, if these were stored in the same database.

The concordance interface for displaying multi-word collocations has similar features. In addition, the user can select the length of collocations to display. Multi-word collocations are then obtained from the previously-extracted binary collocations by using the method described in Section 5.1.[9]

5. Alignment When parallel corpora are available, the target sentence containing the counterpart of the source sentence can be detected and displayed in the alignment interface below the source sentence. The user selects the target language from a list of languages,[10] and specifies the path of the target corpus and the filename transformation rule needed to determine the filename of the target document (i.e., of the translation) from the filename of the source document.[11] Once the target file is found, the sentence that is likely to be the translation of the source sentence is identified using the sentence alignment method which will be presented in Section 6.4. The alignment component works for both for binary collocations and for multi-word collocations.

6. Validation This component provides functionalities that allow the user to create and maintain a list of manually validated collocations from the collocations visualised

[9]Currently, the maximum length allowed is 5, since the procedure that extracts collocations of higher arity is not fully automated.

[10]This list can be easily customised by editing a text file.

[11]Such rules assume that the source folder and the target folder have the same structure, and that the target filename can be obtained from the source filename by replacing the prefix and/or the suffix of the filename (which are assumed to be variable across languages), while keeping the middle part and the file extension constant (e.g., `35.1.001E.txt` can be obtained from `35.1.001F.txt` by replacing the suffix `F` with `E`). The same assumption is made in similar systems such as TransSearch, except that files are manually renamed in a corpus preprocessing stage (Macklovitch et al., 2000).

with the concordance and the alignment interfaces. An entry contains basic information about a collocation (such as the collocation keywords, lexeme indexes for the participating items, syntactic type, score and corpus frequency). A monolingual entry may also contain the source sentence of the currently visualised instance, which provides a naturally-occurring usage sample for the collocation. A bilingual entry stores, in addition, the target sentence found via alignment and the translation proposed for the collocation: the translation can be manually retrieved by the user from the target sentence. Additional information related to the currently visualized collocation instance is stored (namely, the name of the source and target file, the file position of the collocation's items in the source and target files, and the file position of the source and target sentences). Most of this information is automatically (and sometimes, tacitly) filled-in by the system. The entries in the list of collocations validated in a session can be updated, deleted, or saved—completely or in part—by the user in a monolingual and in a bilingual database.[12]

Appendix M presents several screen captures for the components described above.

6.4 Sentence alignment

The sentence alignment method used by our extraction system was designed specifically for the needs of this system. Its main distinguishing feature is that it computes a partial, on-the-fly alignment for the source sentence of the collocation instance which is selected by the user in the visualisation interface.

Given the source file and the file position of one of the items in a collocation, it retrieves the target file[13] and determines the file position of the beginning and the end of the target sentence (i.e., the sentence that is likely to represent the translation of the source sentence in the target language). This output is used directly by the system for displaying the target sentence and is not saved in any way. Since the

[12]The validation component could easily be adapted to the task of collocation annotation carried out in the evaluation experiments described in Section 4.4 by simply adding a new field to the entry, standing for the label associated with a collocation.

[13]The retrieval of the target document is based on the target folder path and on the filename transformation rules specified by the user, as discussed in the previous section.

procedure is robust, it requires no pre-processing of the files. It is very fast and enables the instant visualisation of the target text on the alignment interface.

The alignment is first performed at the paragraph level as described below. The alignment of sentences inside the aligned paragraphs simply assumes a 1:1 match (which, even if not always borne out in practice, is acceptable for the goal of visualising collocations in context in the larger document body).

The rest of this section presents the paragraph alignment method that was implemented. The strategy followed is to choose an initial target paragraph candidate, then to optimize this choice by taking into consideration the relative sizes of neighbouring paragraphs, as well as, to a limited extent, lexical clues.

The choice of the first paragraph (called *pivot*) is based on the file position relative to the file size, by assuming that the positions of the source and target paragraphs inside the corresponding documents are correlated (see Equation 6.1). Given the input position pos_{source}, the pivot is identified as the paragraph found at the position pos_{target} in the target file.

$$\frac{pos_{source}}{length_{source}} \approx \frac{pos_{target}}{length_{target}} \tag{6.1}$$

This initial choice is further refined locally. Let $\mathcal{A}$ be the set of paragraphs around the pivot, identified by their order numbers: $\mathcal{A} = \{pivot - offset, ..., pivot + offset\}$, where *pivot* is the order number for the pivot and *offset* defines the size of the search space (i.e., there are *offset* paragraphs before and after the pivot). The target paragraph (hereafter, TP) is chosen among the paragraphs in $\mathcal{A}$ so that its context is most similar to the context of the source paragraph (SP), with respect to the size of paragraphs in characters.

The graphical representation provided in Figure 6.1 suggests the manner in which the method works, by trying to match the size configuration for the context of SP with a compatible configuration in the target text, as if sliding the context over the space around the pivot in order to find the position where it fits best.

Let s be the paragraph size vector for the source context, which includes c paragraphs before and after SP (the length of s is $2c + 1$). Let t_a be the corresponding

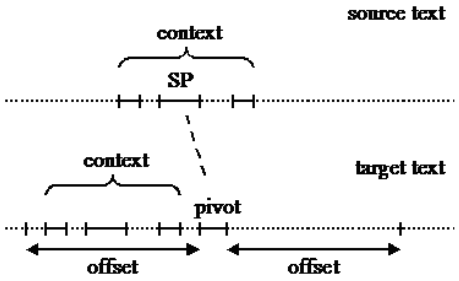


Figure 6.1: Alignment of source and target paragraphs by searching the best match of the context around a pivot.

vector for the target context, with a an index over $\mathcal{A}$ (t_a has the same length as s). The compatibility between the source and target contexts, s and t_a, is measured with the distance formula shown in Equation 6.2, which takes into account the size proportions of adjacent paragraphs.

$$dist(s, t_a) = \sum_{i=1}^{2c} \left| \frac{s_i}{s_{i+1}} - \frac{t_{a_i}}{t_{a_{i+1}}} \right| \tag{6.2}$$

Finally, TP is identified as the argument a that minimizes this distance:

$$TP = \underset{a \in \mathcal{A}}{\operatorname{argmin}}\, dist(s, t_a) \tag{6.3}$$

The assumption made by this method is that the size proportion for two adjacent paragraphs is the same in both the source and the target languages.[14] The default value for both *offset* and c is 5, i.e., TP is sought at a maximum distance of 5 paragraphs from the pivot, by taking into account the size of the 5 preceding and 5 following paragraphs.

After this size-based identification of TP, the match is further validated by a limited analysis of the paragraphs' content, which takes into account the paragraph numbering: the source and target context must have similar numberings.

The alignment method attained an average precision of 90.9% when evaluated on a test set of 800 randomly-chosen sentence alignments obtained on the WTO

[14]Related work on text alignment assumes a correlation between single source and target paragraphs only (Gale and Church, 1993).

corpora for different pairs of languages (French-English, French-Spanish, English-French, and English-Spanish). These corpora are relatively complex, as they contain noisy elements (such as tables), differently-segmented paragraphs and section titles, or deletions of paragraphs.[15]

6.5 Application to collocation translation

The tool described has been successfully used in the task of automatically finding a translation for the extracted collocations (Seretan and Wehrli, 2007). More precisely, the algorithm attempts to locate the translation equivalent of a given collocation in the existing translations, and it does so by using some of the components described in the previous sections.

The strategy used is the following. First, a limited number of corpus sentences in the source language is retrieved for the source collocation, based on its corpus instances detected during extraction. The alignment component is then used for finding, for each source sentence, the corresponding target sentence in the desired target language for which a parallel corpus is available. The target mini-corpus thus obtained is parsed, and collocations are extracted from it using the same method that was applied to the source corpus. Finally, a process of collocation matching takes place, which tries to find, among the extracted collocations, the one that is likely to represent a valid target collocation (that is, a potential translation of the source collocation).

The matching is performed by applying a series of filters on the extracted pairs that gradually reduce their number until a single item is retained, which will be proposed as the target collocation:

- a syntactic filter, which retains only the pairs having a compatible syntactic type. For the time being, the target type allowed is the source type, as it is

[15]State-of-the-art aligners produce almost-perfect results on "normal" texts that match at structural level (98.5%), but their precision degrades sharply on texts with missing fragments (Véronis and Langlais, 2000). As can be inferred from the figure provided in (Véronis and Langlais, 2000, 378), the average precision for the evaluated systems is around 65%.

assumed that collocations preserve their syntactic type across languages (e.g., a verb-object collocation in French corresponds to a verb-object collocation in English, and so on). Although this assumption is too strong, related work has shown that it holds in the majority of cases (Lü and Zhou, 2004). In the future, however, a list of target types could be allowed, provided that syntactic type correspondences would be defined for a given pair of languages to account for differences in the collocation configurations.

- an optional lexical filter, whose application depends on the availability of bilingual dictionaries, and which further reduces the set of candidate translations by choosing only those pairs that contain specific words. This filter is based on the base-collocate dichotomy in collocations, and assumes that the base (i.e., the semantic head of the collocation) preserves its meaning across languages, while the collocate does not. For instance, when the French collocation *gagner argent* is translated into English, the base translates literally (*argent* - *money*), but the collocate changes (*gagner* - *make*; compare *make money* with **win money*, which is an anti-collocation[16]). The algorithm therefore identifies the base word of the source collocation depending on its syntactic type, then searches for translations of that word in the dictionary, and selects the pairs containing the translations found.

- a frequency filter, which chooses among the remaining pairs the most frequent one, which is returned by the algorithm as the translation of the source collocation. Since the target mini-corpus is specific to the current source collocation, the frequency of pairs in this corpus is a good indicator for the target collocation. In case of tied frequency values, no translation is proposed (the experiments run have shown that the inclusion of all of the top pairs affects the precision of the algorithm).

The results obtained for the translation of the top 500 verb-object collocations extracted from English, French, Spanish and Italian corpora part of Europarl (Koehn,

[16]Term introduced by Pearce (2001a).

French-English	**English-French**
accomplir progrès - make progress	breathe life - donner vie
accorder importance - attach importance	bridge gap - combler lacune
accuser retard - experience delay	broach subject - aborder sujet
atteindre but - achieve goal	devote attention - accorder attention
combler fossé - bridge gap	draw conclusion - tirer conclusion
commettre erreur - make mistake	draw distinction - établir distinction
consentir effort - make effort	draw list - dresser liste
constituer menace - pose threat	face challenge - relever défi
demander asile - seek asylum	foot bill - payer facture
donner avis - issue advice	fulfil criterion - remplir critère
effectuer visite - pay visit	give example - citer exemple
établir distinction - draw distinction	give support - apporter soutien
exercer pression - put pressure	have difficulty - éprouver difficulté
faire pas - take step	have reservation - émettre réserve
fixer objectif - set objective	hold discussion - avoir discussion
jeter base - establish basis	hold presidency - assurer présidence
lancer appel - make appeal	learn lesson - tirer leçon
lever obstacle - remove obstacle	lend support - apporter soutien
mener débat - hold debate	make effort - déployer effort
mener discussion - have discussion	make sense - avoir sense
opérer distinction - draw distinction	pose threat - constituer menace
porter fruit - bear fruit	reach compromise - trouver compromis
poser question - ask question	reap benefit - récolter fruit
prendre distance - keep distance	run risk - courir risque
prononcer discours - make speech	set precedent - créer précédent
remplir condition - meet condition	shoulder responsibility - assumer responsabilité
remporter succès - have success	strike balance - trouver équilibre
rendre hommage - pay tribute	take place - avoir lieu
respecter principe - uphold principle	take stand - prendre position
traiter question - address issue	wage war - faire guerre

Table 6.1: Selected translation results, with non-literal collocate equivalent.

2005) showed an average precision of 89.9% and a coverage[17] of 70.9% (the total number of pairs evaluated being 4000). The translation has been done for each of the 12 possible language pairs. Table 6.1 lists some of the translations obtained for 2 language pairs, French-English and English-French. Randomly-sampled results for more language pairs are presented in (Seretan and Wehrli, 2007).

In this experiment, a relatively small number of instances was used for each collocation (50 or less, depending on availability), as that proved to be sufficient for finding a translation. There was no marked difference in precision among the collocations with a high, medium, or low frequency (i.e., between collocations with 31–50 instances, 16-30 instances, and 1-15 instances), as can be seen in Table 6.2 taken from (Seretan and Wehrli, 2007). The translation coverage was, however, affected by the decrease in frequency. It was also found that the impact of the dictionary is reduced, the translation method producing good results even in the absence of bilingual dictionaries (note that monolingual dictionaries are, however, at the heart of our extraction methodology). The average precision for the language pairs for which a bilingual dictionary was used is 92.9%; for the others, 84.5%. The difference in coverage is even lower (71.7% vs. 69.5%).

Language pair	Precision				Coverage				Dictionary
	All	31–50	16–30	1–15	All	31–50	16–30	1–15	
English-French	94.1	95.6	93.3	89.8	71.4	75.8	70.7	58.3	+
English-Italian	85.8	86.2	89.3	75.7	64.8	75.5	57.1	44.0	-
French-English	92.8	94.7	89.3	92.7	72.2	80.0	65.5	59.4	+
French-Italian	92.8	91.8	96.5	87.8	72.2	79.6	66.1	59.4	+
French-Spanish	90.9	92.0	90.9	85.7	75.0	81.5	70.8	60.9	+
Italian-English	82.4	87.6	75.2	74.1	63.6	72.9	58.3	41.5	-
Italian-French	94.1	97.0	88.9	93.1	67.8	79.2	60.0	44.6	+
Italian-Spanish	85.3	89.5	80.0	77.8	80.0	89.8	75.0	55.4	-
Average	89.8	91.8	87.9	84.6	70.9	79.3	65.4	53.0	

Table 6.2: Evaluation results (for the whole translation sets and for different frequency intervals).

Despite its simplicity, the collocation translation method described above produced state-of-the-art results in the task of identifying translation equivalents for

[17]The coverage was measured as the ratio of collocation pairs for which a translation was proposed.

collocations from parallel corpora, as will be seen from the comparison with the related work provided in the next section.

6.6 Related work

Tools that, like ours, enrich the translators' workbench by harnessing existing translations represent "a new generation of translation support tools" (Isabelle et al., 1993). Among the existing sentence-based bilingual concordancers that help the user identify past translations of a given piece of text, the ones that are most similar to our tool are TransSearch (Isabelle et al., 1993) and Termight (Dagan and Church, 1994).

In TransSearch (Isabelle et al., 1993), parallel documents are first aligned at sentence level, then indexed so that they can be searched for specific word forms or for lemmas occurring either in the source text, in the target text, or in both at once. Unlike our tool, TransSearch does not have a collocation extraction capability; still, it can be used for searching pairs of words co-occurring in a delimited text span.[18] Other differences with our system relate to the search mechanism: our tool performs a lemma-based search for items in a collocational relationship and identifies inflected forms and inversions by default; and to the alignment of texts: while in TransSearch complete alignments are computed beforehand, in our system the alignment is done on-the-fly when the user visualises a sentence, and only for that sentence.

Termight (Dagan and Church, 1994) is more similar to our tool in terms of functionalities, as it provides both a list of extracted terms and bilingual concordances for them. The system also proposes a list of potential translations for each term, identified using a word alignment algorithm. The difference with our system is, mainly, in the type of expressions dealt with: noun-phrase terms vs. collocations. The translation method is specific to this kind of rigid expressions, and consists of simply considering the sequence of words delimited by the alignments of the first and last word in an NP. The precision obtained on a test set of 192 English-German correspondences was 40% when the most frequent translation was considered, and 47% when the two most frequent translations were taken into account.

[18] In the current system version, the maximum allowed distance is 25 characters.

Other translation work based on parallel corpora has been carried out, for instance, in (Kupiec, 1993) and (van der Eijk, 1993). Like Termight, these methods identify noun phrase correspondences between two languages. In the first method (Kupiec, 1993), the mapping is done using expectation-maximization, an iterative re-estimation algorithm. The precision reported is 90% for the highest ranked 100 translations obtained from English-French data from the Hansard parallel corpus. The second method (van der Eijk, 1993) takes into account, just as our method does, the frequency of a term in the target mini-corpus, but also its global frequency in the whole target corpus and its expected position within the target sentence. Tested on 100 randomly chosen translations from Dutch into English, it obtained a precision of 68%.

Champollion (Smadja et al., 1996) is the first proper collocation translator. This system, built around Xtract (Smadja, 1993), deals with both rigid and flexible collocations in English, for which it detects a translation in the aligned French sentences from the Hansard corpus with the help of a statistical correlation metric, the Dice coefficient. The method requires an additional post-processing step in which the order of words in a flexible collocation is decided, given that no syntactic analysis is performed on the target side. The system has been evaluated by three annotators and showed a precision of 77% and 61%, respectively, on two different test sets of 300 collocations each (these collocations were randomly selected among the medium-frequency results).

We presume that the improvement obtained by our method over the methods cited above is mainly due to the quality of syntactic information that is available for both the source and target documents.

As for the sentence alignment method implemented for our tool, it is in principle similar to existing work that relies on the size correlation between the source and target text segments, expressed either in characters (Gale and Church, 1993) or in words (Brown et al., 1991a). Unlike methods like (Simard et al., 1992; Kay and Röscheisen, 1993; Chen, 1993), it does not rely on lexical information other than the paragraph numbering (if any). It is a much lighter-weight method that was built specifically for the purpose of easy integration within our tool, and its originality

stems from considering not only the ratio of source and target paragraphs, but also the relative sizes of paragraphs in their neighborhood. Thus, it has the advantage that the source-target proportions do not have to be empirically determined for new language pairs, as in (Gale and Church, 1993).

6.7 Summary

In many translation settings, like the one that motivated the implementation of our tool, translators are not allowed to produce a free translation for a given expression, but are asked to use a precise, "official" translation. Extracting key terms and typical expressions (collocations) from the official documents and consulting the translation archives in order to see how they were translated in the past is therefore a strategy that is likely to alleviate the burden of these translators.

The multilingual dictionaries that exploit parallel corpora constitute the new generation of translation support tools (Isabelle et al., 1993). The system described in this chapter has unique features that distinguishes it from other parallel concordancing tools such as TransSearch (Isabelle et al., 1993) and Termight (Dagan and Church, 1994): collocations are identified using a syntax-based methodology, and they can be searched, filtered, ordered, and saved into a database for further reference. The system requires no preprocessing of the source corpora, so new data is easy to add and analyse. The sentence alignment algorithm is fast and runs only when the user visualizes a specific collocation instance.

Section 6.5 of this chapter illustrated a successful application of the various components built for our tool (i.e., the collocation extractor, the filtering module, and the sentence alignment module) in the task of collocation translation.

Chapter 7

Conclusion

This chapter concludes the thesis by summarising the work undertaken and its major findings, by discussing its shortcomings, and by indicating directions for further development.

7.1 Achievements

The tasks of collocation extraction and syntactic parsing have received considerable attention from computational linguists during the past decades. Significant efforts have been made on both sides, but separately. The work described in this thesis adopts an integrated approach, in which the high level of performance achieved by current parsing technology is exploited in order to boost collocation extraction.

Originally treated as a purely statistical phenomenon that can be modeled with statistical methods quantifying the association strength, word collocation was gradually given a more linguistically-motivated understanding in the literature. More and more researchers posit the existence of a syntactic relationship between the items in a binary collocation, and therefore rely on linguistic analysis of source corpora before the statistical computation, in order to discover syntagmatic associations rather than associations of words from the same semantic field (e.g., *doctor-hospital*). But this analysis is rarely performed with syntactic parsers, despite the recent advances in

the parsing field, at least for certain languages. The existing extraction methodology is still generally confined to shallow levels of analysis based on the identification of (interrupted or uninterrupted) combinations of words of specific POS categories in a limited text span (the window method), or relies, at best, on shallow parsing usually implemented as pattern matching with regular expressions over POS labels. Languages with a richer morphology and a freer word order call, however, for more appropriate tools, because the window method and shallow parsing reach their limits in capturing candidate pairs from a whole range of syntactic constructions that frequently lead to word extraposition.

