Collocations, Collocability and Dictionary

František Čermák
Ústav Českého národního korpusu, filozofická fakulta Univerzity Karlovy
nám. J. Palacha 2, Praha 2
110 00, Czechia
frantisek.cermak@ff.cuni.cz

Abstract
In the first part, basic approaches to collocations and their nature is discussed. In a survey, these deal with (1) corpus-psychological approach, (2) content-grammar words, (3) polysemous-monosemous words, (4) frequent-infrequent words, (5) collocationally large-restricted words, (6) stable-not stable collocations, (7) regular-anomalous combinations. Some attention is paid to relation of polysemy and frequency and role of valency. In the second part, an analysis of selected words from LDCE's letter A is undertaken in their coverage of collocations and confronted with what the British National Corpus has to offer. Finally, major open problems of lexicographical coverage of collocations are listed.

1 Collocations and Dictionaries

Collocations included in dictionaries seem to be a recent and welcome improvement, but they are still far from being found in most of them, apart from exceptions. This event is definitely a tribute to modern large corpora and what they offer which is now hard to ignore, but it is also an acknowledgement of users' practical needs. Linguistically, one may wonder what prompted what. It is certainly true that dictionaries have been (and are meant to be) products of a prominently paradigmatic type, offering all sorts of classification of items included, while it is not generally admitted that at least some syntagmatic information on words could and should be included in them, too. With first attempts to cover this complementary syntagmatic aspect, standing in opposition to paradigmatical ones, one must wonder what kind of observation was taken into consideration and what decisions have been taken.

1.1 Seven Approaches to Collocations and Parameters

The number of approaches to collocations, both in theory and practice, seems to grow (see, for example, recent Partington 1998 and Hoey 2005), but there is still little agreement about some of its aspects. It is evident that identification of collocations is a primary problem. Basically, one may rely on a (**1A**) purely **corpus (computational)** surface **approach** declaring any text combination, if sufficiently frequent, collocation, where, however, the idea of collocation is too broad as it would include combinations that are difficult to accept by any standards. On the other hand, one may go for a (**1B**) purely **psychological approach,** manual and based largely on intuition, but then he/she gets in trouble if a finer type of decision, espe-

cially on border-line cases, has to be made as to what a collocation is and what is not, a primary obstacle being our intuitions's inability to be exact.

In part, the latter type of approach is to be suspected behind some of the few existing dictionaries, too, i.e. either general type of dictionaries as that of Longman (LDCE), or specific dictionaries of collocations, such as Oxford Collocations (2002). In both, however, only some types of words are provided with information on their collocations and it is evident that the authors prefer large word classes only, such as nouns or verbs. As these are traditional content words, one must wonder why other word classes are less covered, or not covered at all, as with Oxford. Though not so prominent, another distinction is to be found here, too, namely that between (**2A**) **content words** and (**2B**) **grammar words**, which get a different treatment.

It is not surprising, especially with learners' dictionaries, that (**3A**) difficult and highly **polysemous words** get more attention, if their collocations are to be covered. Though different, words loaded with pragmatic meanings and functions (especially evaluative ones), could, for simplicity's sake, be subsumed under these, too. Polysemous words stand in contrast to words of (**3B**) relatively **simple and single meaning** (but only relatively so), such as *flour, coal, copper*. This distinction, as well as other, overlaps but is not identical with that of frequency.

Thus, (**4A**) **frequent words** have more collocations, while (**4B**) **less frequent words** have less. Yet it is difficult to accept that dictionaries should pay attention to frequent words only (see in 1.2).

However, little attention is paid to combinatory potential, range or collocability of words, which should not be taken to be identical with (4A-B). It is common experience that (**5A**) the whole **collocational range** (collocability) of most words is and seems to be so **large** that it seems unlimited and is never given in full, i.e. in a list of collocates. In contradistinction to this, there is a (**5B**) group of words that is evidently and **strictly limited in** its **collocational capacity**, where a list of collocates, usually very small, can and has to be given. This last group reverts the view adopted so far and suggests, in fact, an entirely different possibility, namely to view both the head and collocate as a single unit, tantamount, in many approaches, to idioms, cf. *afraid* (*be afraid*), *afoul* (*run afoul*), *amends* (*make amends*).

