Combination, Collocation and Multi-Word Units

František ČERMÁK, Praha, Czech Republic

Abstract

There is a profusion of labels used for the phenomenon of multiword units (MU's), yet they are notoriously underrepresented in handbooks of both grammatical and dictionary type alike. After an illustration of both, problems of criteria which delimit MU's are raised up, together with an number of open issues and necessity of a functional approach is stressed. As three major issues, stableness, typicality and potentiality are discussed. A number of urgent questions waiting for a systematic solutions is listed and a final plea for a more balanced selection and approach is made.

1 Syntagmas, Combinations, Collocations and Other

In the multitude of approaches and handbooks a number of labels, both familiar and less familiar, is used to designate various syntagmas and word combinations that one comes across. If anything may be pointed out in these in general, then it is a lack of consistency of many types to be found here. In the situation of general consensus lacking and conflicting theories and views being proposed, this situation is not surprising. What is surprising is that the same or very similar phenomena are given widely different names even in the same single approach and book.

1.1 Grammatical Approaches

Grammatical approaches often tend to play the multiword units (MU's) down and neglect them. For a first illustration of this, let us have a look, at a representative grammar work, a modern classic now, namely Comprehensive Grammar of the English Language (Longman 1985) by R. Quirk, S. Greenbaum, G. Leech and J. Svartvik. It seems that a number of rather different approaches have been applied here. There are three main areas where multiword units have been observed or, rather, noted to exist, namely, verbs, prepositions and, up to some extent, nouns. The respective verbs are illustrated by such casese as *drink up, dispose of, get away with* and are called multi-word verbs (pp. 1150ff.) here. These are then further subclassified as phrasal verbs, prepositional verbs and phrasal-prepositional verbs, respectively. All of these cases are said to be a class of words which behave as single units. The apparent contradiction is disposed of by a somewhat unorthodox argument:

> "*The term "word" is frequently used, however, not only for a morphologically defined word classes, but also for an item which acts as a single word lexically or syntactically... It is this extented sense of 'verb' as a 'unit which behaves to some extent either lexically or syntactically as a single verb' that we use in labels such as 'prepositional verb'*".

Now, while the argument is true, the label "word" is difficult to accept here for a combination of words. Is is either one word or two words here, tertium non datur. One may only wonder why the term *lexeme* or, rather, *multi-word lexeme* is avoided.

Only a brief attention is paid to idioms here which are said to differ from multi-word verbs in types of transformations they might undergo (pp. 1162ff.). It is difficult to accept the first "semantic" criterion mentioned for the idiomatic status of MU's, namely a frequent existence of a single-word counterpart, e.g. *call for - visit*, as too imprecise, non-systematic and broad. On the other hand, the second criterion, that of the meaning being unpredictable, oddly contrasts with examples given in its support, such as *chatter away, fire away, work away*. Without going into technicalities, it is difficult to see these as really having a fully unpredictable meaning.

Multi-word prepositions, such as *apart from, as for, due to, thanks to*, is an obvious class of of MU's. Yet, it has been given here a different name, namely that of complex prepositions. While the prepositions are said here to be indivisible both syntactically and semantically, their generally open character is acknowledged and the boundary between a complex preposition and a single one is considered to be uncertain and open (pp. 669ff.). No mention is made, however, of idioms in this respect.

A similar situation is to be found with nouns, too, in that no mention is made here of idioms, although their frozen character is acknowledged. Moreover, no multiword character of such nouns is mentioned here either and the phenomenon is handled under the traditional label of compounds only (pp. 280ff.). The examples given here include such as *assistant director, break-down, sit-in*. This terminological inconsistency is further enhanced by a subsequent introduction, without any additional characterization, of the label "parallel structures" for some other types of multiword nouns. These are being illustrated, however, by rather typical idiomatic examples (although the term idiom is not mentioned), such as *arm in arm, side by side*. Somewhat later, finally, and as if as an afterthought, adverbial phrase idioms are mentioned, too, illustrated by *face to face*, but not related to the preceding parallel structures. One cannot escape the impression of neglect, hesitancy and no firm policy here, unfortunately.

It is disconcerting to see how capricious grammarians can get, as, apart from an occasional remark of adverbs such as *by day, by night* (p. 688), there is no mention whatsoever made here of MU's in chapters dealing with other word classes. It might be contended that their inconsistent choice of phenomena they decided to treat is due to the bulk of phnenomena relegated to and covered by dictionaries (although the fact is never mentioned). Yet, one must wonder why grammarians pick up for treatment only s o m e combinations and leave unattended others, equally important, and having the same character?

