Violeta Seretan

TEXT, SPEECH AND LANGUAGE TECHNOLOGY SERIES 44

Syntax-Based Collocation Extraction

Springer

Syntax-Based Collocation Extraction

Text, Speech and Language Technology

VOLUME 44

For further volumes:
http://www.springer.com/series/6636

Syntax-Based Collocation Extraction

by

Violeta Seretan
University of Geneva, Switzerland

Springer

Violeta Seretan
University of Geneva
Department of Linguistics (Office L706)
Rue de Candolle 2
1211 Geneva
Switzerland
violeta.seretan@unige.ch

ISSN 1386-291X
ISBN 978-94-007-0133-5 e-ISBN 978-94-007-0134-2
DOI 10.1007/978-94-007-0134-2
Springer Dordrecht Heidelberg London New York

To Vincenzo

Preface

This book is based on my doctoral dissertation research carried out at the Department of Linguistics, University of Geneva under the supervision of Eric Wehrli. I wish to express my heartfelt gratitude to my thesis committee members, Christian Boitet, Ulrich Heid, Paola Merlo and Jacques Moeschler as well as to the anonymous Springer reviewers for their comments and suggestions. I am especially grateful to Eric Wehrli both for his input on the dissertation and the present volume and for providing me the opportunity to carry out my work in such a stimulating environment during the eight years I spent in Geneva. My work received the support of several organisations, which are gratefully acknowledged: the Swiss Network for International Studies for sponsoring the project that formed the foundation of the present work; the Swiss National Science Foundation for financing my postdoctoral activity; and the Latsis Foundation for awarding me the 2010 Geneva University Latsis Prize for my dissertation.

I am extremely grateful to the many people that have also contributed, directly or indirectly, to making this book a reality: Anamaria Benţea, Chris Biemann, Dan Cristea, Stephanie Durrleman, Thierry Fontenelle, Nikhil Garg, Maria Georgescul, Jean-Philippe Goldman, François Grize, Ileana Ibănescu, Eric Joanis, Alexis Kauffmann, Christopher Laenzlinger, Antonio Leoni, Gabriele Musillo, Luka Nerima, Vivi Năstase, Mar Ndiaye, Simona Orzan, Olivier Pasteur, Sebastian Padó, Genoveva Puskas, Lorenza Russo, Tanja Samardžić, Yves Scherrer, Gabriela Soare, Valentin Tablan, Bernard Testa, Marco Tomassini, Lonneke van Der Plas, among many others. I also wish to thank my editor, Helen van der Stelt, for practical assistance.

I owe a special debt to Livia Polanyi, who cast an expert eye to parts of the manuscript and helped me rephrase my (often too unidiomatic or anti-collocational) turns of phrase. Finally, I wish to express my gratitude to my husband Vincenzo for his constant support, encouragement, advice and original ideas on the most diverse topics, particularly those related to NLP. This book is dedicated to him.

Geneva
June 2010

Violeta Seretan

Contents

Chapter 1
Introduction

1.1 Collocations and Their Relevance for NLP

A large part of the vocabulary of a language is made up of *phraseological units* or *multi-word expressions*, complex lexical items 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 an independent field of research (Cowie, 1998), and in computational linguistics, where growing attention is currently being paid to recognizing and processing multi-word units in various applications (Baldwin and Kim, 2010).

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 authors (Kjellmer, 1987; Howarth and Nesi, 1996; Stubbs, 1995; Jackendoff, 1997; Mel'čuk, 1998; Lea and Runcie, 2002; Erman and Warren, 2000), phraseological units, in general, and, collocations, in particular, are pervasive in texts of all genres and domains, with collocations representing the highest proportion of phraseological units.[1]

While an agreed-upon definition of collocations does not yet exist, they are generally understood as typical combinations of words that differ from regular combinations in that their components co-occur in a short span of text more often than chance would predict. Unlike idioms, collocations have a rather transparent meaning and are easy to decode. Yet, like idioms, they are difficult to encode—in fact, they are said to represent "idioms of encoding" (Makkai, 1972; Fillmore et al., 1988) since