The extraction methods based on syntactic parsing or on syntactically annotated text—such as the ones designed for English (Lin, 1998; Lin, 1999; Orliac and Dillinger, 2003; Pearce, 2001a; Blaheta and Johnson, 2001), German (Zinsmeister and Heid, 2003; Schulte im Walde, 2003), Dutch (Villada Moirón, 2005), or Chinese (Wu and Zhou, 2003; Lü and Zhou, 2004)—remain isolated attempts suffering from various shortcomings, among which: low parsing precision, robustness, or coverage; the limited number of syntactic configurations dealt with; or difficult reproducibility on a different source corpus.

The main objective of our work was the design of a full-fledged extraction method based on syntactic parsing, capable of: i) covering a broad spectrum of collocational phenomena (from binary to longer combinations, from very frequent to infrequent ones, from a restricted to a large set of syntactic configuration); ii) taking into account their syntactic versatility (by handling, in particular, the syntactic constructions in which they occur due to the wide range of grammatical operations they may undergo: passivisation, relativisation, interrogation, clefting, apposition, etc.); and iii) at the same time, working on unrestricted text in multiple languages. The practical motivation at the beginning of this work was the implementation of a system for collocation extraction and concordance in parallel corpora, to be integrated in an actual translation environment, and that enables users to create, from their translation archives, a multilingual collocational database enriched with context samples, which can serve as a reference in future translations.

Some of the requirements stated above are satisfied by the capabilities of the Fips

parser our system uses, and therefore could easily be met: robustness, multilingual support, and broad grammatical coverage. The extraction module was developed around Fips, whose output is directly used for the identification of candidate pairs as the source corpus analysis proceeds. After this syntax-based step, our hybrid extraction method proceeds with the application of the statistical measure—log-likelihood ratios by default—, that produces a ranking of candidates according to their collocational strength. The proposed extraction methodology is not entirely novel, but the results it produces are highly accurate thanks to the syntactic analyses provided by Fips. In particular, the extraction of low-frequency, flexible pairs is facilitated by the grouping of instances under the same type, made possible by the sentence normalisation performed by Fips, which consists in considering the base form for words and their canonical order, irrespective of the surface realization. This result is of key importance, given the high proportion of low-frequency data in text corpora and the problems caused by data sparseness to the performance of association measures (AMs). The absence of a frequency filter on candidates and the application of AMs on syntactically homogeneous material are other distinguishing features of our method, besides the use of the syntactic proximity criterion instead of the linear proximity criterion in selecting candidate pairs.

The most original aspects of our extraction methodology are, first, the extension to collocations made up of more than two lexemes; second, the semi-automatic acquisition of language-dependent syntactic configurations for binary collocations; and, third, using Web data as an alternative to static corpora in order to cope with the problem of data sparseness and to mine for more collocates for a given word. These issues remained typically unaddressed in previous extraction work. The underlying parsing technology allowed us to propose efficient solutions to each of them, the key factors in our endeavour being the possibility to abstract away from particular text realizations, to reduce data sparseness by instance grouping, as well as to retrieve as many instances as possible for the syntactically versatile pairs.

The system developed has been extensively used for collocation extraction in a variety of settings, with the manually validated results being used to populate the lexicon of Fips and, as part of a cyclical process, to guide the parsing itself since

preference is given, in case of ambiguity, to the attachments involving collocations. Moreover, its components, including the sentence alignment method proposed for collocation concordancing in parallel text, have been successfully used in the task of corpus-based collocation translation.

A thorough evaluation of our syntax-based method was provided, its precision being compared against the precision achieved by the window method typically used in the literature, which is, at least in principle, syntactically uninformed. Also, a recall comparison against this method was partly performed by means of several case-study analyses of results at the instance level. In addition, a different case study compared the precision of our method against that of an extraction method based on shallow-parsing. (As far as the extraction extensions are concerned, their output was not evaluated, but, whenever possible, it was contrasted with examples provided in related work or with related resources such as the BBI dictionary).

The precision-based comparison against the window method was performed in two distinct experiments on fairly large corpora in different languages (English, French, Spanish, and Italian), by using manually-produced annotations of different granularity, and by investigating several levels in the output lists. The increase obtained with our method in the grammatical-, multi-word expression (MWE)-, and collocational precision is significant overall. It shows that, despite the inherent parsing errors and the challenges of processing large amounts of natural text, it is worth using a syntax-based approach. Still, it was found that for the high-frequency pairs at the top of the output lists (the first 50 pair types), the difference in MWE- and collocational precision is not always statistically significant; in one case, the window method outperformed our method, although not to a statistically significant extent. The recall comparison pointed out relative strengths and weaknesses of the two methods, and indicated, in particular, that the instance-based precision of the window method can sometimes be very low, because many of the instances it takes into account in ranking a pair type are false positives ("false instances"). Finally, the partial evaluation of the results of a shallow-parsing extractor suggested that its grammatical precision is lower than the precision of our method, and also that uninteresting pairs can be artificially promoted to high-ranked positions due to false positives.

These evaluation results have, first of all, a theoretical importance since they confirm that syntactic information leads to a substantial improvement in the quality of extraction results over the rudimentary techniques in use nowadays. From a practical perspective, the high reliability achieved by our method, in particular at the instance level and for the lower-frequency results situated beyond the very top of the output list, opens new possibilities for both lexicographic investigation and for NLP applications making use of collocational information. In contrast, the output of traditional methods, with the exception of the very top positions, contains a high amount of noise.

The qualitative analysis provided for the results obtained by the two contrasted methods showed that the difference in output pertains more to the content retrieved, and less to the relative order of items in the intersection: while the ranks of common items correlate significantly, the common content covers about 77% of the results of our method and only 17% of the window method. The syntactic filter applied by our method leads to a drastic reduction in the number of candidates by eliminating the noisy pairs, thus alleviating the burden on users that analyse the extraction results and making, at the same time, the statistical computation more tractable. Moreover, the syntactic information facilitates the interpretation of results, both by human users (who can display the pairs grouped according to the syntactic type), and by interested NLP applications (that require this information in order to be able to process the pairs).

As by-products of our work, we mention the suite of 3000 annotated pairs that could be exploited as training data in supervised learning, the syntactic patterns for multi-word collocations that were inferred from the experimental results and that can be used in future extraction, as well as the suggestions of parsing improvement derived from the error analysis of the extraction results (especially for Italian, a language that is currently less developed in Fips). This thesis also offers a concise general description of the AMs typically used in collocation extraction, including their explicit formulae and computation details; we believe that this description could be useful for the interested reader, since the existing descriptions in the literature are rather dispersed, sometimes over-technical, and they often leave out the explicit

computation steps. This thesis also provides a detailed review of the state of the art, organised by language and focussed on the level of text analysis performed.

Finally, the system developed (the main output of the research project at the origin of this work) was made available to the translators of our project partner institution, the WTO headquarters in Geneva. Plans are to make it available in the future to a larger community, probably in the form of a lighter-weight version accessible on-line.

7.2 Shortcomings and future work

The main limitation of the proposed methodology is the relatively limited availability of syntactic parsers at present times. The efficiency of the methods described is conditioned by the ability to perform a preliminary analysis of source texts with a syntactic tool whose robustness, speed, lexical and grammatical coverage are comparable to those of Fips. Nevertheless, the conclusion of our study stands (and was already confirmed for four different languages): whenever parsing is available, a hybrid collocation extraction approach integrating syntactic information is worth taking, as it leads to significantly better results. We argue that syntax-based methods should replace the rudimentary window technique in languages for which the parsing technology is already developed, or for which it will be available in the future.

A related issue is the portability of our extraction method to other parsers, given that some implementation choices were also determined by Fips (e.g., relying on its specific type of syntactic structures, or extracting candidates directly from these structures rather than from syntactically-annotated text). The adaptation to a different parser will certainly require some efforts, depending on how similar the output is to the output of Fips. Other parts of the methodology proposed are, however, easier to transfer (e.g., the extraction of multi-word collocations and the Web-based extraction).

Also, the portability across languages is an issue that has only been partly addressed by our work, as the extraction experiments run so far involved relatively similar languages and used practically the same set of syntactic patterns. Although we defined a strategy for customising the extraction procedure to a new language

(which consists in inducing collocationally relevant syntactic patterns from generic dependency relations and of assigning the most appropriate AM to each), it has not yet been applied on structurally different languages like German, which are now supported by Fips. The multilingual support is one the most imminent developments to be made to our extractor, since Fips is now opened to new languages (e.g., Greek, Romanian, Japanese, Russian and Serbian). Also related to this issue, it would be worth analysing the existing multilingual output—and, particularly, the pairs annotated in our evaluation studies—from a multilingual perspective.

The recall-based evaluation of our extractor is a topic that has not been sufficiently investigated in our present study. The analysis undertaken with the case studies should be carried out further, to gain better insights on what collocational phenomena fail to be captured with a syntactic vs. a non-syntactic approach. But this analysis is hampered by the absence of reference resources, and it requires specific evaluation techniques whose development is currently only at an incipient stage.

Researchers are generally skeptical about whether full parsing is really necessary for high-quality extraction, considering that shallow-parsing techniques may be sufficient. Our small evaluation experiment addressing this issue is not entirely conclusive, because it only looked at V-O and S-V combinations with a single noun. A more thorough evaluation is required in order to check if indeed, as observed, shallow parsing produces a less accurate pair type output and the false positive instances affect the ranking of low-frequency pairs. In any case, we expect recall to be affected more than precision. This evaluation was not a priority in this work, because high-quality shallow parsers capable of detecting predicate-argument relations approximate the behaviour of a parser, both requiring extensive implementation efforts.

Perhaps the broad interpretation of the collocation concept adopted in our work can also be seen as a limitation, as it covers any statistically significant combination of words in a syntactic relationship, including particular subtypes of multi-word expressions that linguists have strived for a long time to tell apart (e.g., support-verb constructions, compounds, idioms, phrasal verbs). It was beyond the scope of this work to propose a classification of these expressions or to distinguish collocations from them. This is an open research issue in the literature that we did not attempt

to solve; nor did we attempt to draw, more generally, a distinction between compositional and non-compositional expressions. Instead, we believe that our work can serve as a basis for future investigations of this kind, although we are aware of the fact that the borderline cases are numerous, and generalisations nearly impossible. What our annotation work indicated was that, despite a significant inter-rater agreement, there are numerous ambiguities between regular combinations and collocations (making up 39% of the total disagreement cases), fewer disagreements between collocations and compounds (11.2%), and only very few between collocations and idioms (less than 2%). In addition, it was found that idioms and compounds constitute only a small fraction of the output pairs marked as multi-word expressions (namely, 2.6% and 15.4%; i.e., 18% altogether), whereas collocations are prevalent among multi-word expressions (75.7%). These findings indicate the critical points to be tackled by future research, while confirming the magnitude of the collocational phenomenon.[1]

Seen from a wider perspective, our work represents only one step towards a better treatment of collocations in a computational framework. The extraction can be further improved by complementary techniques operating at the candidate selection or candidate ranking stage.

For instance, as discussed in Section 3.3.4, more sophisticated text analysis modules (such as anaphora resolution) might prove useful for detecting more collocation types and instances, in view of the fact that low frequency is characteristic of natural language data, which is typically zipfian. Lexical semantic resources (e.g., Wordnet, FrameNet, other thesauri and lexical ontologies) could provide the basis for extraction based on semantic criteria, as initiated by Pearce (2001a).

Novel AMs, more adequate to language data and taking into account the base-collocate dichotomy, are expected to enhance extraction as well, the research on asymmetric AMs being already started (Michelbacher et al., 2007). As an alternative to corpora and the Web, the examples provided by lexicographers in phraseological subentries of classical dictionaries constitute a rich source of collocational information

[1] According to Benson et al. (1986b), collocations are at the same time the most difficult to identify by lexicographers: "The critical problem for the lexicographer has been, heretofore, the treatment of collocations. It has been far more difficult to identify them than idioms or even compounds" (Benson et al., 1986b, 256).

that could be exploited for collocation extraction, and, in particular, for the inference of multi-word collocations.

But collocation extraction is not a goal in itself. Its output is already used in a variety of NLP applications as, mainly, disambiguation clues. Still, the integration of collocations in more challenging tasks dealing with language production, or their proper usage by language learners, cannot succeed without a more adequate contextual description, specifying exactly the allowed (and also, the preferred) degree of morpho-syntactic flexibility (e.g., number variation for nouns, presence or absence of determiners, modification potential, voice for verbs, etc). Such corpus-based analyses are the focus of current research (Tutin, 2004; Villada Moirón, 2005; Ritz, 2006), as are attempts to organise the output material with semantic criteria in order to facilitate lexicographic studies and the practical use in other applications (L'Homme, 2003; Wanner et al., 2006).

Appendix A

Collocation Dictionaries

1. BBI – The BBI Dictionary of English Word Combinations (Benson et al., 1986a)
2. COBUILD – Collins Cobuild English Dictionary (Sinclair, 1995)
3. DAFLES – Dictionnaire d'apprentissage du français langue étrangère ou seconde ('French vocabulary for learners of French as a foreign or second language') (Selva et al., 2002)
4. DC – Dictionnaire de cooccurrences (Beauchesne, 2001)
5. DEC – Dictionary of English Collocations (Kjellmer, 1994)
6. DiCo – Dictionnaire de Combinatoire ('Combinatorics Dictionary') (Polguère, 2000)
7. DOSC - Dictionary of Selected Collocations (Hill and Lewis, 1997)
8. ECD – Dictionnaire explicatif et combinatoire du français contemporain ('The Explanatory-Combinatorial Dictionary of Contemporary French') (Mel'čuk et al., 1984 1988 1992 1999)
9. LAF – Lexique actif du français ('Active lexicon of French') (Polguère, 2000)
10. LDOCE – Longman Dictionary of Contemporary English (Procter, 1987)

11. LLA – Longman Language Activator (Maingay and Tribble, 1993)
12. OCDSE - Oxford Collocations Dictionary for Students of English (Runcie, 2002)

Appendix B

Collocation Definitions

1. (Firth, 1957, 181):

   > Collocations of a given word are statements of the habitual and customary places of that word.

2. (Firth, 1968, 182):

   > Collocations are actual words in habitual company.

3. (Cowie, 1978, 132):

   > [...] the co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern.

4. (Hausmann, 1985)

   > typical, specific and characteristic combination of two words

5. (Cruse, 1986, 40)

   > The term *collocation* will be used to refer to sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent.

6. (Kjellmer, 1987, 133):

   > a sequence of words that occurs more than once in identical form [...] and which is grammatically well structured.

7. (Choueka, 1988):

   > a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components.

8. (Hausmann, 1989, 1010):

   > On appellera collocation la combinaison caractéristique de deux mots dans une des structures suivantes : a) substantif + adjectif (épithète); b) substantif + verbe; c) verbe + substantif (objet); d) verbe + adverbe; e) adjectif + adverbe; f) substantif + (prép.) + substantif."

9. (Benson, 1990):

   > A collocation is an arbitrary and recurrent word combination.

10. (Sinclair, 1991, 170):

    > Collocation is the cooccurrence of two or more words within a short space of each other in a text.

11. (Fontenelle, 1992, 222):

    > The term *collocation* refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure.

12. (Smadja, 1993, 143):

    > recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages.

13. (van der Wouden, 1997, 5):

    > Collocation: idiosyncratic restriction on the combinability of lexical items

14. (Manning and Schütze, 1999, 151):

    > A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things.

15. (McKeown and Radev, 2000, 507)

    > Collocations [...] cover word pairs and phrases that are commonly used in language, but for which no general syntactic and semantic rules apply.

16. (Sag et al., 2002, 7):

    > Institutionalized phrases are semantically and syntactically compositional, but statistically idiosyncratic. [...] We reserve the term *collocation* to refer to any statistically significant cooccurrence, including all forms of MWE [...] and compositional phrases.

17. (Evert, 2004, 9):

    > 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.

18. (Bartsch, 2004, 76)

    > lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other

Appendix C

Detailed AM Formulae

C.1 Chi-square

$$\begin{aligned}\chi^2 &= \sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i,j}\frac{O_{ij}^2}{E_{i,j}} - 2\sum_{i,j}O_{i,j} + \sum_{i,j}E_{i,j} \\ &= \sum_{i,j}\frac{O_{ij}^2}{E_{i,j}} - 2N + \frac{\sum_{i,j}R_iC_j}{N} = \sum_{i,j}\frac{O_{ij}^2}{E_{i,j}} - \frac{2N^2-\sum_{i,j}R_iC_j}{N} \qquad \text{(C.1)}\end{aligned}$$

Since $\sum_{i,j}R_iC_j = a^2+b^2+c^2+d^2+2ab+2ac+2ad+2bc+2bd+2cd = N^2$, by replacing it in C.1 we obtain:

$$\begin{aligned}\chi^2 &= \sum_{i,j}\frac{O_{ij}^2}{E_{i,j}} - N = \frac{Na^2}{(a+b)(a+c)} + \frac{Nb^2}{(a+b)(b+d)} + \frac{Nc^2}{(a+c)(c+d)} + \frac{Nd^2}{(c+d)(b+d)} - N \\ &= \frac{N(a^2(c+d)(b+d)+b^2(a+c)(c+d)+c^2(a+b)(b+d)+d^2(a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)} - \\ &\quad \frac{N(a+b)(a+c)(b+d)(c+d)}{(a+b)(a+c)(b+d)(c+d)} \\ &= \frac{N(a^2d^2+b^2c^2-2abcd)}{(a+b)(a+c)(b+d)(c+d)} = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}\end{aligned}$$

C.2 Log-likelihood ratios

$$
\begin{aligned}
LLR &= -2\log\frac{L(H_0)}{L(H_1)} = -2\log\frac{\mathrm{B}(a;a+b,p)\mathrm{B}(c;c+d,p)}{\mathrm{B}(a;a+b,p_1)\mathrm{B}(c;c+d;p_2)} \\
&= -2\log\frac{\mathrm{B}\left(a;a+b,\frac{a+c}{N}\right)\mathrm{B}\left(c;c+d,\frac{a+c}{N}\right)}{\mathrm{B}\left(a;a+b,\frac{a}{a+b}\right)\mathrm{B}\left(c;c+d;\frac{c}{c+d}\right)} \\
&= -2\log\frac{\binom{a+b}{a}\left(\frac{a+c}{N}\right)^a\left(1-\frac{a+c}{N}\right)^b\binom{c+d}{c}\left(\frac{a+c}{N}\right)^c\left(1-\frac{a+c}{N}\right)^d}{\binom{a+b}{a}\left(\frac{a}{a+b}\right)^a\left(1-\frac{a}{a+b}\right)^b\binom{c+d}{c}\left(\frac{c}{c+d}\right)^c\left(1-\frac{c}{c+d}\right)^d} \\
&= -2(a\log\frac{a+c}{N}+b\log\frac{b+d}{N}+c\log\frac{a+c}{N}+d\log\frac{b+d}{N} \\
&\quad -a\log\frac{a}{a+b}-b\log\frac{b}{a+b}-c\log\frac{c}{c+d}-d\log\frac{d}{c+d}) \\
&= -2(a\log(a+c)-a\log N+b\log(b+d)-b\log N+c\log(a+c) \\
&\quad -c\log N+d\log(b+d)-d\log N-a\log a+a\log(a+b) \\
&\quad -b\log b+b\log(a+b)-c\log c+c\log(c+d)-d\log d+d\log(c+d)) \\
&= -2((a+b)\log(a+b)+(a+c)\log(a+c) \\
&\quad +(b+d)\log(b+d)+(c+d)\log(c+d)-(a+b+c+d)\log N \\
&\quad -a\log a-b\log b-c\log c-d\log d) \\
&= 2(a\log a+b\log b+c\log c+d\log d \\
&\quad -(a+b)\log(a+b)-(a+c)\log(a+c) \\
&\quad -(b+d)\log(b+d)-(c+d)\log(c+d) \\
&\quad +(a+b+c+d)\log(a+b+c+d))
\end{aligned}
$$

Appendix D

Comparative Evaluation - Test Sets (Experiment 1)

Test Set 1 and Test Set 10 used in Experiment 1 (Section 4.4.3). Common items for the two methods are underlined. Ungrammatical items are marked with a star ($\star$).

