Collocations

Kathleen R. McKeown and Dragomir R. Radev

Department of Computer Science

Columbia University

New York, NY 10027

{kathy,radev}@cs.columbia.edu

September 3, 1997

Abstract

This chapter describes a class of word groups that lies between idioms and free word combinations. Idiomatic expressions are these 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 seriosly 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, each individual word in a collocation contributes 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 disambiguation, text generation, and machine translation is also part of this chapter.

1 Introduction

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 can not 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 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 paper, we first overview the linguistic and lexicographic literature on collocations, providing a partial answer for 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 2 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.

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 [8]. 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 part 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 word 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

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: Examples of Collocations

regularities.

Thus, examples of free word combinations include *"put - [object]"* or *"run (i.e., manage) - [object]* 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/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 open collocations, a phrase such as *"explode a myth"* is a true 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 not 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 adj+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.

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 by relatively few linguists in comparison, although early linguistic paradigms which place emphasis on the lexicon are exceptions (e.g., [13, 21]). 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.

Collocations are typically characterized as arbitrary, language (and dialect) specific, recurrent in context, and common in technical language (see overview by Smadja [25]). Arbitrary 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 [3].

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 [3] 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, asking them to mark them as either American English, British English, World English, or unacceptable. The non-native speakers only got 22% of them correct, while the American and British speakers only got 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 [3, 8, 13]. 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 [13]. 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 automatically identify collocations. In addition, rsearchers 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.

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 [21, 13]. Collocations are discussed as one of five means for achieving lexical cohesion in Halliday's work. Repeated use of collocations, among other devices such as repetition and reference, is one way to produce a more cohesive text.

In Mel'čuk's meaning-text model, collocations are positioned within the framework of *lexical functions*. A lexical function (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., A_0 represents nouns and derived adjectivals such as *sun* - *solar*), while others capture the notion of restricted lexical cooccurrence. **Magn** is one example of this type of LF, 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*}. $Oper_1$ is another LF which represents the semantically empty verb which collocates with a given object. Thus, the $Oper_1$ of *analysis* is {*perform*}.

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 [1, 3, 8]. By examining a wide variety of collocates of the same syntactic categories, 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*), *adj - 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*. Often, computational linguists restrict the type of collocations they acquire or use to a subset of these different types (e.g., [7]). 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 [8] 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 [8], 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 betweeen the words of the collocation.

A final, major re-occuring 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 *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 [14]. 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 following list of base-collocator pairs illustrates why this is the case.

- noun (base) – verb (collocator)
- noun (base) – adjective (collocator)

- verb (base) – adverb (collocator)
- adjective (base) – adverb (collocator)
- verb (base) – preposition (collocator)
- verb1 (base) – verb2 in infinitive or -ing form (collocator)

3 Extracting Collocations from Text Corpora

Early work on collocation acquisition was carried out by Choueka, Klein and Neuwitz [6]. They used frequency as a measure to identify a particular type of collocation, a sequence of adjacent words. In their approach, they retrieve 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 is 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, is less sophisticated than later approaches; because it is based on frequency alone, it is sensitive to corpus size.

Church, Hanks, and colleagues [7, **?**] use a correlation based metric to retrieve collocations; in their work, a collocation is defined as a pair of words that appear together more than would be expected by chance. To estimate correlation between word pairs, they use mutual information as defined in information theory [23, 11].

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:

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

Their approach improves over that of Choeuka *et. al.* in that they can retrieve interrupted word pairs, such as subject-verb or verb-object collocations. However, unlike Choeuka, they are restricted to retrieving collocations containing only two words. In addition, the retrieved collocations include words that are semantically related (e.g., *"doctor-nurse", "doctor-dentist"* in addition to true lexical collocations.

Smadja [24, 25, 26] addressed acquisition of a wider variety of collocations than either of the two other approaches. His work features 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.

His system begins by retrieving word pairs using a frequency based metric. The metric computes the *z-score* of a pair, by first computing the average frequency of the words occuring within a ten word radius of a given word and then determing the number of standard deviations above the average frequency for each word pair. Only word pairs with a z-score above a certain threshold are retained. In contrast to Choueka's metric, this metric ensures that the pairs retrieved are not sensitive to corpus size. This step is analogous to the method used by both Choueka and Church, but it differs in the details of the metric.