[1] For example, it is claimed that "in all kinds of texts, collocations are indispensable elements with which our utterances are very largely made" (Kjellmer, 1987, 140); "most sentences contain at least one collocation" (Pearce, 2001a; Howarth and Nesi, 1996); "no piece of natural spoken or written English is totally free of collocation" (Lea and Runcie, 2002, vii); "collocations make up the lion's share of the phraseme inventory, and thus deserve our special attention" (Mel'čuk, 1998, 24). According to Stubbs (1995, 122), "a significant proportion of language use is routinized, conventionalized and idiomatic". Jackendoff (1997) estimates the number of phraseological units in a lexicon of the same order as the number of single words. Similarly, Erman and Warren (2000) estimate that about half of fluent native text is constructed using ready-made units.

V. Seretan, *Syntax-Based Collocation Extraction*, Text, Speech and Language Technology 44, DOI 10.1007/978-94-007-0134-2_1,

Table 1.1 Collocations across languages

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

they are unpredictable for non-native speakers and, in general, do not preserve the meaning of (all of) their components across languages. For illustration, consider the French collocations listed in Table 1.1. The verbal component of these collocations cannot be translated literally to English; a literal translation would lead to unnatural if not awkward formulations, called *anti-collocations* (Pearce, 2001a, 43). Instead, completely different verbs must be used to encode the meaning of these collocations in English.

The past decades have witnessed significant advances in the work of automatic acquisition of collocations from text corpora, most of which aimed at providing lexicographic support. Boosted by the advent of the computer era and the development of corpus linguistics, but also by the present emphasis on the study of words in context—"You shall know a word by the company it keeps!" (Firth, 1957, 179)—this work led to the development of corpus-based dictionaries including collocations; a representative example is COBUILD, the Collins Birmingham University International Language Database (Sinclair, 1995). From the list of dictionaries in Appendix A it is easy to see the recent expansion of lexicographic work devoted to collocations in many languages. As the compilation of such resources is increasingly corpus-based, automatic collocation extraction methods are being heavily used in many lexicographic projects for collecting the raw material to include in dictionaries, for validating the intuition of lexicographers, and for complementing collocation entries with additional corpus-based information such as frequency of use or usage samples.

In Natural Language Processing (NLP), collocational information derived from corpora is crucial for applications dealing with text production. For instance, collocations are considered to be not only useful, but a key factor in producing more acceptable output in machine translation and natural language generation tasks (Heylen et al., 1994; Orliac and Dillinger, 2003). The importance of collocations lies in their prevalence in language,[2] whereas the difficulty in handling them comes, principally, from their ambiguous linguistic status, their equivocal position at the intersection of lexicon and grammar, and the lack of a precise and operational definition.

There is also a marked interest in collocations in NLP from the opposite perspective, that of text analysis, where they also prove useful in a variety of tasks. For

[2] As put by Mel'čuk (2003, 26), "L'importance des collocations réside dans leur omniprésence" [The importance of collocation lies in their omnipresence].

instance, in parsing collocations have been used to solve attachment ambiguities by giving preference to analyses in which they occur (Hindle and Rooth, 1993; Alshawi and Carter, 1994; Ratnaparkhi, 1998; Berthouzoz and Merlo, 1997; Pantel and Lin, 2000; Wehrli, 2000; Volk, 2002). In word sense disambiguation, collocations have been used to discriminate between senses of polysemous words (Brown et al., 1991; Yarowsky, 1995) by relying on 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, collocations have been used in information retrieval (Ballestros and Croft, 1996; Hull and Grefenstette, 1998), text classification (Williams, 2002), and topic segmentation (Ferret, 2002). They are also helpful for a wide range of other applications, from speech recognition and OCR, where homophonic/homographic ambiguity can be resolved by taking their collocates into account (Church and Hanks, 1990), to context-sensitive dictionary look-up where collocations can be used to find the specific dictionary subentry that best matches the context of the target word (Michiels, 2000).

1.2 The Need for Syntax-Based Collocation Extraction

Traditionally, in the absence of more powerful linguistic analysis tools, collocation extraction work relies on the criterion of word proximity in text in order to identify potential candidates. Thus, in so-called *n-gram methods* collocations are modelled as sequences of consecutive words, possibly filtered according to their Part of Speech if POS tagging is available. Alternatively, in *window-based methods* they are modelled as interruptible word pairs found in a short window of text.

