Z A A 51.3 (2003): 235-258
©Stauffenburg Verlag

HANS-JÖRG SCHMID

Collocation: hard to pin down, but bloody useful

Abstract: In the first part of this paper the linguistic phenomenon of collocation is introduced, and its enormous frequency is illustrated. The theoretical questions involved in the definition of the notion of collocation are then discussed. The analysis reveals that on most relevant dimensions such as combined recurrence, predictability and idiomaticity, prototypical collocations have medium rather than extreme values. For example, with regard to combined recurrence, prototypical collocations are half way along the scale between free syntactic combinations and fully fixed expressions. It is argued that this 'mediocrity' of collocations is responsible for the problems that linguists have faced when trying to define the notion of collocation. On the other hand, it is the very same mediocrity that seems to render the phenomenon of collocation so useful to speakers and, as it were, languages. This emerges from a discussion of psycholinguistic, cognitive-linguistic, semantic and pragmatic aspects of collocation in the final section. Collocations are described as half-way entrenched word combinations with a half-way *gestalt* character. While this status arguably reduces the cognitive effort required for their processing, it tends to result in semantic changes and usage restrictions which correspond to their mediocre value on the scale of idiomaticity.

1. Introduction

The availability of computerized corpora and the surge of interest in corpus-linguistic methods have brought back both the phenomenon and the notion of collocation into the linguistic limelight. Despite the large number of recent publications in this area (see the Works Cited), which include explicit attempts at describing its nature (Fontenelle 1994, Bublitz 1996, Herbst 1996, Lehr 1996), the notion of collocation itself has remained somewhat elusive. The aims of this paper are

- to briefly illustrate the ubiquity of collocation (Section 2),
- to clarify what collocation is (Section 3),
- to show that the reason for the difficulty in defining collocation lies in its mediocre nature (Section 4),
- and to propose some psycholinguistic, cognitive-linguistic and pragmatic considerations explaining the utility of collocation (Section 5).

2. The ubiquity of collocation

It is one of the most widely-held assumptions in linguistics that syntax is the grammatical component responsible for the combination of words into larger units. However, looking at any text, for example the passage given below in (1), one finds that there are sequences of words that seem to cohere not only by virtue of the syntactic constructions which they instantiate, but also because their combination somehow seems familiar.

(1) **Don't cry for me Argentina**

(From the musical *Evita* by Sir Andrew Lloyd-Webber)

1 It won't be easy, you'll think it strange
2 When I try to explain how I feel
3 That I still need your love after all that I've done.
4 You won't believe me, all you will see is a girl you once knew,
5 Although she's dressed up to the night at sixes and sevens with you.

6 I had to let it happen, I had to change,
7 couldn't stay all my life down out here,
8 looking out of the window, staying out of the sun.
9 So I chose freedom, running around trying everything new,
10 But nothing impressed me at all. I never expected it, too.

11 Don't cry for me Argentina, the truth is I never left you.
12 All through my wild days, my mad existence,
13 I kept my promise, don't keep your distance.

14 And as for fortune and as for fame,
15 I never invited them in,
16 though it seemed to the world they were all I desired.
17 They are illusions, they're not the solutions they promise to be,
18 The answer was here all the time, I love you and hope you love me.

The string of words / *love you* in the last line of the text, for example, can be considered a syntactic construction consisting of the subject /, the predicate *love* and the object *you.* From the syntactic point of view the speaker is thought to be at liberty to select the next word at any given stage of the sentence at will, restrained only by grammatical rules. John Sinclair (1991: 109) has called this view of language the *open-choice principle.*

On the other hand, it seems that the words *I love you* cohere independently of possible syntactic relationships, simply because they are used so frequently together - at least in this text-type. The probability that the words *I love* will be followed by the word *you* is certainly much higher than would be predicted by the *open-choice principle.* This is carried to the extreme in the sequence at the end of the first stanza. Even if Madonna, at the shooting of the film *Evita,* had failed to remember her text after the words *dressed up at sixes and [...],* she would have known simply by her knowledge of the English language that it could only continue with the word *sevens.* At this point, then, the choice of words is anything but free, because *at sixes and sevens* is a fixed expression or idiom. Linguistic phenomena of this kind have led Sinclair (1991: 109-15) to suggest the so-called *idiom-principle* as a counterpart to the *open-choice principle.* The idiom-principle postulates that language users have a large number of pre-fabricated phrases at their disposal which they use in the production of speech as building-blocks that are larger than words. The difference between the two principles proposed by Sinclair is illustrated in (2):

(2) *Open-choice principle* vs. *Idiom-principle*

all *the* *time* *I* *love* *you* *all the time* *I love you*

The open-choice and the idiom-principle must be seen as complementary rather than mutually exclusive principles. Real idioms like at *sixes and sevens* are very extreme manifestations of the idiom-principle, because they are more or less frozen expressions (cf. e.g. Fraser 1970 and Lipka 1974: 280 for the term *frozen),* which lose their composite meaning when they are altered. Much more frequent than idioms are less rigidly fixed combinations of words which are nevertheless to a large extent predictable: The use of one member of such pairs or sets will allow native speakers of the language to anticipate the use of one or several others. Without yet attempting to give a more precise definition at this stage, I will from now on refer to such predictable combinations of words as *collocations,* and to the linguistic phenomenon as such as *collocation.*[1] The words making up collocations will be referred to as *collocates.*[2]

Possible candidates for collocations in the *Evita* text are the following familiar-sounding word combinations:

– adjective + noun	*wild days* (l.12), *mad existence* (l.12)
– verb + noun	*kept (my) promise* (l.13), *keep (your) distance* (l.13)
– noun (phrase) + verb	*the truth is* (l.11)
– verb + particle	*try to* (l.2), *promise to* (l.17), *dressed up* (l.5) *looking out* (l.8), *staying out* (l.8), *running around* (l.9), *cry for* (l.11), *invited (them) in* (l.15)
– multi-word function words	*at all* (l.10), *as for* (l.14)
– quantifier + determiner + noun	*all the time* (l.18)

[1] It should be emphasized that this conception of the notion of ***collocation*** includes a reference to predictability (for more details see Section 3.4 below). The notion of collocation has also been defined in a more general way, simply as the co-occurrence of two or more words in a text (cf. e.g. Jones & Sinclair 1974: 19, Sinclair 1991: 170, Lehr 1996: 1).

[2] Note that the term ***collocate*** is used here to refer to all words that are part of a collocation. There is a long tradition in collocation studies, which can be traced back as far as Catford (1965: 10) and Sinclair (1966: 415), of distinguishing between the so-called ***node*** of collocations, the word from whose perspective the collocation is being looked at on a particular occasion, and its ***collocates.*** While this is more or less simply a practical distinction (one word can be the node in one analysis and a collocate in another), Hausmann (1984: 401; 1985: 119) mounts the theoretical claim that collocations have a hierarchical internal structure, with one partner, the base (G. ***Basis),*** determining the other, the collocate (G. ***Kollokator).*** Lipka (2002: 181), on the other hand, argues that the term collocation is "neutral with respect to which element is primary or dominant in the relation."

Other frequent types of collocations which are not in evidence in (1) are combinations of verbs and adverbs (e.g. *complain bitterly, amuse thoroughly)* and combinations of adverbs and adjectives (e.g. *sound/fast asleep, closely/intimately acquainted).* In addition, there is a large number of *lexical bundles* (Biber et al. 1999: 990-1024), i.e. common combinations of words regardless of their syntactic structures, such as *the thing is, and things like that, I think so, I don't know, you know what I mean, go and get, what a shame* and so on.[3]

This small text sample should suffice as an illustration - though clearly of course not as evidence - of the claim that collocation is a ubiquitous linguistic phenomenon. This has not remained unnoticed, of course. In fact, as early as 1966 - fairly soon after the concept of collocation had been introduced into linguistics by J. R. Firth in 1951[4] - Halliday and Sinclair began to suggest that collocation be treated as a linguistic level in its own right, side by side with the syntactic level. The implication of this claim is that collocation should be considered part of the systemic component of languages, the *langue* in Saussurean terms, rather than a matter of actual speech. Coseriu (1967: 11) explicitly avoids placing collocation in either of these categories and suggests that it is part of what he calls *norm,* which should be established somewhere in between *langue* and *parole.*

Despite this tradition, linguistic research on collocation, especially in the theoretical domain, is to a large extent still in its infancy, because the phenomenon has proven to be extremely elusive. How can this be accounted for? Many linguists who have tackled the problem of defining collocation have claimed that the difficulty lies in the fact that collocation is, as Herbst puts it, "a classic example of gradience" (1996: 385; see also Stubbs 1995a: 387 and Bublitz 1996: 21). However, as the application of prototype theory to linguistic notions has taught us (cf. e.g. Taylor 1995: chs. 8-12), this is true of linguistic terminology more or less in general; most technical terms in linguistics capture a range of slightly different phenomena and are therefore best defined with reference to prototypical examples. Nevertheless, one does not have the feeling that notions like *word, sentence* or *subject* are especially elusive.

