15. COLLOCATIONS

Kathleen R. McKeown and Dragomir R. Radev
Department of Computer Science
Columbia University, New York
{kathy,radev}@cs.columbia.edu

This chapter describes a class of word groups that lies between idioms and free word combinations. Idiomatic expressions are those in which the semantics of the whole cannot be deduced from the meanings of the individual constituents. Free word combinations have the properties that each of the words can be replaced by another without seriously modifying the overall meaning of the composite unit and if one of the words is omitted, a reader cannot easily infer it from the remaining ones. Unlike free word combinations, a collocation is a group of words that occur together more often than by chance. On the other hand, unlike idioms, individual words in a collocation can contribute to the overall semantics of the compound. We present some definitions and examples of collocations, as well as methods for their extraction and classification. The use of collocations for word sense disambiguation, text generation, and machine translation is also part of this chapter.

15.1 Introduction

While most computational lexicons represent properties of individual words, in general most people would agree that one knows a word from the "company that it keeps" [18]. Collocations are a lexical phenomenon that has linguistic and lexicographic status as well as utility for statistical natural language paradigms. Briefly put, they cover word pairs and phrases that are commonly used in language, but for which no general syntactic or semantic rules apply. Because of their widespread use, a speaker of the language cannot achieve fluency without incorporating them in speech. On the other hand, because they escape characterization, they have long been the object of linguistic and lexicographic study in an effort to both define them and include them in dictionaries of the language.

It is precisely because of the fact that they are observable in language that they have been featured in many statistical approaches to natural language processing. Since they occur repeatedly in language, specific collocations can be acquired by identifying words that frequently occur together in a relatively large sample of language; thus, **collocation acquisition** falls within the general class of corpus-based approaches to language. By applying the same algorithm to different domain-specific corpora, collocations specific to a particular **sublanguage** can be identified and represented.

Once acquired, collocations are useful in a variety of different applications. They can be used for **disambiguation**, including both word-sense and structural disambiguation. This task is based on the principle that a word in a particular sense tends to co-occur with a different set of words than when it is used in another sense. Thus *bank* might co-occur with *river* in one sense and *savings* and *loan* when used in its financial sense. A second important application is **translation**: because collocations cannot be characterized on the basis of syntactic and semantic regularities, they cannot be translated on a word-by-word basis. Instead, computational linguists use statistical techniques applied to aligned, parallel, bilingual corpora to identify collocation translations and semi-automatically construct a **bilingual collocation** lexicon. Such a lexicon can then be used as part of a machine translation program. Finally, collocations have been extensively used as part of language **generation** systems. Generation systems are able to achieve a level of fluency otherwise not possible, by using a lexicon of collocations and word phrases during the process of word selection.

In this chapter, we first overview the linguistic and lexicographic literature on collocations, providing a partial answer to the question "What is a collocation?". We then turn to algorithms that have been used for acquiring collocations, including word pairs that co-occur in flexible variations, compounds that may consist of two or more words that are more rigidly used in sequence, and multi-word phrases. After discussing both acquisition and representation of collocations, we discuss their use in the tasks of disambiguation, translation and language generation. We will limit our discussion to these topics, however we would like to mention that some work has been done recently [3] in using collocational phrases in cross-lingual information retrieval.

15.2 Linguistic and lexicographic views of collocations

Collocations are not easily defined. In the linguistic and lexicographic literature, they are often discussed in contrast with free word combinations at one extreme and idiomatic expressions at the other, collocations occurring somewhere in the middle of this spectrum. A **free word combination** can be described using general rules; that is, in terms of semantic constraints on the words which appear in a certain syntactic relation with a given headword [12]. An **idiom**, on the other hand, is a rigid word combination to which no generalities apply; neither can its meaning be determined from the meaning of its parts nor can it participate in the usual word-order variations. Collocations fall between these extremes and it can be difficult to draw the line between categories. A word combination fails to be classified as free and is termed a collocation when the number of words which can occur in a syntactic relation with a given headword decreases to the point where it is not possible to describe the set using semantic regularities.

Thus, examples of free word combinations include *put*+[object] or *run*+[object]

(i.e. 'manage') where the words that can occur as object are virtually open-ended. In the case of *put*, the semantic constraint on the object is relatively open-ended (any physical object can be mentioned) and thus the range of words that can occur is relatively unrestricted. In the case of *run* (in the sense of 'manage' or 'direct') the semantic restrictions on the object are tighter but still follow a semantic generality: any institution or organization can be managed (e.g. *business*, *ice cream parlor*, etc.). In contrast to these free word combinations, a phrase such as *explode a myth* is a collocation. In its figurative sense, *explode* illustrates a much more restricted collocational range. Possible objects are limited to words such as *belief*, *idea*, *theory.* At the other extreme, phrases such as *foot the bill* or *fill the bill* function as composites, where no words can be interchanged and variation in usage is not generally allowed. This distinction between free word combinations and collocations can be found with almost any pair of syntactic categories. Thus, *excellent/good/useful/useless dictionary* are examples of free word adjective+noun combinations, while *abridged/bilingual/combinatorial dictionary* are all collocations. More examples of the distinction between free word combinations and collocations are shown in Table 1.