Although many researchers (Smadja, 1993; Heid, 1994; Pearce, 2002; Krenn, 2000a; Evert, 2004b) postulate that collocation extraction should ideally rely on the syntactic analysis of the source corpora, the syntactic structure of the source text has been taken into account in extraction work only in rare cases; mostly for languages, like German, for which traditional methods are recognized to be inefficient due to systematic long-range dependencies. Usually, syntax-based alternatives are discarded because researchers argue that large-scale parsers are unavailable for many languages and that these methods lack robustness and deliver insufficient precision on unrestricted text, as well as being computationally intensive. Even though significant advances in parsing have meanwhile been achieved, most collocation extraction work still relies on conventional syntax-free methods. However, we believe that the time is right for a methodological shift. We will demonstrate why we think so by means of a series of examples.[3]

Consider, first, the pair *[to] solve – problem* occurring in the corpus excerpt shown in Example 1.

(1) The *problem* is therefore, clearly a deeply rooted one and cannot be *solved* without concerted action by all parties.

[3] As all the examples provided thorough this book, the following are naturally-occurring (rather than invented) examples.

As illustrated by this pair, collocations allow for virtually unrestricted morphological and syntactic transformations, leading to surface realizations that are substantially divergent from the base word form and the expected word order (e.g., the verb preceding the object in an SVO language like English). Syntactic analysis is necessary to capture the collocation instances that occur in text in a different form, or for cases in which the collocated words are not found in the immediate vicinity of each other.

Furthermore, even if some pairs of words have a tendency to co-occur in the same form and within a short span of text, a syntactic analysis is required to ensure that they are actually syntactically related, and that they do not constitute extraction noise. For example, a pair like *human – organisation*, although apparently well-formed, is not a valid result if extracted from contexts like the one in (2); only *human rights* and *human rights organisation* are valid collocation candidates in this example.

(2) *human* rights *organisations*

Finally, we argue that syntactic analysis is a real necessity if the extraction results are to be used in other applications. Without explicit syntactic information, extraction results are highly ambiguous and difficult to interpret outside their original context. For instance, for a pair like *question – asked*, both a passive interpretation ("the question asked by somebody") and an active interpretation are possible ("the question asks if"). In Example 3, deep parsing is necessary to identify that *question* and *asked* are in a subject-verb relation rather than in a verb-object relation. Shallow parsers typically fail to analyse such pairs correctly because, lacking a global interpretation for the whole sentence, they may favour wrong local attachments.

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

Although it may be argued that the goal of automatic extraction is to identify collocation types rather than particular instances, we argue that it is important to identify instances accurately for several reasons: (a) most linguistic phenomena observed in a corpus are infrequent, therefore for each collocation type it is important to detect the maximum number of instances to allow for statistical inference; (b) some collocations, particularly for languages exhibiting a high degree of word order freedom, occur systematically in long-range dependencies (Goldman et al., 2001); therefore, capturing these dependencies is the only means to capture a collocation type; and (c) ignoring difficult extraction cases prevents the description of results in terms of morpho-syntactic variation potential, which is actually one of the main aims pursued in theoretical and lexicographic work devoted to collocations.

1.3 Aims

The main objective of the work described in this book is to take advantage of recent advances achieved in syntactic parsing to propose a collocation extraction methodology that is more sensitive to the morpho-syntactic context in which

collocations occur in the source corpora. Given the encouraging results obtained by syntax-based approaches to other NLP tasks—for instance, term extraction (Maynard and Ananiadou, 1999), semantic role labelling (Gildea and Palmer, 2002) and semantic similarity computation (Padó and Lapata, 2007)—we will demonstrate the extent to which such an approach is feasible and appropriate for the task of collocation extraction.