Defining *collocation,* on the other hand, is more challenging than pinning down these notions. The next section of this paper has a two-fold function: It will be discussed how collocation should be defined and demonstrated why it is so difficult to do so. One reason why definitions of collocation given in previous publications on the topic are on the whole less clear than is desirable is that not all criteria are explicitly mentioned but some are presupposed because they seem so obvious. I will try to avoid this pitfall by starting from scratch and proceeding slowly and carefully in a step-by-step fashion.

The most comprehensive survey of types of collocations in English I am aware of - which does not include ***lexical bundles*** in the sense mentioned above, however - is given in Benson, Benson & Ilson (1997). Other dictionaries of collocations are Cowie & Mackin (1975), Cowie et al. (1983) and Kjellmer (1994).

Bublitz (1996: 2, Fn.1) mentions that Firth borrowed the term from H.E. Palmer, but he gives no references for this claim.

3. A step-by-step definition of collocation

3.1 More than one word

The first criterion needed for a definition of *collocation* is self-evident: At least two words must be involved. One word alone cannot form a collocation, because there is nothing for it to collocate with. Collocations can of course consist of more than two words, although it is probably true to say that both the majority and the prototypical instances of collocations are two-word combinations.

3.2 Adjacency

Second, the words in question should be adjacent, as has been the case in most of the examples given so far. Problems arise with collocations like *keep one's promise* and *keep one's distance* in the *Evita*-text, in which a collocation between the verb *keep* and the nouns *promise* and *distance* can be suspected, although they are not adjacent. This problem will be dealt with in the later parts of the next subsection.

3.3 Combined recurrence

Words that are adjacent in a given text are only eligible for the status of collocations if they do not occur next to each other by mere chance but because they are frequently used in this particular combination. In short, collocations should be recurrent word combinations. As mentioned in Section 2, there are linguists who do not apply this criterion but define *collocations* simply as combined occurrences (rather than recurrent combinations). Like Herbst (1996: 383), I believe that definitions of this type somehow miss the point. For one thing, the *langue*-related relevance of collocations is not captured by such definitions - even though it can of course be included by introducing subcategories such as *usual* and *unusual* collocations. And second, it is not clear what is gained by calling co-occurrences of words *collocations,* when the term *combination,* or indeed *co-occurrence* itself, covers the same range of phenomena.

This does not mean that the idea of recurrence is unproblematic, though. Theoretically clear and plausible as this criterion is, it is also extremely difficult to operationalize. In the provisional assessment of potential collocations in the *Evita*-text given above I relied on my intuition with respect to whether they were recurrent or not. But this subjective method is of course completely inadequate. Fortunately, however, since the advent of corpus linguistics research has benefited from more objective possibilities of dealing with the question of recurrence in word combinations.[5]

One of the larger corpora is the *British National Corpus* (see Leech 1993 for a description), a comparatively balanced corpus covering a wide range of different registers and text-types. It consists of 90 million words of written texts and 10

On computerized corpora and their use in collocation studies, see e.g. Sinclair (1991), Clear (1993), Smadja (1993), Stubbs (1995a and 1995b), Biber (1996), Bublitz (1996,1998), Esser (1999).

million words of transcribed spoken conversation. The BNC was used to check the combined recurrence of some of the potential collocations of the *Evita-text* by searching for frequencies of a key word of the collocation. The results of these queries are summarized in (3).

(3) Frequency counts in the *British National Corpus*

Single word	Frequency in BNC	Potential collocation	Frequency in BNC
love (VB)	4,074	*I love you*	712
truth	8,250	*the truth is*	427
days	33,071	*wild days*	5
existence	6,592	*mad existence*	2
distance	6,829	*keep your distance*	12
promise (N)	2,325	*kept my promise*	3
dressed	3,747	*dressed up*	297
looking	26,638	*looking out*	539
promise (VB)	762	*promise to*	598
as	670,445	*as for*	681
for	899,331	*as for*	681

My intuitions as to most of the potential collocations in the lyrics of *Don't cry for me Argentina* are confirmed by these findings. For example, no fewer than 712 out of the 4,074 examples of the base-form of the verb *love* are instances of the combination *I love you.* This amounts to a proportion of almost 20 per cent. Not surprisingly, 598 instances of the total number of 762 of the base-form of the verb *promise* occur in the collocation *promise to,* a proportion of roughly three quarters. Even though for most other potential collocations both the absolute and the relative number of occurrences are much lower,[6] all proved to be recurrent in a 100-million word sample of English. (Perhaps somewhat surprisingly, the expressions *wild days* and *mad existence* can muster no more than 5 and 2 occurrences respectively, and in each case one of the occurrences is in fact the line from the song.)[7]

It is tempting to lean back complacently at this point and regard the problem of defining the notion of *collocation* as being solved. Collocations would thus be definable as recurrent combinations of two or more adjacent words; the question as to whether a combination is recurrent or not could be decided by consulting a large corpus. The only thing to be determined would be how often a combination of words would have to recur for it to be accepted as a collocation. The right choice for such a threshold clearly depends on the size of the corpus used.[8]

The likelihood that two or more words occur next to each other in a corpus is contingent; it depends on the total number of words in the corpus and on the number of occurrences of the single words (highly frequent words are more likely to occur next to any other word than very rare ones). For more details, see Church & Hanks (1990), Clear (1993), and Stubbs (1995b), who argues that raw frequencies are worth looking at despite the obvious need for relative scores (39-40).

It should be mentioned that there is of course a much more interesting and illuminating way of checking the recurrence of a potential collocation in a corpus than a simple frequency count: looking at so-called KWIC-concordances (Key Word In Context) which list the manifestations of potential collocations in the corpus in their immediate contexts (see e.g. Sinclair 1991: 32-35, 150-53).

Unfortunately, however, this definition is still not as straightforward as it sounds. One problem is related to the theoretical status of the phenomenon of collocation. Clearly, if collocations are actually regarded as belonging to the system of a language, or at least to Coseriu's *norm,* one should not only look for recurrent combinations of *word-forms* as parts of actual speech, but for combinations of *lexemes.*[9] The ideal collocational corpus query would thus not consist of specific word-forms like *kept (my promise),* but of lemmas such as KEEP as standing for the forms *keep, keeps, keeping* and *kept.* Fortunately, the programmers of corpus query languages have found ways of solving this problem and offer users the possibility to retrieve all forms of a lexeme wholesale, as it were, just by keying in the baseform and some additional symbol.

The second problem, which has already been mentioned (see 3.2), concerns collocations whose parts are not adjacent. Among the collocations from the *Evita*-text *keep your distance* and *kept my promise* can serve as examples of such discontinuous collocations. Intuitively thinking, one might assume that these collocations are more frequent than the scores in (3) suggest, but due to the words in between the collocates it is difficult to locate them automatically. This can be solved by using wild-cards in corpus queries, for example in order to search for all instances of *keep* and *distance* with one word intervening. The results for the two improved corpus queries are given in (4).

