English Today, 40, Vol. 10, No. 4 (October 1994).

What on Earth are Collocations?

An assessment of the ways in which certain words co-occur and others do not

Thierry Fontenelle

My interest in collocations dates back to the time when I started teaching English to French-speaking advanced students. On the occasion of one of the small tests which spice every teacher's life, my students had to translate the following sentence (among many others) from French into English: "la chaleur avait fait tourner le lait". The main purpose of such an exercise was to check that the students were aware that the structure faire + infinitive in French does not necessarily correspond to *make* + infinitive in English. Among the many possibilities that had been produced, one attracted my attention immediately, namely "The heat had made the milk rotten". Since I expected something like "the heat had turned the milk" or "the heat had turned the milk sour", I concluded that the student had produced that deviant construction because he or she lacked knowledge of the appropriate adjective to render the idea of absence of freshness.

The problem was that I had to explain to the student that the sentence was syntactically (i.e. grammatically) correct and that any native speaker of English would most certainly understand the meaning of a combination of words such as "rotten milk", but that this very combination was likely to elicit some kind of mocking smile. I then started to think about the reason why we can say that an egg is rotten, bad or addled, while milk can go or turn sour and butter become rancid. The adjectives *bad*, *rotten*, *addled*, *sour* or *rancid* can all be combined with nouns denoting foodstuffs but are by no means interchangeable. This means that some words are more likely to combine with specific items to form natural-sounding combinations while other types of combinations are simply not found, even though they would be possible and understandable, at least theoretically. The aim of this article is to outline the properties of these accepted combinations and to clarify this phenomenon, generally known as *collocation*.

Basic issues

Knowing that the noun *milk* combines with the verb *turn* and the adjective *sour* is certainly most interesting, but it tells us little, if at all, about the very nature of such a combination. After all, an expression such as "to lick somebody's boots" can also be said to illustrate a particular type of combination; yet all linguists would agree that this expression is not a collocation proper, but an idiom. What characterizes idioms is the fact that they constitute a single semantic entity, and

that their meaning is not tantamount to the sum of the meanings of the words they are made up of. In the above example, there is no actual licking whatsoever and the expression is not about boots either. Moreover, idioms are not variable and cannot be submitted to various standard syntactic manipulations such as the following ones (the asterisk indicates that the sentences are not grammatical):

a) *Passivization* *The teacher's boots had been licked by one of his pupils.
b) *Pronominalization* *The student had licked my colleagues' boots but hesitated to lick mine.
c) *Cleft sentence* *It was my boots that he had tried to lick.
d) *Insertion of material* *The student tried to lick the teacher's leather boots.

Of course, the variations described above are possible if one wants to sound jocular or if the sentence is taken literally, in its non-idiomatic sense: consider (a) or (d). Moreover, there are idioms which may undergo only some transformations while excluding others (consider "bury the hatchet", which can be passivized, which indicates that, as Michiels (1977) puts it, there is a *frozenness hierarchy*). However, it is admitted that an idiom is basically a fixed multi-word unit whose meaning cannot be computed from the meanings of its components. As such, it is easy to see that the combinations "sour milk" or "the milk turned" cannot be considered as idioms since we are primarily concerned with milk. These combinations are not subject to the same types of syntactic constraints as idioms.

This is typical of what Aisenstadt (1979) calls *restricted collocations*, i.e. word combinations whose constituents are restricted in their commutability. Unlike idioms, restricted collocations do not form one single semantic unit and display some variability. Unlike totally free phrases, however, the elements they are made of are not freely interchangeable, which explains why we cannot say that the milk was rotten or that the egg was rancid or sour. As Aisenstadt notes, such collocations are not limited to adjective-noun or subject-verb combinations. We also find adverb-adjective (*stark naked*, *dead drunk*) or verb-object collocations (*to command admiration*, *to pay attention*, *to make a mistake*). Moreover, as noted by Carter and McCarthy (1988: 35), the concept of collocation is independent of grammatical categories: the relationship which holds between the verb *argue* and the adverb *strongly* is the same as that holding between the noun *argument* and the adjective *strong*.