In addition to the metric, however, Smadja adds 3 additional filters based on linguistic properties. These filters are used to ensure to increase the accuracy of the retrieved collocations by removing any which are not true lexical collocates. First, he removes any collocations of a

given word where the collocate can occur equally well in any of the 10 positions around the given word. This filter removes 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 adj-noun collocation would more often occur with the adjective several words before the noun). A second filter notes patterns of interest, identifying whether a word pair is always used rigidly with the same distance between words or whether there is more than one position. Finally, he uses syntax to remove collocations where a given word does not occur significantly often with words of the same syntactic function. Thus, verb-noun pairs are filtered to remove those that do not consistently occuring the same syntactic relation. For example, a verb-noun pair that occurs equally often in subject-verb and verb-object relations would be filtered.

After retrieving word pairs, Smadja uses a second stage to identify words which co-occur significantly often with identified collocations. This way, he accounts for the recursive property of collocations noted by Cowie [8]. In this stage, Xtract produces all instances of appearance of the two words (i.e., concordances) and analyzes the distributions of words and parts of speech in the surrounding positions, retaining only those words around the collocation that occur with probability greater than a given threshold. This stage produces rigid compounds (i.e., adjacent sequences of words that typically occur 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 *"the Dow Jones industrial average"* while an example of a phrasal template is *"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. Four thousand 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 collocations. This evaluation dramatically illustrates 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 lexicogrpher in the first step of output, the syntactic filter retained 94% of those collocations.

4 Using Collocations for Disambiguation

One of the 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 one 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 nineties [5] applies a statistical method using as parameters the context in which the ambiguous word appears. Seven factors are 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's a verb or the first verb to the left of the word otherwise. The system developed indicates that the use of collocational information results in a 13% increase in performance over the conventional trigram model.

Work by Dagan and Itai [9] takes the ideas set forth Brown et al.'s work further. They augment the use of statistical translation techniques with lingustic information such as syntactic relations between words. By using a bilingual lexicon and a monolingual corpus of one language,

they have been able to avoid the manual tagging of text and the use of aligned corpora.

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

5.1 Text generation using phrasal templates

Several early text generation systems use canned phrases as sources of collocational information to generate phrases. One of them is UC, or the Unix consultant, developed at UC Berkeley [16]. The system responds to user questions related to the UNIX operating system and uses text generation to convey the answers. Another such system is Ana, developed by Karen Kukich at the University of Pittsburgh [18] which generates reports of activity at the stock market. The underlying paradigm behind generation of collocations in these two systems is related to the reuse of canned phrases, such as the following from [18]:

- "opened strongly"
- "picked up momentum early in trading"
- "got off to a strong start"

On one hand, Kukich's approach is computationally tractable, as there is no processing involved in the generation of the phrases, while on the other, it doesn't allow for the flexibility

that a text generation system requires in the general case. For example, she needs to have separate entries in her grammar for two quite similar phrases: "opened strongly" and "opened weakly".

Another system that makes extensive use of phrasal collocations is FOG [4]. This is a highly successful system which generates bilingual (French and English) weather reports that contain a multitude of canned phrases such as "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.

5.2 Text generation using automatically acquired collocational knowledge

Smadja and McKeown [27] have discussed the use of (automatically retrieved) collocations in text generation.

Smadja's system, Xtract, uses statistical techniques to extract collocations from free text **??**. The output of Xtract is then fed to a separate program, Cook [27], which uses a functional unification paradigm (FUF, [17, 10]) 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 can 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 represents 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 contains the noun phrase "stock prices", some possible values for the {SV-collocates} are "reach", "chalk up", and "drift". Using such representation, Cook is able to generate a sentence such as this one [27]:

X chalked up strong gains in the morning session

Smadja's lexicalization algorithm consists of six steps:

- lexicalize topic.
- propagate collocational constraints
- lexicalize subtopics
- propagate collocational constraints
- select a verb
- 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 keeps full phrases with slots that can be filled by words obeying certain constraints, whereas Cook keeps only the words in the collocation and thus avoids a combinatorial explosion when several constraints (of collocational or other nature) need to be combined.