(4)		
	keep/keeps/keeping/kept _ distance:	127 occurrences in BNC
	keep/keeps/keeping/kept _ promise:	85 occurrences in BNC

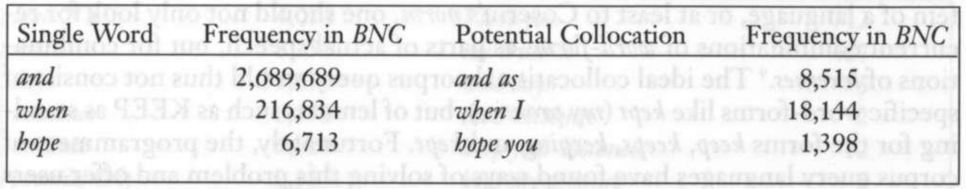
However, it would not be uncommon for even more words to intervene between two potential collocates, as for instance in *he kept his terrible promise that [...].* Even for this problem corpus linguists have found a solution: They do not confine their interest to those words immediately to the right and left of the node, but take the next three or even five words to the right and left into account as well. If another word is significantly more frequent within this span of the word

Kjellmer (1982: 26), for example, uses of a threshold of "more than once" in the one-million word Brown Corpus, and Clear (1993: 277) three occurrences in a corpus of 25 million words. For Jones & Sinclair collocations are recurrent when they occur together "more often than their respective frequencies and the length of text in which they appear would predict" (1974: 19). See also Stubbs (1995a) and Bublitz (1996: 6) on this question.

Lehr proposes the following terminological reflection of this distinction: "Um im weiteren zwischen aktualen Kollokanten als Phänomenen der syntagmatischen Ebene und virtuellen Kollokanten, die als Stellvertreter für eine ganze Reihe von als identisch angesehenen, syntagmatischen Kollokanten fungieren, differenzieren zu können, muß eine im Kontextualismus nicht vorgesehene terminologische Unterscheidung getroffen werden: Virtuelle Kollokanten, die Abstraktionen über einem oder mehreren, in singulären Sprachereignissen existenten Kollokanten sind, sollen von nun an ***Kollokanteme (Kntm)*** heißen." (Lehr 1996: 40) in question, it is also considered to enter into a collocation with it.[10] This helps to identify discontinuous collocations. The strength of the collocational ties between words within a certain span is given by means of such statistical measures as *T-score* and *Mutual Information score}'*

A third source of complication concerns the practical application of the notion of recurrence. Consider the frequency scores given in the table in (5).

(5)



Single Word	Frequency in *BNC*	Potential Collocation	Frequency in *BNC*
and	2,689,689	*and as*	8,515
when	216,834	*when I*	18,144
hope	6,713	*hope you*	1,398

On the basis of the data in table (3) above, it was assumed that the combination *as for* could pass for a collocation, because it occurred 681 times in the corpus. However, the combination *and as,* which is found in the same place in the *Evita*-text, is not exactly infrequent in the corpus either, occurring no fewer than 8,515 times. This is largely due to the fact that with a frequency of over two-and-a-half million, *and* is likely to occur much more often before and after any other word whatsoever - and still more often, of course, in a 4:4-span around a word - than words which are less frequent. Still higher frequencies can be observed for other combinations from the *Evita*-text, which seem to be equally accidental and are therefore not at all manifestations of *collocation* as this notion is understood here. So the statistical criterion of recurrence is not the final solution to the problem of defining the concept. Herbst does seem to be right when he claims that

> any attempt to define collocation in this narrow sense [i.e. with reference to statistical significances, HJS] can thus only be aiming at defining a kind of prototype of collocation, recognizing the gradience character of the distinction between collocation and free combination. (Herbst 1996: 385)

3.4 Mutual expectancy and predictability

Basically, there are three ways out of the dilemma into which the notion of recurrence has manoeuvred us. First, recurrent combinations of the type *and as* can simply be regarded as collocations as well, an approach which recommends itself when the ultimate aim is to preserve objectivity in the definition of collocation.

[10] As for the size of the span, it has been argued that a window of four words to the left and the right of the node tends to be sufficient for the discovery and proof of collocations (Jones & Sinclair 1974: 21; Sinclair 1991: 170). As Bublitz (1996: 6, 14-15) has shown, however, the size of the span depends on the collocations investigated.

[11] See Church & Hanks (1990), Clear (1993) and Biber (1996) for descriptions of useful computational tools and statistical measures, and Stubbs (1995b) for a critical survey.

However, this would extend the concept of collocation to comprise something which should not really fall into it, for it does not capture the experimentally testable fact (see this section further down) that some co-occurrences of words are not only chance products but 'belong together' for some reason or other. It is for these that the notion of collocation should be reserved.

A second option is to stipulate that the parts of a collocation must also be parts of a grammatical structure (cf. e.g. Kjellmer 1982: 26). This view is also problematic because it is not clear what sort of grammatical structure should be required. The words *they* and *are* in *they are illusions,* for instance, are obviously linked grammatically, but it does not seem sensible to consider them collocations. If'grammatical structure' were confined to units smaller than clauses such as phrases, prime examples of collocations like those consisting of verbs and objects (e.g. *commit a crime, hatch a plot)* would be excluded, unless one regards them as constituting one verb phrase. But even then there would still be difficulties with rather typical collocations of the type *dog - bark* und *horse - neigh,* which represent subjects and verbs in clauses.

A third, more promising, way out of the dilemma was already envisaged by Firth a long time before the problem with the operationalization of recurrence could begin to bother linguists, because at that time they had not yet begun to work with computers and had not even tried to prove recurrence objectively.[12] In his first rather vague remarks on what collocation was and to what extent it was relevant for linguistics Firth wrote that collocation had to do with "mutual expectancies of words" (1957: 195). Applying this criterion, one will be able to exclude such combinations as in table (5) from the notion of collocation, even though they are recurrent as far as objective statistics are concerned. Clearly the words in question do not 'expect each other.'

The notion of mutual expectancy has been re-interpreted as *predictability* by other linguists, for example by Greenbaum (1970), Bublitz (1996) and Herbst (1996: 389). The idea is that when two words 'mutually expect each other,' native speakers of the language will be able to predict with some degree of certainty the occurrence of one word when they encounter the other. In a way, predictability is thus the pragmatic counterpart to mutual expectancy: The former looks at word combinations from the language users' perspective and the latter from the language-immanent perspective of the words themselves.[13]

Mutual expectancy and predictability are highly relevant and valuable criteria because they seem to capture the psychological essence of the phenomenon of collocation, viz. the associative relation between syntagmatically related words (Lipka 2002: 181-82). The snag about these notions is that they are highly subjective and of little reliability. So we are almost back to square one, since the ad-

It should be added, however, that not least due to their interest in collocations, Halliday (1966: 159) and Sinclair (1966) were among the first linguists who envisaged the creation of large-scale corpora as a basis for linguistic research.

That the words making up collocations are in fact predictable is also shown by the fact that deliberate, creative floutings of collocations - and even of less strong, mainly evaluative associations between words called semantic prosodies - can be exploited to create irony and other types of non-literal meaning (cf. Louw 1993 and Bublitz 1996: 25-26).

vantages of corpus examination concerning objectivity are lost if we apply these criteria. It is true that in principle, mutual expectancy and predictability can be verified by experiments, for example by completion tests and association tests. In the former type, a classic application of which is Greenbaum's (1970) investigation of the relation between intensifying adverbs and the choice of verbs,[14] participants are confronted with beginnings or other parts of sentences and asked to supply likely continuations. For example, when asked to continue the words *I badly [...],* 65 per cent of Greenbaum's subjects responded with the verb *need* and 28 per cent with *want* (1970: 36). In association experiments subjects are presented with a stimulus word and asked to respond with the first word(s) that spring to their minds (cf. Aitchison 1994: 24, 84-86). Collocates are among the most frequent responses is such experiments, for example *dark* as a response to *night,* or *water* as a response to *salt.* This result is interpreted as suggesting that there is a fairly close associative network in our minds for words forming collocations. The degree to which words mutually expect each other is measured in terms of the frequency with which particular words are named and the speed of the response - the rationale being that frequent and rapidly produced words reflect entrenched patterns of spreading activating in the mental lexicon.[15] The problem with all experimental tests of collocations, however, is that they require a lot of effort while being restricted to relatively small sections of the lexicon.

3.5 Relations between the criteria

At this stage the proposed definition of collocation consists of four criteria:

1. the requirement that two or more words are involved;
2. the adjacency of these words, at least within a certain span;
3. their combined recurrence;
4. and their mutual expectancy or predictability.