Phraseology and language teaching

In an influential paper entitled "Words shall be known by the company they keep", Mackin (1978) tackles the problem of how to teach collocations to foreign language learners. He stresses the fundamental distinction between production

and understanding and argues that collocations do not pose any serious problems in the understanding process. Any non-native speaker is likely to recognize and understand a collocation but the converse is far from being true. Using collocations and selecting the appropriate term is much more difficult and may even be considered as one of the most serious stumbling blocks in language learning. This is also why both Aisenstadt and Mackin argue for the compilation of specialized dictionaries, since it is generally admitted that collocations cannot be accounted for in terms of grammatical rules. It is therefore natural to consider them as an element of our lexical knowledge.

Their unpredictable (linguists would say "idiosyncratic") nature makes them particularly well-suited for inclusion in a special type of dictionary designed not so much for decoding, i.e. understanding, text (like most traditional dictionaries) as for encoding text. The problem is that we first have to address the question of deciding on what elements to include in such a dictionary. Defining collocations as non-idiomatic expressions on the one hand and as non-free combinations on the other hand enables us to discard the two extremes of a continuum, but the area we are eventually left with is still too fuzzy and we immediately feel the need for further clarification.

Cowie (1986) distinguishes between *free* (or *open*) *collocations* on the one hand and *restricted collocations* on the other. The former allow substitution of either of their elements without semantic change in the other elements. For example, one can eat rice, pudding, cake or chocolate. The range of possible direct objects for the verb *eat* is practically infinite. Similarly, *eat* can be replaced by a series of synonyms such as *devour*, *munch*, *gobble* and lots of others. In "restricted collocations", one element is used in a figurative or specialized sense (one can blow a fuse, in which case the verb *blow* has acquired a figurative meaning, absent in "blow a horn" or "blow a trumpet").

In the same paper, Cowie focuses his attention on *overlapping collocations*, which are likely to pose numerous problems for the language student. For example, the verb *quench* collocates with the nouns *fire* and *thirst*. The verb *extinguish* can also be used in combination with *fire*, but it does not collocate with *thirst*. Conversely, the verb *slake* does occur in collocation with *thirst*, while *slake fire* is not an acceptable combination.

Since these collocations are hardly predictable, it is essential that they should be dealt with in specialized dictionaries. Interestingly, the collocational dictionary should provide answers to questions such as "What can be done to the noun X?" or "Which adjectives can typically modify the noun X?" This means that the basic unit the collocational dictionary starts from is most often the noun in the case of verb-noun or adjective-noun collocations. A collocational dictionary should therefore be able to inform the user that the noun *bachelor* can be used in conjunction with the adjective *confirmed*. Here, it is the noun that determines the sense of the adjective, and not the reverse. For encoding purposes, it then makes sense to put that collocation under the noun

entry since a foreign student is most likely to wonder how he or she can express that someone is a bachelor to a "high degree", for example (note that French speakers call such a person "un célibataire endurci", literally "a hardened bachelor").

Linguists and lexicographers have paid particular attention to a sub-class of restricted collocations called *delexical collocations*. They typically consist of a grammaticalized verb and a direct object. The verb belongs to a closed class including highly frequent items such as *have*, *make*, *do*, *take*, *get*, *give*, etc. Examples are "to do somebody a disservice", "to make a mistake", "to have a drink", "to make a claim", "to give a sigh" or "to take/have a bath". Following Gross (1981), many linguists call these collocations "support verb constructions" because the verb's sole role is to "support" the noun with which it co-occurs, by establishing a link between this noun and the subject of the sentence, conveying information on tense, person and aspect. Such collocations are often a nightmare to language students (just imagine the great pains students have to take with the distinction between *make* and *do* in English!).

Grammatical collocations

All the examples given in the preceding sections involve two items belonging to open (non-finite) classes, for instance a verb and a noun or an adjective and a noun. These collocations are frequently referred to as ***lexical collocations***, as opposed to ***grammatical collocations***. Unlike the former, grammatical collocations involve one element from an open class and an element from a closed class, typically, but not necessarily, a preposition. For example, the verb *depend* collocates with the preposition on and not with of. Similarly, we say 'on the stock exchange' and not 'at the stock exchange'. The very title of this paper illustrates a grammatical collocation since the noun *earth* holds a privileged relationship with the preposition *on* (unlike *hell* which collocates with the definitive article *the* or with the preposition *in* in American English; another title for this paper could then have been 'What the/in hell are collocations?' and certainly not 'What on hell…').

