František ČERMÁK, The Institute of the Czech National Corpus, Faculty of Philosophy, Charles University, Prague.

Linguistic Units and Text Entities: Theory and Practice

Abstract

Both linguistic and lexicographic aspects of the system and text units are examined and the question of their mutual mapping is taken up. From a semiotic view and against the background of F. de Saussure's ideas about the language entity and unit, three levels of units and entities suggested for inclusion in the lexicographer's description are suggested. Current preoccupation with typical collocations seems to be too narrow to include other units and rare syntagmas. Alternative views on some common collocations are examined and a need for other criteria is voiced. A detailed example of a functional classification of collocational idioms is offered for the Czech language followed by a survey of basic combinatory phenomena found in the text and system.

Keywords: (multiword) units, entities, idioms, system, collocation, corpus morpheme, abstract nouns

1. Introduction

There has always been a paradox. In linguistic theory, the notion of various familiar **system units**, such as **lexeme** is taken for granted, usually. However, what is being used as material basis and presented in dictionaries is far from clear, both theoretically and practically. Hence, there are too many "practical" ad hoc solutions. Moreover, in some lexicographic approaches, no attempt or pretension is made to veil their refusal to come to grips with this mapping of system units into text and vice versa. What should this relation be, if any? To put it differently, one can ask such general questions as:

(1) are there two **different sets of entities**, for the language system and the text, respectively, or
(2) is there some **overlapping** between them only, or
(3) is there a weaker or stronger **mapping** of the former onto the latter?

Ever since Chomskyan beginnings, it has become fashionable in the profusion of theories and countertheories, to stop using in linguistics such general Saussurean terms as **unit** or **entity**, as if they did not have any substance, while more specific terms, such as **word, lexeme, idiom** etc. are used here. This has been accompanied by a strong dislike to do any exhaustive material mapping (in the sense referred to above). Hence, there is a multitude of transformational grammars or rather theories and proposals of various denominations along transformational lines, but not a single "transformational" dictionary. This disregard for the units of the system has had a broad impact, unfortunately. On the other hand, corpus linguistics, true to its meticulous search for any type of form or configuration of forms in text which might seem significant (Sinclair 1991 and others), has shifted the focus from the system units to their textual forms and collocations while always stressing the exhaustive type of approach. Hence the corpus-based dictionaries approach is advocated here, since obviously, a considerable amount of data has to be made available first and only then, a useful selection can take place.

Now, the corpus linguistics is basically on track, but what exactly is to be understood by **significant**, i.e. by significant forms and their combinations? Should it mean **typical** only, as is often the case nowadays, then a lot of what is **rare** in the corpus and still represents some kind of the system unit, becomes a drop-out, automatically. Delving a bit deeper, one has to admit that the reason for this is in the insufficient size of the corpus used which is to be partly blamed.

In general, if we disregard any hasty planning of a dictionary, lack of expertise and/or theoretical background, then primarily, two basic sorts of problems relating to units have to be singled out:

(1) *system-text mapping* and
(2) *quality and quantity of data* available to us.

Clearly, solutions to these problems should be evaluated on a scale according to the type of compromise, where some compromises would be better, some worse, and some hardly acceptable. In what follows, both problems will be commented on, briefly.

2. Units and Entities

De Saussure regarded **units** of *la langue* as being identical with language **entities** (de Saussure 1982, 145), which are defined in his *Cours* as being based on oppositions (145) and differences (168) within the sum total of their mutual associations in the system (la langue, 189). To put it differently, these entities exist only through their meaning ("sens") and function (191, 149). However, what he did not mention so explicitly, but what is recorded in the manuscript edition of *Cours* by Engler (1967-1974), is his emphasis on the necessary condition that any unit, for it to be a unit, must be felt as such (Engler 2152 B). This, clearly, implies its shared stability and fixedness in the system. De Saussure's usage of the term **entity** is, unfortunately, somewhat misleading. It seems that by the **concrete entity** he generally meant both what one would, today, call system units as against text ones (cf. 145), i.e., for example, **lexeme** and its **manifestations** (if any), while the term **abstract entities** was used, primarily, for such notions as genitive case, word class, word order etc. (190), which he viewed as being always based, in a final analysis, on the concrete entities (190). Thus, having distinguished between **class** (category) and its **members** (units), he has not made any systematic attempt here to distinguish between the **type-token** levels of the unit's existence. Semiotically, an entity may exist, in his well-known view, only thanks to its association of the signified (*signifié*) and the signifier (*signifiant*, 144).