Another text generation system that makes use of a specific type of collocations is SUMMONS [20, 22]. 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 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.

5.3 Other techniques

An interesting technique used by Iordanskaja et al. [15] in the GOSSiP system involves the modification of the structure of a semantic network prior to generation in order to choose a more fluent wording of a concept. For example, instead of generating the generic "use compilers and editors" their approach chooses to say "run compilers and editors".

DIOGENES (McCardell 1988; Nirenburg et al. 1988) uses numerical values for constraints in order to choose one over another.

McCardell gives as an example the collocation "pitch dark" in which the use of "pitch" over "extremely" or "very" indicates the application of the function "intensity" to the word "dark".

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 (source) language will use a collocation in another (target) language. Let's 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 "os-

ushtestvljat'". 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 by no means are they sufficiently exhaustive.

Luckily for natural language researchers, there exist a large number of bilingual and multilingual aligned (i.e., each sentence in one of the languages corresponds to a known sentence in the other language) corpora. Such bodies of text are an invaluable resource in machine translation in general, and in the translation of collocations and technical terms in particular.

Frank Smadja and his collaborators [28] 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 Hansards corpus [2]. The Hansards corpus is pre-aligned but it contains a number of sentences in one of the languages that don't have a direct equivalent in the other. Champollion's approach includes three stages:

- identify syntactically/semantically meaningful units in the source language
- decide whether the units represent constituents or flexible word pairs
- 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 machine translation systems. 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.

[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 Hansards corpus contains bilingual report of debates and proceedings of the Canadian parliament.

Kupiec [19] describes an algorithm for the translation of a specific kind of collocations - namely noun phrases. He also makes use of the Canadian Hansards corpus. The algorithm involves three steps:

- tag sentences in the (aligned) corpora
- use finite-state recognizers to find noun phrases in both languages
- use iterative re-estimation to establish correspondences between noun phrases

Some examples retrieved are **Atlantic Canada Opportunities Agency - Agence de promotion économique du Canada atlantique** and **late spring - fin du printemps**. An evaluation of his algorithm has shown that 90 of the 100 highest ranking correspondences are correct.

A tool for semi-automated translation of collocations, **termight**, is described in [**?**]. It is used to facilitate translators in finding technical term correspondencies in bilingual corpora. The method proposed by Dagan and Church uses word alignment allows the identification of infrequent terms that would otherwise be missed due to their low significance.

The reader should not remain with the impression that only 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, and German have also been investigated. Fung [12] uses 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).

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 three such dictionaries: the Collins Cobuild Dictionary, the BBI Combinatory Dictionary of English (BBI) and NTC's Dictionary of Phrasal Verbs and Other Idiomatic Verbal Phrases.

Cobuild is the largest collocational dictionary [3] 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.

The BBI dictionary [2] is geared towards learners of English and focuses on lexical and grammatical collocations.

NTC's dictionary covers 2,796 verbs and 13,870 definitions or paraphrases of their collocational usage with different prepositions.

8 Conclusion

References

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

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

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

[3] http://titania.cobuild.collins.co.uk/collscd.html

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

[5] 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.

[6] 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.

[7] K. Church and P. 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.

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

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

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

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

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

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

[14] F. Haussmann. Kollokationen im deutschen woerterbuch: ein beitrag zur theorie des lexicographischen beispiels'. In H. Bergenholtz and J. Mugdon, editors, *Lexikographie und Grammatik.* Niemeyer, Turgen, FRG, 1985.

[15] Lidija Iordanskaja, Richard Kittredge, and Alain Polguère. Lexical selection and paraphrase in a meaning-text model, 1989.

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

[17] M. Kay. Functional grammar. In *Proceedings of the 5th Annual Meeting of the Berkeley Linguistic Society*, 1979.

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

[19] 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.

[20] Kathleen R. McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. In *Proceedings, 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, Seattle, Washington, July 1995.

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

[22] Dragomir R. Radev and Kathleen R. McKeown. Building a generation knowledge source using internet-accessible newswire. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, April 1997.

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

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

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

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

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

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