Idioms	Collocations	Free word combinations
to kick the bucket	to trade actively	to take the bus
dead end	table of contents	the end of the road
to catch up	orthogonal projection	to buy a house

Table 1: Collocations vs. idioms and free word combinations.

Because collocations fall somewhere along a continuum between free-word combinations and idioms, lexicographers have faced a problem in deciding when and how to illustrate collocations as part of a dictionary. Thus, major themes in lexicographic papers address the identification of criteria that can be used to determine when a phrase is a collocation, characteristics of collocations, and representation of collocations in dictionaries. Given the fact that collocations are lexical in nature, they have been studied primarily by lexicographers and by relatively fewer linguists, although early linguistic paradigms which place emphasis on the lexicon are exceptions (e.g. [22, 30]). In this section, we first describe properties of collocations that surface repeatedly across the literature. Next we present linguistic paradigms which cover collocations. We close the section with a presentation of the types of characteristics studied by lexicographers and proposals for how to represent collocations in different kinds of dictionaries.

15.2.1 Properties of collocations

Collocations are typically characterized as arbitrary, language- (and dialect-) specific, recurrent in context, and common in technical language (see overview by [37]). The notion of **arbitrariness** captures the fact that substituting a synonym for one of the words in a collocational word pair may result in an infelicitous lexical combination. Thus, for example, a phrase such as *make an effort* is acceptable, but *make an exertion* is not; similarly, *a running commentary*, *commit treason*, *warm greetings* are all true collocations, but *a running discussion*, *commit treachery*, and *hot greetings* are not acceptable lexical combinations [5].

This arbitrary nature of collocations persists across languages and dialects. Thus, in French, the phrase *régler la circulation* is used to refer to a policeman who *directs traffic*, the English collocation. In Russian, German, and Serbo-Croatian, the direct translation of *regulate* is used; only in English is *direct* used in place of *regulate*. Similarly, American and British English exhibit arbitrary differences in similar phrases. Thus, in American English one says *set the table* and *make a decision*, while in British English, the corresponding phrases are *lay the table* and *take a decision*. In fact, in a series of experiments, Benson [5] presented non-native English speakers and later, a mix of American English and British English speakers, with a set of 25 sentences containing a variety of American and British collocations. He asked them to mark them as either American English, British English, World English, or unacceptable. The non-native speakers got only 22% of them correct, while the American and British speakers got only 24% correct.

While these properties indicate the difficulties in determining what is an acceptable collocation, on the positive side it is clear that collocations occur frequently in similar contexts [5, 12, 22]. Thus, while it may be difficult to define collocations, it is possible to **observe** collocations in samples of the language. Generally, collocations are those word pairs which occur frequently together in the same environment, but do not include lexical items which have a high overall frequency in language [22]. The latter include words such as *go*, *know*, etc., which can combine with just about any other word (i.e. are **free word combinations**) and thus, are used more frequently than other words. This property, as we shall see, has been exploited by researchers in natural language processing to identify collocations automatically. In addition, researchers take advantage of the fact that collocations are often domain specific; words which do not participate in a collocation in everyday language often do form part of a collocation in technical language. Thus, *file* collocates with verbs such as *create*, *delete*, *save* when discussing computers, but not in other sublanguages.

Stubbs [42] points out some other interesting properties of collocations. For example, he indicates that the word *cause* typically collocates with words expressing negative concepts, such as *accident*, *damage*, *death*, or *concern*. Conversely, *provide* occurs more often with positive words such as *care*, *shelter*, and *food*.

15.2.2 Halliday and Mel'čuk

Many lexicographers point back to early linguistic paradigms which, as part of their focus on the lexicon, do address the role of collocations in language [30, 22]. Collocations are discussed as one of five means for achieving lexical cohesion in Halliday and Hasan's work. Repeated use of collocations, among other devices such as repetition and reference, is one way to produce a more cohesive text. Perhaps because they are among the earliest to discuss collocations, Halliday and Hasan present a more inclusive view of collocations and are less precise in their definition of collocations than others. For them, collocations include any set of words whose members participate in a semantic relation. Their point is that a marked cohesive effect in text occurs when two semantically related words occur in close proximity in a text, even though it may be difficult to systematically classify the semantic relations that can occur. They suggest examples of possible relations, such as complementarity (e.g., *boy, girl*), synonyms and near-synonyms, members of ordered sets (e.g., *Monday, Tuesday*; *dollars, cents*), part-whole (e..g, *car, brake*), as well as relations between words of different parts of speech (e.g., *laugh, joke*: *blade, sharp*; *garden, dig*). They point to the need for further analysis and interpretation of collocations in future work; for their purpose, they simply lump together as collocations all lexical relations that cannot be called referential identity or repetition.