D.1 Test Set 1

Parse-based method				**Window method**			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
premier ministre	A-N	4317.6	1047	Monsieur président	N-N	21138.1	2680
bloc québécois	N-A	3946.1	429	premier ministre	A-N	5571.0	1293
discours de trône	N-P-N	3894.0	426	madame présidente	N-N	5279.2	419
vérificateur général	N-A	3796.7	460	*Monsieur présider	N-V	3804.6	385
parti réformiste	N-A	3615.0	474	vérificateur général	N-A	3403.9	447
gouvernement fédéral	N-A	3461.9	860	bloc québécois	N-A	3124.3	407
missile de croisière	N-P-N	3147.4	323	parti réformiste	N-A	3083.0	462
Chambre de commune	N-P-N	3083.0	430	campagne électoral	N-A	2905.3	306
livre rouge	N-A	2536.9	215	livre rouge	N-A	2773.5	272
secrétaire parlementaire	N-A	2524.7	283	discours trône	N-N	2574.7	413
question adresser	S-V	2460.9	321	gouvernement fédéral	N-A	2395.2	896
opposition officiel	N-A	2294.2	217	milliard dollar	N-N	2364.8	483

Parse-based method				Window method			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
programme social	N-A	2165.7	394	missile croisière	N-N	2292.5	327
jouer rôle	V-O	1909.5	199	secrétaire parlementaire	N-A	2287.0	285
poser question	V-O	1877.1	282	question s'adresser	N-V	2086.2	313
milliard de dollar	N-P-N	1846.1	259	maintien paix	N-N	2054.4	435
créer emploi	V-O	1709.5	261	ressource humain	N-A	2053.9	318
développement de ressource	N-P-N	1626.4	200	million dollar	N-N	1954.6	460
prendre décision	V-O	1607.4	278	monsieur président	N-N	1909.7	282
création de emploi	N-P-N	1552.3	245	tenir compte	V-N	1869.2	282
défense national	N-A	1444.4	201	opposition officiel	N-A	1772.0	215
adresser à ministre	V-P-N	1397.1	238	programme social	N-A	1763.2	562
petit entreprise	A-N	1378.3	204	petit entreprise	A-N	1588.0	269
consentement unanime	N-A	1333.8	101	créer emploi	V-N	1528.6	301
million de dollar	N-P-N	1258.5	215	s'adresser ministre	V-N	1521.9	270
maintien de paix	N-P-N	1250.4	189	consentement unanime	N-A	1379.4	108
président suppléant	N-A	1216.4	115	poser question	V-N	1341.7	256
députer honorable	V-O	1194.0	118	Chambre commune	N-N	1313.4	406
remercier député	V-O	1186.9	200	création emploi	N-N	1175.3	296
rapport de vérificateur	N-P-N	1185.4	140	*petite entreprendre	N-V	1164.1	123
comité permanent	N-A	1184.0	157	soin santé	N-N	1125.9	225
taux de chômage	N-P-N	1180.4	102	logement social	N-A	1107.7	226
compte public	N-A	1170.3	143	essai missile	N-N	1103.2	200
régler problème	V-O	1091.6	161	défense national	N-A	1075.0	175
code criminel	N-A	1081.1	90	dernier année	A-N	1058.4	237
essai de missile	N-P-N	1073.9	158	président suppléant	N-A	1047.8	109
vote libre	N-A	1040.3	90	présidente suppléant	N-A	1027.6	88
chef de opposition	N-P-N	1032.8	145	sécurité social	N-A	1010.1	226
marché de travail	N-P-N	1018.3	120	jouer rôle	V-N	998.8	135
solliciteur général	N-A	985.2	128	personne âgé	N-A	997.9	128
ministre de Affaires	N-P-N	974.1	78	comité permanent	N-A	988.3	163

Parse-based method				Window method			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
étrangères							
Conseil de trésor	N-P-N	971.3	93	code criminel	N-A	963.7	88
député d'en face	N-A	970.3	109	secteur privé	N-A	936.7	146
prendre mesure	V-O	951.1	228	solliciteur général	N-A	936.6	124
dernier année	A-N	950.0	233	*essai croisière	N-N	928.9	168
développement régional	N-A	936.3	127	certain nombre	A-N	925.4	157
prendre parole	V-O	931.6	182	vote libre	N-A	925.2	98
gouvernement précédent	N-A	930.0	196	parti libéral	N-A	905.4	193
croissance économique	N-A	908.9	142	*développement humain	N-A	884.5	179
mesure législatif	N-A	892.3	141	prendre décision	V-N	851.2	205

D.2 Test Set 10

Parse-based method				Window method			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
autre ministre	A-N	146.6	9	force susceptible	N-A	193.7	29
chômage élevé	N-A	146.5	18	avoir raison	V-N	193.7	114
député de opposition	N-P-N	145.9	64	*programme président	N-N	193.7	4
redonner à Canadien	V-P-N	145.9	29	ministre Canada	N-N	193.5	108
revenu moyen	N-A	145.4	23	*fleuve Saint-Laurent	N-N	193.2	18
présence de casque	N-P-N	144.9	17	*nouveau démocratique	N-A	193.1	29
processus de consultation	N-P-N	144.6	23	Chambre ajourner	N-V	192.8	28
bureau de régie	N-P-N	144.5	13	trouver moyen	V-N	192.7	49
modifier constitution	V-O	144.2	20	entrer vigueur	V-N	192.4	21
jeu de tueur	N-P-N	144.2	15	avoir honneur	V-N	192.2	81
valoir peine	V-O	143.7	14	autre pays	A-N	191.6	131
*président de élection	N-P-N	143.5	25	*femme enfant	N-N	190.7	58

Parse-based method				Window method			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
ministre de commerce	N-P-N	142.9	37	Conseil privé	N-A	190.7	33
payer part	V-O	141.9	25	*Québec Chambre	N-N	190.4	4
système de guidage	N-P-N	141.3	9	vaste consultation	A-N	190.0	27
problème grave	N-A	141.3	23	être fois	V-N	189.5	45
vote de confiance	N-P-N	141.1	14	bref délai	A-N	189.3	18
manutention de grain	N-P-N	140.9	14	dette déficit	N-N	188.9	55
emploi à long terme	N-A	140.8	24	ministère finance	N-N	188.7	59
viser objectif	V-O	140.5	25	ministre pays	N-N	188.7	19
bout de ligne	N-P-N	139.9	14	M. Gaston	N-N	188.7	26
poser question	V-O	139.7	20	*emploi Chambre	N-N	188.6	1
amendement constitutionnel	N-A	139.4	16	accord nord-américain	N-A	188.2	23
dizaine de millier	N-P-N	139.3	17	réduction déficit	N-N	188.2	53
soutenir concurrence	V-O	139.3	13	exprimer point de vue	V-N	188.1	31
entrer en vigueur	V-P-N	139.1	16	soldat canadien	N-A	187.0	55
déposer projet de loi	V-O	139.1	27	*chambrer séance	V-N	186.6	20
petit entrepreneur	A-N	138.8	23	justice pénal	N-A	186.1	20
entendre député	V-O	138.6	33	*député député	N-N	185.9	96
aide à rénovation	N-P-N	138.4	16	service jeunesse	N-N	185.9	45
économie canadien	N-A	138.1	59	comité examiner	N-V	185.8	48
force de dissuasion	N-P-N	138.0	9	reprendre étude	V-N	185.7	31
secteur de activité	N-P-N	137.6	13	fin froid	N-A	185.7	24
dire mot	V-O	137.3	21	*étape deuxième	N-A	185.3	22
fonds de investissement	N-P-N	137.0	20	*Canada Canadien	N-N	185.1	31
poser question	V-O	136.7	23	bureau poste	N-N	185.1	40
débattre question	V-O	136.4	28	dépense gouvernemental	N-A	185.1	70
former comité	V-O	136.3	23	développement gouvernement	N-N	184.7	12
adresser à ministre	V-P-N	136.3	18	assistance social	N-A	184.0	46
poser geste	V-O	135.7	19	*région gouvernement	N-N	183.7	5
commettre crime	V-O	135.6	11	tenir promesse	V-N	183.7	34

Parse-based method				**Window method**			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
confier mandat	V-O	135.6	15	*Québec ministre	N-N	183.1	16
mise en chantier	N-P-N	135.5	9	frai partagé	N-A	182.9	17
acquisition de arme à feu	N-P-N	135.4	13	faire part	V-N	182.8	50
gagner vie	V-O	135.2	19	représentant élu	N-A	182.6	21
retombée économique	N-A	135.1	24	*gouvernement travail	N-N	182.5	22
assurer sécurité	V-O	134.9	19	*dollar ministre	N-N	182.5	3
rester à feuilleton	V-P-N	134.9	13	réforme social	N-A	182.5	72
aborder sujet	V-O	134.2	18	revenu garanti	N-A	180.5	25
service de police	N-P-N	134.2	10	leader solliciteur	N-N	180.5	31

Appendix E

Comparative Evaluation - Annotations (Experiment 1)

Annotations for Test Set 1 and Test Set 10 used in Experiment 1 (Section 4.4.3). Common items for the two methods are underlined. Stars (*) mark ungrammatical items; diamonds (◇) – true positives (MWEs), and dashes (-) – complete disagreements. Unmarked items are regular combinations.

E.1 Annotations for Test Set 1

Parse-based method			Window method		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
◇	premier ministre	1-1-1	◇	Monsieur président	2-1-1
◇	bloc québécois	1-1-1	◇	premier ministre	1-1-1
◇	discours de trône	1-1-1	◇	madame présidente	2-1-1
◇	vérificateur général	1-1-2	*	Monsieur présider	2-0-0
◇	parti réformiste	1-1-1	◇	vérificateur général	1-1-1
	gouvernement fédéral	2-2-1	◇	bloc québécois	1-1-1
◇	missile de croisière	1-1-1	◇	parti réformiste	1-1-1
◇	Chambre de commune	1-1-1	◇	campagne électoral	1-1-1
◇	livre rouge	1-1-2	◇	livre rouge	1-1-1
◇	secrétaire parlementaire	1-1-1	◇	discours trône	1-2-1

Parse-based method			Window method		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
◇	question adresser	1-1-1	◇	gouvernement fédéral	1-1-1
	opposition officiel	1-2-2		milliard dollar	2-1-2
◇	programme social	1-1-1	◇	missile croisière	1-1-1
◇	jouer rôle	1-1-1	◇	secrétaire parlementaire	1-2-1
◇	poser question	1-1-1	◇	question s'adresser	1-1-1
	milliard de dollar	2-2-2	◇	maintien paix	1-1-1
◇	créer emploi	1-1-1	◇	ressource humain	1-1-1
◇	développement de ressource	1-1-2		million dollar	2-2-2
◇	prendre décision	1-1-1	◇	monsieur président	2-1-1
◇	création de emploi	1-1-1	◇	tenir compte	1-1-1
◇	défense national	1-1-1		opposition officiel	2-2-1
	adresser à ministre	2-1-2	◇	programme social	1-2-1
◇	petit entreprise	1-1-2	◇	petit entreprise	1-1-1
◇	consentement unanime	1-1-1	◇	créer emploi	1-2-1
	million de dollar	2-2-2		s'adresser ministre	2-2-2
◇	maintien de paix	1-1-1	◇	consentement unanime	1-2-1
◇	président suppléant	2-1-1	◇	poser question	1-1-1
-	députer honorable	0-1-2	◇	Chambre commune	1-1-1
	remercier député	2-1-2	◇	création emploi	1-2-1
	rapport de vérificateur	2-2-2	*	petite entreprendre	0-2-0
◇	comité permanent	1-1-1	◇	soin santé	1-1-1
◇	taux de chômage	1-1-1	◇	logement social	1-1-1
◇	compte public	1-1-1	◇	essai missile	1-1-1
◇	régler problème	1-1-1	◇	défense national	1-1-1
◇	code criminel	1-1-1		dernier année	2-2-1
	essai de missile	2-2-2		président suppléant	2-2-1
	vote libre	1-2-2		présidente suppléant	2-2-1
◇	chef de opposition	1-1-1	◇	sécurité social	1-1-1
◇	marché de travail	1-1-1	◇	jouer rôle	1-1-1
◇	solliciteur général	1-1-2	◇	personne âgé	1-1-1
◇	ministre de Affaires étrangères	1-1-1	◇	comité permanent	2-1-1

Parse-based method			Window method		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
◇	Conseil de trésor	1-1-1	◇	code criminel	1-1-1
	député d'en face	2-1-2	◇	secteur privé	1-1-1
◇	prendre mesure	1-1-1	◇	solliciteur général	1-2-1
	dernier année	2-1-2	*	essai croisière	0-0-0
	développement régional	1-2-2		certain nombre	2-2-1
◇	prendre parole	1-1-1	◇	vote libre	1-1-1
	gouvernement précédent	2-1-2	◇	parti libéral	1-2-1
◇	croissance économique	1-1-1	*	développement humain	0-0-1
◇	mesure législatif	1-1-1	◇	prendre décision	1-1-1

E.2 Annotations for Test Set 10

Parse-based method			Window method		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
	autre ministre	2-2-2		force susceptible	2-2-1
	chômage élevé	2-2-0	◇	avoir raison	1-1-1
◇	député de opposition	2-1-1	*	programme président	0-0-2
	redonner à Canadien	2-2-0		ministre Canada	2-0-2
◇	revenu moyen	1-1-1	*	fleuve Saint-Laurent	0-0-2
	présence de casque	2-2-2	*	nouveau démocratique	0-1-0
	processus de consultation	2-2-1	◇	Chambre ajourner	1-2-1
◇	bureau de régie	1-1-2	◇	trouver moyen	1-1-1
	modifier constitution	2-2-1	◇	entrer vigueur	1-1-1
	jeu de tueur	2-2-2	◇	avoir honneur	1-1-1
◇	valoir peine	1-1-1		autre pays	2-2-2
*	président de élection	1-0-0	*	femme enfant	2-0-0
◇	ministre de commerce	1-1-1	◇	Conseil privé	1-2-1
	payer part	1-2-2	*	Québec Chambre	0-0-0
◇	système de guidage	1-1-1	◇	vaste consultation	1-2-1

Parse-based method			**Window method**		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
◇	problème grave	1-2-1		être fois	2-2-0
◇	vote de confiance	1-1-1	◇	bref délai	1-1-1
	manutention de grain	1-2-2	-	dette déficit	2-0-1
◇	emploi à long terme	1-1-1		ministère finance	1-2-2
◇	viser objectif	1-1-1		ministre pays	2-0-2
◇	bout de ligne	1-1-1		M. Gaston	2-1-2
◇	poser question	1-1-1	*	emploi Chambre	0-0-0
◇	amendement constitutionnel	1-2-1		accord nord-américain	1-2-2
	dizaine de millier	2-2-2	◇	réduction déficit	1-2-1
	soutenir concurrence	2-1-2	◇	exprimer point de vue	1-1-1
◇	entrer en vigueur	1-1-1		soldat canadien	2-2-2
◇	déposer projet de loi	1-1-1	*	chambrer séance	0-0-0
◇	petit entrepreneur	1-1-2	◇	justice pénal	1-1-1
	entendre député	2-2-2	*	député député	0-0-0
	aide à rénovation	2-2-2	◇	service jeunesse	1-2-1
	économie canadien	2-2-2		comité examiner	2-2-0
◇	force de dissuasion	1-1-1	◇	reprendre étude	1-2-1
◇	secteur de activité	1-1-1		fin froid	0-2-2
◇	dire mot	1-1-2	*	étape deuxième	0-2-0
◇	fonds de investissement	1-1-1	*	Canada Canadien	0-0-2
◇	poser question	1-1-1	◇	bureau poste	1-1-1
◇	débattre question	1-1-2	◇	dépense gouvernemental	1-2-1
◇	former comité	1-2-1	-	développement gouvernement	1-0-2
	adresser à ministre	2-2-2	◇	assistance social	1-1-1
◇	poser geste	1-1-1	*	région gouvernement	0-0-2
◇	commettre crime	1-1-1	◇	tenir promesse	1-1-1
◇	confier mandat	1-1-1	*	Québec ministre	0-0-0
◇	mise en chantier	1-1-1	◇	frai partagé	1-2-1
	acquisition de arme à feu	2-2-2	◇	faire part	1-2-1
◇	gagner vie	1-1-1		représentant élu	2-2-1
◇	retombée économique	1-1-1	*	gouvernement travail	0-0-2

Parse-based method			Window method		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
◇	assurer sécurité	1-1-1	*	dollar ministre	0-0-0
	rester à feuilleton	2-2-2	◇	réforme social	1-2-1
◇	aborder sujet	1-1-1	◇	revenu garanti	1-2-1
◇	service de police	1-1-1		leader solliciteur	2-1-2

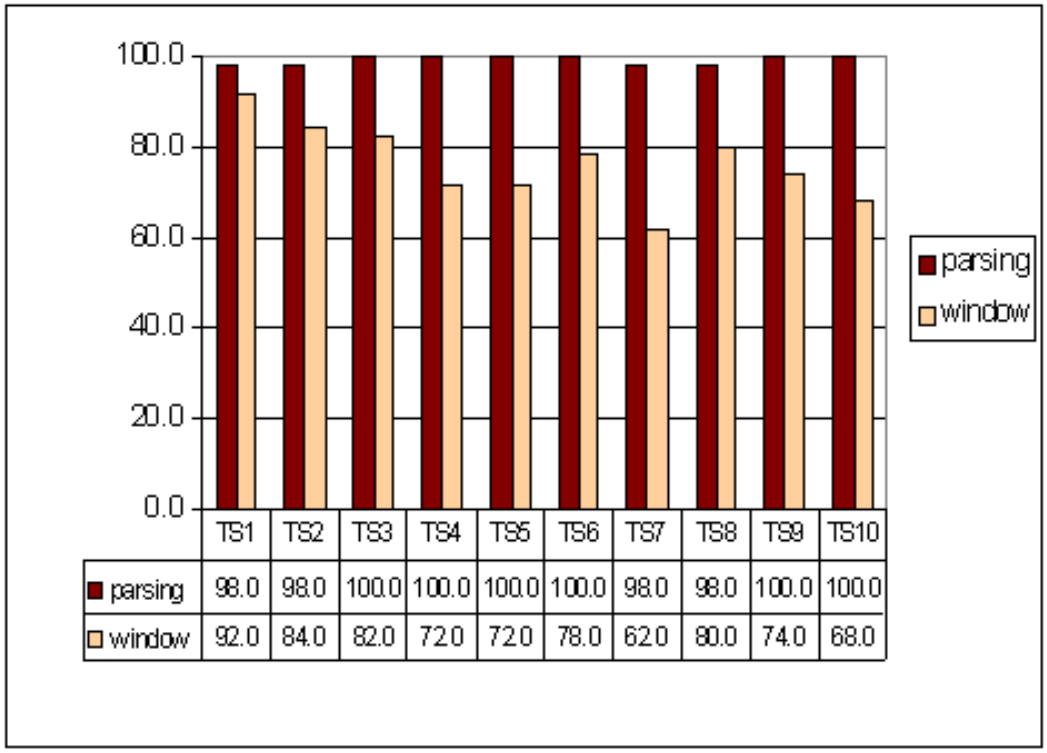
Appendix F

Comparative Evaluation - Result Charts (Experiment 1)

Graphical display of the results obtained in Experiment 1 (Section 4.4.3).