1.2 Polysemy and Frequency

In a corpus-based approach (1A), oriented to content words primarily (2A), there seem to prevail certain **correlations** among the remaining three parameters mentioned above, namely meaning (3), frequency (4) and collocability (5) of words. Since no serious research is available (but see Čermák 2002), it appears that polysemy and frequency of words are rather similar in their distribution and are more or less directly proportional to their collocability. Hence, it seems that the higher (or broader) collocability, i.e. array of collocates, of a word is the higher its frequency and polysemy, and vice versa. Despite well-known hesitation as to how polysemy should be conceived, it is easy to see, using any standard desk dictionary, that the correlation suggested above basically holds. In the following, data from Longman will be used to illustrate this; since the dictionary does not give exact frequency, figures from BNC (British National Corpus) will be used instead. This can be illustrated by some words taken

from the letter A of the dictionary (the first number referring to BNC frequency, the second (bold) to the number of senses given in the dictionary), thus

(**A**) *as* 653 137 (**20**, given under two headwords), *any* 122 151 (**7**, under two headwords), *around* 43 391 (**14**), *ask* 18 642 (**15**, under two headwords), *approach* 16 020 (**11**, under two headwords), *arm* 8621 (**15**, under two headwords), *aim* 6394 (**7**, under two headwords), *aid* 8847 (**9**, under two headwords), *arrive* 2836 (**7**),

(**B**) *assurance* 1811 (**3**), *attendance* 1802 (**4**), *auction* 1318 (**2**, under two headwords), *approve* 1037 (**2**), *apology* 663 (**5**), *applaud* 183 (**2**), *assimilation* 287 (**2**), *absently* 198 (**1**), *abreast* 190 (**3**), *abseil* 92 (1), *attire* 111 (**1**), *abrogate* 16 (**1**), *attestation* 14 (**1**), etc.

The general tendency seems to hold, showing that the lower the frequency of a word is, the fewer its senses are. It is evident that some of the sense divisions are questionable since sense numbers are sometimes given to a single collocation or even an idiom, which is not a „standard" sense. The problem here is that the sense is attributed to the idiom or collocation only (or, rather, a single constituent only) and does not exist outside of it (this is the case of two senses of *abreast*, as in *keep/stay a. of sth*, and *walk/ride* etc. *a. of*). Both collocability and polysemy are rather different with different word classes and, within nouns, between concrete (such as *arm*) and abstract nouns (most of examples here). Finally, cases with the highest frequency of occurrence in the corpus, such as *as, any, around*, illustrate the difference between grammar and content words, their sense division being, in an alternative view, a division of functions rather than of meanings.

1.3 Valency

All of the above types are related to meaning, basically, and their collocates tell us which members of certain types of words collocate with the head word, such as *any, ask, approach.* In contrast, most basic words (nouns, verbs, adjectives, etc.) enter their syntactical constructions only thanks to combination with certain compulsory words, namely valency-markers, such as prepositions, often required by verbs, cf. ***ask** for (sth)/that (CL)/about (sth)/how (CL)* (where *sth* stands for an in/animate noun, *CL* for a clause) etc. (but not, e.g. *ask on*).

Limiting **valency** to simple markers may not be enough, however. Though no dictionary views cases such as *in accordance with* as a noun with valency markers, linguistically there is no reason not to do so, despite the standard practice, and to view the whole combination as a completely independent item. In fact, for this multi-word idiomatic preposition, BNC gives a number of collocates, mostly verbs, such as *act, appoint, assign, behave, be, carry out, compile, determine, demand, do, make, prepare, propose, reach,* etc. Only on its surface these collocates may seem to be an open and endless class; in fact, they can easily be subsumed under the most frequent head (should we revert the view), namely *be*, and two generic ones *do/make.* This is corroborated by (static and other) verbs that do not accept *in accordance with* as its valency marker, such as *lie, exist, sleep*, etc.

Longman dictionary does not, in fact, mention valency at all, although it heavily employs it, often for discrimination of meaning. As an illustration of this, a couple of the initial senses

(numbered and commented) of ***to ask*** may serve, where a sense or subsense is not defined but is given as a phrase (such as ***ask about something*** below), accompanied by an example.