As a next step, it is vital to recognize that the third criterion, the tendency to recur, and the fourth, mutual expectancy/predictability, are not independent of each other. In fact, there is a relation of direct proportionality between the two criteria: A collocation will be the more predictable the more frequent it is. This relationship between the two dimensions is represented as a diagram in Figure 1, where the black bar represents the extension of collocation. Out of the whole population of possible word combinations which are charted by the two axes, it captures those which show a tendency to occur together and mutually expect each other to roughly the same extent. It must be added that there are different reasons why the areas outside the black bar do not fall under the notion of collocation. The area on the right-hand side below the black bar - highly recurrent but hardly predictable combinations - misses out on the criterion of predictability. It is here that frequent but adventitious combinations like *and as, when I* or *hope you* (see (5) above) are located. The area on the left-hand side above the black bar, on the other hand, is not really excluded from the domain of collocation by a definitorial criterion; it is just not very likely that combinations of this type exist, because if a combination is not recurrent, it can hardly be predictable. Possible combinations of this type are obsolete collocations, which are no longer recurrent but still predictable for a small section of the speech community. Attention must also be drawn to the fact that the black area does not reach down to the origin of the co-ordinate system. This is because combinations with a very low tendency of recurrence, whose predictability will consequently also be low, are not collocations, but free combinations, i.e. accidental, syntactically motivated co-occurrences.

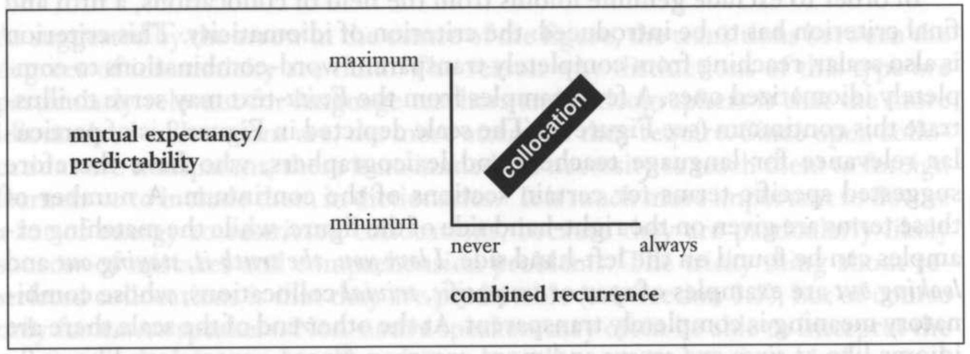


Fig. 1: The relation between mutual expectancy/predictability and recurrence

In the right-hand top corner of the co-ordinate system there are word combinations which always, or almost always, co-occur and are thus highly predictable. At this end one must also be very cautious because additional aspects come into play here. On the one hand, combinations of lexical and grammatical items like *promise to* or *try to* are located in this area, which I would certainly like to treat as collocations. These are so-called *grammatical collocations* (Benson, Benson & Ilson 1997: XV-XXIX), which straddle the boundary between lexical and grammatical relations. Combinations with a very high degree of predictability consisting of lexical words only, on the other hand, show a tendency that has so far only been mentioned in passing, the tendency towards idiomatization. This means that there is a high probability that the composite meaning of such combinations is not equal to the sum, or some sort of conflation, of the meanings of its components but more than that, or even something completely different

[14] On verb-intensifier collocations, see also Greenbaum (1988) and Bublitz (1998), on adjective intensification Lorenz (1999).

[15] Herbst (1996:384) discusses the problem of whether such collocations are linguistic phenomena at all, or whether the association of ***dark*** with ***night*** "must be attributed to certain facts in the world." From a psychological and cognitive-linguistic perspective, this issue is less important because the strict traditional distinction between linguistic and encyclopedic knowledge has more or less been eroded.

(more about this will be said in Section 6 below).[16] This is clearly illustrated by the combination *at sixes and sevens,* the overall meaning of which is not deducible from the meanings of its components. What is at stake here, then, is the differentiation between collocations and idioms, which is just as important as the one at the other end of the scales of predictability and recurrence, the one between collocations and free combinations.

3.6 Idiomaticity

In principle, there is no reason why idioms should not be regarded as extreme forms of collocations, as was suggested at the beginning of this paper. This view has been adopted, for example, by Robins (1971) and Palmer (1981). However, this does not do justice to either of the two phenomena, neither from a theoretical nor from a psycholinguistic point of view, as has been noted by Mitchell (1971), Cowie (1981), Bublitz (1996: 4) and others. The two phenomena are certainly distinct enough to merit a terminological separation.

In order to exclude genuine idioms from the field of collocations, a fifth and final criterion has to be introduced: the criterion of idiomaticity. This criterion is also scalar, reaching from completely transparent word-combinations to completely idiomatized ones. A few examples from the *Evita-text* may serve to illustrate this continuum (see Figure 2). The scale depicted in Figure 2 is of particular relevance for language teachers and lexicographers, who have therefore suggested specific terms for certain sections of the continuum. A number of these terms are given on the right-hand side of the figure, while the matching examples can be found on the left-hand side. *I love you, the truth is, staying out* and *looking out* are examples of *open* or *unspecific, trivial* collocations, whose combinatory meaning is completely transparent. At the other end of the scale there are idioms like *at sixes and sevens* and more common frozen expressions like *as for* and *at all,* whose meanings cannot be deduced from the meanings of their components. Between these extreme poles different degrees of idiomaticity are manifested in *keep your distance, kept my promise, dressed up* and *running around.* The main characteristic of these *restricted collocations* is that one of the component words - here *distance, promise, dress* and *around* - more or less preserves its literal meaning while the other is not used with a meaning akin to its literal or basic one (cf. Glaser 1986: 38-41).

[16] One reason for this tendency towards idiomaticity is that the meaning of one collocate can 'rub off (Bublitz 1996: 13) onto the meaning of the other, a process known as ***contagion*** since Bréal (cf.Ullmann 1962: 189).

Examples	**Degree of idiomaticity**	**Terms**
I love you, the truth is *staying out, looking out*	completely transparent	*open collocation* or *unspecific, trivial combination* (Hausmann 1984, 1985)
keep your distance *kept my promise* *dressed up* *running around*	↕	*restricted collocation* (Aisenstadt 1981, Cowie 1981, Gläser 1986: 38-41, Carter 1987: 62-65)
as for *at all* *at sixes and sevens*	completely idiomatized	*idiom*

Fig. 2: The scale of idiomaticity

As suggested by the arrow in the centre of the figure, the transitions between the degrees of idiomaticity are fluid. The reason why distinctions of this type are particularly relevant for language teachers and lexicographers is that the more idiomatized collocations are, the more attention they require. Since open collocations are transparent, there is no immediate necessity to teach them to foreign learners or to include them in dictionaries.[17] It is much more important to devote a lot of energy to restricted collocations, because they are particularly likely sources of mistakes and comprehension problems. The tricky thing about restricted collocations is that they are predictable (see Section 3.5), but of course only for *native* speakers. Non-native speakers may often be able to interpret the relation underlying such collocations when they encounter them; in language production, however, they are victims of the fact that the choice of the precise words that make up a collocation is to a large extent arbitrary. In this sense, collocations are *not* predictable at all, but dependent on the use or *norm* of a language (see Section 3.3; cf. also Herbst 1996: 386).

Having made this clear we can go one step farther and relate the criterion of idiomaticity to the third and fourth criteria, recurrence and predictability. This is shown in Figure 3.

[17] Of course it is true that the frequent active use of open collocations is a sign of near-nativeness, and therefore open collocations do indeed require attention. In earlier stages of proficiency, however, being able to understand and use restricted collocations may be important to foreign learners of a language.