In later years, Mel'čuk provided a more restricted view of collocations. In the meaning–text model, collocations are positioned within the framework of **lexical functions** (LFs). An LF is a semantico-syntactic relation which connects a word or phrase with a set of words or phrases. LFs formalize the fact that in language there are words, or phrases, whose usage is bound by another word in the language. There are roughly 50 different simple LFs in the meaning–text model, some of which capture semantic relations (e.g. the LF **anti** posits a relation between antonyms), some of which capture syntactic relations (e.g. $\mathbf{A}_0$ represents nouns and derived adjectivals such as *sun–solar*), while others capture the notion of restricted lexical co-occurrence. The LF **magn** is one example of this, representing the words which can be used to magnify the intensity of a given word. Thus, **magn**(*need*) has as its value the set of words {*great*, *urgent*, *bad*}, while **magn**(*settled*) has the value {*thickly*}, and **magn**(*belief*) the value {*staunch*}. **O***per*$_1$ is another LF which represents the semantically empty verb which collocates with a given object. Thus, the **O***per*$_1$ of *analysis* is {*perform*}.

15.2.3 Types of collocations

In an effort to characterize collocations, lexicographers and linguists present a wide variety of individual collocations, attempting to categorize them as part of a general scheme [2, 5, 12]. By examining a wide variety of collocates of the same syntactic category, researchers identify similarities and differences in their behavior, in the process coming a step closer to providing a definition. Distinctions are made between grammatical collocations and semantic collocations. **Grammatical collocations** often contain prepositions, including paired syntactic categories such as verb+preposition (e.g. *come to*, *put on*), adjective+preposition (e.g. *afraid that*, *fond of*), and noun+preposition (e.g. *by accident*, *witness to*). In these cases, the open-class word is called the **base** and determines the words it can collocate with, the **collocators**2. Often, computational linguists restrict the type of collocations they acquire or use to a subset of these different types (e.g. [11]). **Semantic collocations** are lexically restricted word pairs, where only a subset of the synonyms of the collocator can be used in the same lexical context. Examples in this category have already been presented.

Another distinction is made between **compounds** and **flexible word pairs**. Compounds include word pairs that occur consecutively in language and typically are immutable in function. Noun+noun pairs are one such example, which not only occur consecutively but also function as a constituent. Cowie [12] notes that compounds form a bridge between collocations and idioms, since, like collocations, they are quite invariable, but they are not necessarily semantically opaque. Since collocations are recursive (*ibid.*), collocational phrases, including more than just two words, can occur. For example, a collocation such as *by chance* in turn collocates with verbs such as *find*, *discover*, *notice*. Flexible word pairs include collocations between subject and verb, or verb and object; any number of intervening words may occur between the words of the collocation.

15.2.4 Collocations and dictionaries

A final, major recurring theme of lexicographers is where to place collocations in dictionaries. Placement of collocations is determined by which word functions as the base and which functions as the collocator. The base bears most of the meaning of the collocation and triggers the use of the collocator. This distinction is best illustrated by collocations which include "support" verbs: in the collocation *take a bath*, *bath* is the base and the support verb *take*, a semantically empty word in this context, the collocator. In dictionaries designed to help users encode language (e.g. generate text), lexicographers argue that the collocation should be located at the base [23]. Given that the base bears most of the meaning, it is generally easier for a writer to think of the base first. This is especially the case for learners of a language. When dictionaries are intended to help users decode language, then it is more appropriate to place the collocation at the entry for the collocator. The base–collocator pairs listed in Table 2 illustrate why this is the case.

Base	Collocator	Example
noun	verb	*set the table*
noun	adjective	*warm greetings*
verb	adverb	*struggle desperately*
adjective	adverb	*sound asleep*
verb	preposition	*put on*

Table 2: Base–collocator pairs

15.3 Extracting collocations from text corpora

Early work on collocation acquisition was carried out by Choueka *et al.* [8]. They used **frequency** as a measure to identify a particular type of collocation, a sequence of adjacent words. In their approach, they retrieved a sequence of words that occurs more frequently than a given threshold. While they were theoretically interested in sequences of any length, their implementation is restricted to sequences of two to six words. They tested their approach on an 11 million word corpus from the *New York Times* archives, yielding several thousand collocations. Some examples of retrieved collocations include *home run*, *fried chicken*, and *Magic Johnson*. This work was notably one of the first to use large corpora and predates many of the more mainstream corpus based approaches in computational linguistics. Their metric, however, was less sophisticated than later approaches; because it was based on frequency alone, it is sensitive to corpus size.

Church, Hanks, and colleagues [11, 10] used a **correlation-based metric** to retrieve collocations; in their work, a collocation was defined as a pair of words that appear together more than would be expected by chance. To estimate correlation between word pairs, they used **mutual information** as defined in Information Theory [34, 17].

If two points (words) x and y have probabilities $P(x)$ and $P(y)$, then their mutual information $I(x, y)$ is defined to be [9]:

$$I(x, y) = log_2 \frac{P(x, y)}{P(x) \cdot P(y)}$$

In the formula, $P(x, y)$ is the probability of seeing the two words x and y within a certain window.