1 QUESTION

[intransitive and transitive] to speak or write to someone in order to get an answer, information, or a solution

ask who/what/where etc I just asked him where he lived.

ask somebody something We'll have to ask someone the way to the station.

ask somebody if/whether Go and ask Tom whether he's coming tonight.

ask (somebody) about something Visitors usually ask about the history of the castle.

ask around (=ask in a lot of places or ask a lot of people) *I'll ask around, see if I can find you a place to stay.*

2 FOR HELP/ADVICE ETC

[intransitive and transitive] to make a request for help, advice, information etc *If you need anything, you only have to ask.*

ask somebody to do something Ask John to mail those letters tomorrow.

ask to do something Karen asked to see the doctor.

ask for Some people find it difficult to ask for help.

ask somebody for something He repeatedly asked Bailey for the report.

ask (somebody) if/whether you can do something Ask your mom if you can come with us.

ask that Was it too much to ask that he be allowed some privacy?

3 PRICE

[transitive]to want a particular amount of money for something you are selling *How much is he asking? They're asking a fortune for that house.*

4 INVITE

[transitive usually + adverb/preposition] to invite someone to your home, to go out with you etc

ask somebody to do something Let's ask them to have dinner with us some time.

ask somebody out (=ask someone, especially someone of the opposite sex, to go to a film, a restaurant etc with you) *Jerry's too scared to ask her out.*

ask somebody in (=invite someone into your house, office etc) *Don't leave them standing on the doorstep – ask them in!*

ask somebody over/round (=invite someone to come to your home) *We must ask our new neighbours over for a drink.*

5 DEMAND

[transitive] if you ask something of someone, you want them to do it for you *It would be better if he cooperated, but perhaps I'm asking too much.*

ask something of somebody You have no right to ask anything of me.

6
be asking for trouble
to do something that is very likely to have a bad effect or result
Saying that to a feminist is just asking for trouble.

7
ask yourself something
to think carefully and honestly about something *You have to ask yourself where your responsibilities really lie.*
etc ...

Not numbered:
ask after somebody phrasal verb
if you ask after someone, you want to know whether they are well, what they are doing etc
I spoke to James today. He was asking after you.
ask for somebody phrasal verb
if you ask for someone, you want to speak to them There's someone at the door asking for Dad.

It is largely irrelevant how valency is handled in dictionaries, provided it is handled at all. In this approach, valency markers are viewed as a phenomenon per se; in a broader, corpus-based approach, however, valency amounts to a type of physical neighbours of the given head, i.e. it may appear as a type of collocate. The only difference, in a simplified view, is the grammatical nature of valency, based on very broad classes it relates to. Collocability as a set of collocates, on the other hand, may be viewed as having a lexico-semantic character. Despite the difference between the grammatical and the lexical, both are complementary syntagmatic features of the lexeme with some overlapping.

1.4 Stability and Regularity (Remaining Two Parameters)

Despite problems with delimitation of collocations (in a broad sense), at least two more familiar distinctions should briefly be observed. A basic one is that between (**6A**) **stable** and (**6B**) **not stable** combinations, where stable ones are part of the language system as combinations and any collocation belonging here should be viewed as a unit, while non-stable combinations are textual and do not form a higher unit, lexeme. It is evident that in their evidence and description it is the former that should be given more care and listed (ideally) in full, while the latter cannot often be listed exhaustively. It is generally recognised that to draw an exact line between the two may be difficult in practical (and lexicographical) life.

Finally, a useful distinction is that of (**7A**) **regular** and (**7B**) **anomalous** collocations (and higher combinations), i.e. of those based on semantic and formal rules and those that are not based on some of the (expected) rules. The former corresponds to the bulk of collocations of most types, while the latter may be identified with idioms and phrasemes mostly. Although idioms have been given attention for a long time, it is their satisfactory recognition as such and description that calls for more care.

Should one take the seven binary criteria (1A-7B) for collocations and their classification as a starting point, recognizing that some important types have been left out, such as multi-word terms (more in Čermák 2001, 2002), one should also take a consequent decision for their balanced treatment, too. In the following, a brief illustration of the current Longman practice (LDCE) will be offered and discussed, concentrating on some points only.