1.2 Dictionary Approaches

It is not surprising, then, that also dictionaries, almost invariably, give an equal impression of being inconsistent in their attention paid to MU's. There have been two basic approaches, forming opposite ends of a scale: one listing MU's under the chosen sense of the single-word lemma, and another one listing them separately at the end of the dictionary article. The first approach, perhaps an older one, tries to pin down, or rather, guess which sense of the single-word lemma might correspond to a particular MU, grouping, then, MU's there. This seems to

be an attempt almost metaphysical in its very root. As a consequence, it is difficult to find MU's in any reliable way as this guessing game is rather difficult to follow (and appreciate). This is patently obvious in particularly large dictionary articles.

The second approach, not pretending to know how to classify MU's according to senses, operates at the end of the dictionary article. Generally speaking, it is difficult to accept the underlying conviction that each idiom (MU) has somehow a sense directly attributable to a sense of the single word lemma. Hence, two alternative variant approaches of this can be found, one simply listing MU's under a common label at the end of the article and one, also at the end of the article, listing MU's in some kind of an order giving each a separate number. The former variant approach has been adopted by, for example, New Oxford Dictionary of English (1998), the latter by Random House Websters Unabridged Dictionary (1996).

As a rule, there is no particularly reliable criterion applied in the ordering of MU's in the first approach, neither in their functional labelling and the user has no way of knowing, except for their definition, how the MU's are used. An odd decision has been applied in the case of some nominal MU's which are not handled in this section (headed as PHRASES) but as separate lemma's in the alphabet (e.g. *blue shark, blue shift, blue chipper*). Disregarding the fact that the order introduced in the second approach is equally a problem, a major inconsistency is in the false impression evoked by the sequential numbering given both to individual senses of single-word lemma and to individual MU's. It is difficult to accept the underlying impression that each MU represents a separate sense of the headword.

Also here, no attempt is made to distinguish any further, such as to offer a functional labelling to MU's. It may be useful for the user to know, next to what the MU means, also how it is used. One of the ways how to achieve this, is to provide it with a functional information on, for example, word class type. It would help the user to realize in what ways a MU can and should be incorporated into a sentence. Elsewhere (Čermák 1998), it has been shown that idioms seem to be copying, in their behaviour, the word class function of the single word lexemes and that there are as many functional types of idioms as the number of word classes. Thus it is not difficult to adopt a unified approach to MU's and provide them with this information, too. Consider, for example, such cases, as being idioms or, rather, MU's, as multiword conjunctions, prepositions, particles, interjections and adverbs (to leave aside the more notorious ones, such as verbs and nouns), as illustrated by *as if, even though* (: conjunctions); *as to, as long as* (: prepositions); *all right, as well* (: particles); *All right!, For God's sake!* (: interjections); *for good, on the other hand, from head to foot* (: adverbs) etc.

2 Problems of Criteria and Open Questions

2.1 Syntagmatic aspects

Providing MU's in dictionaries with functional labels may not be the most important problem to be solved. There are other problems as well. Most dictionaries do not even attempt to strike some sort of balance between coverage of their paradigmatic and syntagmatic aspects and it is the syntagmatic aspects which are sadly lacking in what might be called a decent representation. While some dictionaries do not even attempt to state that the noun *attention* collocates, among

other things, with both the verbs *pay* and *give*, some hide the fact in sentential examples, while some other merely list the combinations. Listing them is about most what one can expect to get from even a large dictionary. Even specialized works, such as The BBI Combinatory Dictionary of English (BBI 1986), never tell one what the difference in the use and meaning of both might be and may be somewhat misleading. Thus, *The Word Bank* corpus gives 79 combinations of *attention* with forms of *pay*, while the number of its combinations with *give*, i.e. 35, is 50 per cent smaller. Yet it is given much more prominence in the BBI being listed independently under a separate number. To take a different example, on the basis of the collocation *in broad daylight*, a foreign learner of English might be tempted by the lurking analogy to try to combine *broad* with its opposite, too, making it *broad night*. Yet this is exactly the kind of information which is never given reliably in dictionaries and the user is left to his or her own devices here. Obviously, the problem, probably the most serious for a lexicographer, is the degree of stableness of combinations since stableness is closely linked to acceptability.