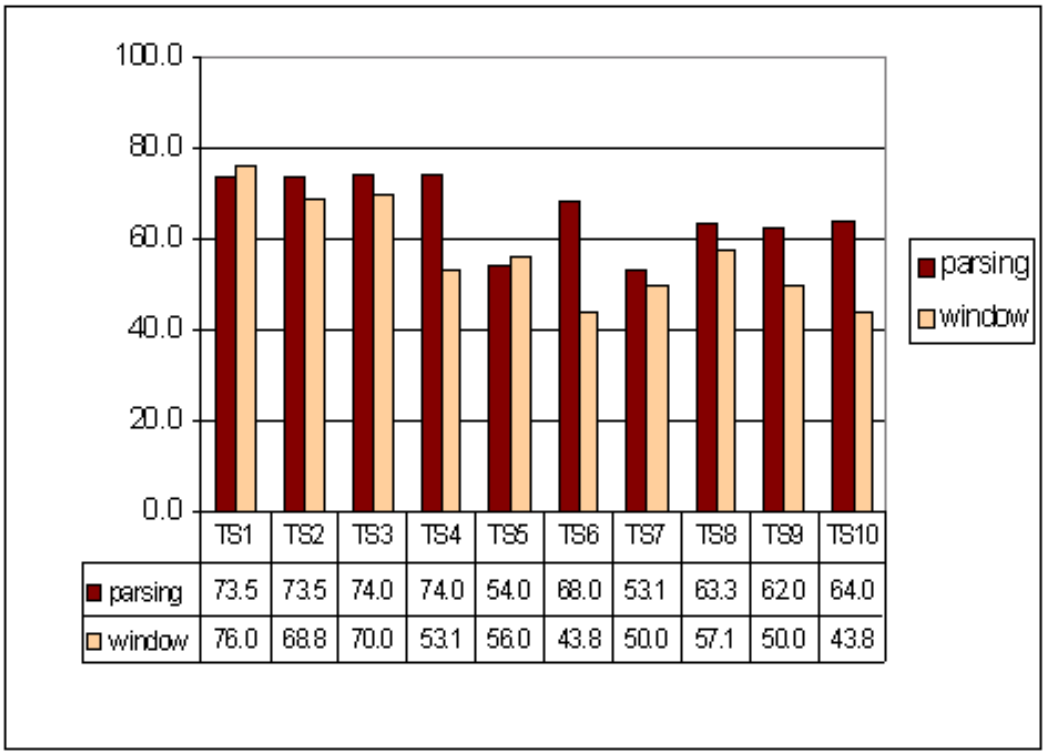
F.1 Grammatical precision

F.2 MWE precision



	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10
parsing	98.0	98.0	100.0	100.0	100.0	100.0	98.0	98.0	100.0	100.0
window	92.0	84.0	82.0	72.0	72.0	78.0	62.0	80.0	74.0	68.0

Figure F.1: Comparative evaluation results for Experiment 1: Grammatical precision.



	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10
parsing	73.5	73.5	74.0	74.0	54.0	68.0	53.1	63.3	62.0	64.0
window	76.0	68.8	70.0	53.1	56.0	43.8	50.0	57.1	50.0	43.8

Figure F.2: Comparative evaluation results for Experiment 1: MWE precision.

Appendix G

Comparative Evaluation - Test Sets (Experiment 2)

Test Set 1 and Test Set 2 used in Experiment 2 (Section 4.4.4) for English data. Common items for the two methods are underlined. Ungrammatical items are marked with a star ($\star$).

G.1 English - Test Set 1

Parse-based method				**Window method**			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
take place	V-O	5254.0	852	lady gentleman	N-N	25176.7	2036
next item	A-N	4887.1	530	Mr. president	N-N	18141.5	4217
amendment no	N-N	4754.2	718	Mr. President	N-N	13939.8	2845
same time	A-N	4511.0	803	madam president	N-N	9466.7	1408
honourable member	A-N	3805.4	379	to take place	V-N	4794.4	976
Swedish presidency	A-N	3413.0	427	same time	A-N	4622.2	868
close debate	V-O	3230.8	411	next item	A-N	4149.1	533
Belgian presidency	A-N	3090.8	356	honourable member	A-N	4123.4	502
candidate country	N-N	3012.0	398	debate to close	N-V	3624.0	408
play role	V-O	3005.4	340	Swedish presidency	A-N	3578.2	453
internal market	A-N	2949.6	373	to take account	V-N	3348.5	647

Parse-based method				**Window method**			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
adopt resolution	V-O	2945.7	422	amendment no	N-N	3300.4	877
sustainable development	A-N	2875.5	338	sustainable development	A-N	3219.8	474
court of auditor	N-P-N	2844.5	264	Belgian presidency	A-N	3182.5	374
item be	S-V	2640.6	518	internal market	A-N	2997.2	476
draw attention	V-O	2528.8	225	medium to size	N-V	2994.4	233
madam president	N-N	2441.8	238	white paper	A-N	2898.8	275
vote take	S-V	2181.3	325	*member to state	N-V	2873.9	534
rule of procedure	N-P-N	2098.9	293	item to be	N-V	2775.5	526
united nation	A-N	2000.6	193	*commission commission	N-N	2716.9	78
common position	A-N	1966.0	317	to play role	V-N	2628.5	333
rule of law	N-P-N	1955.5	209	to adopt resolution	V-N	2354.3	431
civil society	A-N	1943.8	205	*human to right	N-V	2319.9	231
European council	A-N	1925.8	456	civil society	A-N	2270.7	305
court of justice	N-P-N	1860.5	180	*member to state	N-V	2218.9	372
question no	N-N	1820.7	324	to draw attention	V-N	2190.4	241
European commission	A-N	1791.2	365	vote to take	N-V	2188.1	408
great deal	A-N	1761.6	229	*country commission	N-N	2166.3	19
commission proposal	N-N	1676.5	311	common position	A-N	2144.8	416
apply to	V-P	1637.6	592	court auditor	N-N	2095.9	314
legislative resolution	A-N	1623.2	186	united nation	A-N	2002.6	218
table amendment	V-O	1538.9	223	madam President	N-N	1950.6	409
free movement	A-N	1506.7	142	amendment to table	N-V	1916.7	321
point of order	N-P-N	1499.9	135	developing country	A-N	1913.0	339
right direction	A-N	1485.8	159	*commission country	N-N	1844.3	66
Kyoto protocol	N-N	1478.5	156	to bear mind	V-N	1832.5	190
foot-and-mouth disease	N-N	1466.6	145	parliament to adopt	N-V	1819.9	502
take into	V-P	1451.1	496	motion resolution	N-N	1819.1	350
solve problem	V-O	1406.9	182	to traffic human being	V-N	1785.8	164
legal basis	A-N	1402.7	222	third country	A-N	1727.9	417
agree with	V-P	1399.7	392	next year	A-N	1710.9	417
cut speaker	V-O	1378.5	97	legal basis	A-N	1699.3	318

Parse-based method				Window method			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
enter into	V-P	1365.4	235	common policy	A-N	1634.7	551
green paper	A-N	1360.7	106	candidate country	N-N	1578.6	639
*like president	V-O	1335.3	173	great deal	A-N	1577.6	242
suspend sitting	V-O	1317.7	107	small medium	A-N	1573.2	182
take account	V-O	1315.6	242	mutual recognition	A-N	1562.1	149
crime organize	S-V	1304.7	116	intergovernmental conference	A-N	1557.8	164
oral amendment	A-N	1232.6	136	to accept amendment	V-N	1550.6	334
reach agreement	V-O	1226.1	191	illegal immigration	A-N	1537.9	178

G.2 English - Test Set 2

Parse-based method				Window method			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
cancellation of debt	N-P-N	62.1	5	union action	N-N	73.1	15
listen to voice	V-P-N	62.1	7	to begin work	V-N	73.1	30
definitive solution	A-N	62.1	10	*point service	N-N	73.1	2
country national	N-N	62.1	27	service point	N-N	73.1	2
wheel vehicle	N-N	62.0	5	*service matter	N-N	73.1	4
European research	A-N	62.0	42	*budget decision	N-N	73.1	12
lay rule	V-O	62.0	14	*measure house	N-N	73.1	5
seal of approval	N-P-N	62.0	5	confidence Mr.	N-N	73.1	1
number of death	N-P-N	62.0	22	to like thank you	V-N	73.1	17
open door	V-O	61.9	8	*report fight	N-N	73.1	11
fulfil mandate	V-O	61.9	9	*change debate	N-N	73.0	1
achieve level	V-O	61.9	22	animal infected	N-A	73.0	9
particular concern	A-N	61.9	17	*environment regulation	N-N	73.0	1
mine operation	V-O	61.9	6	*commission opposition	N-N	73.0	1

Parse-based method				**Window method**			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
farming community	N-N	61.9	17	*year requirement	N-N	73.0	3
criminal act	A-N	61.9	9	*member competition	N-N	73.0	3
increase premium	V-O	61.9	11	future action	N-N	73.0	2
renewal of agreement	N-P-N	61.8	11	*European Union dialogue	N-N	73.0	10
key player	A-N	61.8	11	second phase	A-N	73.0	17
face dilemma	V-O	61.8	7	convention to ratify	N-V	73.0	16
maintain position	V-O	61.8	22	to want go	V-N	73.0	14
deserve recognition	V-O	61.7	9	community law	N-N	73.0	202
have chance	V-O	61.6	34	medium term	N-N	73.0	40
face difficulty	V-O	61.6	14	*position parliament	N-N	73.0	54
sugar regime	N-N	61.6	9	report nation	N-N	73.0	2
chair by	V-P	61.6	17	stability country	N-N	73.0	16
partnership agreement	N-N	61.6	17	energy issue	N-N	73.0	9
research fund	N-N	61.6	17	*European Parliament protection	N-N	73.0	9
dramatic event	A-N	61.5	9	*European Union consultation	N-N	73.0	2
entail cost	V-O	61.5	9	economic stability	A-N	73.0	34
contain in	V-P	61.5	148	valentine day	N-N	73.0	9
be debate	V-O	61.5	258	*strategy directive	N-N	73.0	5
scientific assessment	A-N	61.5	11	to raise point	V-N	73.0	39
set priority	V-O	61.4	16	*presidency people	N-N	72.9	1
Coptic Christian	N-N	61.4	4	*mechanism Mr.	N-N	72.9	2
know as	V-P	61.4	16	*reform number	N-N	72.9	1
balanced participation	A-N	61.4	8	*transport union	N-N	72.9	2
combat exclusion	V-O	61.4	8	*research measure	N-N	72.9	5
Finnish presidency	A-N	61.4	10	*take European	N-N	72.9	2
switch to euro	V-P-N	61.4	6	*change community	N-N	72.9	8
affected area	A-N	61.4	9	*council limit	N-N	72.9	3
present system	A-N	61.4	25	*limit council	N-N	72.9	3

Parse-based method				Window method			
Key1 + Prep + Key2	**Type**	**LLR**	**F**	**Key1 + Key2**	**Type**	**LLR**	**F**
range of measure	N-P-N	61.4	17	to exceed limit	V-N	72.9	11
thorough analysis	A-N	61.3	9	*government interest	N-N	72.9	8
put emphasis	V-O	61.3	13	*amendment fund	N-N	72.9	4
main priority	A-N	61.3	16	*resolution committee	N-N	72.9	28
⋆gross domestic	A-N	61.3	4	clean energy	A-N	72.9	12
lose credibility	V-O	61.3	9	*policy business	N-N	72.9	9
defend against	V-P	61.3	17	other amendment	A-N	72.9	71
hit nail	V-O	61.3	4	*problem public	N-N	72.9	3

Appendix H

Comparative Evaluation - Annotations (Experiment 2)

Annotations for Test Set 1 and Test Set 2 used in Experiment 2 (Section 4.4.4) for English data. Common items for the two methods are underlined. The following marks are used: stars (*) – ungrammatical items; diamonds (◇) – true positives (collocations); NE – named entities; CP – compounds; ID – idioms; dashes (-) – complete disagreements. Unmarked items are regular combinations.

H.1 English - Annotations for Test Set 1

Parse-based method			Window method		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
◇	take place	3 – 3		lady gentleman	1 – 1
	next item	1 – 1		Mr. president	1 – 1
	amendment no	1 – 1		Mr. President	1 – 1
CP	same time	4 – 4		madam president	1 – 1
-	honourable member	1 – 3	◇	to take place	3 – 3
	Swedish presidency	1 – 1	CP	same time	4 – 4
◇	close debate	3 – 3		next item	1 – 1
	Belgian presidency	1 – 1	-	honourable member	1 – 3
	candidate country	1 – 1	◇	debate to close	3 – 3

Parse-based method			Window method		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
◇	play role	3 – 3		Swedish presidency	1 – 1
◇	internal market	3 – 3	◇	to take account	3 – 3
◇	adopt resolution	3 – 3		amendment no	1 – 1
◇	sustainable development	3 – 3	◇	sustainable development	3 – 3
NE	court of auditor	2 – 2		Belgian presidency	1 – 1
	item be	1 – 1	◇	internal market	3 – 3
◇	draw attention	3 – 3	CP	medium to size	4 – 4
	madam president	1 – 1	CP	white paper	4 – 4
	vote take	1 – 1	*	member to state	0 – 0
NE	rule of procedure	2 – 2		item to be	1 – 1
NE	united nation	2 – 2	*	commission commission	0 – 0
	common position	1 – 1	◇	to play role	3 – 3
-	rule of law	3 – 4	◇	to adopt resolution	3 – 3
-	civil society	4 – 1	*	human to right	0 – 0
NE	European council	2 – 2	-	civil society	4 – 1
NE	court of justice	2 – 2	*	member to state	0 – 0
	question no	1 – 1	◇	to draw attention	3 – 3
NE	European commission	2 – 2	◇	vote to take	3 – 3
CP	great deal	4 – 4	*	country commission	0 – 0
	commission proposal	1 – 1		common position	1 – 1
◇	apply to	3 – 3	NE	court auditor	2 – 2
	legislative resolution	1 – 1	NE	united nation	2 – 2
◇	table amendment	3 – 3	-	madam President	1 – 2
◇	free movement	3 – 3	◇	amendment to table	3 – 3
CP	point of order	4 – 4	-	developing country	3 – 4
◇	right direction	3 – 3	*	commission country	0 – 0
NE	Kyoto protocol	2 – 2	◇	to bear mind	3 – 3
-	foot-and-mouth disease	3 – 4		parliament to adopt	1 – 1
-	take into	1 – 3	-	motion resolution	3 – 0
◇	solve problem	3 – 3	◇	to traffic human being	3 – 3
◇	legal basis	3 – 3	-	third country	4 – 3

Parse-based method			Window method		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
◇	agree with	3 – 3		next year	1 – 1
-	cut speaker	5 – 3	◇	legal basis	3 – 3
◇	enter into	3 – 3		common policy	1 – 1
NE	green paper	2 – 2		candidate country	1 – 1
*	like president	0 – 0	CP	great deal	4 – 4
◇	suspend sitting	3 – 3	CP	small medium	4 – 4
◇	take account	3 – 3		mutual recognition	1 – 1
-	crime organize	3 – 0		intergovernmental conference	1 – 1
	oral amendment	1 – 1	◇	to accept amendment	3 – 3
◇	reach agreement	3 – 3		illegal immigration	1 – 1

H.2 English - Annotations for Test Set 2

Parse-based method			Window method		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
◇	cancellation of debt	3 – 3	-	union action	0 – 1
◇	listen to voice	3 – 3		to begin work	1 – 1
-	definitive solution	1 – 3	*	point service	0 – 0
-	country national	4 – 3		service point	1 – 1
-	wheel vehicle	3 – 4	*	service matter	0 – 0
-	European research	2 – 1	*	budget decision	0 – 0
◇	lay rule	3 – 3	*	measure house	0 – 0
◇	seal of approval	3 – 3		confidence Mr.	1 – 1
	number of death	1 – 1	-	to like thank you	0 – 1
-	open door	3 – 5	*	report fight	0 – 0
◇	fulfil mandate	3 – 3	*	change debate	0 – 0
◇	achieve level	3 – 3		animal infected	1 – 1
◇	particular concern	3 – 3	*	environment regulation	0 – 0

Parse-based method			Window method		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
-	mine operation	1 – 0	*	commission opposition	0 – 0
	farming community	1 – 1	*	year requirement	0 – 0
◇	criminal act	3 – 3	*	member competition	0 – 0
-	increase premium	1 – 3		future action	1 – 1
◇	renewal of agreement	3 – 3	*	European Union dialogue	0 – 0
-	key player	4 – 3		second phase	1 – 1
◇	face dilemma	3 – 3	◇	convention to ratify	3 – 3
◇	maintain position	3 – 3	-	to want go	1 – 0
◇	deserve recognition	3 – 3	CP	community law	4 – 4
◇	have chance	3 – 3	-	medium term	3 – 4
◇	face difficulty	3 – 3	*	position parliament	0 – 0
	sugar regime	1 – 1		report nation	1 – 1
	chair by	1 – 1	◇	stability country	3 – 3
-	partnership agreement	1 – 3		energy issue	1 – 1
◇	research fund	3 – 3	*	European Parliament protection	0 – 0
◇	dramatic event	3 – 3	*	European Union consultation	0 – 0
◇	entail cost	3 – 3	-	economic stability	3 – 1
-	contain in	3 – 1	NE	valentine day	2 – 2
	be debate	1 – 1	*	strategy directive	0 – 0
-	scientific assessment	1 – 3	◇	to raise point	3 – 3
◇	set priority	3 – 3	*	presidency people	0 – 0
NE	Coptic Christian	2 – 2	*	mechanism Mr.	0 – 0
◇	know as	3 – 3	*	reform number	0 – 0
-	balanced participation	1 – 3	*	transport union	0 – 0
-	combat exclusion	1 – 3	*	research measure	0 – 0
	Finnish presidency	1 – 1	*	take European	0 – 0
	switch to euro	1 – 1	*	change community	0 – 0
-	affected area	1 – 3	*	limit council	0 – 0
-	present system	1 – 3	*	council limit	0 – 0
-	range of measure	1 – 3	◇	to exceed limit	3 – 3
◇	thorough analysis	3 – 3	*	government interest	0 – 0

Parse-based method			**Window method**		
Mark	**Key1 + Prep + Key2**	**Annot**	**Mark**	**Key1 + Prep + Key2**	**Annot**
◇	put emphasis	3 – 3	*	amendment fund	0 – 0
-	main priority	1 – 3	*	resolution committee	0 – 0
*	gross domestic	0 – 0	◇	clean energy	3 – 3
◇	lose credibility	3 – 3	*	policy business	0 – 0
◇	defend against	3 – 3		other amendment	1 – 1
ID	hit nail	5 – 5	*	problem public	0 – 0

Appendix I

Comparative Evaluation - Result Charts (Experiment 2)

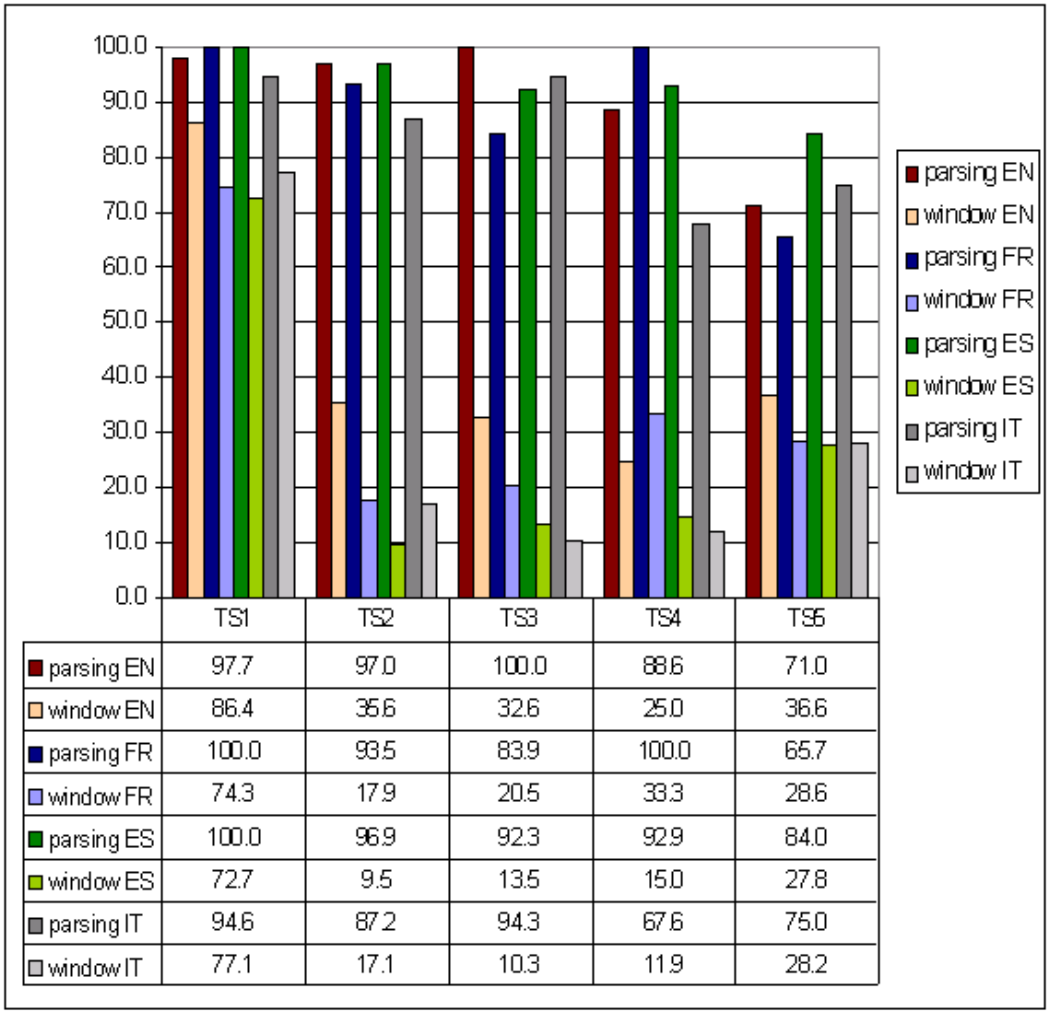
Graphical display of the results obtained in Experiment 2 (Section 4.4.4).