Among other things, these statements imply that any alleged entity, where the link between its form and function or meaning (or its signified, *signifié*) is not clear, stable or documented at all, must be viewed with some suspicion and may be discarded, if a further analysis does not help in clearing it up. This is, in a nutshell, what linguists and specifically lexicographers do while trying to single out, from the text continuum, meaningful entities. However, on the basis of this Saussurean conclusion some uncomfortable questions can also be formulated. Given the obvious and patently true hypothesis that words do not collocate (combine) mutually in the same degree everywhere in the lexicon, a continuum of combination or collocation types must be presumed; then, along a scale starting with very obvious and frequent combinations through less frequent combinations, ending in highly infrequent and hypothetical (potential) ones.

Now, if the notion of **collocation** (as in Sinclair, 1991, for a survey see, e.g., Heid, 1994) is to be taken to mean *any* type of meaningful combination, should the term not be discarded, since

the term **lexical syntagmatics** would do the job? Or should it, on the other hand, be taken to designate some section of the scale only? How does one define it, by valency, syntactic, semantic, pragmatic criteria, or any combination of these? Hence, the widespread confusion about the term **collocation** and its ill-defined nature. Although there is no doubt about the relevance of **functional and semantic criteria** on which any sensible lexical **syntagma** (collocational and other) is based, there are still precious few statistical tools, such as MI-Score and t-score, which might help us to shed some light into the vast combinatory realms of lexical units. It has to be borne in mind that most of the approaches so far have dealt exclusively with combinatory phenomena which are common and typical. Clearly, only a **minor** part of the field has been paid some attention.

While I do not propose to go into this matter any further here, I want to point to an obvious dividing line in the combinatory continuum which may be found in de Saussure's argumentation: where a stable and clear link between both form and function/meaning may be repeatedly found, these might be seen as fixed units, **lexemes**, made up of both single and multiple word forms and forming part of the language system. The rest is different in its not being (quite) stable and fixed. There is no doubt that the solution of the **fixity or stability** problem may be sought, at least in part, just at this relation. But is the boundary between both parts of the combinatorial continuum really clear? Often it is not, and where it is, are there any specific areas to be delimited in what lies behind "the border line"? The obvious first candidates for fixed units, stable entities, found both in the system and the text, are **idioms** and **(technical) terms** of all sorts. But what about the rest?

3. Levels and Entities

3.1. Above the Word

The rest, largely unexplored, seems to be a tangle of problems. Turning over to the clearer and better-explored part of the combinatory continuum, how should we view such combinatory cases as *black coffee* and *white coffee*? Provided that we admit the combination *green coffee* as acceptable, too, even though it is not quite the same thing, and if, perhaps, *pale coffee* might be passed for an expression of a criticism, how acceptable are such combinations, i.e. of a colour adjective + the noun *coffee*, such as **blue coffee, *violet coffee, *scarlet coffee, *yellow coffee...?* There is no excluding these from our consideration on any ground whatsoever, but their probability of occurrence seems to be quite low and the present size of corpora is simply not sufficiently large to be of much help here. Declaring *black/white coffee* to be collocations does not solve much as, in doing so, only the **frequency** (and, perhaps, **habituality**) aspect would be stressed, leaving the other aspects unattended, and we do not know whether that alone can be sufficient as a criterion. Should the size of corpora be made tenfold, would not the scope and horizon of frequent combinations grow and a new selection and reclassification would become feasible? Even if we accept this frequency criterion with some reservation, we still have no way of knowing where to stop and draw some sort of line (lines?) in the realm of less frequent and rare combinations. Some of the problems to be dealt with here include the following:

(A) Do we, by calling these combinations **collocations**, solve the problem of classifying them as system entities, **multiword lexemes**, or as something else? The oversimplification is rather obvious.