To this end, we rely on detailed syntactic information provided by a multilingual syntactic parser[4] to design an extraction method in which collocation candidates are identified in text according to their syntactic relatedness. By using the syntactic proximity criterion instead of the linear proximity criterion in choosing candidate pairs, we will show that a substantial improvement can be gained in the quality of extraction results. We test this hypothesis by evaluation experiments performed for several languages which compare the precision obtained against a traditional syntax-free method. In addition, we show 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 to rank the candidate pairs according to their likelihood to constitute collocations. We will argue that improvement in collocation identification can thus be achieved, by applying association measures on syntactically homogeneous material, and by providing these measures with accurate frequency information on pairs selected by syntax-based methods, as shown by a series of case-study evaluations in which we compare the ranks proposed by the two methods and investigate the causes that lead the syntax-free method to artificially promote erroneous pairs to high positions in the results list at the expense of interesting pairs.

In addition to collocation pairs, our work focuses on the extraction of collocations made up of more than two words. We will show that our syntax-based method which extracts binary collocations can be extended to efficiently identify complex collocations (i.e., collocations containing nested collocations, like *draw a clear distinction*, *reach a joint resolution*, *proliferation of weapons of mass destruction*). We also attempt to broaden as much as possible the set of syntactic configurations (patterns) allowed for the extraction of binary collocations, and, for this purpose, we provide a way to detect all collocationally relevant patterns in a language. In addition to patterns like verb-object or adjective-noun which are the most representative for collocations, other patterns involving functional categories are also relevant and arguably very important—for instance, patterns including prepositions, determiners, and conjunctions (compare the preposition-noun collocation *on page* with the anti-collocation **at page*).[5]

Another practical investigation described in this book is directed towards the acquisition of bilingual collocation resources for integration into a rule-based machine translation system (Wehrli et al., 2009). We propose an efficient method

[4] The parser Fips developed at the Language Technology Laboratory of the University of Geneva (Wehrli, 1997, 2007).

[5] See Baldwin et al. (2006) for a detailed account of the idiosyncratic syntax and semantics of preposition-noun expressions in particular.

for finding translation equivalents for collocations in parallel corpora, and, to this end, we employ our syntax-based collocation extraction method on both the source and target versions of the corpus.

1.4 Chapters Outline

In this chapter, we introduced *word collocation*, the central concept to which the book is devoted. We discussed the relevance of word collocation for NLP and provided arguments for a syntax-based approach to collocation extraction. We also outlined the main research directions pursued in our study. The remainder of the book is organised as follows.

Chapter 2: On Collocations

Chapter 2 looks further into the complex phenomenon of collocation and guides the reader through the maze of the numerous, and often conflicting, descriptions that have been provided in the literature on this topic. We identify the most salient defining features, which will serve as the basis for the discussion in the rest of the book.

Chapter 3: Survey of Extraction Methods

Chapter 3 sets the stage for the practical explorations described in following chapters. We start with a discussion of the extent to which theoretical descriptions have been taken into account in the practical work on collocation extraction. We then describe the basics of extraction methodologies relying on statistical association measures, discuss the role of linguistic preprocessing of source corpora, and provide an extensive review of existing extraction work.

Chapter 4: Syntax-Based Extraction

In Chapter 4, we first take a closer look at the existing syntax-based extraction work then state the specific requirements that our method satisfies. We continue by presenting the syntactic parser which is used in our work and by describing and evaluating our extraction method. We present a monolingual and a cross-lingual evaluation experiment in which we compare the syntax-based approach against the traditional syntax-free approach represented by the window method. We also provide a qualitative analysis of the results and a comparison of the two approaches at a more abstract level.

Chapter 5: Extensions

Chapter 5 extends the proposed extraction methodology in three different directions. The first two aim to ensure that a broader spectrum of collocational phenomena in the source text is covered—thus, we propose solutions for the extraction of complex collocations, and for the detection of all syntactic configurations appropriate to collocations in a given language. The third direction explores the topic of automatic acquisition of bilingual collocation correspondences. The solution proposed relies on the application of the monolingual extraction method on both the source and target versions of a parallel corpus.

Chapter 6: Conclusion

In the last chapter, we summarize the main findings of our work and point to directions for further research, including the portability of the proposed methodology to new languages and parsing tools; the exploration of the complex interplay between syntactic parsing and collocation extraction; and the use of complementary resources and tools for improving extraction results and their subsequent processing by human users or NLP applications.