2 Situation of a Dictionary: an Analysis

Above (1.3), an illustration of LDOC's verb ***to ask*** has been given, partly. Such a basic entry is usually accompanied by data in its Phrase bank and Example bank boxes, standing, roughly, for collocations and sentence-type exemplification.

2.1 Collocations in LDOC

Examples from the former include collocations:

(*Dictionary phrases*) ask $50/$1,000 etc for sth, ask (sb) about sth, ask (sb) if/whether you can do sth, ask after sb, ask around, ask for, ask for sb, ask sb for sth, ask sb if/whether, ask sb in, ask sb out, ask sb over/round, ask sb sth, ask sb to do sth, ask sth of sb, ask that, ask to do sth, ask who/what/where etc, ask yourself sth, asking ... questions, be asking for it, be asking for trouble, be sb's for the asking, don't ask, don't ask me., if you ask me

(*Phrases from other entries*) a big ask- see ask, n, ask ... a favor- see favour, n, ask ... advice- see advice, n, ask for the moon- see moon, n, ask out for a meal- see meal, n, ask permission- see permission, n, ask ... question- see question, n, ask/beg sb's pardon (for sth)- see pardon, n, ask/beg/pray etc, for (sb's) forgiveness- see forgiveness, n, ask/tell sb flat out- see flat, adv, ask/tell/show sb the way- see way, n, asked ... difficult questions- see, question, n, asked ... outright- see outright, adv, asking for trouble- see trouble, n, attempt/do/ask etc the impossible- see impossible, n, be yours for the taking/asking- see yours_pron, forgive me for asking/saying etc sth- see forgive, v, forgive my asking/saying etc- see forgive, v, grant/obtain/ask/seek etc leave (to do sth)- see leave, n, I hate to ask/interrupt/disturb etc- see hate, v, if you don't mind my saying so/if you don't, ind me asking- see mind, v, may well ask- see may_modal verb, might I say/ask/add etc- see might_modal verb, might well ask- see might_modal verb,

need I ask/need I say more/ need I go on etc?- see need, v, pardon me for interrupting/asking/saying- see pardon, v, say/add/ask etc pointedly- see pointedly, adv, seek/ask for clarification- see clarification, n, speak/ask/answer etc directly- see directly, adv, who to ask/contact/blame etc- see who_pron

(*Words used with ask*) PREPOSITION: about, NOUNS: advice, permission, question, ADVERBS: for, how, why

On a closer look, all of the collocations under *Dictionary phrases* are those that have already been given in the lexicographical entry proper and it is hard to see this as being of much use. However, though this may be due to an interpretation, it is difficult to accept that only one preposition (*about*) is given here (under *Words used with ask*). Likewise, it is difficult to understand why only three nouns have been singled out, namely *advice, permission, question,* out of which only *question* has an importantly high frequency in BNC, but none that typically follow *ask* with a preposition or a question word (*for, about...*), a case which is very frequent indeed, such as *ask for advice, help, money,* etc. Moreover, there is no mention

of a rather frequent collocation *ask questions* whose high profile is evident from BNC. This has been an example of how collocations are handled for one of the highly frequent words (*to ask* being among the first 1000 words). If one looks in the opposite direction, at words on the frequency margin of the vocabulary (as recorded by LDCE), a picture may look like this:

abeyance
in abeyance something such as a custom, rule, or system that is in abeyance is not being used at the present time
fall into abeyance (=no longer be used)

Here, the Phrase bank box gives

(Dictionary phrases) fall into abeyance, in abeyance,
(Words used with: abeyance) PREPOSITIONS: in, into, VERBS: fall, hold

Thus, in contrast to the entry proper, the box gives only a single additional information, namely the collocate *hold*.

There is a number of such low-frequency words that, at the same time (see above), have a restricted collocability or collocation range and their use is therefore particularly difficult to deduce from the dictionary information.