Whenever $P(x, y) = P(x) \cdot P(y)$, the value of $I(x, y)$ becomes 0 which is an indication that the two words x and y are not members of a collocation pair. If $I(x, y) < 0$, then the two words are in complementary distribution [9]. Other metrics for computing strength of collocations are discussed in [32].

Church *et al.*'s approach was an improvement over that of Choueka *et al.* in that they were able to retrieve interrupted word pairs, such as subject+verb or verb+object collocations. However, unlike Choueka et al., they were restricted to retrieving collocations containing only two words. In addition, the retrieved collocations included words that are semantically related (e.g. *doctor–nurse*, *doctor–dentist* in addition to true lexical collocations.

Smadja and his colleagues [36, 37, 38] addressed acquisition of a wider variety of collocations than either of the two other approaches. This work featured the use of a several **filters** based on linguistic properties, the use of several **stages** to retrieve word pairs along with compounds and phrases, and an **evaluation** of retrieved collocations by a lexicographer to estimate the number of true lexical collocations retrieved.

Their system **Xtract** began by retrieving word pairs using a frequency-based metric. The metric computed the **z-score** of a pair, by first computing the average frequency of the words occurring within a ten-word radius of a given word and then determining the number of standard deviations above the average frequency for each word pair. Only word pairs with a z-score above a certain threshold were retained. In contrast to Choueka *et al.*'s metric, this metric ensured that the pairs retrieved were not sensitive to corpus size. This step is analogous to the method used by both Choueka *et al.* and Church *et al.*, but it differs in the details of the metric.

In addition to the metric, however, Xtract used three additional filters based on linguistic properties. These filters were used to ensure an increase in the accuracy of the retrieved collocations by removing any which were not true lexical collocates. First, Xtract removed any collocations of a given word where the collocate can occur equally well in any of the ten positions around the given word. This filter removed semantically related pairs such as *doctor–nurse*, where one word can simply occur anywhere in the context of the other; in contrast, lexically constrained collocations will tend to be used more often in similar positions (e.g. an adjective+noun collocation would more often occur with the adjective several words before the noun). A second filter noted patterns of interest, identifying whether a word pair was always used rigidly with the same distance between words or whether there is more than one position. Finally, Xtract used syntax to remove collocations where a given word did not occur significantly often with words of the same syntactic function. Thus, verb+noun pairs were filtered to remove those that did not consistently present the same syntactic relation. For example, a verb+noun pair that occurred equally often in subject+verb and verb+object relations would be filtered out.

After retrieving word pairs, Xtract used a second stage to identify words which co-occurred significantly often with identified collocations. This way, the recursive property of collocations noted by [12] was accounted for. In this stage, Xtract produced all instances of appearance of the two words (i.e. concordances) and analyzed the distributions of words and parts of speech in the surrounding positions, retaining only those words around the collocation that occurred with probability greater than a given threshold. This stage produced **rigid compounds** (i.e. adjacent sequences of words that typically occurred as a constituent such as noun compounds) as well as **phrasal templates** (i.e. idiomatic strings of words possibly including slots that may be replaced by other words). An example of a compound is (1), while an example of a phrasal template is (2).

(1) the Dow Jones industrial average

(2) The NYSE's composite index of all its listed common stocks fell *NUMBER* to *NUMBER*.

Xtract's output was evaluated by a lexicographer in order to identify precision and recall. 4,000 collocations produced by the first two stages of Xtract, excluding the syntactic filter, were evaluated in this manner. Of these, 40% were identified as good collocations. After further passing these through the syntactic filter, 80% were identified as good. This evaluation dramatically illustrated the importance of combining linguistic information with syntactic analysis. Recall was only measured for the syntactic filter. It was noted that of the good collocations identified by the lexicographer in the first step of output, the syntactic filter retained 94% of those collocations.

15.4 Using collocations for disambiguation

One of the more common approaches to word-sense disambiguation involves the application of additional constraints to the words whose sense is to be determined. Collocations can be used to specify such constraints. Two major types of constraints have been investigated. The first uses the general idea that the presence of certain words near the ambiguous one will be a good indicator of its most likely sense. The second type of constraint can be obtained when pairs of translations of the word in an aligned bilingual corpus are considered.

Research performed at IBM in the early 1990s [7] applied a statistical method using as parameters the context in which the ambiguous word appeared. Seven factors were considered: the words immediately to the left or to the right to the ambiguous word, the first noun and the first verb both to the left and to the right, as well as the tense of the word in case it was a verb or the first verb to the left of the word otherwise. The system developed indicated that the use of collocational information results in a 13% increase in performance over the conventional trigram model in which sequences of three words are considered.

Work by Dagan and Itai [14] took the ideas set forth in Brown *et al.*'s work further. They augmented the use of statistical translation techniques with linguistic information such as syntactic relations between words. By using a bilingual lexicon and a monolingual corpus of one language, they were able to avoid the manual tagging of text and the use of aligned corpora.

Other statistical methods for word sense disambiguation are discussed in Chapter YZX.