Chapter 2
On Collocations

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, who actually popularized the term *collocation*, derived from the Latin word *collocare* ("to place together, to assemble"). But long before that, pedagogical studies on first and second language acquisition were already concerned with collocations, seen as language chunks which are memorized by speakers as whole units and which constitute the major means for achieving language fluency (Pawley and Syder, 1983). According to some researchers, amongst whom Gitsaki (1996), collocations have even been known and studied by the ancient Greeks.

At the beginning of the twentieth century, Harold Palmer, who pioneered the study of English as a Foreign Language (EFL), 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, to include collocational information in the dictionaries from the series that he initiated with the *Idiomatic and Syntactic English Dictionary* (1942) and that continued with *A Learner's Dictionary of Current English* (Hornby et al., 1948b), *The Advanced Learner's Dictionary of Current English* (Hornby et al., 1952), and the *Oxford Advanced Learner's Dictionary* (Hornby et al., 1948a), reprinted multiple times. This pedagogical trend was continued, most notably, by Anthony P. Cowie, Peter Howarth, and Michael Lewis. As for Lewis (2000, 173), he considers collocations as the "islands of reliability" of speakers' utterances. The recent years have shown a continued interest in studying collocations in the context of Foreign Language and Teaching (Meunier and Granger, 2008), parallelled by sustained efforts of compiling collocation dictionaries for many languages (see Appendix A).

Thus, it can be stated that collocations unveiled primarily from 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.

V. Seretan, *Syntax-Based Collocation Extraction*, Text, Speech and Language Technology 44, DOI 10.1007/978-94-007-0134-2_2,

2.2 A Survey of Definitions

The most general understanding of the term *collocation*—as introduced in the framework of contextualism or described in earlier linguistic studies—is that of a relation of *affinity* which holds between words in a language, and which is revealed by the typical co-occurrence of words, i.e., by the recurrent appearance of words in the context of each other. Contextualists consider that in characterizing a word, its context plays the most important role: "You shall know a word by the company it keeps!" (Firth, 1957, 179). Earlier, Bally (1909) used the expression "groupements usuels" ("usual phrases") to refer to words that show an affinity for each other, while preserving their autonomy: "conservent leur autonomie, tout en laissant voir une affinité évidente qui les rapproche" (Bally, 1951, 70–72).[1] In order to describe this affinity, Coseriu (1967) later used the metaphor "lexical solidarity".

The 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. As Mel'čuk (1998) points out,

> [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, these definitions remain quite vague, as 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. The confusion was only augmented by the examples provided by various researchers, which are highly inconsistent and reflect the divergence of points of view.

As stated many times in the collocation literature (Hausmann, 1989; Bahns, 1993; Lehr, 1996), the understanding of the term varied with researchers' point of view. In NLP, the precise understanding is often subject to the desired usage of collocations in an application: "the definition of collocations varied across research projects" (McKeown and Radev, 2000, 523); "the practical relevance is an essential ingredient of their definition" (Evert, 2004b, 17).

As Bahns (1993, 57) points out, "collocation is a term which is used and understood in many different ways". 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.

[1] Bally distinguishes between "groupements passagers" (free combinations), "groupements usuels" (collocations), and "séries phraséologiques" (idioms).

2.2.1 Statistical Approaches

Both the pedagogical and contextualist definitions of collocation mentioned so far imply a statistical component. In order to be acquired by speakers through memorisation in block, collocations must be identifiable on the basis of their frequency—that is, they must recur enough times to be perceived as usual word combinations. Similarly, in contextualism collocations are described in terms of typical word co-occurrence, or as words that show the "tendency to occur together" (Sinclair, 1991, 71). The notions of *frequency, typicality* or *tendency* refer to features that are modelled in statistics.

As a matter of fact, the majority of collocation definitions adopt a statistical view. Although the phenomenon described has an implicit linguistic connotation, the linguistic aspects are often ignored; thus, in the purely statistical approaches to collocations, definitions are almost exclusively given in statistical terms. For instance, Firth (1957) gives the following definition:

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

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

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 represents 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 exclusively given 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).

(6) We reserve the term *collocation* to refer to any statistically significant cooccurrence (Sag et al., 2002, 7).