I.1 Grammatical precision

I.2 MWE precision

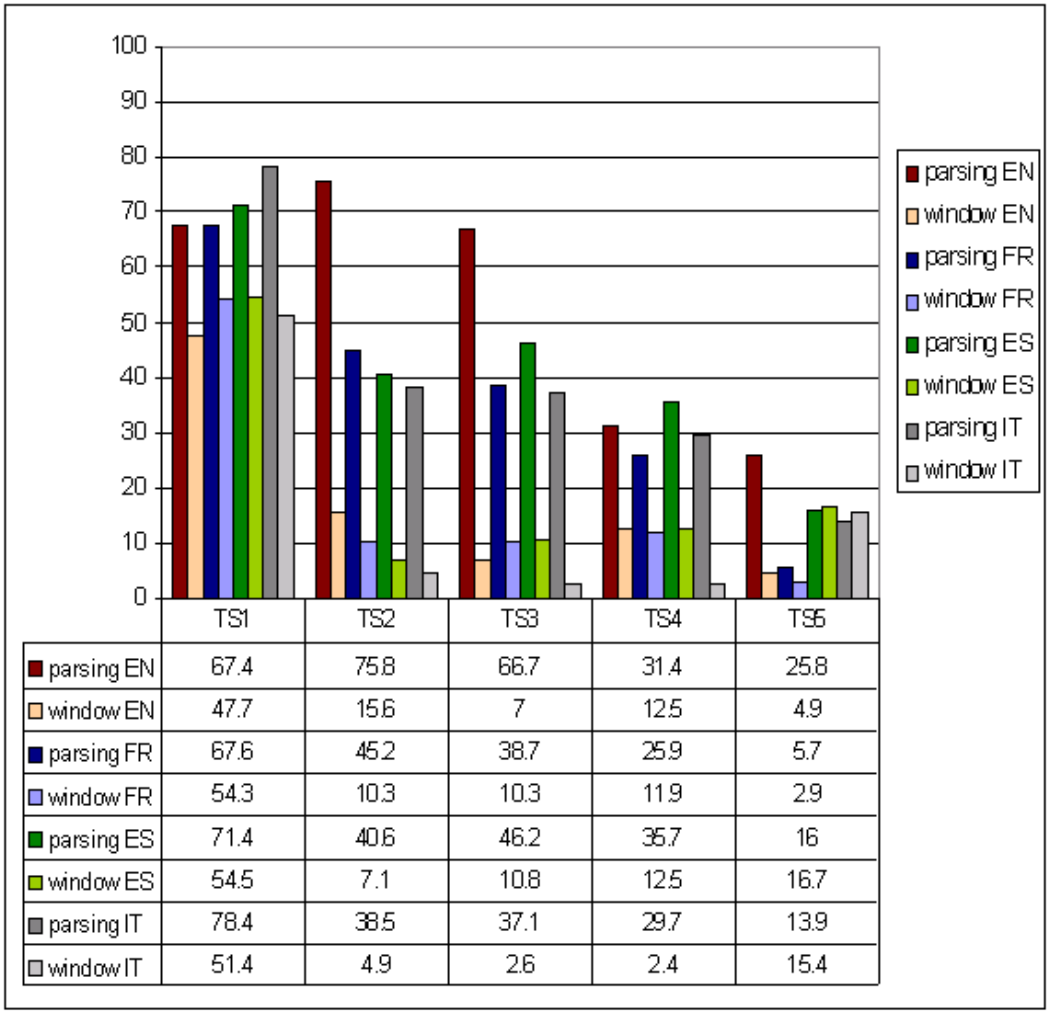
I.3 Collocational precision

I.4 Overall results



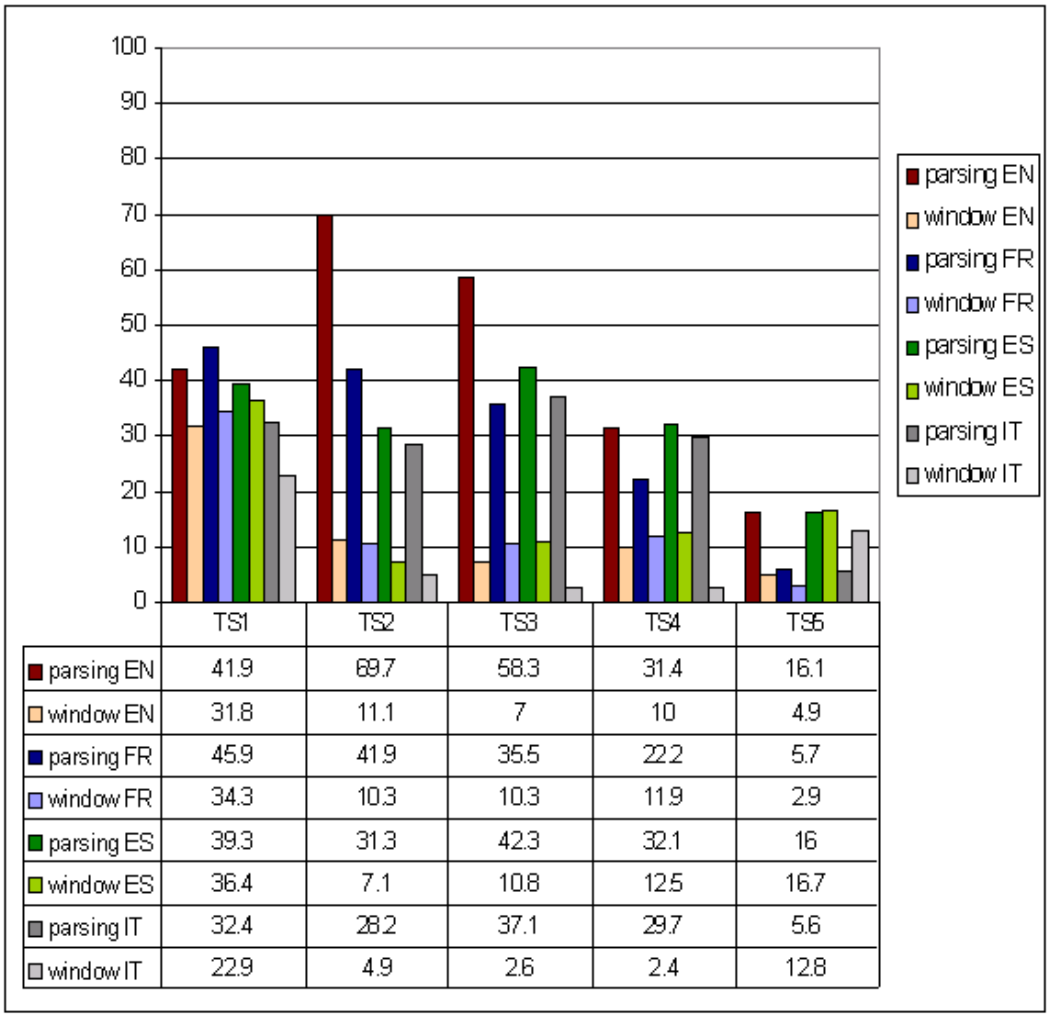
	TS1	TS2	TS3	TS4	TS5
parsing EN	97.7	97.0	100.0	88.6	71.0
window EN	86.4	35.6	32.6	25.0	36.6
parsing FR	100.0	93.5	83.9	100.0	65.7
window FR	74.3	17.9	20.5	33.3	28.6
parsing ES	100.0	96.9	92.3	92.9	84.0
window ES	72.7	9.5	13.5	15.0	27.8
parsing IT	94.6	87.2	94.3	67.6	75.0
window IT	77.1	17.1	10.3	11.9	28.2

Figure I.1: Comparative evaluation results for Experiment 2: Grammatical precision.



	TS1	TS2	TS3	TS4	TS5
parsing EN	67.4	75.8	66.7	31.4	25.8
window EN	47.7	15.6	7	12.5	4.9
parsing FR	67.6	45.2	38.7	25.9	5.7
window FR	54.3	10.3	10.3	11.9	2.9
parsing ES	71.4	40.6	46.2	35.7	16
window ES	54.5	7.1	10.8	12.5	16.7
parsing IT	78.4	38.5	37.1	29.7	13.9
window IT	51.4	4.9	2.6	2.4	15.4

Figure I.2: Comparative evaluation results for Experiment 2: MWE precision.



	TS1	TS2	TS3	TS4	TS5
parsing EN	41.9	69.7	58.3	31.4	16.1
window EN	31.8	11.1	7	10	4.9
parsing FR	45.9	41.9	35.5	22.2	5.7
window FR	34.3	10.3	10.3	11.9	2.9
parsing ES	39.3	31.3	42.3	32.1	16
window ES	36.4	7.1	10.8	12.5	16.7
parsing IT	32.4	28.2	37.1	29.7	5.6
window IT	22.9	4.9	2.6	2.4	12.8

Figure I.3: Comparative evaluation results for Experiment 2: Collocational precision.

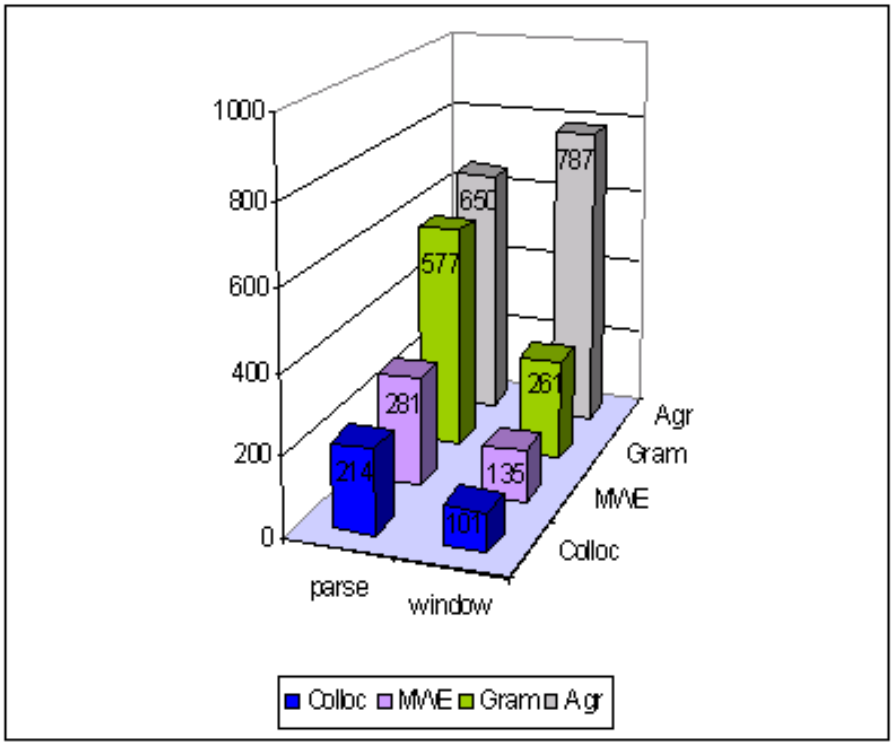


Figure I.4: Comparative evaluation results for Experiment 2: Total number of items by category (collocational pairs, MWE pairs, grammatical pairs, and pairs agreed upon).

Appendix J

Output Comparison: Intersection and Rank Correlation

Output comparison for the parse-based method (P) and mobile-window method (W) in Experiment 2 (Section 4.4.4). Column 3 and 4 (*Perc. P*, *Perc. W*) show the percentage of pair types in the output of method P and W, respectively, that satisfy the criterion in column 1. Column 5 lists the number of common pair types satisfying that criterion. The numbers in columns 6 and 7 represent the ratio of common items (the intersection) relative to the whole output of methods P and W (*Perc. C. P*, *Perc. C. W*). Finally, column 8 displays the Spearman's ρ correlation coefficient computed on the intersection.

	Lang.	**Perc. P**	**Perc. W**	**Common**	**Perc. C. P**	**Perc. C. W**	ρ
rank $\leq$ 1000	EN	0.32%	0.07%	315	0.10%	0.02%	0.597
	FR	0.32%	0.07%	293	0.09%	0.02%	0.710
	IT	0.31%	0.07%	300	0.09%	0.02%	0.476
	ES	0.33%	0.07%	308	0.10%	0.02%	0.708
	avg	0.32%	0.07%	304.0	0.10%	0.02%	0.623
rank $\leq$ 2000	EN	0.64%	0.14%	535	0.17%	0.04%	0.612
	FR	0.64%	0.14%	541	0.17%	0.04%	0.486
	IT	0.61%	0.15%	545	0.17%	0.04%	0.441
	ES	0.65%	0.15%	551	0.18%	0.04%	0.550
	avg	0.64%	0.14%	543.0	0.17%	0.04%	0.522

	Lang.	**Perc. P**	**Perc. W**	**Common**	**Perc. C. P**	**Perc. C. W**	ρ
rank $\leq$ 5000	EN	1.59%	0.35%	1088	0.35%	0.08%	0.532
	FR	1.61%	0.35%	1195	0.38%	0.08%	0.509
	IT	1.53%	0.37%	1327	0.41%	0.10%	0.592
	ES	1.63%	0.36%	1159	0.38%	0.08%	0.421
	avg	1.59%	0.36%	1192.3	0.38%	0.09%	0.513
rank $\leq$ 10000	EN	3.18%	0.70%	2030	0.65%	0.14%	0.406
	FR	3.22%	0.70%	2318	0.75%	0.16%	0.553
	IT	3.06%	0.73%	2601	0.79%	0.19%	0.918
	ES	3.26%	0.73%	2133	0.70%	0.16%	0.418
	avg	3.18%	0.71%	2270.5	0.72%	0.16%	0.574
all	EN	100.00%	100.00%	237565	75.59%	16.53%	0.477
	FR	100.00%	100.00%	228373	73.48%	16.09%	0.521
	IT	100.00%	100.00%	260434	79.58%	19.07%	0.471
	ES	100.00%	100.00%	247983	80.81%	18.02%	0.480
	avg	100.00%	100.00%	243588.8	77.36%	17.43%	0.487
LLR $>$ 1000	EN	0.02%	0.01%	40	0.01%	0.00%	0.734
	FR	0.04%	0.01%	57	0.02%	0.00%	0.792
	IT	0.04%	0.02%	71	0.02%	0.01%	0.841
	ES	0.04%	0.01%	77	0.03%	0.01%	0.868
	avg	0.03%	0.01%	61.3	0.02%	0.00%	0.809
LLR $>$ 100	EN	0.62%	0.63%	1206	0.38%	0.08%	0.692
	FR	0.84%	0.69%	1496	0.48%	0.11%	0.621
	IT	0.81%	0.86%	1704	0.52%	0.12%	0.482
	ES	0.74%	0.74%	1426	0.46%	0.10%	0.658
	avg	0.75%	0.73%	1458.0	0.46%	0.10%	0.613
LLR $>$ 10	EN	16.00%	13.14%	22352	7.11%	1.56%	0.594
	FR	16.98%	13.13%	23457	7.55%	1.65%	0.655
	IT	15.51%	12.90%	24507	7.49%	1.79%	0.611
	ES	14.57%	12.79%	21033	6.85%	1.53%	0.627
	avg	15.76%	12.99%	22837.3	7.25%	1.63%	0.622
f $>$ 100	EN	0.12%	0.06%	187	0.06%	0.01%	0.870
	FR	0.15%	0.08%	224	0.07%	0.02%	0.834

	Lang.	**Perc. P**	**Perc. W**	**Common**	**Perc. C. P**	**Perc. C. W**	ρ
	IT	0.12%	0.11%	279	0.09%	0.02%	0.900
	ES	0.15%	0.07%	247	0.08%	0.02%	0.849
	avg	0.13%	0.08%	234.3	0.07%	0.02%	0.863
f $>$ 10	EN	2.55%	2.08%	5212	1.66%	0.36%	0.796
	FR	2.78%	2.19%	5542	1.78%	0.39%	0.747
	IT	2.78%	2.86%	6698	2.05%	0.49%	0.749
	ES	3.10%	2.56%	6204	2.02%	0.45%	0.731
	avg	2.80%	2.42%	5914.0	1.88%	0.42%	0.756
f $>$ 3	EN	9.35%	9.34%	20639	6.57%	1.44%	0.662
	FR	10.21%	9.66%	22036	7.09%	1.55%	0.641
	IT	10.02%	11.65%	25280	7.72%	1.85%	0.640
	ES	10.97%	11.19%	24251	7.90%	1.76%	0.614
	avg	10.14%	10.46%	23051.5	7.32%	1.65%	0.639

Appendix K

Multi-Word Collocations - Random Results

Randomly selected 3-grams and 4-grams obtained by applying the multi-word collocation extraction method (Section 5.1) to the output of the Experiment 1 (Section 4.4.3).

K.1 3-grams

Rank	Keys	Types	Freq	Score
87	loi (sur) protection (de) pêche	N-P-N, N-P-N	13	184
91	ancien gouvernement conservateur	A-N, N-A	18	180
98	base (de) force canadien	N-P-N, N-A	12	175
106	rehausser crédibilité (de) parlement	V-O, N-P-N	10	169
127	rapport (de) comité permanent	N-P-N, N-A	15	151
128	être bon chose	V-O, A-N	12	149
136	réintégrer marché (de) travail	V-O, N-P-N	10	145
162	économiser million (de) dollar	V-O, N-P-N	12	131
166	processus (de) négociation collectif	N-P-N, N-A	8	125
220	créer emploi à long terme	V-O, N-A	10	105
224	député exprimer opinion	S-V, V-O	9	104
234	faire objet (de) débat	V-O, N-P-N	10	102
240	connaître opinion sur	V-O, V-P	7	100

Rank	Keys	Types	Freq	Score
249	réduction (de) taxe (sur) cigarette	N-P-N, N-P-N	8	98
256	programme (de) nutrition prénatal	N-P-N, N-A	5	97
273	faire objet (de) examen	V-O, N-P-N	8	93
304	faire (de) choix difficile	V-P-N, N-A	6	89
307	nomination (de) nouveau président	N-P-N, A-N	6	89
319	créer comité spécial	V-O, N-A	8	87
330	faire bref commentaire	V-O, A-N	6	85
347	traverser période difficile	V-O, N-A	5	83
357	député retirer candidature	S-V, V-O	5	81
386	gouvernement prendre responsabilité	S-V, V-O	8	78
391	région (de) capitale national	N-P-N, N-A	4	78
410	prêter oreille attentif	V-O, N-A	4	76
457	ouvrir processus budgétaire	V-O, N-A	4	71
478	parti réformiste être	N-A, S-V	10	69
485	adresser (à) ministre (de) développement	V-P-N, N-P-N	15	68
487	commission (de) libération conditionnel	N-P-N, N-A	4	68
488	élaboration (de) politique gouvernemental	N-P-N, N-A	5	68

K.2 4-grams

Rank	Keys	Types	Freq	Score
28	Chambre reprendre interrompre étude	S-V, V-O, V-O	10	112
59	programme (de) aide (à) remise (en) état	N-P-N, N-P-N, N-P-N	4	75
73	afficher liste (de) candidat dans	V-O, N-P-N, V-P	4	70
91	nomination (de) nouveau président (de) société	N-P-N, A-N, N-P-N	4	64
97	gouvernement être gouvernement méchant	S-V, V-O, N-A	4	62
99	représentant (de) bureau (de) régie interne	N-P-N, N-P-N, N-A	4	62
134	gouvernement devoir assumer responsabilité	S-V, S-V, V-O	4	56
138	bon accès (à) marché mondial	A-N, N-P-N, N-A	3	56

Rank	Keys	Types	Freq	Score
153	féliciter ministre (de) revenu national	V-O, N-P-N, N-A	3	53
176	rétablir confiance (de) Canadien dans	V-O, N-P-N, V-P	3	51
181	gouvernement devoir agir de	S-V, S-V, V-P	3	50
183	honorable député avoir parole	A-N, S-V, V-O	4	50
183	honorable député avoir parole	A-N, S-V, V-O	4	50
190	extension (de) régime (de) accession (à) propriété	N-P-N, N-P-N, N-P-N	3	49
232	gouvernement prendre engagement envers	S-V, V-O, V-P	3	45
236	ministre (de) Affaires étrangères aborder avec	N-P-N, S-V, V-P	3	44
237	féliciter honorable ministre pour	V-O, A-N, V-P	3	44
249	jouer rôle (de) premier plan dans	V-O, N-P-N, V-P	4	43
276	proposer (à) loi (sur) protection (de) pêche	N-P-N, N-P-N, N-P-N	3	42
352	trouver juste milieu entre	V-O, A-N, V-P	2	39
360	premier ministre pouvoir donner	A-N, S-V, S-V	3	39
414	être (à) prise (avec) taux (de) chômage	N-P-N, N-P-N, N-P-N	2	37
422	groupe demander révocation (de) député	S-V, V-O, N-P-N	2	37
453	être excellent joueur pour	V-O, A-N, V-P	2	37
460	injecter million (de) dollar (dans) économie	V-O, N-P-N, N-P-N	2	37
461	ministre avoir rencontre avec	S-V, V-O, V-P	2	37
471	trouver bon solution possible	V-O, A-N, N-A	2	37
479	lacune (dans) système (de) justice pénal	N-P-N, N-P-N, N-A	2	36
482	présider forum national sur	V-O, N-A, V-P	2	36
483	réforme (de) système (de) justice pénal	N-P-N, N-P-N, N-A	2	36

Appendix L

Tool - Association Measures

List of AMs used by the collocation extraction and visualisation tool (Chapter 6). The AMs that were not already introduced in Section 3.2.4 are documented, for instance, in (Evert, 2004).