Church *et al.* [9] have suggested a method for word disambiguation in the context of Optical Character Recognition (OCR). They suggest that collocational knowledge helps choose between two words in a given context. For example, the system may have to chooose between *farm* and *form* when the context is either:

(3) federal . . . credit

or

(4) some . . . of

In the first case, the frequency of *federal* followed by *farm* is 0.50, while the frequency of *federal* followed by *form* is 0.039. Similarly, the frequencies of *credit* following either *farm* or *form* are 0.13 and 0.026, respectively. One can therefore approximate the probabilities for the trigram *federal farm credit*, which is $(0.5 \times 10^{-6}) \times (0.13 \times 10^{-6}) = 0.065 \times 10^{-12}$ and for *federal form credit*, which is $(0.039 \times 10^{-6}) \times (0.026 \times 10^{-6}) = 0.0010 \times 10^{-12}$. Since the first of these probabilities is 65 times as large as the second one, the OCR system can safely pick *farm* over *form* in (3). Similarly, *form* is 273 times more likely than *farm* in (4). Church *et al.* also note that syntactic knowledge alone would not help in such cases, as both *farm* and *form* are often used as nouns.

15.5 Using collocations for generation

One of the most straightforward applications of collocational knowledge is in natural language generation. There are two typical approaches applied in such systems: the use of phrasal templates in the form of canned phrases and the use of automatically extracted collocations for unification-based generation. We will describe some of the existing projects using both of these approaches. At the end of this section we will also mention some other uses of collocations in generation.

15.5.1 Text generation using phrasal templates

Several early text-generation systems used **canned phrases** as sources of collocational information to generate phrases. One of them was UC (Unix consultant), developed at Berkeley by Jacobs [25]. The system responded to user questions related to the Unix operating system and used text generation to convey the answers. Another such system was Ana, developed by Kukich [27] at the University of Pittsburgh, which generated reports of activity on the stock market. The underlying paradigm behind generation of collocations in these two systems was related to the reuse of canned phrases, such as the following from [27]: *opened strongly*, *picked up momentum early in trading*, *got off to a strong start*.

On the one hand, Kukich's approach was computationally tractable, as there was no processing involved in the generation of the phrases, while on the other, it did not allow for the flexibility that a text generation system requires in the general case. For example, Ana needed to have separate entries in its grammar for two quite similar phrases: *opened strongly* and *opened weakly*.

Another system that made extensive use of phrasal collocations was FOG [6]. This was a highly successful system which generated bilingual (French and English) weather reports that contained a multitude of canned phrases such as (5).

(5) Temperatures indicate previous day's high and overnight low to 8 a.m.

In general, canned phrases fall into the category of **phrasal templates**. They are usually highly cohesive and the algorithms that can generate them from their constituent words are expensive and sophisticated.

15.5.2 Text generation using automatically acquired collocational knowledge

Smadja and McKeown [39] have discussed the use of (automatically retrieved) collocations in text generation.

The Xtract system mentioned above (Section 15.3) used statistical techniques to extract collocations from free text. The output of Xtract was then fed to a separate program, Cook [39], which used a functional unification paradigm FUF [26, 16]) to represent collocational knowledge, and more specifically, constraints on text generation imposed by the collocations and their interaction with constraints caused by other components of the text generation system. Cook could be used to represent both compound collocations (such as *the Dow Jones average of 30 Industrial Stocks*) and predicative collocations (such as *post — gain* or *indexes — surge*).

Cook represented collocations using attribute–value pairs, such as **Synt-R** (the actual word or phrase in the entry), **SV-collocates** (verbal collocates with which the entry is used as the subject), **NJ-collocates** (adjectival collocates that can modify the noun), etc. For example, if **Synt-R** contained the noun phrase *stock prices*, some possible values for the **SV-collocates** would be *reach*, *chalk up*, and *drift*. Using such representations, Cook was able to generate a sentence such as (6).

(6) X chalked up strong gains in the morning session.

Cook's lexicalization algorithm consisted of six steps:

1. Lexicalize topic.
2. Propagate collocational constraints.
3. Lexicalize subtopics.
4. Propagate collocational constraints.
5. Select a verb.
6. Verify expressiveness.

A comparison between the representation of collocations in Ana and Cook will show some of the major differences in the two approaches: whereas Ana kept full phrases with slots that could be filled by words obeying certain constraints, Cook kept only the words in the collocation and thus avoided a combinatorial explosion when several constraints (of collocational or other nature) needed to be combined.

Another text generation system that makes use of a specific type of collocations is SUMMONS [33]. In this case, the authors have tried to capture the collocational information linking an entity (person, place, or organization) with its description (pre-modifier, apposition, or relative clause) and to use it for generation of referring expressions. For example, if the system discovers that the name *Ahmed Abdel-Rahman* is collocated with *secretary-general of the Palestinian authority*, a new entry is created in the lexicon (also using FUF as the grammar for representation) linking the name and its description for later use in the generation of references to that person.