It should be noted that, for some linguists, grammatical collocations also include what is commonly considered to be the realm of complementation, viz. the various types of syntactic structures a given lexical item subcategorizes for. For instance, verbs such as *avoid* or *suggest* require an *–ing* from while the verbs *offer* or *decide* require a *to*-infinitive. *That*-clauses can also be part of grammatical collocations, as illustrated by the fact that an adjective such as *adamant* can be used with this type of clause (consider the sentence 'She was adamant that I should come with her').

Traditionally, British learners' dictionaries such as the *Longman Dictionary of Contemporary English*, the *Oxford Advanced Learner's Dictionary* or the *Collins Cobuild English Language Dictionary* are very good at capturing

grammatical collocations. Foreign students are used to consulting these reference works to find out, for example, which specific preposition a given lexical item requires. A sophisticated (and sometimes rather obscure) system of grammar codes is proposed to help the learner select the correct syntactic construction in which an item has to be inserted. These dictionaries, however, often provide little help when it comes to supplying a user with information on lexical collocations. As was pointed out above, the need for specialized dictionaries has led various teams of linguists and lexicographers to develop new generations of collocational resources. In the following section, I would like to introduce two such dictionaries which aim at capturing the various types of collocations I have described above.

Two dictionaries – two methodologies

The *BBI Combinatory Dictionary of English: A Guide to Word Combinations* (Benson *et al*. 1986) is a monolingual dictionary of English that is designed to tell users which words go together. For a noun entry, for example, it will give the set of verbs that take this noun as subject or as direct object, the adjectives that typically modify it or the other nouns which form with it a collocation. Like a learner's dictionary, it also mentions the various grammatical collocations the items enter. Panel 1 illustrates the entry for the noun *impression*. As can be seen, the first three categories show the verbs which collocate with this noun (from which any user should infer that 'to do an impression on someone' is out in English). The fourth category contains the adjectives that can co-occur with the noun to form lexical collocations. Finally, the fifth and sixth categories contain information on the grammatical collocations.

First specimen:
The BBI Combinatory Dictionary

impression n. 1. to create an ~ 2. to make an ~ on, upon 3. to gain an ~ 4. an accurate; deep, indelible, lasting, profound, strong, erroneous, false, inaccurate, wrong; excellent; favorable; first; fleeting; general; good; painful; personal; pleasant; unfavorable; unpleasant; vivid ~ 5. an ~ that + clause (she created the erroneous ~ that her family is wealthy) 6. under an ~ (I was under the ~ that you would come)

Second specimen: *The BBI Combinatory Dictionary*

Bee I n. ['insect'] 1. to keep ~s 2. a killer; queen; worker ~ 3. ~s buzz, hum; sting; swarm 4. a cluster; colony; swarm of ~s 5. (misc.) as busy as a ~

One first thing to note is that the dictionary contains no semantic information whatsoever. This means that a user should be ready to use an explanatory dictionary next to the BBI in order to check the meaning of the words mentioned in the entry. It is easy to see that the collocates are classified as a function of their part of speech, not of their meaning. A user who did not have sufficient knowledge of the English language would be at a loss when trying to select the adequate collocate to express a given meaning. Despite this drawback, however, it has to be admitted that the information provided by the dictionary is most often of crucial importance in an encoding task.

Panel 2 illustrates the combinatory properties of the noun *bee* in the BBI. As can be noted, special attention has been paid to capture noun-verb but also noun-noun collocations. This means that the information mentioned here enables the user to find answers to the following questions: which verbs express the typical noise made by bees? Which term refers to a 'group' of bees?

The introduction to the BBI is very interesting because the authors detail the various types of collocations they deal with in the body of their dictionary. They coin the terms *EN* and *CA* to refer to what they call *Eradication/Nullification collocations* and *Creation/Activation collocations* respectively. They usually consist of a transitive verb and a direct object and, in the former case, refer to the destruction of something (e.g. 'to abrogate/repeal/do away with a law'; 'to annul a marriage'; 'to delete/erase a file'). The latter refer to the creation or activation of something, as in 'to reach a verdict', 'to make an impression' or 'to commit suicide'. Such distinctions are extremely important for language teaching but, unfortunately, the authors do not apply them in their dictionary. Using such codes as *CA* or *EN* could have been one first step towards a more semantics-oriented lexicon, which is, I think, what foreign students most need.