(B) Should we, following the Prague school scholars, view this problem also as made up of **complex denominations** (Mathesius 1975, Filipec-Čermák 1985)? Due to its psychological, i.e. external starting-point, this approach does not solve much in this context. It may be objected that rare combinations, dwindling gradually into mere **potentialities**, may not be of immediate concern for many practical lexicographers (yet), but they are very much the concern of theoretical linguists. To make matters more complicated, let us notice that neither *black coffee* nor *white coffee* is really black or white, respectively (explanations such as Mel'čuk's non-standard lexical function being in operation here are of ad hoc value only indicating that a more general approach applicable elsewhere is needed, Mel'čuk 1995).

(C) Should this aspect, then, be used for declaring these combinations to be **idioms**, since the familiar definition of idioms based on non-compositionality still seems to prevail? Or,

(D) Taking into account, at least in culinary circles, the existence of some sort of definition (cooking instructions, recipes) being available for them, should *black/white coffee* be declared to be **terms**?

A corpus analysis (Cobuild 1995) of the collocations of *coffee* has, in fact, revealed only that there is only one combination of it with a colour adjective, namely *black coffee* (occurring twice in the five-million corpus). The rest is different; combinations such as *fresh coffee, freshly ground coffee, instant coffee, strong coffee* etc. are hardly acceptable as stable and fixed combinations (in contradistinction to collocations in some approaches), perhaps with the exception of *instant coffee* which is a term. It is clear that moving from what is (proto)typical and easily recognizable to the less typical brings one into the realm of the **vague**, non-determinate phenomena burdened with growing scarcity of criteria. I am afraid that neither of the above criteria used offers any generally acceptable approach. This being so, all sorts of ad hoc approaches and partial classifications emerge. The "*coffee*" example used is a very simple one indicating its clear limitation due to the closed set of colour terms; but vast realms of words and their combinations are not like that at all.

There is something counterintuitive in the concept of **multiword units**, i.e. in viewing a discrete combination as a **whole**. Accordingly, this term is not very much employed, except in the most obvious cases, such as typical idioms and terms. It is quite difficult, for instance, to find a grammar of English which would offer an **exhaustive list** of multiword prepositions, such as *as to, as for, in connection with*, conjunctions etc. In fact, they are not found in grammars of most languages; yet, for example, the latest count for the Czech language has come up with no less than some 400 of such prepositions. While they usually do appear in dictionaries, they may not be labeled there as prepositions at all and are often presented in some other way (in boldface etc.) as some further unspecified combinations. This approach is specifically hard to accept in grammars, which almost always pretend to be so very much exhaustive, but neither is any ideal solution to be found in dictionaries.

3.2. Word Forms

But there are also entities of the opposite order, lying below the level of single word lexemes, namely **word forms**. Traditionally, they have been used to demonstrate that there is a difference between the system entities and text ones, although perhaps in degree only, not principally. It does not automatically follow that between such forms as, for example, *go* and *went* there is the **type-token** difference. On the other hand, it is a familiar experience that some word forms, to

the exclusion of others, have a different meaning (sense) not only contextually but always (repeatedly); this is a fact very much stressed nowadays by corpus linguistics. Beside irregular plural forms (such as the English *mice*) and suppletive forms (see *go-went* above), such (personal inflectional) forms as *I wonder (why...)* contrasted with other forms of the verb, such as *he wondered, I wondered (why...)* etc., may appear to be a typical case, especially if a translation into another language is attempted: while, in the Czech language, *I wonder* is best translated as *rád bych věděl* (i.e., with the Conditional Mood, literally "I would like to know"), the other forms of this verb correspond to different equivalents, such as *být zvědavý, divit se* and there is no possibility for a mutual substitution. However, our awareness and attention paid to the existence and number of such cases has been only marginal and not systematic, thus far.