AM	**Explicit formula**
Chi-square	$\frac{(a+b+c+d)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$
Dice coefficient	$\frac{2a}{2a+b+c}$
Geometric mean	$\frac{a}{\sqrt{(a+b)(a+c)}}$
Jaccard	$\frac{a}{a+b+c}$
Lindell	$\frac{ad-bc}{(a+c)(b+d)}$
Log-likelihood ratios (LLR)	$2(a\log a + b\log b + c\log c + d\log d -$ $(a+b)\log(a+b) - (a+c)\log(a+c) -$ $(b+d)\log(b+d) - (c+d)\log(c+d) +$ $(a+b+c+d)\log(a+b+c+d))$
Mutual information (MI)	$\log_2 \frac{a(a+b+c+d)}{(a+b)(a+c)}$

AM	**Explicit formula**
Odds ratio	$\log \frac{ad}{bc}$
Poisson-Stirling	$a\left(\log a - \log \frac{(a+b)(a+c)}{(a+b+c+d)} - 1\right)$
Relative frequency	$\frac{a}{a+b+c+d}$
Relative risk	$\log \frac{a(b+d)}{b(a+c)}$
Salience	$\log_2 \frac{a(a+b+c+d)}{(a+b)(a+c)} \log_2 a$
t-score	$\frac{a(a+b+c+d)-(a+b)(a+c)}{(a+b+c+d)\sqrt{a}}$
z-score	$\frac{a(a+b+c+d)-(a+b)(a+c)}{\sqrt{a+b+c+d}\sqrt{(a+b)(a+c)}}$

Appendix M

Tool - Screen captures

M.1 Corpus selection component

M.2 Collocation filter component

M.3 Concordanding component

M.4 Alignment component

M.5 Validation component

M.6 Web-based extraction component

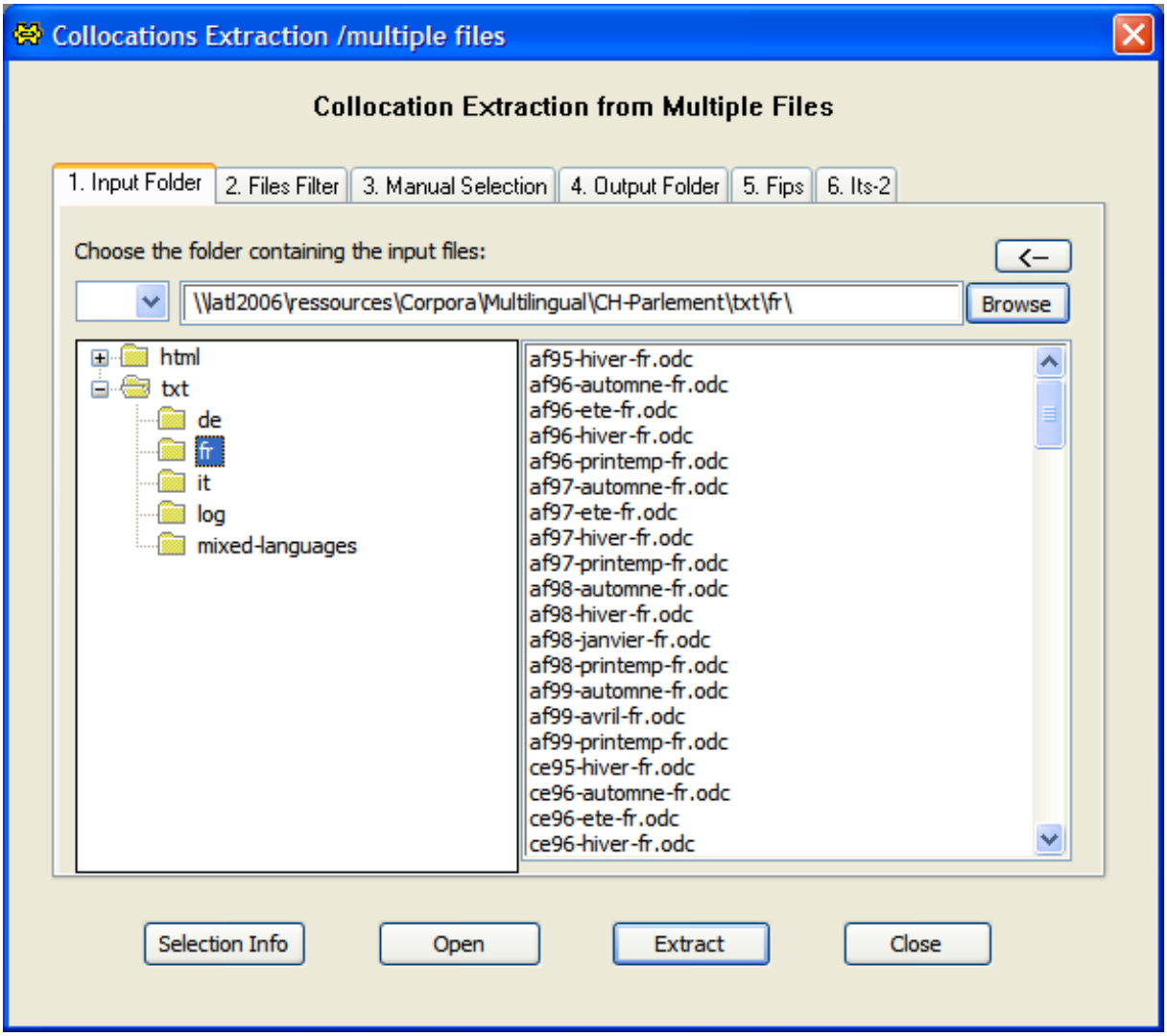


Figure M.1: Corpus selection component (input folder).

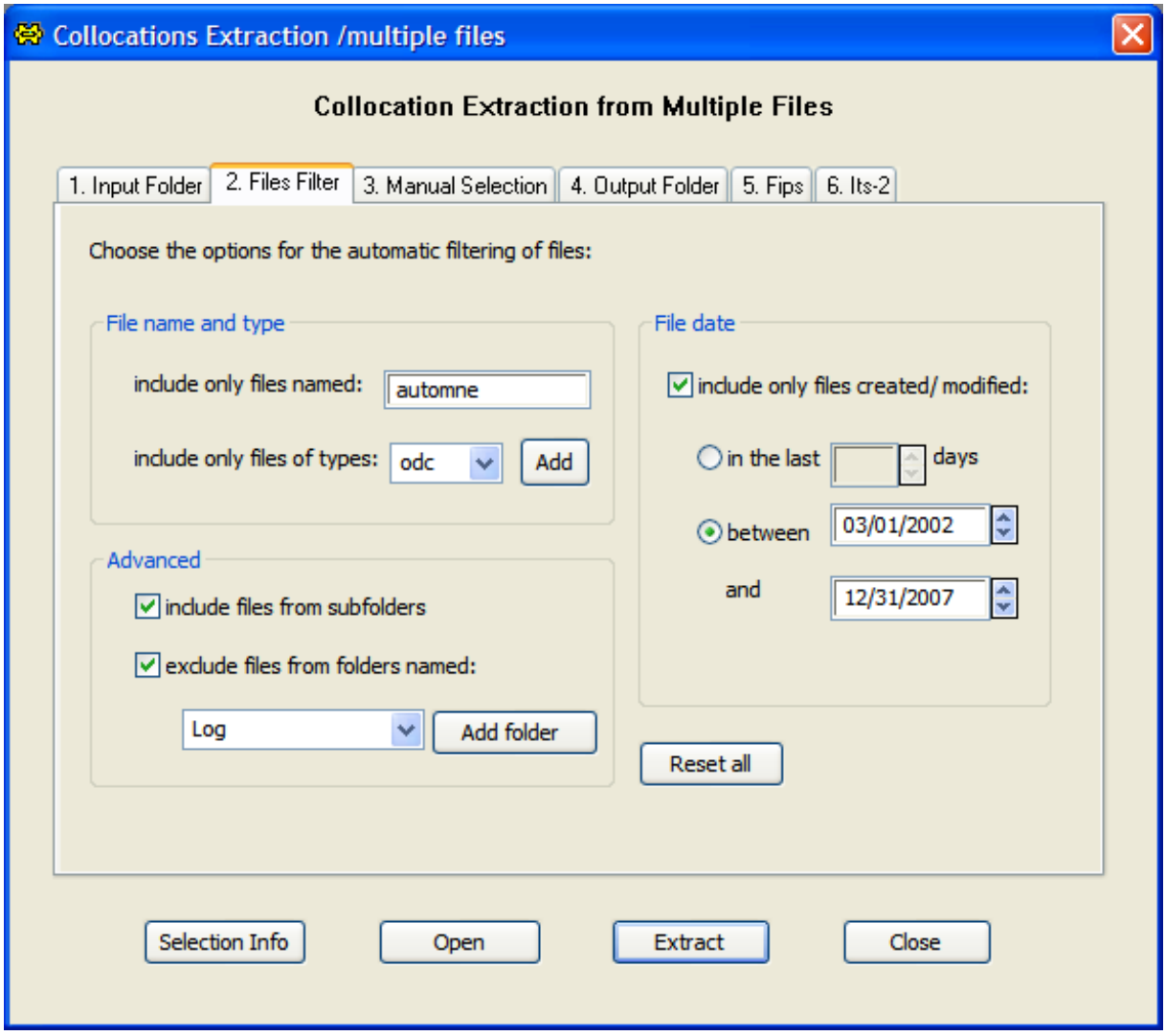


Figure M.2: Corpus selection component (automatic filter).

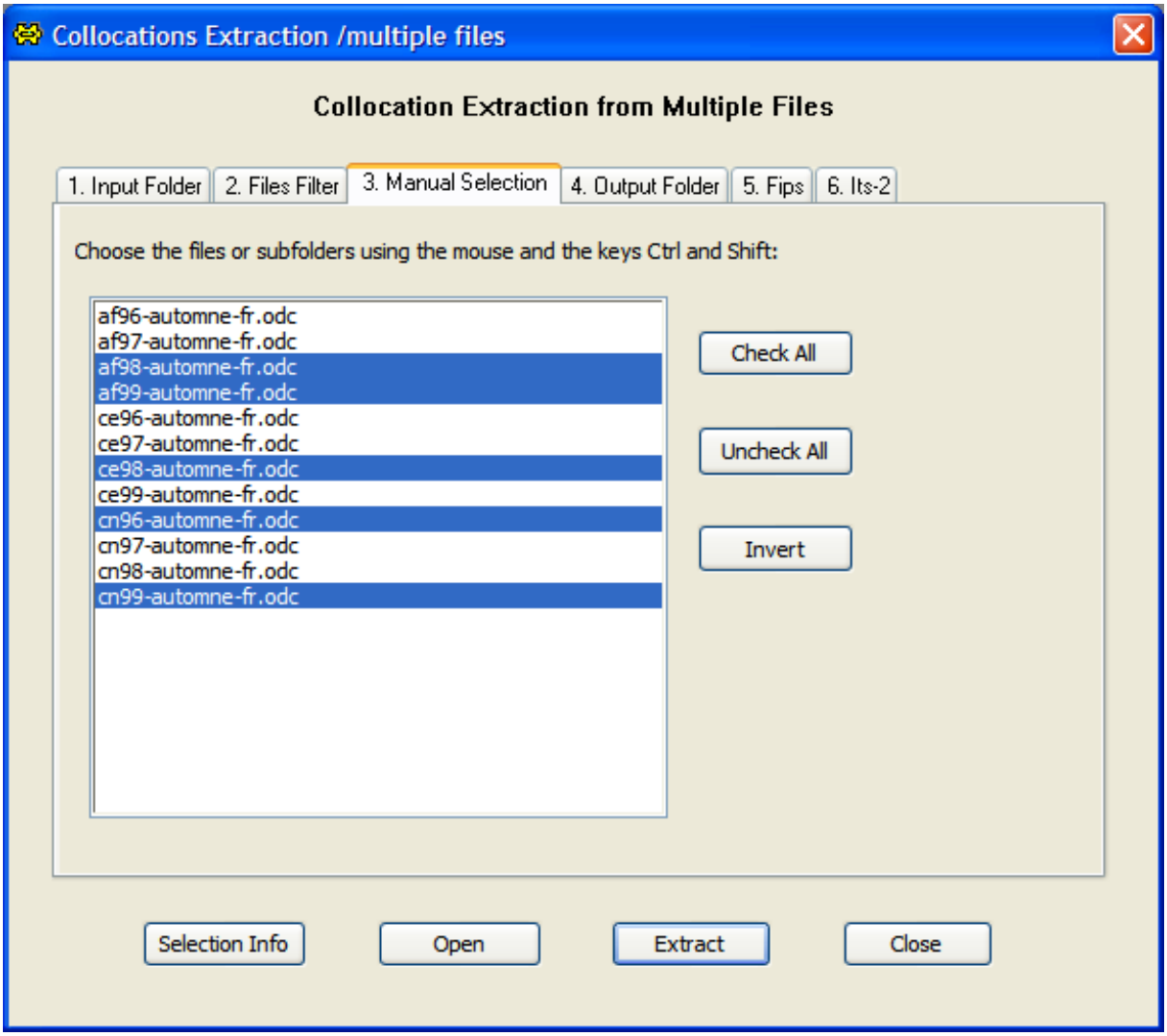


Figure M.3: Corpus selection component (manual selection).

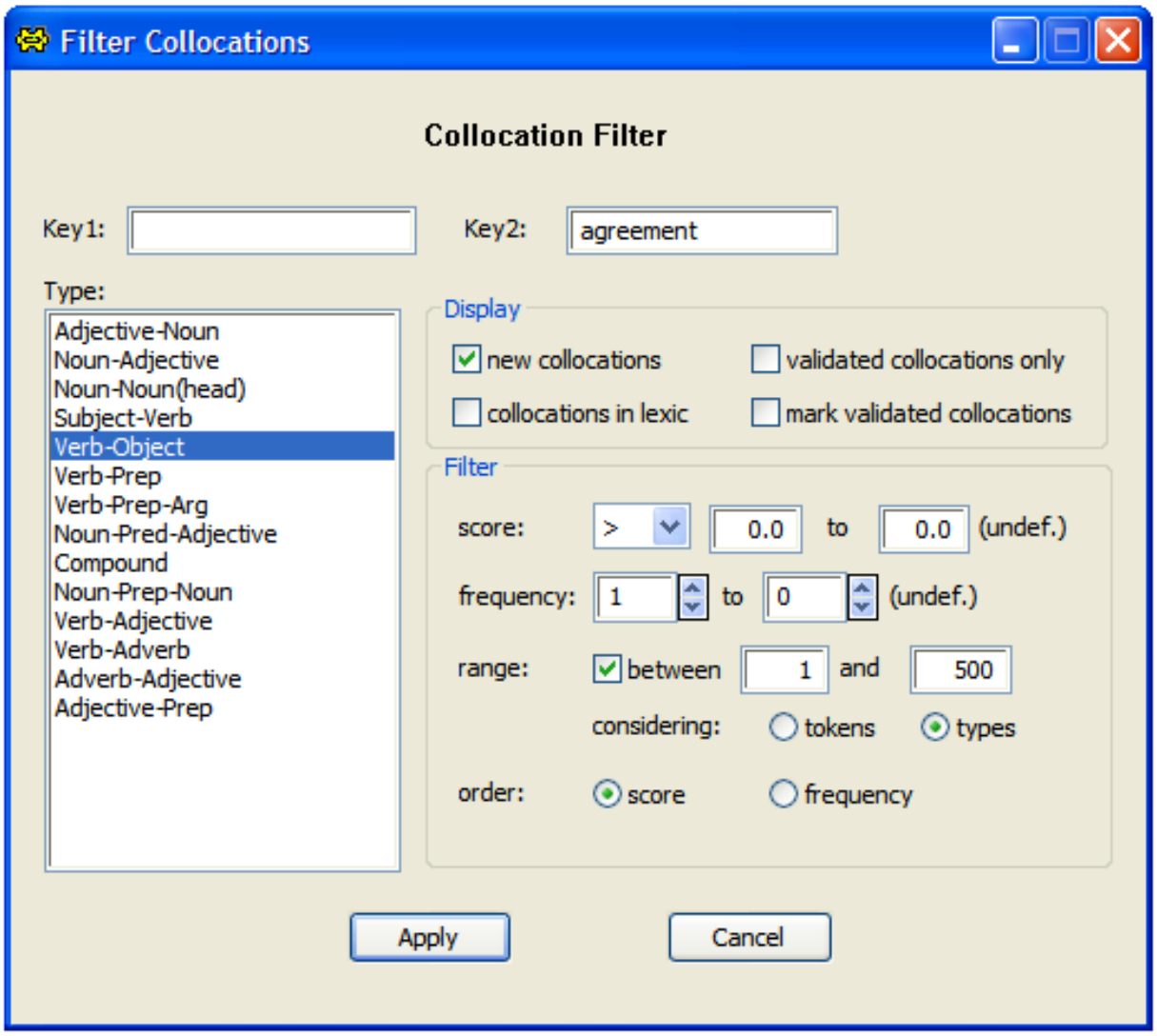


Figure M.4: Collocation filter component.