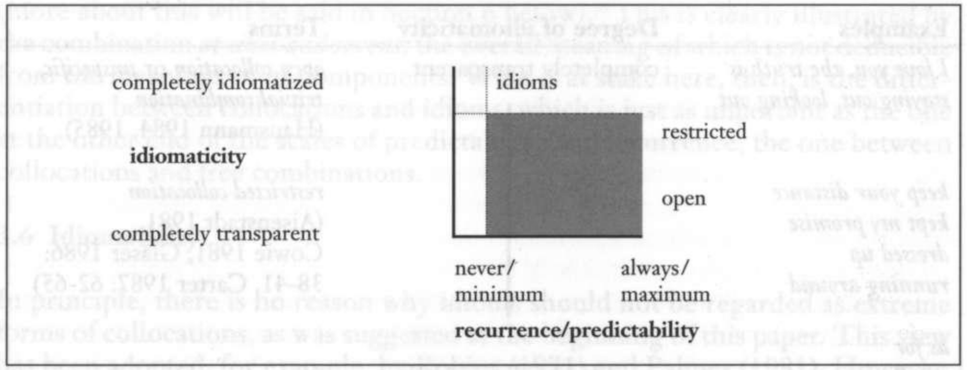


Fig. 3: The relation between recurrence/predictability and idiomacity (Version I)

In analogy to Figure 1, the combination of the criteria can be said to chart an area for which the term *collocation* is useful. With respect to the complex dimension of recurrence/predictability, the area of non-recurrent and hardly-ever-recurrent combinations must be excluded just as in Figure 1. And with respect to the criterion of idiomaticity, highly idiomatized expressions must also be excluded and treated as *idioms.* Restricted collocations start from a certain degree of idiomaticity.

On closer examination of Figure 3, it becomes apparent that certain parts of this extension will probably occur only very rarely, because there is, once again, a relation between recurrence/predictability on the one hand and idiomaticity on the other. These areas are excluded by lines a, b and c in Figure 4. First, it is highly unlikely that combinations of words which are rarely used together will acquire an idiomatic meaning, because frequent common usage is clearly a prerequisite for idiomaticity. So the upper left-hand area of the graph, which charts combinations that are more highly idiomatized than is likely according to their degree of recurrence, is excluded by line a. It should be borne in mind that this is not to mean that such combinations would not fall under the extension of the term collocation, but that it is unlikely that large numbers of such collocations exist.

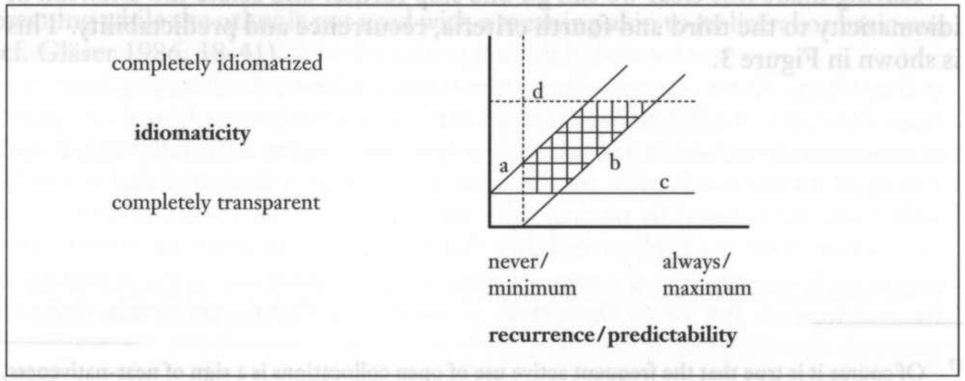


Fig. 4: The relation between recurrence/predictability and idiomaticity (Version II)

The same goes for highly recurrent but scarcely idiomatized collocations, which are excluded by line b. A number of frequent collocations like *the truth is that* [...] may not seem to be affected by idiomatization. However, as a closer study of such potentially open collocations suggests, even innocuous-looking expressions like *the truth is that [...]* or *the thing is that [...]* have acquired more specific meanings. (I will return to this claim in Section 5 below.) Line c excludes word combinations with a very low degree of idiomaticity from the prototypical core of collocations, again based on the rationale that typical collocations are not completely transparent. Finally, as in the previous figure, the domain of idioms - which are recurrent but completely idiomatized - must be excluded, and this is represented by line d. Even though the precise location of these three lines is admittedly to a large extent arbitrary, the hatched area in Figure 4 suggests where most prototypical collocations are to be found with respect to the charted dimensions.

4. The mediocrity of collocation[18]

The most frequent and prototypical cases of collocation cover the central areas on all three dimensions. Since, as already mentioned, it is reasonable to anchor definitions in (proto-)typical cases, *collocations* must be defined as 'combinations of lexemes exhibiting a medium degree of observable recurrence, mutual expectancy and idiomaticity.' What this unsightly definition makes clear is that collocation is not only a gradual, probabilistic phenomenon but, worse still, it is an entirely mediocre one, and it is this mediocrity that causes the problems with the definition of collocation.

In order to reveal the implications of mediocrity, it will be useful to first consider how mediocre things are dealt with in everyday life in general: They are not particularly exciting or even interesting, and consequently they do not attract a lot of attention. Why is there no Guinness Book of mediocrities? Because it would be an utterly boring book. When people watch a sports event, say a ski or bicycle race, what they are mainly interested in is of course the winner and perhaps the runner-up. Some people may also be interested in who came last, but the potentially large number of mediocre finishers between the triumphant winners and the pitiful losers do not normally attract a lot of attention. The same is true of all similar competition-like things such as prizes, awards or exams. Extremes tend to arouse our interest, mediocrities tend to go unnoticed.

Mediocrities are not only less interesting than extremes, but also less tangible and more difficult to describe. This tendency is even reflected in language itself. On many scales, the English language only provides words - especially only short and simple ones - to label the poles, whereas the sections of scales between

It should be made clear that the words ***mediocre*** and ***mediocrity*** are used here in the neutral meaning 'the quality or condition of being intermediate' (Simpson & Weiner 1989, s.v. ***mediocrity*** 1) deriving from Aristotle, i.e. without any negative connotations (cf. Simpson & Weiner 1989, s.v. ***mediocrity*** 5: 'the quality or condition of being mediocre [...] now chiefly with disparaging implication, in contrast with excellence or superiority').

the extremes, the mediocrities, often require more elaborate ways of encoding. Cases in point are *old - young* vs. *middle-aged, tall - short* vs. *medium-height,* and *rich - poor, clever - stupid* and a large number of other pairs of polar opposites where the scale in between is hardly covered by simplex words at all.

As has been shown, the notion of collocations also serves to capture such non-extreme, mediocre phenomena. The poles on the dimensions discussed here are occupied by other concepts or principles. The poles on the dimension of recurrence/predictability are described in terms of the free combinability of words according to the rules of syntax on the one hand, and extreme examples of frozen expressions like *kith and kin* on the other, whose parts hardly ever occur otherwise. And on the scale of idiomaticity, one pole is reflected in the default assumption that sentence meaning is the sum or interaction of word meanings, while the other is seen as belonging to the fields of idioms and phraseology. It is somewhere between these poles that one must look for collocations. But unlike the poles of these scales, their larger mediocre sections are more difficult to reify and grasp and, as a consequence, more difficult to describe terminologically. Arguably, this is the reason why it is so difficult to define the notion of collocation, and it is presumably also the reason why it did not for a long time attract the amount of interest it deserves. It is mainly due to the emergence of corpus-linguistic procedures which enable us to assess at least the recurrence of potential collocations with some degree of objectivity that the phenomenon of collocation is now being taken more seriously.

Whether a linguistic phenomenon is easy or hard to describe and define does not of course affect the use of the phenomenon. It is a meta-linguistic problem, not one related to language as such. And that collocation is indeed a very frequent feature of language has been illustrated in Section 2. The final section of this paper proposes a number of reasons why collocation is so ubiquitous.

5. The utility of collocation

In what way do language users - and, in a manner of speaking, languages - profit from the existence of collocation? In order to answer this question, I would like to introduce a second example passage, this time taken from spontaneous spoken discourse. The passage presented in (6) was originally produced by a Californian youth, who was asked what he liked about a former friend of his who had stolen 400 dollars from him. This speaker tends to encode his ideas in stereotypical chunk-like phrases of uncertain grammatical status rather than syntactically complete clauses and sentences. These are highlighted by bold print in (6).