15.5.3 Other techniques

An interesting technique used by [24] in the GOSSiP system involved the modification of the structure of a semantic network prior to generation in order to choose a more fluent wording of a concept. Their work makes of Meaning Text Theory [30], using lexical functions to represent colloctions in the generation lexicon. Lexical functions allow very general rules to perform certain kinds of paraphrases. For example, this approach allows for the generation of *Paul made frequent use of Emacs Friday.* as a paraphrase of *Paul used Emacs frequently Friday.*, where the support verb *made* and the noun collocate replace the verb *use*. Other generation systems have also explored the use of Meaning Text Theory to handle generation of collocations [31, 29].

15.6 Translating collocations

Since collocations are often language-specific and cannot be translated compositionally in most cases, researchers have expressed interest in statistical methods which can be used to extract bilingual pairs of collocations for parallel and non-parallel corpora.

Note that one cannot assume that a concept expressed by way of a collocation in one language will use a collocation in another language. Let us consider the English collocation *to brush up a lesson*, which is translated into French as *repasser une leçon* or the English collocation *to bring about* whose Russian translation is the single word verb осуществлять. Using only a traditional (non-collocational) dictionary, it is hard to impossible to find the correct translation of such expressions. Existing phraseological dictionaries contain certain collocations but are by no means sufficiently exhaustive.

Luckily for natural language researchers, there exist a large number of bilingual and multilingual aligned corpora (see Chapter XYZ). Such bodies of text are an invaluable resource in machine translation in general, and in the translation of collocations and technical terms in particular.

Smadja *et al.* [40] have created a system called Champollion[1] which is based on Smadja's collocation extractor, Xtract. Champollion uses a statistical method to translate both flexible and rigid collocations between French and English using the Canadian Hansard corpus[2]. The Hansard corpus is pre-aligned but it contains a number of sentences in one of the languages that do not have a direct equivalent in the other. Champollion's approach includes three stages:

1. Identify syntactically/semantically meaningful units in the source language.

2. Decide whether the units represent constituents or flexible word pairs.

[1] The French egyptologist Jean-François Champollion (1790–1832) was the first to decipher the ancient Egyptian hieroglyphs using parallel texts in Egyptian, demotic, and Greek found on the Rosetta stone.

[2] The Canadian Hansard corpus contains bilingual reports of debates and proceedings of the Canadian parliament.

3. Find matches in the target languages and rank them, assuming that the highest-ranked match for a given source-language collocation is its translation in the target language.

Champollion's output is a bilingual list of collocations ready to use in a machine translation system. Smadja *et al.* indicate that 78% of the French translations of valid English collocations were judged to be good by the three evaluations by experts.

Kupiec [28] describes an algorithm for the translation of a specific kind of collocation, namely noun phrases. He also made use of the Canadian Hansard corpus. The algorithm involved three steps:

1. Tag sentences in the (aligned) corpus.
2. Use finite-state recognizers to find noun phrases in both languages.
3. Use iterative re-estimation to establish correspondences between noun phrases.

Some examples retrieved are shown in Table 3. An evaluation of his algorithm has shown that 90 of the 100 highest ranking correspondences are correct.

English collocation	French collocation
late spring	fin du printemps
Atlantic Canada Opportunities Agency	Agence de promotion économique du Canada atlantique

Table 3: Translations extracted from the Canadian Hansard Corpus

A tool for semi-automatic translation of collocations, *Termight*, is described in [13]. It is used to aid translators in finding technical term correspondences in bilingual corpora. The method proposed by Dagan & Church used extraction of noun phrases in English and word alignment to align the head and tail words of the noun phrases to the words in the other language. The word sequence between the words corresponding to the head and tail is produced as the translation. *Termight* was implemented within a full editor environment which allows for practical use of its results as an aid to translators. Because it didn't rely on statistical correlation metrics to identify the words of the translation, it allowed the identification of infrequent terms that would otherwise be missed due to their low statistical significance.

The reader should not remain with the impression that French and English are the only two languages for which research in translation of collocations has been done. Language pairs involving other languages, such as Japanese, Chinese, Dutch, and German have also been investigated. Fung [19] used a pattern-matching algorithm to compile a lexicon of nouns and noun phrases between English and Chinese. The algorithm has been applied on the Hong

Kong government bilingual corpus (English and Cantonese). Fu and Xia [45] also compute a bilingual Chinese-English lexicon, although they have less of a focus on the inclusion of terms. They use Estimation-Maximization [15] to produce word alignment across parallel corpora and then apply various linguistic filtering techniques to improve the results.

Researchers have investigated translation of Japanese phrases using both parallel and non-parallel corpora. Fung [20] uses morpho-syntactic information in addition to alignment techniques to find correspondences between English and Japanese terms. Her algorithm involves tagging both the English and Japanese texts of a parallel corpus, extracting English NPs from the English side and aligning a subset of Japanese translations to the English side manually for training. Frequency information and learning through consideration of unknown words are used to produce the full lexicon. Exploratory work in the use of non-parallel corpora for translation, a potentially much larger resource than parallel corpora, exploits the use of correlations between words to postulate that correlations between words in one text are likely to also appear in the translations. Tanaka and Iwasaki [43] demonstrate how to use non-parallel corpora to choose the best translation among a small set of candidates, while Fung [20] uses similarities in collocates of a given word to find its translation in the other language.