3.3. Below the Word

In some languages where there is typologically no inflection, this word form type of entity may assume a different appearance and include only a **part** of the word, or, rather, of a single **morpheme (root, affix)**. This is the case, primarily but not exclusively, of verbal prefixes in Slavonic languages, of some compositional roots in Germanic, Uralic languages or Sino-Tibetan type of languages and of other phenomena, such as those which are related to the phenomenon of incorporation and polysynthesis in many other types of languages.

Thus, for example, such Slavonic (mainly Czech, Slovak, Polish) **prefixes** as *vy-/wy-* (out of), *z-* (used for perfectivization) are specimens of a boundary phenomenon between the **affix** with a limited distribution (which belongs to a limited set of words, each of them being recorded in dictionary) and a free element, enabling an independent formation of new words. These are quite acceptable; even though they are never written in isolation, some of them not being found in dictionaries. To list them in a grammar only would greatly limit uses of a good dictionary. Such **suffixes** as the Indonesian *-kah* or not quite identical Finnish *-ko/-kö*, used for signalizing questions, might represent a similar case. Some elements in Chinese compounds, such as *wù* in *ren wù* "person/man (← man + thing)" are hardly ever used outside their respective compounds, and, for a number of reasons, a dictionary should record them, too. Some **compound elements** may appear as a surprisingly useful element in a bilingual dictionary between a language which abounds in compounds and one which does not use them much, cf. *house-, huis-, hus-, koti-, ház-* in English, Dutch, Swedish, Finnish and Hungarian and the corresponding Czech and Polish full-fledged adjectives *domácí, domowy* etc. The dictionary compiler creates here, for the benefit of the user, an entry artificially based on a decomposition in such languages as these five.

From the point of view of lexical language denomination, there are basically four possible ways how to best express such common (simple) notions in various languages. These are to be seen in the equivalent expressions for the same thing, namely *RAILWAY*, in, for example, Czech, Finnish and French, cf.

(1) a derivative *železnice*,
(2) a compound *rautatie*, and
(3) a collocation *chemin de fer*;

the fourth case, a simple root, is not available here. To return back to Chinese and other polytonic languages, one must take into account that it is the syllable that is the dominant unit here, overlapping largely with the morpheme (tantamount to root) and words; due to the

Chinese way of writing (to be found in other languages, too) placing characters next to each other, it is difficult to distinguish all of the boundaries here. Thus, is the Chinese *tiělù* (=iron+way) a compound or a collocation? Accordingly, the distinction between the **compound** and **collocation**, that one is so confidently accustomed to recognize in European languages, is extremely blurred and problematic to draw here. It does cast some shadow on the wisdom of a one-sided preoccupation with collocations only.

The same principle of writing, which may not be so apparent from Chinese due to its use of characters, is much more clear in Vietnamese which did use Chinese characters a short time ago, too. Here (and elsewhere), still another phenomenon to be mentioned is a number of dependent **function words** which serve purely grammatical and pragmatic functions, i.e. they are never used independently as words, such as *à* (a question particle) or *các* (non-definite noun plural particle). It could be argued that some of these words (Vietnamese and other) merit their inclusion into a dictionary (a bilingual one, primarily), since they do not seem to have any obvious counterpart, but that is only one of the reasons. Rather, they are specimens of the non-existence of any sharp boundary between lexis and grammar. Many languages have very common and extremely frequent words used for grammar purposes solely, which a dictionary must include, see yet another example of such words for signalling questions in the initial Polish *czy* or alternating Malay/Indonesian *b(id)ak* and *(bu)kan*. In all of these cases, we have to deal with rather specific syntagmatic combinations, not typical collocations.