(6)

> **cool dude, you know**, catch women, **this and that**. But he, must get **his nose wide open**, behind some other girl, and **this and that**, and get [unintelligible] not **making money and shit, lose his heart**, and **lay in the crib** and **rip me off**. Instead of **going up there**, take it from somebody else, he don't know **or something**, [unintelligible] he had, intentions on doing that, until he **lost it all, falling in love**, with **his damn nose wide open**, scared he would get **knocked down**.
>
> (Gumperz, Kaltmann & O'Connor 1984: 15; punctuation, which indicates prosody, taken over from the original; other signals of intonation omitted; emphasis added)

The style of this passage seems fairly typical of American inner-city youth slang. It is characterized by colloquial collocations such as *cool dude* ('guy'), *his nose wide open* (a reference to cocaine),[19] *making money, lose his heart* (either 'fall in love' or 'lose the courage to be out in the streets'),[20] *lay in the crib* ('stay at home'),[21] *rip me off* ('steal my money'), *going up there* ('go to the richer parts of the city'), *lost it all, falling in love, knocked down.* Other characteristics are the abundance of collocations with extremely vague meanings, e.g. *this and that* (twice), *and shit, or something,* and the use of discourse markers *(you know).*

It is revealing to consider what is going on here from a psycholinguistic point of view. Very much in line with Sinclair's *idiom principle* (see Section 1), the speaker does not plan and utter every single word individually but relies on prefabricated chunks instead. This style of speech has some fairly obvious advantages. Arguably, it requires less cognitive energy to retrieve several word-forms from the mental lexicon in one go as one chunk, rather than each one of them individually. This way collocations help to reduce the mental effort required for language processing. Evidence supporting this can be found in language acquisition and patholinguistics: Thus it is claimed (cf. e.g. Stubbs 1995a: 379) - and can easily be observed - that in the early stages of language acquisition toddlers make ample use of collocation-like holophrases, like *all gone, go home, shoe off, no more, have dinner* or *give a kiss.* Similar stereotypical, stock-like phrases are also reported to remain surprisingly stable in the otherwise muddled jargon of people suffering from Wernicke's aphasia (Crystal 1980: 147; 1987: 271). Neither of these phenomena would be possible if collocations required higher processing skills and more processing capacity than free syntactic combinations.

The reduction of processing load does not only apply to speech production, the perspective from which I have described it so far, but also to comprehension. For the predictability of collocations also reduces the cognitive cost necessary for understanding utterances: When hearers or readers come across the first part of a collocation, they can devote less attention to the later collocates - especially if it is a fairly restricted collocation - than is the case for non-recurrent syntactic combinations. As a consequence, the frequent and appropriate use of collocations in speech or writing facilitates the ease with which a text can be comprehended. It is not an uncommon experience to be confronted with speeches or texts by highly advanced foreign learners which are immaculate from a gram-

Nose is explained as 'cocaine, which one inhales' in Green (1984); cocaine is also frequently referred to as ***nose candy*** (cf. Green 1984, Beale 1989, Ayto & Simpson 1992). In Green (1984), ***wide open*** is explained as 'vulnerable, undefended.'

See ***heart*** '(street gang use) courage, bravery, spirit' in Green (1984).

Crib '3. house, apartment, anywhere one lives' (Green 1984).

matical point of view but nevertheless somewhat difficult to understand. This can be caused by a shortage and/or violation of common collocations, since common collocations are, as it were, the right things to say, and if things are said the right way they are easier to understand.

A corollary of this is that collocations support and enhance the connections created by syntax, thus contributing to clause-internal coherence (cf. Halliday 1966: 152, Sinclair 1966:411 and Bublitz 1996: 26-27). As certain passages in example (6) show, collocations can in fact replace syntax as the major source of coherence on the local level (i.e. within clauses or intonation units). While many of the chunks produced by the young man are rather suspect from a syntactic point of view, comprehension is to a large extent safeguarded by their collocational coherence.

Since processing capacity limitations are far more severe in spontaneous spoken speech - which is an on-going, on-line as it were, process - than in written language, one would expect collocation to be more frequent in spoken discourse than in written. Unfortunately, I cannot offer any systematic data to substantiate this. What seems to be quite clear in the first place is that the types of collocations vary according to medium: There are probably typically spoken and typically written collocations, with a considerable degree of overlap between the two classes. Furthermore, within both major media one would expect to find specific inventories of collocations cropping up particularly frequently in certain text-types or registers (in the written medium e.g. academic prose, business letters, personal letters, news items etc.), which would contribute to the identifiable style of these registers. While there is surprisingly little material on this in textbooks on linguistic style (e.g. Esser 1993), a wealth of information can be found in Biber et al. (1999: 990-1024), where lexical bundles typical of written registers and spoken conversation are discussed.

So far I have only been concerned with the advantageous effects of collocations on the on-going processing of language. I will now go one step further and consider more permanent effects related to the mental lexicon. Since collocations consist of several words, it could be argued that the processing advantages are outweighed by the greater effort required for their acquisition and storage in the mental lexicon. This does not seem to be convincing, however. For one thing, the evidence from first language acquisition mentioned earlier is clearly at odds with this view. Second, the evidence from association and priming experiments mentioned above (see Section 3.4) indicates that at least retrieval seems to be facilitated. In addition, the fact that collocations are *per definitionem* **recurrent** combinations of words (see Section 3.5 above) has an effect on their entrenchment in the mental lexicon: According to cognitive linguists (Langacker 1987: 59, see also Schmid forthcoming: Section 2.1), the degree of entrenchment of concepts and other linguistic units correlates with the frequency of their activation. Now if the words making up a collocation are indeed frequently co-activated, their joint entrenchment in the mental lexicon should be facilitated as well.[22]

As it can be assumed that non-recurrent word-combinations do not become entrenched, one can conclude that those that ***are*** to some degree entrenched as combinations ***must*** be recurrent. (This supports the claims made in Section 3.5 above.)

In Cognitive Linguistics, the process of *entrenchment* is not only held responsible for the storage of elements in the mental lexicon and the relative ease of their retrieval, but also for the formation of linguistic units. By way of a habitualization or automatization effect, entrenchment leads to routines for the activation of certain conceptual complexes represented linguistically as words and, more importantly in this context, as multi-word complexes or even syntactic constructions. Fully-entrenched complexes and constructions acquire unit status, or, in terms derived from perceptual psychology, the status of linguistic *gestalts* (Lakoff 1977; see also Schmid forthcoming: Section 3).[23] If we translate the claims made in Section 4 into this cognitively-oriented terminology, it is the hallmark of collocations (as opposed to idioms) that they are not fully entrenched as linguistic units or *gestalts* but only partially.

This half-way entrenchment status of collocations seems to be cognitively advantageous and to go some way towards explaining the ubiquity of collocation. To see in what way, it will be helpful to once more compare collocations with free syntactic combinations, on the one hand, and fully idiomatized expressions, on the other. In contrast to free syntactic combinations, collocations have acquired some degree of entrenchment as units and are thus to some extent manipulable as *gestalts.* If it is true, however, that entrenchment as a holistic *gestalt* is so helpful then the question of course is why collocations do not become further entrenched. Why are idioms, whose distinction it is that they are firmly entrenched as holistic units, not much more frequent than they are? Why do collocations tend to remain at the level of partial entrenchment rather than move towards full entrenchment as idioms? Of course, some collocations have done so, but this is not the main point here. The answer to these questions lies in the nature of *gestalts.* When a complex of linguistic elements acquires *gestalt* status, this has two main implications: First, the parts of *gestalts* become less salient in comparison to the whole (Langacker 1987: 59), so that firmly entrenched units (of any size) can be processed and manipulated as single, holistic chunks of information. And second, the information conveyed by the whole chunk tends to differ from the information conveyed by the sum of its components. This is one of the basic insights Gestalt Psychology has brought to our attention. Thus, the more entrenched a multi-word expression becomes as one holistic unit, the less likely it is to preserve its composite meaning, i.e. the meaning of a simple sum or combination of its collocates. This means that the coalescence of several words into one linguistic *gestalt* tends to result in a modification of meaning - the process traditionally described as idiomatization. More often than not the semantic change is specialization rather than generalization: the meanings of fully idiomatic expressions tend to be quite specific. And the more specific the meaning of an expression is, the less widely it can be applied. So a greater degree of entrenchment of multi-word units than the partial entrenchment of prototypical collocations would result in a reduction of their applicability, and hence also their frequency. Since idiomatization also tends to go hand in hand with restric-