2.2 Stableness

It is not very often that the problem of stableness (or stabilization) of lexeme combinations is brought up, let alone helpfully tackled. People may not even realize that the crucial question to be first asked, to put it linguistically, is "does one deal here, in the given combination, with a phenomenon of la langue already or still of la parole?". All of the stable combinations, MU's, are part of la langue and merely reproduced in speech, whereas combinations which are not stable, fixed (some prefer to call them preconstructed, frozen etc.) are part of la parole, the speech, and are formed ad hoc, again and again in each case. So far the theory. The problem with such a nice clear-cut theory is that it does not always work. Since, obviously, there is a multitude of combinations which share characteristics of both, one has to use a cline or scalarapproach here with as many grades between the two opposites as one may find and discern. Yet the necesssity to use this scalar approach does not invalidate the distinction. Rather, it views it prototypically, in an ideal case, and it is up to the lexicographer to draw the line. However, the enormous lexicographer's aid in today's corpora has its pitfalls, too, in the current overemphasis on typicality, on inclusion into the dictionary only of what has been sufficiently and amply attested.

One may wonder what kind of benefit could be derived from, for example, a careful selection and inclusion of current and yet not so typical cases of lexeme combinations stretching the norm somewhat whereby also less typical but not uncommon collocations would thus indicate a further and possible use.

2.3 Typicality

In the quest for criteria, if any (with most approaches offering only pragmatical ad hoc solutions), a kind of answer is sought in typicality. While the best solution is seen in the Mutual Information and t-score so far where the quest for criteria has rested, the problem is far from being solved. In a recent contribution (Trap-Jensen 1996), a distinction was made between two kinds of typicality, one provided by the corpus and one by data from a simple association test in Danish. However, examples such as *white minority rule, white coat* (coming from the Danish

corpus) as against *white snow, white sheet (of paper)* (coming from an association test) seem to show two different phenomena rather than two kinds of typicality.

While the former examples stand for typical collocations, it is rather problematic to call the latter standard collocations (whatever that might mean). Admittedly, these are difficult to come by in a corpus of any size so far. However, a slightly more common form of the latter of the type, i.e. *snow/paper is white*, being used for classification of more than one kind, signals rather special (but hardly typical) occasions when such cases might be used and the particular quality (white colour) of *snow* or *paper* explicitly mentioned. It is obvious that *whiteness* is a definitional quality which might be one of more, perhaps many, and used, for example, for identification. It is very much part of la langue; de Saussure would undoubtedly call it part of valeur, value of the lexeme. There might be, ultimately, also a cline here without any sharp dividing line between the two. Yet, it seems, for the time being, that an alternative, "temporal" way of how to view the two kinds of syntagmas or combinations lies in the stability of the existence of their denotata, as these are reflected by the language system. There is hardly any doubt about the constant, "timeless" quality of the colour of *snow*, but there is nothing constant about, say, the colour of the *minority rule.* Contrary to what statistics and frequency shows, the degree of stability of the former in the language is far greater than that of the latter. It is evident that, despite an obvious interrelationship, such a type of stability is not to be confused with the stableness or, rather, fixity of a combination in the sense referred to above.

There is, however, room for a possible misunderstanding, too. It would hardly do to consider either kind to be outside of either la langue or la parole, of either the lexicon as part of la langue, or the area of semantico-grammatical rule application in texts. There is nothing in between (contrary to Trap-Jensen 1996, 283) and seeming transitions may turn out to be due to a lack of data or differences in norm applied by different users.

2.4 Potentiality

Next to positive facts, units and their combinations, there are also rules in the lexicon. It is often forgotten, while speaking about rules, either semantic or grammatical, that there are other potential rules, too, in the lexicon, whose existence is directly due to a prior existence of the former kind and which represent a kind of extension of the former. Disregarding for the moment that even large corpora are far from ideal, one may become aware of the opposite end of a new scale here, of obvious potentiality to create, on the basis of a strong analogy, new combinations which are (almost) perfectly acceptable. While much may be objected against the recent Chomskyan overblown emphasis on the language creativity, this is not quite the same thing. Especially in cases where common sense suggests a possibility, speakers do not hesitate to form a new combination and use it. This potentiality, contrasted to probability, however, has not found its treatment in dictionaries, so far. After all, not all users wish to speak and appreciate only what is typical and pre-patterned and they might wish to strike a subjectively new balance between the new and the stereotype.