The ultimate semantics-oriented collocational dictionary which covers the whole of a language has never yet been compiled, but the fragments described by Mel'čuk in his so-called *Explanatory Combinatory Dictionary of Modern French* are perhaps what comes closest to it. The *ECD*, of which three volumes have been produced, covering approximately 300 French items, is not a collocational dictionary proper. However, one of the main features of this revolutionary tool is that it includes, besides a comprehensive account of the meaning and syntactic patterns of every lexeme, a thorough description of the combinatorial (i.e. collocational) properties of words. To do so, it resorts to the

concept of *lexical function*, which can be defined as a meaning relation between a keyword and other words or phraseological combinations of words. The general form of such a function is f(X) = Y, where X is the keyword (the headword in the ECD) and Y is the collocate which has to be selected in order to express the meaning denoted by f(X). An example may be in order here to clarify this concept. The lexical function *Magn* means "very much", "to a high degree". When applied to the noun *smoker*, it yields the adjective *heavy* because a heavy smoker is basically someone who smokes a lot and *heavy* holds a privileged relationship with the noun *smoker* (other adjectives such as *big*, *high* or *great* are excluded). This relationship can be noted as follows:

Magn (smoker) = heavy

Similarly, one may say that *Magn* (bachelor) = confirmed or *Magn* (rain) = torrential.

Mel'čuk's contention is that it is possible to describe standard co-occurrence knowledge in terms of approximately 50 lexical functions, ranging from **Son** (French, to denote the typical sound or cry made by something), **Oper** (to denote the 'empty' transitive verb which takes the keyword as object) or **Liqu** (to refer to the destruction of something) to **Mult** (which denotes the multitude or aggregate of something) or Sing (the converse function which denotes a single entity). The following sets of examples illustrate the potential richness of this description:

Son (dog) = bark
(horse) = neigh
(elephant) = trumpet

Oper (pressure) = exert
(attention) = pay
(suicide) = commit

Liqu (file) = delete/erase
(law) = abrogate/abolish
(marriage) = annul

Mult (fish) = school/shoal
(bee) = swarm
(dog) = pack

Sing (glass) = splinter
(advice) = piece
(dust) = speck

Although many lexical functions are meant to capture collocational properties, some of them are actually a shorthand notation for the basic lexical-semantic relationships such as synonymy, antonymy or hyponymy. Lexical functions then cover the whole of the paradigmatic (vertical) and syntagmatic (horizontal) relations of any entry word. This makes every entry in the dictionary very lengthy and, unfortunately, hardly decipherable for the layman since the very use of the ECD presupposes profound knowledge of the sophisticated system of lexical functions.

It is common knowledge that hardly any users ever read the preface of a dictionary, apart perhaps from those rather odd and freakish linguists who spend most their time closely scrutinizing dictionaries to try to improve them. One may then reasonably wonder whether Mel'čuk's dictionaries are ever likely to become a popular tool among translators, editors, journalists and teachers, as is argued on the back cover. It is undeniable, however, that Mel'čuk's contribution is sure to pave the way for better collocational dictionaries in the future.

Collocations and computers

The advent of computer technology has made it possible to gather huge corpora, or collections of authentic texts, and to develop techniques which enable lexicographers to rely on large bodies of evidence when compiling their dictionaries instead of drawing exclusively on their intuition. Software packages have also been designed to extract from texts of several million words in length the various types of co-occurrence knowledge discussed above.

These computer programs resort to complex statistical calculations to identify collocations. The contention of many computational lexicographers is that collocations are a purely statistical phenomenon: potential collocates are found by examining and counting the *span* of words that appear to the left and the right of a given item (the so-called *node*). If some words appear more frequently together than in isolation (i.e. if the frequency of co-occurrence is greater than chance would predict), they form a significant grouping and are considered collocations.

Since it is of paramount importance to identify such combinations in constructing a lexicon for, say, machine translation or for any other automatic language processing task, it goes without saying that computerized tools for extracting, representing and manipulating collocational knowledge are likely to become the norm rather than the exception in lexicographical circles. It should be noted, however, that this alternative definition of collocations casts the net somewhat wider since there is no restriction to the number of words involved. Some authors (e.g. Smadja 1993) would even go so far as to classify as a phrase-long collocation a whole sentence such as 'The Dow Jones average of 30

industrials fell X points to Y', where X and Y are empty slots that must be filled in by a number. Other linguists (e.g. Nattinger & DeCarrico 1992) would not consider such a combination as a collocation proper, but as a lexical phrase.