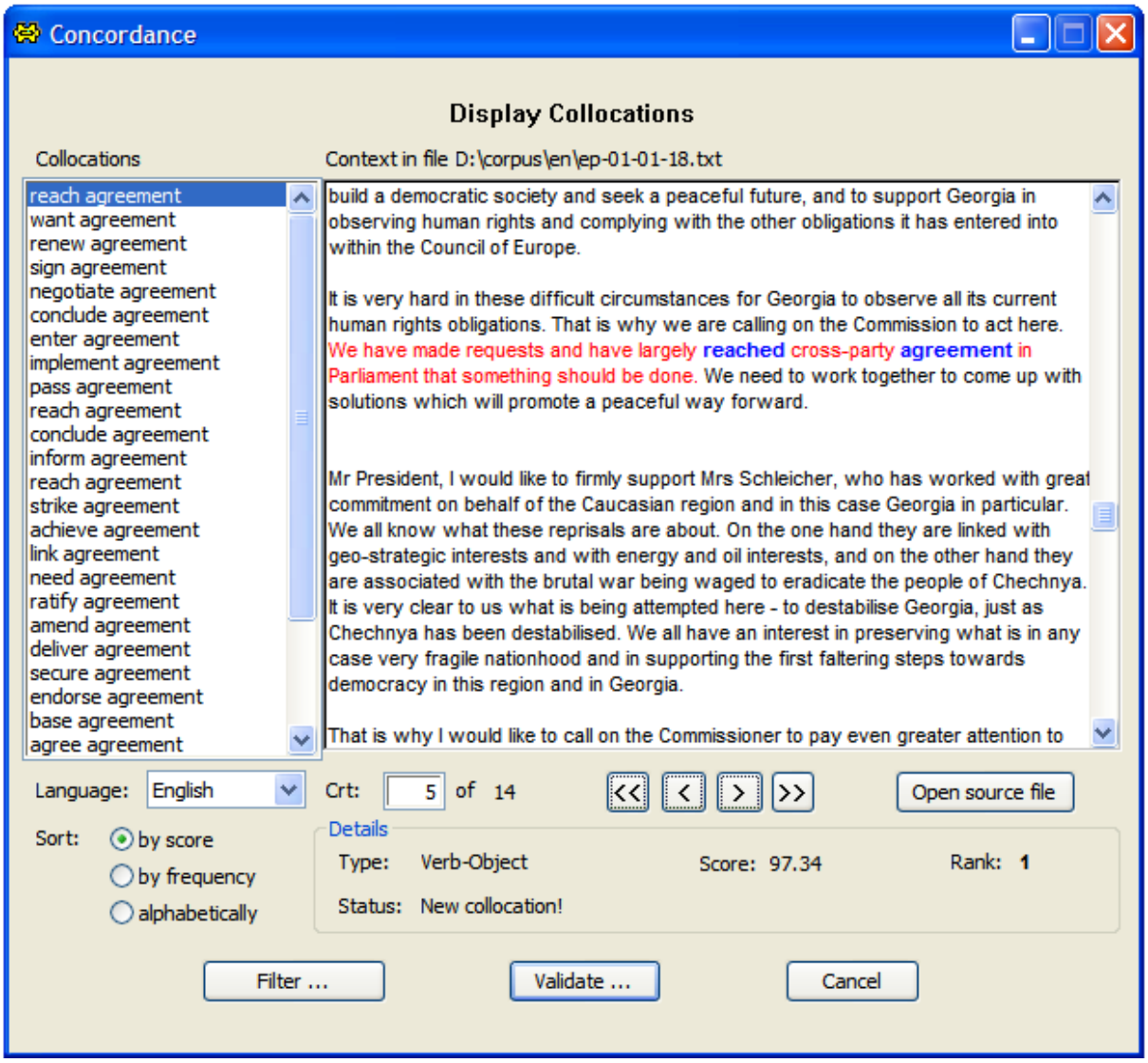


Figure M.5: Concordanding component.

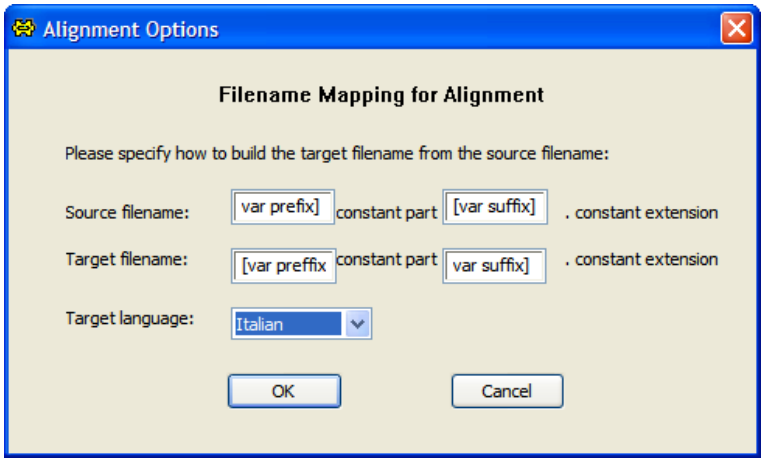


Figure M.6: Alignment component (filename transformation rules).

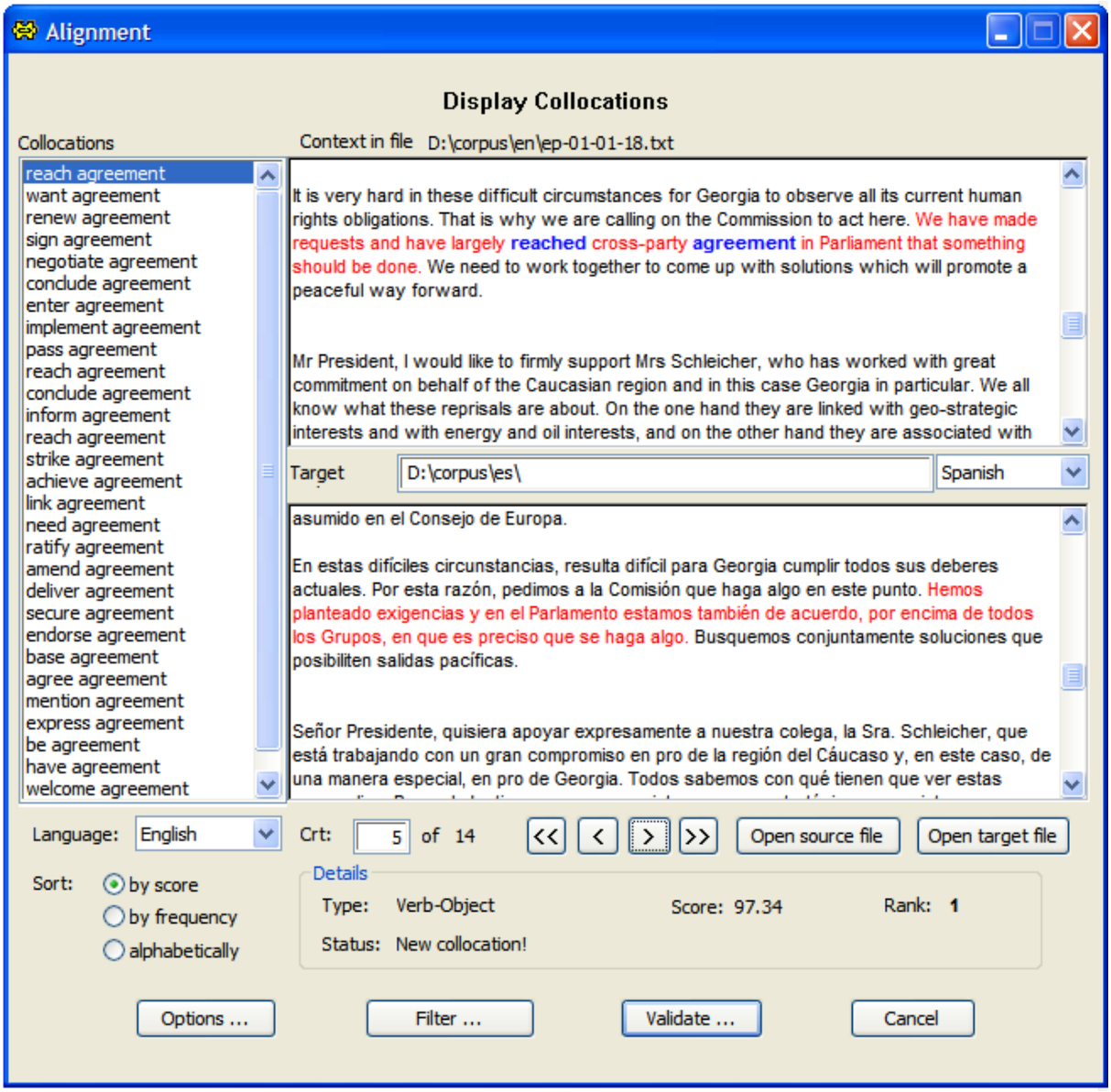


Figure M.7: Alignment component.

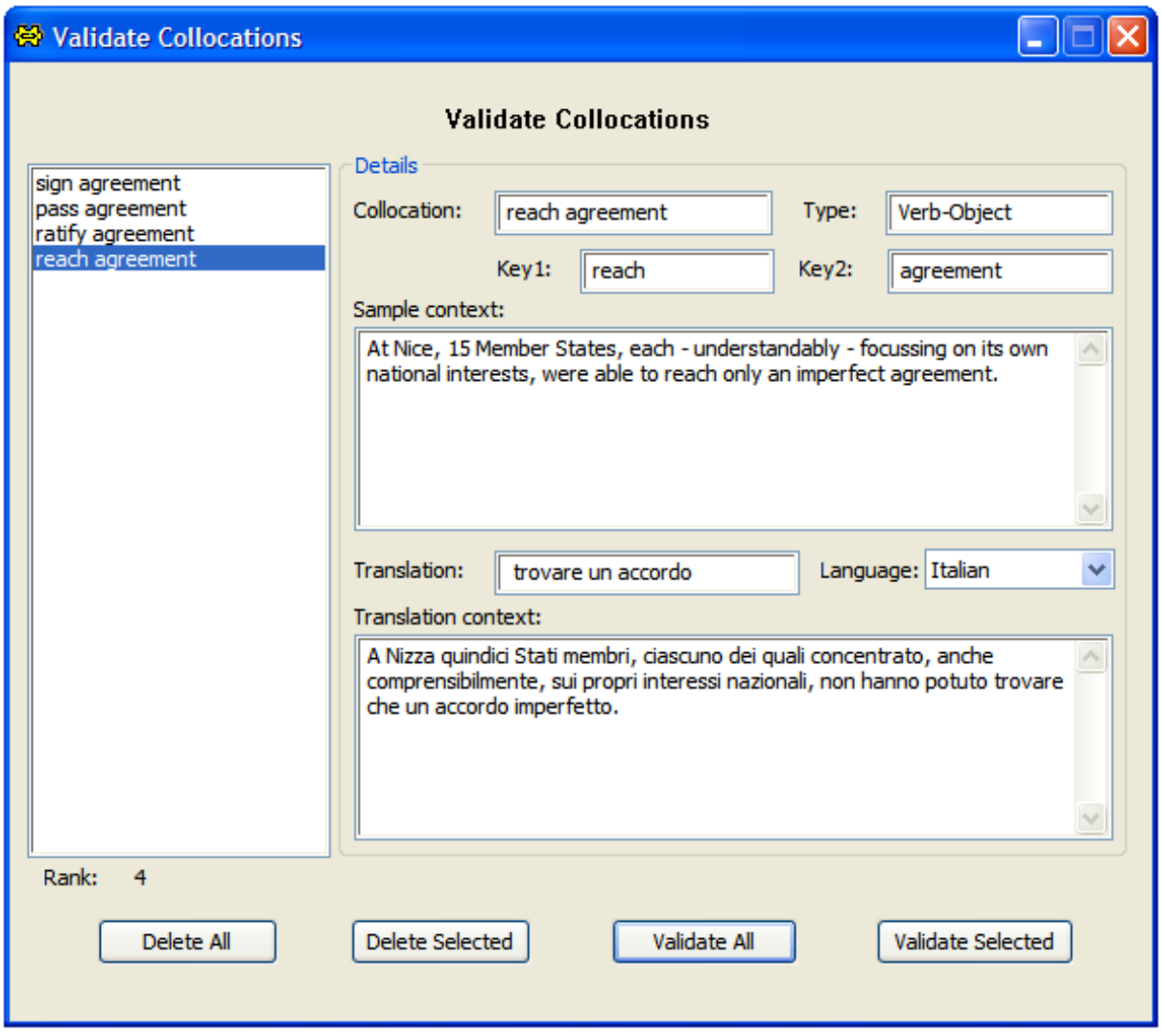


Figure M.8: Validation component.

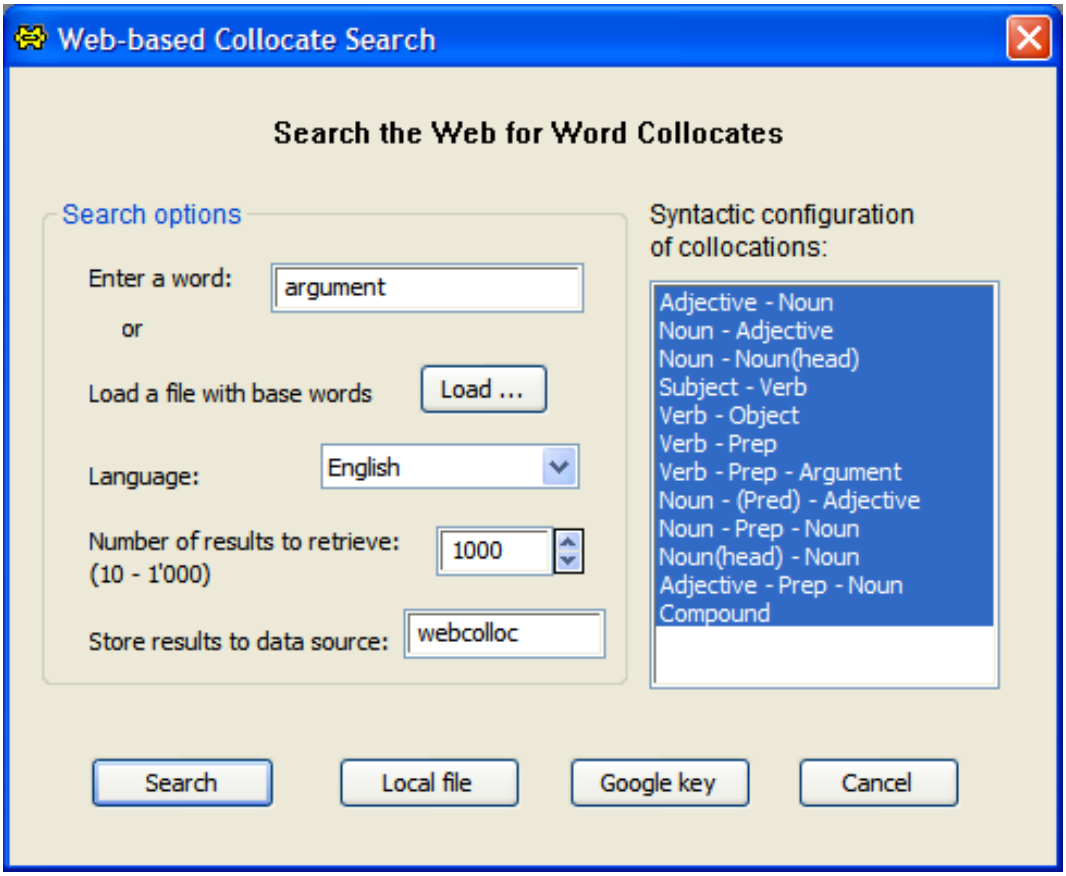


Figure M.9: Extraction of collocations from the Web.

Bibliography

Margarita Alonso Ramos and Agnès Tutin. 1996. A classification and description of lexical functions for the analysis of their combinations. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 147–167. Benjamins, Amsterdam/Philadelphia.

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

Vincent Archer. 2006. Acquisition semi-automatique de collocations à partir de corpus monolingues et multilingues comparables. In *Proceedings of Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2006)*, April.

Jens Bahns. 1993. Lexical collocations: a contrastive view. *ELT Journal*, 1(47):56–63.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, pages 86–90, Montreal, Canada.

Lisa Ballestros and W. Bruce Croft. 1996. Dictionary-based methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference in Database and Expert Systems Applications*, pages 791–801.

Charles Bally. 1909. *Traité de stylistique française*. Klincksieck, Paris.

Satanjeev Banerjee and Ted Pedersen. 2003. The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

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

Sabine Bartsch. 2004. *Structural and Functional Properties of Collocations in English. A Corpus Study of Lexical and Pragmatic Constraints on Lexical Cooccurrence*. Gunter Narr Verlag, Tübingen.

Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. 1994. A "not-so-shallow" parser for collocational analysis. In *Proceedings of the 15th conference on Computational linguistics*, pages 447–453, Kyoto, Japan.

Jacques Beauchesne. 2001. *Dictionnaire des cooccurrences*. Guérin, Montréal.

Morton Benson, Evelyn Benson, and Robert Ilson. 1986a. *The BBI Dictionary of English Word Combinations*. John Benjamins, Amsterdam/Philadelphia.

Morton Benson, Evelyn Benson, and Robert Ilson. 1986b. *Lexicographic Description of English*. John Benjamins, Amsterdam/Philadelphia.

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

Godelieve L. M. Berry-Rogghe. 1973. The computation of collocations and their relevance to lexical studies. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith, editors, *The Computer and Literary Studies*, pages 103–112. Edinburgh.

Cathy Berthouzoz and Paola Merlo. 1997. Statistical ambiguity resolution for principle-based parsing. In Nicolas Nicolov and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing: Selected Papers from RANLP'97*, Current Issues in Linguistic Theory. John Benjamins, Amsterdam/Philadelphia.

Don Blaheta and Mark Johnson. 2001. Unsupervised learning of multi-word verbs. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, pages 54–60, Toulouse, France.

Christian Boitet, Mathieu Mangeot, and Gilles Sérasset. 2002. The PAPILLON Project: Cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons. In *Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002)*, Taipei, Taiwan.

Didier Bourigault. 1992a. LEXTER, vers un outil linguistique d'aide à l'acquisition des connaissances. In *Actes des 3èmes Journées d'acquisition des Connaissances*, Dourdan, France, April.

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

Elisabeth Breidt. 1993. Extraction of V-N-collocations from text corpora: A feasibility study for German. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, U.S.A.

Joan Bresnan. 2001. *Lexical Functional Syntax*. Blackwell, Oxford.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991a. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL 1991)*, pages 169–176, Berkeley, California.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991b. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL 1991)*, pages 264–270, Berkeley, California.

Nicoletta Calzolari and Remo Bindi. 1990. Acquisition of lexical information from a large textual Italian corpus. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 54–59, Helsinki, Finland.

Simon Charest, Éric Brunelle, Jean Fontaine, and Bertrand Pelletier. 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)*, pages 283–292, Toulouse, France, June.

Stanley Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Columbus, Ohio, USA, June.

Noam Chomsky. 1995. *The Minimalist Program*. MIT Press, Cambridge, Mass.

Yaacov Choueka, S.T. Klein, and E. Neuwitz. 1983. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing*, 4(1):34–38.

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

Kenneth W. Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, B.C. Association for Computational Linguistics.

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

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1989. Parsing, word associations and typical predicate-argument relations. In *Proceedings of the International Workshop on Parsing Technologies*, pages 103–112, Pittsburgh. Carnegie Mellon University.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum, Hillsdale, NJ.

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

Eugenio Coseriu. 1967. Lexikalische solidaritäten. *Poetica*, (1):293–303.

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

Anthony P. Cowie. 1998. *Phraseology. Theory, Analysis, and Applications*. Claredon Press, Oxford.

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

Peter Culicover and Ray Jackendoff. 2005. *Simpler Syntax*. Oxford University Press, Oxford.

Ido Dagan and Kenneth Church. 1994. *Termight*: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP)*, pages 34–40, Stuttgart, Germany.

Béatrice Daille. 1994. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Université Paris 7.

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

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

Stefan Evert and Hannah Kermes. 2002. The influence of linguistic preprocessing on candidate data. In *Proceedings of Workshop on Computational Approaches to Collocations (Colloc02)*, Vienna, Austria.

Stefan Evert and Hannah Kermes. 2003. Experiments on candidate data for collocation extraction. In *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 83–86, Budapest, Hungary.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.

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

Stefan Evert, Ulrich Heid, and Kristina Spranger. 2004. Identifying morphosyntactic preferences in collocations. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 907–910, Lisbon, Portugal.

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

Afsaneh Fazly. 2007. *Automatic Acquisition of Lexical Knowledge about Multiword Predicates*. Ph.D. thesis, University of Toronto.

Olivier Ferret and Michael Zock. 2006. Enhancing electronic dictionaries with an index based on associations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 281–288, Sydney, Australia, July.

Olivier Ferret. 2002. Using collocations for topic segmentation and link detection. In *Proceedings of the 19th International Conference on Computational linguistics (COLING 2002)*, pages 260–266, Taipei, Taiwan.

Olivier Ferret. 2003. Filtrage thématique d'un réseau de collocations. In *Proceedings of TALN 2003*, pages 347–352, Batz-sur-Mer, France, June.

Charles Fillmore, Paul Kay, and Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language*, 64(3):501–538.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul.

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

John R. Firth. 1968. A synopsis of linguistic theory, 1930–55. In F.R. Palmer, editor, *Selected papers of J. R. Firth, 1952–1959*, pages 168–205. Indiana University Press, Bloomington.

Joseph L. Fleiss. 1981. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Thierry Fontenelle. 1992. Collocation acquisition from a corpus or from a dictionary: a comparison. *Proceedings I-II. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere*, pages 221–228.