2.5 Rules and Regularity

A question often, though rather implicitly, asked by the lexicographer is about regularity of a given combination. Yet, what is regular, is in fact related to a number of very different rules,

some of which holding for a particular and rather narrow domain only, but related also to an endless debate should semantic rules be included. Thus the seeming commonsense approach saying that any irregular combination is of (primary) interest to the lexicographer does not say much, as the distinction regular-irregular is far from being clear.

A different kind of regularity is to be found in usage where standard use is rule-based, of course. However, every kind of usage means following certain rules to the exclusion of other; there are always some rules which do not apply. One may, then, wonder, if a kind of guidance to the usage may not be offered, at least in some instances, by indication, for example, the impossible or "forbidden" cases. While this interesting problem would ultimately lead to a discussion of what is feasible, let me point out at least one positive example here instead. On the Czech dictionary of idioms (Čermák 1994) this possibility has been succesfully employed by explicitly indicating a set of (morphological and syntactical) categories or, rather, transformations, which a given idiom does not normally undergo.

2.6 Outstanding questions

The number of questions waiting for an answer and solution seems to be endless. In connection to the distinction just made, some of the outstanding question might, briefly, be:

1- How relevant for the identification of an acceptable combination is its length, extension (in number of words, usually), as well as its discontinuous character occurring rather uncomfortably often? It is natural that most illustrations one finds in dictionaries are simple binary adjacent combinations. But one should expect more.

2- How are certain stereotypes and other prebabricated combinations to be identified and represented if these have a rather unstable, multivariate form? An answer of a kind is linked to the size of corpora used. It would seem to be just impossible to find (often) two major variants with an equal frequency. Thus, an invariant is indicated.

3- How are exocentric types of combinations to be identified and covered? Their character does not yield the impression of an entity so readily as that of endocentric constructions. Undoubtedly, there is always the lexicographer’s intuition at play while choosing one example and discarding another one. And intuition seems to be rather in favour of endocentric combinations.

4- To what extent should combinations, collocations etc. in their treatment be viewed typologically? The fact that, for example, polysynthetic languages rely heavily on compounds where other languages use combinations should not prevent one from the same effort in their description. Obviously, discrete character of combinations is due to the typological character of the language in question. It is difficult to see any difference in criteria and approach between collocations and compounds, which are, basically, of the same character.

3 Desiderata and Solutions?

It is obvious, as many people would argue, that dictionaries should ideally include a more balanced selection of all system units on a multiword level. However, since there is no such thing as an ideal dictionary offering this desirable state of affairs one must assume that real, practical solutions are always based on some kind of non-ideal selection. And it is up the user to either accept it in good faith (realizing, perhaps, that it is difficult to do better) or (being more suspicious of money-oriented publishers) to distrust it. However, without answers to the above and other questions it is still very much a wishfull thinking and the practice of ad hoc solutions will not be abandoned. Feasible solutions have to be sought in the identification of criteria which follow from these questions. Since most of the problems outlined have a scalar character, a specific kind of answer here, specifically that of selection, is to be found, among other things, in their frequency profiles. Resulting varying degrees of selection on a scale could be attributed to various sizes and types of dictionaries.

References

[1] Benson, M. Benson, E., and Ilson, R. (1986). *The BBI Combinatory Dictionary of English. A Guide to Word Combinations.* John Benjamins Publishing Comp., Amsterdam/Philadelphia.

[2] Čermák, František (1994). Czech Idiom Dictionary. in *Euralex 1994-Proceedings,* eds. W. Martin, W. Meijs et al., Euralex Amsterdam, pp. 426-431.

[3] Čermák, František (1998). Linguistic Units and Text Entities: Theory and Practice, in *Actes EURALEX'98 Proceedings,* Th. Fontenelle, Ph. Hiligsmann, A. Michiels, A. Moulin, S. Theissen (eds.). Université de Liege, Liege, pp. 281-290.

[4] *The New Oxford Dictionary of English* (1998). Oxford University Press, Oxford.

[5] *Random House Websters Unabridged Dictionary* (1996). Random House, New York.

[6] Trap-Jensen, Lars (1996). Word Relations: Two Kinds of Typicality and Their Place in the Dictionary. in *Euralex '96 Proceedings I-II.* University of Göteborg, G"oteborg, pp. 283-291.