Most languages do have such words and so does English. Words such as the following may look familiar: *kith and kin, nary, nelly, nick, niggling, nitty-gritty, nub, spitting, scot-free, searing, shipshape, sidesplitting, slake, sleight, smithereens, snit, snook, splitting, standstill, stash, staunch, stave, stir-crazy, stumbling, sub-zero, tit for tat, to and fro.*

2.2 Collocations in LDCE and BNC

However, to stick to LDCE's letter A, these include

abeyance, abjure, ablaze, aboard, abreast, abrogate, absently, absolution, accede, accessory, acclaim, acclamation, accolade, accordance, according to, accursed, accusatory, acerbic, acidly, adenoidal, admittance, adoptive, aflame, afloat, afoot, afoul, afterthought, aground, alec, alight, allay, aloof, amends, amiss, amok, anew, apace, aplomb, aright, arrant, arrears, askance, astray, astride, asunder, auspices, avail.

Let us have a look at some of these words and compare their profiles in LDCE and BNC, as far as their collocations go. Needless to stress that with such low-frequency words knowledge of their collocations is vital for any successful use. These include

abeyance (*fall in, hold,* while BNC suggests also *be, keep*), **abreast** (*draw, keep, stay, walk,* Phrase bank giving only *keep, stay* while BNC adds *go, come, pass* etc.), **abrogate** (*treaty, act, clause, privilege,* the last three being in BNC only), **absently** (*gaze, nod, ask, smile, caress, nod, wach, notice, stroke, repeat, look, ask, lit the gas*, where only *gaze* and *nod* are given in the dictionary), **absolution** (*bring, administer, give, have, pronounce, receive, seek,* where LDCE mentions only *give*), **abundance** (*of, in,* while *have, there is* comes from BNC), **accede** (*to, demand, request, throne, un/conditionally* where the last adverbial collocate comes from BNC), **acclaim** (*win, international, great, popular, public, general, at, by,* where both prepositions are mentioned by BNC only), **acclamation** (*elect by*, BNC

only), **accordance** (*in a. with*, since BNC records less that 1 per cent of independent use of *accordance*, i.e. outside of *in accordance with*, the very entry is highly misleading), **according to** (only *a. to*), **acerbic** (*wit, comments, humour, style, voice* where only *wit* is recorded by LDCE), **acidly** (*say, remark, think, smile* where LDCE gives only *say*), **aflame** (*face, eyes, body, a. with a drink*, only BNC), **afloat** (*keep, stay,* while BNC adds *be, get, go*), **afoot** (*be, plans, moves, changes,* without, however, mentioning its exclusive syntactic position, BNC adding *sabotage, tomfoolery*), **afterthought** (*as, add,* which is misleading since BNC shows that 90% of use consists in *as an a.*, not mentioned in LDCE), **agog** (only BNC *be, keep, peer*), **alec** (only as *smart a.*, thus a set collocation and no isolated word), **alight** (*with, set,* BNC adding *be, catch*), **allay** (*fear, concern, suspicion,* BNC adding *pain, nostalgia, criticism*), **aloof** (*hold, keep, remain, stay,* BNC adding *be, stand*, but also, importantly, its attributive use), **amends** (*make, try, for, in* which suggests a broader collocability than the almost exclusive *make a.*), **amiss** (*with, in, come, go, be,* BNC suggesting also *take it, anything a.?*), **amok** (only with *run,* hence no collocate but a set collocation), **anew** (*begin, start,* where BNC adds *build, consider* etc, but implies, as the LDCE definition would suggest, that it is not possible to combine it with *do, dream, irritate, ponder, worry*), **annul** (*election, marriage, result,* BNC adding *order, law, elections, decision, effect*), **apace** (*continued,* but BNC broadens to *grow, proceed, develop, increase, continue!*), **aright** (*be, set things,* while BNC adds *hear, lead, understand, set*), **arrant** (*nonsense!,* where BNC adds *hypocrisy, rudeness, sexism*), **arrears** (*in, be, fall, get,* with BNC adding *define, claim*, though it is almost exclusively used with only *be, fall in*), **asleep** (*be, fall, fast, sound,* BNC adding *lie*), **astride** (*sitting* BNS adding *sit, legs, horse, stool, motorbike*), **asunder** (*torn, split, rent,* while BNC offers also *blow, rend, tear*), **avail** (*be of/to no a.,* supplemented by BNC's *of little a.* suggesting, again, a set collocation).