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

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

> 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) co-occurs with a word that is more frequent, and *downward* collocations, in which it combines with a less frequent word (Sinclair, 1991, 116). For example, when the word *back* is examined, *back from* is an upward collocation since *from* is more frequent than *back*, while *bring back* is a downward collocation, since *bring* is less frequent than *back*.

2.2.2 Linguistic Approaches

While in the contextualist (and similar) approaches the structural relation between items in a collocation is 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. As compared to the statistical account, the linguistically-motivated one adopts a more restrictive view. In this view, collocations are seen, first of all, as expressions of a language. This account emphasizes the linguistic status of collocations, considering them as syntactically-motivated combinations; consequently, the participating words must be related syntactically. This structural condition prevails over the proximity condition requiring them to appear within a short space of each other. The definitions below (which are less popular in the NLP community) emphasize the condition that collocations are syntactically well-formed constructions:

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

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

(9) 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).

(10) 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).

(11) 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).

(12) 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 linguistic definitions of collocations has been provided by Mel'čuk in the framework of the Meaning-Text Theory, by means of the lexical functions formalism (Mel'čuk, 1998, 2003). This definition, presented later in Section 2.4.3, also considers the collocation as a syntactically-bound word combination.

In the linguistically-motivated approaches, the condition for the participating words to occur in the context of each other is no longer explicitly stated. Obviously, some proximity limitation persists, since the syntactic well-formedness criterion implies that the collocational span is the phrase, clause or, at most, the sentence containing these words. The statistical component is still present in the linguistically-motivated definitions, and is expressed, for instance, by attributes like "conventional", "characteristic", or "recurrent".

In contrast with the statistical approaches, the collocation is seen here as a directed (asymmetrical) relation, in which the role played by the participating words is uneven, and is mainly determined by the syntactic configuration of the collocation. Thus, as will be later discussed in Section 2.5.1, Hausmann (1979, 1985) and Mel'čuk (1998, 2003) use distinct terms, such as *base* and *collocate*, in order to account for the distinct role played by the items in a collocation pair. Hausmann (2004) further specifies that the base is *autosemantic*, whereas the collocate is *synsemantic*, i.e., it can only be interpreted with reference to the whole collocation. In addition, Kjellmer (1991) introduces the notion of *left* and *right predictive collocations* to indicate that an item in a collocation is predicted by the other.

Several authors have attempted to provide a more precise characterisation of collocations from a linguistic point of view, i.e., to capture their intrinsic syntactic and semantic properties in order to distinguish them from other phraseological units and to allow for their more adequate processing in various NLP applications. These attempts are reviewed later in Section 2.5.

2.2.3 Collocation vs. Co-occurrence

As indicated in Section 2.2.1, the term *collocation* has originally been used in a broad sense, for describing the general event of recurrent word co-occurrence.

This purely statistical view was later contrasted by a more restricted, linguistically-motivated view, which explicitly states that the items in a collocation are syntactically related. The second view has recently gained in popularity, and some authors have suggested to use distinct terms to distinguish between the two understandings.[2]

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-motivated approach. For example, Manning and Schütze (1999) and Evert (2004b) 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,[3] I refer to the distributional notion as *cooccurrences* (...) I reserve the term *collocation* for an intensionally defined concept (Evert, 2004b, 17).

The distinction between *co-occurrences* and *collocations* seems to be nowadays unanimously accepted (Bartsch, 2004), and will also be adopted in our work.

2.3 Towards a Core Collocation Concept

As emerges from the review in Section 2.2, a multitude of collocation definitions exist in the literature; some of the most well-known are presented in Appendix B. These are often divergent and may therefore lead to confusion, in spite of the fact that a main distinction can be drawn according to the underlying approach (i.e., a purely statistical one vs. a linguistically-motivated one). This section describes our attempt 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 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 (2004b), we consider that the variations brought to the basic definition may 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, 2004b, 17).

---

[2] For instance, Wanner et al. (2006, 611) notes that the second notion "is different from the notion of collocation in the sense of Firth (1957) (...) who define a collocation as a high probability association of lexical items in the corpus".

[3] A statistical approach, called *distributional*, and a phraseological (linguistic) approach, called *intensional* (Evert, 2004b).