In other work, van der Eijk [44] has compared several methods for automatic extraction and translation of technical terminology in Dutch and English. He achieves best results under the assumption that technical terms are always NPs and therefore candidate terms can be pinpointed using a combination of a pattern matcher and a part of speech tagger. Some examples of the terms retrieved by his system are shown in Table 4.

Dutch term	English term
hardnekkige weerzin	persisting aversion
vroegtijdige standaardisatie	early standardisation
wisselwerking ... produkten	inter-working ... products

Table 4: Dutch-English technical term pairs

15.7 Resources related to collocations

Two classes of resources might be of interest to researchers interested in the extraction or translation of collocations. Several dictionaries of collocations exist either on paper or in a CD-ROM format. We would like to note four such dictionaries: the *Collins Cobuild Dictionary*, the BBI *Combinatory Dictionary of English*, NTC's *Dictionary of Phrasal Verbs and Other Idiomatic Verbal Phrases*, and the *Dictionary of Two-Word Verbs for Students of English*.

Cobuild [35] is the largest collocational dictionary whose CD-ROM version gives access to 140,000 English collocations and 2,600,000 examples of how these collocations are used. The collocations and examples are extracted from the 200-million word Bank of English corpus [1]. *Cobuild* provides an on-line service, Cobuild*Direct* that provides access to both concordances and collocations from its corpus. Cobuild*Direct* is available from (http://titania.cobuild.collins.co.uk/-direct_info.html).

The BBI dictionary [4] is geared towards learners of English and focuses on lexical and grammatical collocations, including nineteen verb patterns. Since the goal of the BBI is to make it easy for learners to find collocations, collocations are placed within the entry for the base (see Section 2.4). The BBI was evaluated for ease of use through two experiments. In the first, non-native speakers were asked to fill in a missing word in a set of sentences, where each contained a blank. Their performance consistently improved when they used the BBI in the task (from 32% accuracy to 93% accuracy). In the second task, Russian speakers were given a list of Russian collocations along with the associated English collocations, each of which was missing a word. Again, accuracy improved with the use of the BBI (from 40-45% to 100%).

NTC's dictionary [41] covers 2,796 verbs and 13,870 definitions or paraphrases of their collocational usage with different prepositions. Even though the focus of this dictionary is primarily on idiomatic collocations (whose meaning cannot be extracted from the meanings of the constituent words), since its primary audience includes learners of English as a second language, it also includes a large number of commonly used collocations.

The *Dictionary of Two-Word Verbs for Students of English* [21] specializes in phrasal verbs such as *add up*, *keep on*, and *pile up*. The dictionary includes some 700 such collocations along with examples and transitivity information.

In addition to dictionaries, researchers can use the software package TACT (Text Analysis Computing Tools) which consists of 16 programs for text retrieval and analysis of literary texts, also contains a component, *Usebase* which allows retrieval of collocations from a corpus. TACT is accessible from (http://-www.chass.utoronto.ca/cch/tact.html).

15.8 Summary

While definitions of collocations have varied across research projects, the fact that they are observable in large samples of language has led to successes in their use in various statistical applications. Sophisticated approaches to collocation acquisition for representation in a lexicon now exist. These semi-automatically developed phrasal lexicons have been used for the tasks of language generation, machine translation, and to some extent to information retrieval. In addition, identification and integration of collocations and lexical context have also played a central role in tasks such as statistical approaches to word sense disambiguation.

In addition to the original research works cited in this chapter, we would like to bring to the attention of readers two overviews of collocations in [42] and [32].

References

[1] The Bank of English. http://titania.cobuild.collins.co.uk/boe_info.html.

[2] D.J. Allerton. Three or four levels of co-occurrence relations. *Lingua*, 63:17–40, 1984.

[3] Lisa Ballesteros and W. Bruce Croft. Dictionary-based methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, pages 791–801, 1996.

[4] M. Benson, E. Benson, and R. Ilson. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins, Amsterdam and Philadelphia, 1986.

[5] Morton Benson. The structure of the collocational dictionary. *International Journal of Lexicography*, 2:1–14, 1989.

[6] L. Bourbeau, D. Carcagno, E. Goldberg, R. Kittredge, and A. Polguère. Bilingual generation of weather forecasts in an operations environment. In *Proceedings of the 13th International Conference on Computational Linguistics*. COLING, 1990.

[7] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264–270, Berkeley, California, 1991.

[8] Y. Choueka, T. Klein, and E. Neuwitz. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal for Literary and Linguistic computing*, 4:34–38, 1983.

[9] K. Church, W. Gale, P. Hanks, and D. Hindle. Parsing, word associations and typical predicate-argument relations. In *Proceedings of DARPA Speech and Natural Language Workshop, (October Meeting)*. Morgan Kaufman: New York, 1989.