In view of LDCE's being still somewhat isolated, but also a pioneering presentation of collocations in an almost systematic way, one has to see it as such; in general, however, a serious attempt has been made here to cover a hitherto neglected and important field. The small sample based on a single letter (A) suggests that much more can be found in the dictionary as a whole.

3 Questions and a Summary

Leaving aside the substance and precise delimitation of what collocations are and concentrating, rather, on recurrent and joint appearances of word combinations, it is evident that, despite this limited view, a number of questions remains open. Some of these are specifically relevant as desiderata for dictionaries, too. However, this summary, rather declaratory in its nature, does not aim at anything resembling a discussion, which is not possible here.

Obviously, it might (should) be a lexicographer's **aim** to give the user feeling of being sure that his/her use in selecting a word and its collocates is secure and correct.

It is, however, no easy task, even for the lexicographer who is aware of aspects 1-7 above, to find specific **criteria** for his/her selection, let alone dictionary presentation, of appropriate collocates and collocations. There, at least three criteria seem to be obvious.

Admittedly, these should respect (a) **usefulness** of selected data and information, which will naturally differ according to the dictionary's size and purpose (aiming at passive only or also active use).

Data and collocations should be (b) primarily **typical and regular** with a careful and principled choice of those that are not typical or regular.

Finally, (c) all combinations attested by the usage as being **set, stable** and firm should be recorded as such; hereby one drops, in fact, the idea of covering words and their collocates only and moves over to multi-word lexemes.

It may seem rather obvious and easy to say what has been said so far. Should one accept and follow this, one has to solve the ensuing consequences which are far from clear, however. Here, too, at least three major related **problems** may be mentioned. Lexicography and, in fact, linguistics too, does not have reliable tools for discrimination of

(**a**) what is **stable** from what is not (so) stable (and less stable); intuition and frequency likewise may not always be a help;

(**b**) what is (very) **typical** from (less typical and) marginal.

At least linguists should also insist that all cases of closed collocational classes should be given in full and this should be made clear to the user.

In addition to this, good dictionaries might aim at giving some information on

(**c**) **potentiality** of (not-recorded but obvious) use of dictionary items; these potential (or creative) uses suggest themselves rather easily with words having a fairly large number of collocates, for example.

Obviously, this contribution has concentrated mostly on one extreme of the collocation range of lexemes and its reflexion in dictionaries. Others, such as very large and medium-size type of collocability will have a different nature and will require a different approach.

References

A. Dictionaries

Longman Dictionary of Contemporary English, (ed.) Della Summers, 3rd ed. 2003, Longman, Edinburgh.

Oxford Collocations Dictionary for Students of English, 2002, (ed.) M. Runcie Oxford.

B. Other Literature

Čermák, F. (2000), 'Combination, Collocation and Multi-Word Units', in *Proceedings of The Ninth Euralex International Congress EURALEX 2000*, Heid, U., Evert, S., Lehmann, E., Rohrer, C. (eds), Institut für Maschinelle Sprachverarbeitung Universität Stuttgart, Stuttgart, pp. 489-495.

Čermák, F. (2002), 'Types of Language Nomination: Universals, Typology and Lexicographical Relevance', in Braasch A., Povlsen C., *Proceedings of the Tenth EURALEX International Congress EURALEX 2002*, Center for Sprogteknologi, Copenhagen, pp. 237-247.

Čermák, F. (2001), '*Syntagmatika slovníku: typy lexikálních kombinací*', in Čeština – univerzália a specifika 3, (eds.) Hladká, Z., Karlík, P., Brno 2001, pp. 223-232 (Lexicon's Syntagmatics: Types of Lexical Combinations).

Heid, U. (1994), 'On Ways Words Work Together – Topics in Lexical Combinatorics', in Martin, W. et al., *Proceedings of the VIth Euralex International Congress,* Amsterdam 1994, pp. 226-257.

Hoey, M. (2005), *Lexical Priming. A new theory of words and language*, London, Routledge.

Partington, A. (1998), *Patterns and Meaning. Using Corpora for English Language Research and Teaching*, Amsterdam, John Benjamins.