For more information on the notion of ***gestalt,*** see Ungerer & Schmid (1996: 33-37) and their references (1996: 52).

tions on syntactic variability (i.e. with higher degrees of *frozenness),* it is not only semantic versatility that decreases but also syntactic.[24]

One further twist must be added here, as it is of course not universally true that a greater degree of entrenchment always results in greater semantic specificity. A good counterexample is the collocation *the fact that [...]* which has arguably lost the epistemic meaning residing in the lexeme *fact* and is used as a general-purpose device for anchoring clauses in a noun phrase (see Schmid 2000: 99-100). In spoken language in particular, there is in fact a host of firmly entrenched collocations that are similarly unspecific from a semantic point of view. Examples in (6) are *this and that, and shit* as well as *or something.* Lexical bundles like *the question is, the thing is, I think, I'll tell you what* and many others belong here, and also such discourse markers as *you know, I mean* or *you see.*

There are several ways of explaining these collocations while adhering to the more general claims on idiomatization. First, whereas many of these expressions may be unspecific from a strictly semantic point of view, their freedom of occurrence is still severely limited. These restrictions are perhaps not as drastic as those for the use of formulaic greetings (e.g. *good morning* or *goodbye)* but they can be compared to the latter. This means that these expressions are restricted pragmatically rather than semantically. Discourse markers and most of the other lexical bundles mentioned above do not occur as integral parts of the *bodies* of utterances - in Biber et al.'s (1999: 1072-82) terminology - but are either inserted as parentheses or attached to the body as *prefaces,* i.e. fronted elements, or as *tags,* i.e. postposed ones. Furthermore, as already hinted at in Section 3.6 above, even apparently neutral prefaces like *the truth is [...], the thing is [...]* or *the question is* [...] have acquired usage restrictions that cannot be derived from their composite meaning. The expression *the thing is [...]* is explained in Summers (1995: s.v. *thing* 27) as being 'used when explaining a problem or the reason for something'; Biber et al. (1999: 1075) gloss the use of *the question is [...]* as "presenting an issue in an explicit, forceful way." In Schmid (2000: 243-44, 336), it is shown that the collocations *the fact/truth/reality/certainty is that [...]* are used predominantly in contrastive contexts as strong emphasizers of epistemic necessity.

Second, some of the semantically most unspecific collocations typical of spontaneous spoken discourse, e.g. *this and that, and something, and stuff, you know (what I mean),* but also the more specific prefaces like *the thing is [...], the question is [...]* can be explained as linguistic habits. Many people do not seem to be aware of the extent to which they use such expressions. Presumably it is no coincidence that these routines tend to occur either before or after the bodies of utterances. While the latter convey most of the propositional content and are thus planned consciously, prefaces and tags often serve the function of gaining time for the planning of the utterance proper (in the case of prefaces) or of rounding off or extending one's turn, for example when no other speaker is willing to take over (in the case of tags; cf. Biber et al. 1999: 1080-82). It would clearly be detrimental to burden processing, which at least in the case of prefaces is to be facilitated by this strategy, with linguistic material that has to be planned and requires cognitive energy. Therefore it is a fairly safe guess that these collocations require very little processing effort. In addition, their semantic generality does not cause any communicative problems, since neither speakers nor hearers really expect any contribution to propositional content. Expressions of this type operate on the interpersonal rather than the ideational level of language.

6. Conclusion

The concept of *collocation* shares with other linguistic notions such as *sentence, word, text* or *subject* the problem that it is does not capture a neatly delimited set of entities or phenomena, but a range of somehow similar ones. For the other concepts, however, prototypical manifestations which can be described in terms of maximum manifestations of certain criteria can often be found. Prototypical English subjects, for example, are at the very beginning of clauses, they represent human beings or other typical agents at the top of our empathy hierarchy (Langacker 1991: 306-09), encode themes/topics and convey given information. The snag with the notion of collocation is that the prototypical cases do not correspond to the poles of the relevant dimensions, but to mediocre sections. In contrast to the other linguistic concepts - which lend themselves quite readily to definitions in terms of prototypes - collocation is less tangible, because it is very difficult to conceptualize something like 'combinations of lexemes exhibiting a medium degree of observable combined recurrence, mutual expectancy and idiomaticity' as a category prototype.

While this may be a nuisance on a meta-linguistic level, it seems to have beneficial effects on the use of collocation. It has been argued here that it is the very essence of collocations as partially entrenched linguistic gestalts that accounts for their utility and ubiquity. This status seems to be cognitively advantageous in that it strikes the balance between the welcome reduction of processing load achieved by chunk-like storage and retrieval, on the one hand, and the less welcome narrowing of the range of applicability caused by complete idiomatization.

The grammatical theory that provides the most advanced account of the phenomena discussed here is probably Construction Grammar (cf. Fillmore 1988, 1997, Fillmore et al. 1988, Goldberg 1995, 1996). According to Goldberg, "a construction is defined to be a pairing of form with meaning/use such that some aspects of the form or some aspect of the meaning/use is not strictly predictable from the component parts or from other constructions already established to exist in the language. On this view, phrasal patterns [...] are given theoretical status" (Goldberg 1996: 68). This means that collocations are integrated in the regular section of grammar side by side with grammatical constructions as meaning-bearing elements. It is the aim of Construction Grammar to capture specific semantic and/or pragmatic properties of lexico-syntactic combinations without necessarily focussing on their syntactic properties alone. Deviations from the predictable behaviour of constructions can be found in meaning and/or use; this corresponds to my claim that degrees of entrenchment of collocations correlate with semantic and/or pragmatic usage restrictions. In Construction Grammar, collocations are treated as a subclass of idioms (Fillmore 1997: 8), where the words have "some special conventional association" with each other.

Works Cited

Aisenstadt, Ester (1981). "Restricted Collocations in English Lexicology and Lexicography." In: *ITL. Review of Applied Linguistics* 53, 53-61.

Aitchison, Jean (1994). *Words in the Mind: An Introduction to the Mental Lexicon*. 2nd ed., Oxford: Blackwell.

Ayto, John & John Simpson (1992). *The Oxford Dictionary of Modern Slang*. Oxford: Oxford UP.

Bahns, Jens (1997). *Kollokation und Wortschatzarbeit im Englischunterricht*. Tübingen: Narr.

Bazell, Charles C., John C. Catford, Michael A.K. Halliday & Robert H. Robins, eds. (1966). *In Memory of J.R. Firth*. London: Longman.

Beale, Paul, ed. (1989). *A Concise Dictionary of Slang and Unconventional English*. Berlin: Langenscheidt.

Benson, Morton, Evelyn Benson & Robert Ilson (1997). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Rev. ed., Amsterdam/Philadelphia: Benjamins.

Biber, Douglas (1996). "Investigating Language Use through Corpus-Based Analyses of Association Patterns." *International Journal of Corpus Linguistics* 1, 171-97.

Biber, Douglas, Susan Conrad, Geoffrey Leech, Stig Johansson & Edward Finegan (1999). *Longman Grammar of Spoken and Written English*. London etc.: Longman.

Bublitz, Wolfram (1996). "Semantic Prosody and Cohesive Company: 'somewhat predictable.'" *Leuvense Bijdragen* 85, 1-32.

Bublitz, Wolfram (1998). "'I entirely dot dot dot.' Copying Semantic Features in Collocations with Up-Scaling Intensifiers." In: Rainer Schulze, ed., *Making Meaningful Choices in English*. Tübingen: Niemeyer, 11-22.

Carter, Ronald (1987). *Vocabulary*. London etc.: Allen & Unwin.

Catford, John C. (1965). *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. London: Oxford UP.

Church, Kenneth W. & Patrick Hanks (1990). "Word Association Norms, Mutual Information & Lexicography." *Computational Linguistics* 16, 22-29.

Clear, Jeremy (1993). "From Firth Principles. Computational Tools for the Study of Collocation." In: Mona Baker, Gill Francis & Elena Tognini-Bonelli, eds., *Text and Technology. In Honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins, 271-92.