[10] Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 116–164. Erlbaum, 1991.

[11] Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th meeting of the ACL*, pages 76–83. Association for Computational Linguistics, 1989.

[12] A.P. Cowie. The treatment of collocations and idioms in learner's dictionaries. *Applied Linguistics*, 2(3):223–235, 1981.

[13] Ido Dagan and Kenneth Church. TERMIGHT: Identifying and translating technical terminology. In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing*, Stuttgart, Germany, October 1994. Association for Computational Linguistics.

[14] Ido Dagan and Alon Itai. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596, December 1994.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society series B*, 39:1–38, 1977.

[16] Michael Elhadad. *Using argumentation to control lexical choice: a unification-based implementation*. PhD thesis, Computer Science Department, Columbia University, 1993.

[17] R. Fano. *Transmission of Information: A statistical Theory of Information*. MIT Press, Cambridge, MA, 1961.

[18] J. R. Firth. The technique of semantics. *Transactions of the Philological Society*, pages 36–72, 1935.

[19] Pascale Fung. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, pages 236–233, Boston, Massachusetts, June 1995.

[20] Pascale Fung. *Using Word Signature Features for Terminology Translation from Large Corpora*. PhD thesis, Computer Science Department, Columbia University, New York, NY, 1997.

[21] Eugene J. Hall. *Dictionary of two-word verbs for students of English*. Minerva Books, Ltd., New York, 1982.

[22] M.A.K Halliday and Ruqaiya Hasan. *Cohesion in English*. English Language Series. Longman, London, 1976.

[23] F. J. Hausmann. Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels. In H. Bergenholtz and J. Mugdan, editors, *Lexikographie und Grammatik, Akten des Essener Kolloquiums zur Grammatik im Wörterbuch*, pages 118–129. Niemeyer, 1985.

[24] Lidija N. Iordanskaja, Richard Kittredge, and Alain Polguère. Lexical selection and paraphrase in a meaning-text generation model. In Cécile L. Paris,

William R. Swartout, and William C. Mann, editors, *Natural language generation in artificial intelligence and computational linguistics.* Kluwer Academic Publishers, July 1991. Presented at the Fourth International Workshop on Natural Language Generation. Santa Catalina Island, California, July, 1988.

[25] Paul S. Jacobs. *A knowledge-based approach to language production.* PhD thesis, University of California, Berkeley, 1985.

[26] Martin Kay. Functional grammar. In *Proceedings of the 5th Annual Meeting of the Berkeley Linguistic Society*, pages 142–158, Berkeley, CA, February 1979.

[27] Karen Kukich. *Knowledge-based report generation: a knowledge engineering approach to natural language report generation.* PhD thesis, University of Pittsburgh, 1983.

[28] Julian Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, 1993.

[29] Rita McCardell Doerr. *A Lexical-semantic and Statistical Approach to Lexical Collocation Extraction for Natural Language Generation.* PhD thesis, University of Maryland Baltimore County, Baltimore, MD, USA, 1995.

[30] Igor A. Mel'čuk and N. V. Pertsov. *Surface-syntax of English, a formal model in the Meaning-Text Theory.* Benjamins, Amsterdam/Philadelphia, 1987.

[31] Sergei Nirenburg. Lexicon building in natural language processing. In *Program and abstracts of the 15th International Conference of the Association for Literary and Linguistic Computing*, Jerusalem, Israel, 1988.

[32] Michael Oakes. *Statistics for Corpus Linguistics.* Edinburgh University Press, Edinburgh, United Kingdom, 1998.

[33] Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3), September 1998.

[34] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[35] John M. Sinclair (editor in chief). *Collins COBUILD English Language Dictionary.* Collins, London, 1987. Web site: http://titania.cobuild.collins.co.uk/.

[36] Frank Smadja. *Retrieving Collocational Knowledge from Textual Corpora. An Application: Language Generation.* PhD thesis, Computer Science Department, Columbia University, New York, NY, 1991.

[37] Frank Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, March 1993.

[38] Frank Smadja and Kathleen McKeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th annual meeting of the ACL*, pages 252–259, Pittsburgh, PA, June 1990. Association for Computational Linguistics.

[39] Frank Smadja and Kathleen R. McKeown. Using collocations for language generation. *Computational Intelligence*, 7(4), December 1991.

[40] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, March 1996.

[41] Richard Spears. *NTC's Dictionary of Phrasal Verbs and Other Idiomatic Expressions*. NTC Publishing Group, Lincolnwood, Illinois, 1996.

[42] Michael Stubbs. *Text and Corpus Analysis*. Blackwell Publishers, Oxford, United Kingdom, 1996.

[43] Kumiko Tanaka and Hideya Iwasaki. Extraction of lexical translations from non-aligned corpora. In *16th International Conference on Computational Linguistics (COLING'96)*, Copenhagen, Denmark, July 1996.

[44] Pim van der Eijk. Automating the acquisition of bilingual terminology. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, the Netherlands, April 1993.

[45] Dekai Wu and Xuanyin Xia. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213, Columbia, Maryland, October 1994.