The statistical approach makes it possible to discover another type of collocations, namely what has come to be known as negative polarity items. Such items can exclusively appear in a non-assertive context, e.g. the combination of the verbs *stand*, *help* or *bear* with the auxiliary *can* in 'She couldn't help laughing', or 'I can't stand being alone'. Similarly, the verb *bother* can only appear in a non-assertive context, as testified by the ungrammaticality of 'She bothered to do it'.

One implication of the use of statistical techniques is that some of the pairs of items retrieved by these systems are not considered collocations at all by some linguists. Consider for instance the pair of nouns *doctor* and *nurse* which are found to co-occur significantly in texts. Do they qualify as collocations because they appear together more often than expected, or are they simply words that are often found in the same contexts because they belong to the same subject field and their meanings are related? Both positions are defensible but they have different implications for dictionary making and for language teaching.

Conclusions

The aim of this paper was not to suggest yet another definition of collocations but rather to try to sketch the various facets of these quaint combinations of words. It should now be clear that there is no such thing as a clear, non-controversial and all-embracing definition of a collocation. This very notion should be conceived as a rather fuzzy area along a cline ranging from totally free combinations on the one hand to completely fixed multi-word units on the other.

Various factors need to be taken into account to get a clear picture of collocations, from purely linguistic factors to more statistical phenomena. It should be realized that the various competing schools would probably not reach any agreement as to the exact place a given combination of items occupies on the continuum mentioned above. If, however, I have managed to convince the reader that collocations, whatever they are, are worth investigating and that they are crucial both in practical and theoretical lexicography and in language teaching, I will have reached the goal I had set at the beginning of this paper.

References

Aisenstadt, E. 1979. "Collocability Restrictions in Dictionaries." In *ITL*, 45-46, pp. 71-74.

Benson, M., Benson, E. & Ilson, R. 1986. *The BBI Combinatory Dictionary of English*. Amsterdam & Philadelphia: John Benjamins Publishing.

Carter, R. 1987. *Vocabulary: Applied Linguistics Perspectives*. London: Unwin Hyman.

Carter, R. & McCarthy, M. 1988. *Vocabulary and Language Teaching*. Harlow: Longman.

Church, K. & Hanks, P. 1990. "Word Association Norms, Mutual Information and Lexicography." *Computational Linguistics*, Vol 16 (3), pp. 22-29.

Cowie, A. P. 1986. "Collocational Dictionaries – A comparative view." In Murphy (ed.) *Fourth Joint Anglo-Soviet Seminar*, London: British Council, pp. 61-69.

Cowie, A. P. (ed.) 1989. *Oxford Advanced Learner's Dictionary of Current English*. Oxford: OUP.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge: CUP.

Gross, M. 1981. "Les Bases Empiriques de la Notion de Prédicat Sémantique." *Langages*, 63, pp. 7-52.

Howarth, P. 1993. "A Phraseological Approach to Academic Writing". In G. Blue (ed.) *Language, Learning and Success: Studying through English*. London: Macmillan.

Mackin, R. 1978. "On Collocations: Words shall be known by the company they keep". In P. Strevens (ed.) *In Honour of A. S. Hornby*. Oxford: OUP, pp. 149-165.

Mel'čuk, I. *et al*. 1984. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches Lexico-Sémantiques*, Vol. I Montreal: Les Presses de l'Université de Montréal.

Michiels, A. 1977. "Idiomaticity in English." In *Revue des Langues Vivantes*. XLIII, 2, pp. 184-199.

Nattinger, J.R. & DeCarrico, J.S. 1992. *Lexical Phrases and Language Teaching*. Oxford: OUP.

Sinclair, J. 1987. *Collins COBUILD English Language Dictionary*. Glasgow: HarperCollins.

Smadja, F. 1993. "Retrieving collocations from text: Xtract." *Computational Linguistics*, Vol 19 (1), pp: 143-177.

Summers, D. 1987. *Longman Dictionary of Contemporary English*. Harlow: Longman.