Coseriu, Eugenio (1967). "Lexikalische Solidaritäten." *Poetica* 1, 293-303.

Cowie, Anthony P. (1981). "The Treatment of Collocations and Idioms in Learners' Dictionaries." *Applied Linguistics* 2, 223-35.

Cowie, Anthony & Ronald Mackin (1975). *Oxford Dictionary of Current Idiomatic English, vol. 1: Verbs with Prepositions and Particles*. Oxford: Oxford UP.

Cowie, Anthony P., Ronald Mackin & Isabel R. McCaig (1983). *Oxford Dictionary of Current Idiomatic English, vol. 2: Phrase, Clause and Sentence Idioms*. Oxford: Oxford UP.

Crystal, David (1980). *Introduction to Language Pathology*. London: Edward Arnold.

Crystal, David (1987). *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge UP.

Esser, Jürgen (1993). *English Linguistic Stylistics*. Tübingen: Niemeyer.

Esser Jürgen (1999). "Collocation, Colligation, Semantic Preference and Semantic Prosody: New Developments in the Study of Syntagmatic Word Relations." In: Wolfgang Falkner & Hans-Jörg Schmid, eds., *Words, Lexemes, Concepts – Approaches to the Lexicon. Studies in Honour of Leonhard Lipka*. Tübingen: Gunter Narr, 155-65.

Fillmore, Charles J. (1988). "The mechanisms of 'Construction Grammar.'" In: Shelley Axmaker, Annie Gaisser & Helen Singmeister, eds., *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*. Berkeley: U of California P, 35-55.

Fillmore, Charles J. (1997). "Idiomaticity." Lecture posted in the Internet. [Available at http://www.icsi.berkeley.edu/ ~kay/bcg/lec02.html]

Fillmore, Charles J., Paul Kay & M. Catherine O'Connor (1988). "Regularity and Idiomaticity in Grammatical Constructions: The Case of *let alone*." *Language* 64: 501-538.

Firth, John R. (1957; 1951). "Modes of meaning." In: John R. Firth, *Papers in Linguistics 1934-1951*. London: Oxford UP, 190-215. [Originally published in *Essays and Studies*, The English Association, 1951].

Fontenelle, Thierry (1994). "What on earth are collocations?" *English Today* 10 (4), 42-48.

Fraser, Bruce (1970). "Idioms Within a Transformational Grammar." *Foundations of Language* 6, 22-42.

Gläser, Rosmarie (1986). *Phraseologie der englischen Sprache*. Leipzig: VEB Verlag Enzyklopädie.

Goldberg, Adele (1995). *Constructions: A Construction Grammar Approach to Argument Structure Constructions*. Chicago: UP of Chicago.

Goldberg, Adele (1996). "Construction Grammar." In: Keith Brown & Jim Miller, eds., *Concise Encyclopedia of Syntactic Theories*. Oxford: Pergamon, 68-71.

Green, Jonathan (1984). *The Dictionary of Contemporary Slang*. London: Pan Books.

Greenbaum, Sidney (1970). *Verb-Intensifier Collocations in English: An Experimental Approach*. Den Haag: Mouton.

Greenbaum, Sidney (1988). "Some Verb-Intensifier Collocations in American and British English." In: Sidney Greenbaum, ed., *Good English and the Grammarian*. London: Longman, 113-24.

Gumperz, John J., Hannah Kaltman & Mary C. O'Connor (1984). "Cohesion in Spoken and Written Discourse: Ethnic Style and the Transition to Literacy." In: Deborah Tannen, ed., *Coherence in Spoken and Written Discourse*. Norwood/NJ: Ablex, 3-19.

Halliday, Michael A.K. (1966). "Lexis as a Linguistic Level." In: Bazell et al. (1966), 148-162.

Hausmann, Franz Josef (1984). "Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen." *Praxis des neusprachlichen Unterrichts* 31, 395-406.

Hausmann, Franz Josef (1985). "Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels." In: Henning Bergenholtz & Joachim Mugdan, eds., *Lexikographie und Grammatik: Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.-30.6.1984*. Tübingen: Niemeyer, 118-29.

Herbst, Thomas (1996). "What are Collocations: Sandy Beaches or False Teeth." *English Studies* 4, 379-93.

Jones, Susan & John M. Sinclair (1974). "English Lexical Collocations." *Cahiers de Lexicologie* 24, 15-61.

Kjellmer, Göran (1982). "Some Problems Relating to the Study of Collocations in the Brown Corpus." In: Stig Johansson, ed., *Computer Corpora in English Language Research 1975-1981*. Bergen: Norwegian Computer Centre for the Humanities, 25-33.

Kjellmer, Göran (1987). "Aspects of English Collocations." In: Willem Meijs, ed., *Corpus Linguistics and Beyond. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi, 133-40.

Kjellmer, Göran (1994). *A Dictionary of English Collocations Based on the Brown Corpus*. 3 vols. Oxford: Clarendon.

Lakoff, George (1977). "Linguistic gestalts." In: *Papers From the Thirteenth Regional Meeting, Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, 321-34.

Langacker, Ronald W. (1987). *Foundations of Cognitive Grammar, vol.1: Theoretical Prerequisites*. Stanford: Stanford UP.

Langacker, Ronald W. (1991). *Foundations of Cognitive Grammar, vol. 2: Descriptive Application*. Stanford: Stanford UP.

Leech, Geoffrey (1993). "100 Million Words of English." *English Today* 33 (9), 9-15.

Lehr, Andrea (1996). *Kollokationen und maschinenlesbare Korpora: Ein operationales Analysemodell zum Aufbau lexikalischer Netze*. Tübingen: Niemeyer.

Lipka, Leonhard (1974). "Probleme der Analyse englischer Idioms aus struktureller und generativer Sicht." *Linguistik und Didaktik* 20, 274-85.

Lipka, Leonhard (2002). *English Lexicology: Lexical Structure, Word Semantics & Word-Formation*. Tübingen: Narr.

Lorenz, Gunter R. (1999). *Adjective Intensification – Learner Versus Native Speaker: A Corpus Study of Argumentative Writing*. Amsterdam/Atlanta: Rodopi.

Louw, Bill (1993). "Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies." In: Mona Baker, Gill Francis & Elena Tognini-Bonelli, eds., *Text and Technology. In Honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins, 157-76.

Mitchell, T.F. (1971). "Linguistic 'goings-on': Collocations and Other Matters on the Syntagmatic Record." *Archivum Linguisticum* 2, 35-69.

Palmer, Frank R. (1981). *Semantics*. 2nd ed. Cambridge: Cambridge UP.

Robins, Robert H. (1971). *General Linguistics. An Introductory Survey*. 2nd ed., London: Longman.

Schmid, Hans-Jörg (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Berlin: Mouton de Gruyter.

Schmid, Hans-Jörg (forthcoming). "Entrenchment, Salience and Basic Levels." In: Dirk Geeraerts & Hubert Cuyckens, eds., *Handbook of Cognitive Linguistics*. Oxford: Oxford UP.

Sinclair, John McH. (1966). "Beginning the Study of Lexis." In: Bazell et al. (1966), 410-430.

Sinclair, John McH. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford UP.

Smadja, Frank (1993). "Retrieving Collocations from Text: Xtract." *Computational Linguistics* 19,143-77.

Stubbs, Michael (1995a). "Collocations and Cultural Connotations of Common Words." *Linguistics and Education* 7, 379-90.

Stubbs, Michael (1995b). "Collocations and Semantic Profiles. On the Cause of the Trouble with Quantitative Studies." *Functions of Language* 2, 23-55.

Summers, Della, ed. (1995). *The Longman Dictionary of Contemporary English*. 3rd ed., Harlow: Longman.

Simpson, John A. & Edmund S.C. Weiner (1989). *The Oxford English Dictionary*. 2nd ed., Oxford: Oxford UP.

Taylor, John R. (1995). *Linguistic Categorization: Prototypes in Linguistic Theory*. 2nd ed., Oxford: Clarendon.

Ullmann, Stephen (1962). *Semantics. An Introduction to the Science of Meaning*. Oxford: Basil Blackwell.

Ungerer, Friedrich & Hans-Jörg Schmid (1996). *An Introduction to Cognitive Linguistics*. London: Longman.