Thierry Fontenelle. 1997a. *Turning a bilingual dictionary into a lexical-semantic database*. Max Niemeyer Verlag, Tübingen.

Thierry Fontenelle. 1997b. Using a bilingual dictionary to create semantic networks. *International Journal of Lexicography*, 10(4):276–303.

Thierry Fontenelle. 1999. Semantic resources for word sense disambiguation: a *sine qua non*? *Linguistica e Filologia*, (9):25–43. Dipartimento di Linguistica e Letterature Comparate, Università degli Studi di Bergamo.

Thierry Fontenelle. 2001. Collocation modelling: from lexical functions to frame semantics. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, pages 1–7, Toulouse, France.

Gill Francis. 1993. A corpus-driven approach to grammar: Principles, methods and examples. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 137–156. John Benjamins, Amsterdam.

Katerina T. Frantzi and Sophia Ananiadou. 1996. Extracting nested collocations. In *Proceedings of the 15th International Conference on Computational linguistics (COLING'96)*, pages 41–46, Copenhagen, Denmark.

Katerina T. Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 2(3):115–130.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102. Special Issue on Using Large Corpora.

Christina Gitsaki. 1996. *The Development of ESL Collocational Knowledge*. Ph.D. thesis, University of Queensland.

Jean-Philippe Goldman, Luka Nerima, and Eric Wehrli. 2001. Collocation extraction using a syntactic parser. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, pages 61–66, Toulouse, France.

Gregory Grefenstette and Simone Teufel. 1995. Corpus-based method for automatic identification of support verbs for nominalizations. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–103, Dublin, Ireland, March.

Maurice Gross. 1984. Lexicon-grammar and the syntactic analysis of French. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 275–282, Morristown, NJ, USA.

Gaston Gross. 1996. *Les expressions figées en français*. OPHRYS, Paris.

Jan Hajič. 2000. Morphological tagging: Data vs. dictionaries. In *Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference*, pages 94–101, Seattle, Washington.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Peter Hargreaves. 2000. Collocation and testing. In Michael Lewis, editor, *Teaching Collocations*. Language Teaching Publications, Hove.

Franz Josef Hausmann. 1979. Un dictionnaire des collocations est-il possible? *Travaux de littérature et de linguistique de l'Université de Strasbourg*, 17(1):187–195.

Franz Iosef Hausmann. 1985. Kollokationen im deutschen wörterbuch. ein beitrag zur theorie des lexikographischen beispiels. In Henning Bergenholtz and Joachim Mugdan, editors, *Lezikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch*, Lexicographica. Series Major 3, pages 118–129.

Franz Josef Hausmann. 1989. Le dictionnaire de collocations. In F. J. Hausmann et al., editor, *Wörterbücher: Ein internationales Handbuch zur Lexicographie. Dictionaries, Dictionnaires*, pages 1010–1019. de Gruyter, Berlin.

Ulrich Heid and Sybille Raab. 1989. Collocations in multilingual generation. In *Proceeding of the Fourth Conference of the European Chapter of the Association for Computational Linguistics EACL'89)*, pages 130–136, Manchester, England.

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

Dirk Heylen, Kerry G. Maxwell, and Marc Verhagen. 1994. Lexical functions and machine translation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, pages 1240–1244, Kyoto, Japan.

Jimmie Hill and Michael Lewis, editors. 1997. *Dictionary of Selected Collocations*. Language Teaching Publications, Hove.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Michael Hoey. 1991. *Patterns of Lexis in Text*. Oxford University Press, Oxford.

Michael Hoey. 1997. From concordance to text structure: New uses for computer corpora. In J. Melia and B. Lewandoska, editors, *Proceedings of Practical Applications of Language Corpora (PALC 1997)*, pages 2–23, Lodz, Poland.

Michael Hoey. 2000. A world beyond collocation: New perspectives on vocabulary teaching. In Michael Lewis, editor, *Teaching Collocations*. Language Teaching Publications, Hove.

Peter Howarth and Hilary Nesi. 1996. The teaching of collocations in EAP. Technical report, University of Leeds, June.

Chu-Ren Huang, Adam Kilgarriff, Yiching Wu, Chih-Ming Chiu, Simon Smith, Pavel Rychly, Ming-Hong Bai, and Keh-Jiann Chen. 2005. Chinese Sketch Engine and the extraction of grammatical collocations. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 48–55, Jeju Island, Korea.

David A. Hull and Gregory Grefenstette. 1998. Querying across languages: A dictionary-based approach to multilingual information retrieval. In Karen Spark Jones and Peter Willett, editors, *Readings in Information Retrieval*, pages 484–492. Morgan Kaufmann, San Francisco.

Susan Hunston and Gill Francis. 1998. Verbs observed: A corpus-driven pedagogic grammar. *Applied Linguistics*, 19(1):45–72.

Susan Hunston, Gill Francis, and Elizabeth Manning. 1997. Grammar and vocabulary: Showing the connections. *English Language Teaching Journal*, 3(51):208–215.

Satoru Ikehara, Satoshi Shirai, and Tsukasa Kawaoka. 1995. Automatic extraction of uninterrupted collocations by n-gram statistics. In *Proceedings of first Annual Meeting of the Association for Natural Language Processing*, pages 313–316.

P. Isabelle, Dymetman M., Foster G., Jutras J-M., Macklovitch E., Perrault F., Ren X., and Simard M. 1993. Translation analysis and translation automation. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-93)*, Kyoto, Japon.

Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of the 35th annual meeting on Association for Computational Linguistics*, pages 24–31, Morristown, NJ, USA.

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

Sylvain Kahane and Alain Polguère. 2001. Formal foundations of lexical functions. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, pages 8–15.

Martin Kay and Martin Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142. Special Issue on Using Large Corpora.

Frank Keller and Mirella Lapata. 2003. Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Hannah Kermes and Ulrich Heid. 2003. Using chunked corpora for the acquisition of collocations and idiomatic expressions. In F. Kiefer and J. Pajzs, editors, *Proceedings of 7th Conference on Computational Lexicography and Corpus Research*, Budapest, Hungary. Research Institute for Linguistics, Hungarian Academy of Sciences.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the Web as corpus. *Computational Linguistics*, 29(3):333–347. Special issue on Web as corpus.

Adam Kilgarriff and David Tugwell. 2001. WORD SKETCH: Extraction and display of significant collocations for lexicography. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, pages 32–38, Toulouse, France.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116, Lorient, France.

Adam Kilgarriff. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition*, pages 33–40, Sussex, U.K., April.

Seonho Kim, Zooil Yang, Mansuk Song, and Jung-Ho Ahn. 1999. Retrieving collocations from Korean text. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 71–81, Maryland, U.S.A.

Seonho Kim, Juntae Yoon, and Mansuk Song. 2001. Automatic extraction of collocations from Korean text. *Computers and the Humanities*, pages 273–297.

Mihoko Kitamura and Yuji Matsumoto. 1996. Automatic extraction of word sequence correspondences in parallel corpora. In *Proceedings of the 4th Workshop on Very Large Corpora*, pages 79–87, Copenhagen, Denmark, August.

Göran Kjellmer. 1987. Aspects of English collocations. In Willem Meijs, editor, *Corpus Linguistics and Beyond*, pages 133–140. Rodopi, Amsterdam.

Göran Kjellmer. 1990. Patterns of collocability. In J. Aarts and W. Meijs, editors, *Theory and practice in Corpus Linguistics*, pages 163–178. Amsterdam.

Göran Kjellmer. 1991. A mint of phrases. In Karin Aijmer and Bengt Altenberg, editors, *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, pages 111–127. Longman, London/New York.

Göran Kjellmer. 1994. *A Dictionary of English Collocations*. Claredon Press, Oxford.

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

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

Brigitte Krenn. 2000a. Collocation mining: Exploiting corpora for collocation idenfication and representation. In *Proceedings of KONVENS 2000*, Ilmenau, Germany.

Brigitte Krenn. 2000b. *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*, volume 7. German Research Center for Artificial Intelligence and Saarland University Dissertations in Computational Linguistics and Language Technology, Saarbrücken, Germany.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 17–22, Columbus, Ohio, U.S.A.

Christopher Laenzlinger and Eric Wehrli. 1991. Fips, un analyseur interactif pour le français. *TA informations*, 32(2):35–49.

Pierre Lafon. 1984. *Dépouillements et statistiques en lexicométrie*. Slatkine – Champion, Genève/Paris.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Andrea Lehr. 1996. *Germanistische Linguistik: Kollokationen und maschinenlesbare Korpora*, volume 168. Niemeyer, Tübingen.

Michael Lewis. 2000. *Teaching Collocations. Further Developments in the Lexical Approach*. Language Teaching Publications, Hove.

Marie-Claude L'Homme. 2003. Combinaisons lexicales spécialisées (CLS) : Description lexicographique et intégration aux banques de terminologie. In Francis Grossmann and Agnès Tutin, editors, *Les collocations: analyse et traitement*, pages 89–103. Editions "De Werelt", Amsterdam.

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

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324, Morristown, NJ, USA.

Bill Louw. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 157–176. John Benjamins, Amsterdam.

Yajuan Lü and Ming Zhou. 2004. Collocation translation acquisition using monolingual corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pages 167–174, Barcelona, Spain, July.

Qin Lu, Yin Li, and Ruifeng Xu. 2004. Improving Xtract for Chinese collocation extraction. In *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pages 333–338.

Anke Lüdeling, Stefan Evert, and Marco Baroni. 2007. Using Web data for linguistic purposes. In Nadjia Nesselhauf Marianne Hundt and Caroline Biewer, editors, *Corpus linguistics and the Web*, pages 7–24. Rodopi, Amsterdam.

Elliott Macklovitch, Michel Simard, and Philippe Langlais. 2000. TransSearch: A free translation memory on the World Wide Web. In *Proceedings of the Second International Conference On Language Resources and Evaluation (LREC 2000)*, pages 1201–1208, Athens, Greece, June.

Susan Maingay and Chris Tribble. 1993. *Longman Language Activator Workbook*. Longman, Harlow, England.

Adam Makkai. 1972. *Idiom Structure in English*. Mouton, The Hague.

Mathieu Mangeot. 2006. Papillon project: Retrospective and perspectives. In *Proceedings of the LREC 2006 Workshop on Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine*, Genoa, Italy, May.

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

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

I. Dan Melamed. 1997. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Conference of the Association for Computational Linguistics (ACL'97)*, pages 305–312, Madrid, Spain.

Igor Mel'čuk et al. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches léxico-sémantiques*. Presses de l'Université de Montréal, Montréal.

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

Igor Mel'čuk. 2003. Collocations: définition, rôle et utilité. In Francis Grossmann and Agnès Tutin, editors, *Les collocations: analyse et traitement*, pages 23–32. Editions "De Werelt", Amsterdam.

Lukas Michelbacher, Stefan Evert, and Hinrich Schütze. 2007. Asymmetric association measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, Borovetz, Bulgaria, September.

Archibald Michiels. 2000. New developments in the DEFI Matcher. *International Journal of Lexicography*, 13(3):151–167.

Rosamund Moon. 1998. *Fixed expressions and idioms in English: A corpus-based approach.* Claredon Press Oxford, Oxford.

Luka Nerima, Violeta Seretan, and Eric Wehrli. 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)*, pages 131–134, Budapest, Hungary.

Michael P. Oakes. 1998. *Statistics for Corpus Linguistics.* Edinburgh University Press, Edinburgh.

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

Brigitte Orliac. 2006. Un outil d'extraction de collocations spécialisées basé sur les fonctions lexicales. *Terminology*, 12(2):261–280.

Andrew Pawley and Frances H. Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J.C. Richards and R.W. Schmidt, editors, *Language and communication*, pages 191–227. Longman, London.

Darren Pearce. 2001a. Synonymy in collocation extraction. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 41–46, Pittsburgh, U.S.A.

Darren Pearce. 2001b. Using conceptual similarity for collocation extraction. In *Proceedings of the 4th UK Special Interest Group for Computational Linguistics (CLUK4)*, pages 34–42, Sheffield, U.K.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Third International Conference on Language Resources and Evaluation*, pages 1530–1536, Las Palmas, Spain.

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

Ted Pedersen. 1996. Fishing for exactness. In *Proceedings of the South Central SAS User's Group Conference (SCSUG-96)*, pages 188–200, Austin, TX, October.

Alain Polguère. 2000. Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*, pages 517–527, Stuttgart, Germany.

P. Procter, editor. 1987. *Longman Dictionary of Contemporary English*. Longman, Harlow and London.

Antoinette Renouf and John Sinclair. 1991. Collocational frameworks in English. In Karin Aijmer and Bengt Altenberg, editors, *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. Longman, London/New York.

Philip Resnik and Aaron Elkiss. 2005. The Linguists Search Engine: An overview. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)*, pages 33–36, Ann Arbor, Michigan, June.

Julia Ritz. 2006. Collocation extraction: Needs, feeds and results of an extraction system for German. In *Proceedings of the workshop on Multi-word-expressions in a multilingual context at the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 41–48, Trento, Italy, April.

Moira Runcie, editor. 2002. *Oxford Collocations Dictionary for Students of English*. Oxford University Press.

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

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing*, pages 100–108, Pittsburgh, U.S.A.

Sabine Schulte im Walde. 2003. A collocation database for german verbs and nouns. In F. Kiefer and J. Pajzs, editors, *Proceedings of 7th Conference on Computational Lexicography and Corpus Research*, Budapest, Hungary. Research Institute for Linguistics, Hungarian Academy of Sciences.

Thierry Selva, Serge Verlinde, and Jean Binon. 2002. Le DAFLES, un nouveau dictionnaire électronique pour apprenants du français. In Anna Braasch and Claus Povlsen, editors, *Proceedings of the Tenth Euralex International Congress (EURALEX 2002)*, pages 199–208, Copenhagen, Denmark.

Gilles Sérasset. 2004. A generic collaborative platform for multilingual lexical database development. In Gilles Sérasset et al., editor, *Proceeding of the Workshop on Multilingual Linguistic Ressources (MLR2004)*, pages 73–79, Geneva, Switzerland, August.

Violeta Seretan and Eric Wehrli. 2006a. Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 953–960, Sydney, Australia, July.

Violeta Seretan and Eric Wehrli. 2006b. Multilingual collocation extraction: Issues and solutions. In *Proceedings of COLING/ACL Workshop on Multilingual Language Resources and Interoperability*, pages 40–49, Sydney, Australia, July. 2006.

Violeta Seretan and Eric Wehrli. 2007. Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, pages 401–410, Toulouse, France, June.

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

Violeta Seretan, Luka Nerima, and Eric Wehrli. 2004a. Multi-word collocation extraction by syntactic composition of collocation bigrams. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003*, Current Issues in Linguistic Theory, pages 91–100. John Benjamins, Amsterdam/Philadelphia.

Violeta Seretan, Luka Nerima, and Eric Wehrli. 2004b. A tool for multi-word collocation extraction and visualization in multilingual corpora. In *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*, pages 755–766, Lorient, France.

Violeta Seretan, Luka Nerima, and Eric Wehrli. 2004c. Using the Web as a corpus for the syntactic-based collocation identification. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1871–1874, Lisbon, Portugal, May.

Violeta Seretan. 2005. Induction of syntactic collocation patterns from generic syntactic relations. In *Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 1698–1699, Edinburgh, Scotland, July.

Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. 1997. Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 476–481, Madrid, Spain.

Max Silberztein. 1993. *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Masson, Paris.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, Montréal, Canada.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

John Sinclair. 1995. *Collins Cobuild English Dictionary*. Harper Collins, London.

Frank Smadja, Kathleen McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1):1–38.

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

Matthew Stone and Christine Doran. 1996. Paying heed to collocations. In *Proceedings of the Eighth International Workshop on Natural Language Generation*, pages 91–100, Herstmonceux, Sussex, England.

Michael Stubbs. 1995. Corpus evidence for norms of lexical collocation. In G. Cook and B. Seidlhofer, editors, *Principle & Practice in Applied Linguistics. Studies in Honour of H.G. Widdowson*. Oxford University Press, Oxford.

Michael Stubbs. 2002. *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell, Oxford.

Aristomenis Thanopoulos, Nikos Fakotakis, and George Kokkinakis. 2002. Comparative evaluation of collocation extraction metrics. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 620–625, Las Palmas, Spain, May.

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

Pim van der Eijk. 1993. Automating the acquisition of bilingual terminology. In *Proceedings of the Sixth Conference on European chapter of the Association for Computational Linguistics*, pages 113–119, Utrecht, The Netherlands, April.

Ton van der Wouden. 1997. *Negative Contexts. Collocation, polarity, and multiple negation*. Routledge, London and New York.

Ton van der Wouden. 2001. Collocational behaviour in non content word. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, pages 16–23, Toulouse, France.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Relative compositionality of multi-word expressions: A study of verb-noun (V-N) collocations. In *Natural Language Processing IJCNLP 2005*, volume 3651 of *Lecture Notes in Computer Science*, pages 553–564. Springer, Berlin/Heidelberg.

Jean Véronis and Philippe Langlais. 2000. Evaluation of parallel text alignment systems: The ARCADE project. In Jean Véronis, editor, *Parallel text processing: Alignment and use of translation corpora*, Text, Speech and Language Technology Series, pages 369–388. Kluwer Academic Publishers, Dordrecht.

María Begoña Villada Moirón. 2005. *Data-driven identification of fixed expressions and their modifiability*. Ph.D. thesis, University of Groningen.

Martin Volk. 2002. Using the Web as a corpus for linguistic research. In R. Pajusalu and T. Hennoste, editors, *Catcher of the Meaning. A festschrift for Professor Haldur im.* Publications of the Department of General Linguistics 3, University of Tartu, Estonia.

Leo Wanner, Bernd Bohnet, and Mark Giereth. 2006. Making sense of collocations. *Computer Speech & Language*, 20(4):609–624.

Leo Wanner. 1997. *Exploring lexical resources for text generation in a systemic functional language model*. Ph.D. thesis, University of the Saarland, Saarbrücken.

Eric Wehrli. 1997. *L'analyse syntaxique des langues naturelles: Problèmes et méthodes*. Masson, Paris.

Eric Wehrli. 2000. Parsing and collocations. In D. Christodoulakis, editor, *Natural Language Processing*, pages 272–282. Springer Verlag.

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

Eric Wehrli. 2007. Fips, a "deep" linguistic multilingual parser. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 120–127, Prague, Czech Republic, June.

Joachim Wermter and Udo Hahn. 2004. Collocation extraction based on modifiability statistics. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 980–986, Geneva, Switzerland.

Joachim Wermter and Udo Hahn. 2006. You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 785–792, Sydney, Australia, July.

Geoffrey Williams. 2002. In search of representativity in specialised corpora: Categorisation through collocation. *International Journal of Corpus Linguistics*, 7(1):43–64.

Hau Wu and Ming Zhou. 2003. Synonymous collocation extraction using translation information. In *Proceeding of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 120–127, Sapporo, Japan.

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*, pages 80–87, Las Cruces (New Mexico), U.S.A.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of ARPA Human Language Technology Workshop*, pages 266–271, Princeton.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)*, pages 189–196, Cambridge, MA.

Diana Zaiu Inkpen and Graeme Hirst. 2002. Acquiring collocations for lexical choice between near-synonyms. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 67–76, Philadephia, Pennsylvania.

Rémi Zajac, Elke Lange, and Jin Yang. 2003. Customizing complex lexical entries for high-quality MT. In *Proceedings of the Ninth Machine Translation Summit*, pages 433–438, New Orleans, U.S.A.

Henk Zeevat. 1995. Idiomatic blocking and the Elsewhere principle. In Martin Everaert, Erik-Jan van der Linden, André Schenk, and Rob Schreuder, editors, *Idioms: Structural and Psychological Perspectives*, pages 301–316. Lawrence Erlbaum Associates, Hillsdale, New Jersey and Hove, UK.

Heike Zinsmeister and Ulrich Heid. 2002. Collocations of complex words: Implications for the acquisition with a stochastic grammar. In *Proceedings of Workshop on Computational Approaches to Collocations (Colloc02)*, Vienna, Austria.

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

Heike Zinsmeister and Ulrich Heid. 2004. Collocations of complex nouns: Evidence for lexicalisation. In *Proceedings of KONVENS 2004*, Vienna, Austria.