It is evident that there are, then, at least **three levels of entities** of some interest, for both the linguist and the lexicographer. Note that due to the existence of both types of entities, which are either larger or smaller than the word, those languages where the term *dictionary* is based on the root used for the WORD, as in the Slavonic or Germanic languages, it may no longer be quite fitting. In its pointing to words only, it may, in fact, become somewhat misleading, cf., for example, Czech *slovník*, Finnish *sanakirja*, or Swedish *ordbok* etc. (where *slovo, sana* and *ord* stand for "word").

4. Multi-Word Entities and Units.

It is difficult to discern in the diversity of **sub-word entities** any system or at least classes and languages differ here very much. The other, opposite type consisting of **multi-word units** is somewhat different. To show some of the regularities and systematic aspects of this area of entities, examples from a single language will be, briefly, used (SČFI 1-3, 1983, 1988, 1994, Filipec-Čermák, 1985, Čermák 1994b).

In many respects, the criterion of the formal class membership (e.g. word classes) seems to be useful for a formal classification of word combinations and idioms; also it has long been used in the Czech linguistic practice. Yet, on a closer look, it is a mere **external** type of criterion covering the type of component parts used in their construction, which may be particularly suitable for **formal** statistical and combinatory-based approaches, but to a smaller degree for other approaches. Should we want to use, on the other hand, meaning and (nominative) **function** as the criterion here, a different and more revealing approach would be implied. An important aspect of **idioms** is that **functionally** they may be viewed as **extensions** of all standard word classes. Since we are dealing with units on and around the word level, sentence type idioms seem to be excluded, but that is not quite true. There are, traditionally, ten word classes recognized in the Czech language and there are as many functional classes

corresponding to these in the region of the Czech collocational (i.e. sub-sentential) idioms; thus, every single word class has a kind of extension in functionally equivalent idioms. These may, then, be illustrated by such simple cases as:

1 **NOUNS:** *zlatý důl* ("a gold mine"), *teplé místečko* ("a cushy job")
2 **ADJECTIVES:** *neslaný nemastný* ("wishy-washy"), *stalí mladí* ("young and old alike")
3 **VERBS:** *číst někomu levity* ("read someone the riot-act"), *umět se narodit* ("have the luck of the devil")
4 **ADVERBS:** *široko daleko* ("widely, in many places"), *kolem dokola* ("round and about")
5 **PRONOUNS:** *nějaký ten* ("some, quite a few"), *moje maličkost* ("my humble self")
6 **NUMERALS:** *jeden dva* ("one or two, some"), *jeden za druhým* ("one after another")
7 **PREPOSITIONS:** *co do* ("as for"), *s ohledem na* ("with regard to"), *až na* ("with the exception of")
8 **CONJUNCTIONS:** *ba i* ("even, to be sure"), *kór když* ("let alone")
9 **PARTICLES:** *co když* ("what about/if"), *kdyby tak* ("I wish")
10 **INTERJECTIONS:** *jen klid!* ("(take it) easy"), *to zrovna!* ("no idea!"); this class includes an important part of sentence idioms, too.

There are some tendencies to be observed here which have their implications for the job of the dictionary compilation, too. Let us mention some of them.

a-As a rule, the number of grammar idioms (5-9, with the obvious exception of numerals) is far greater than the number of single-form grammar words in each of the respective classes. This makes them a serious and important extension of the single-item grammar word stock which otherwise does not really grow any longer.

b-A peculiar feature is a pronounced scarcity of functionally adjectival idioms.

c-The Verb-Noun relation differs statistically from the relation of these two word classes in the vocabulary: there are several times more of the verbal idioms (of many types) than noun idioms. This relation, however, has its important counterweight in **terms**, which belong to another area of multi-word units. Clearly, with the majority of multi-word terms in the area of functional nouns, terms can be viewed as a kind of opposite end on the scale whose one end is taken up by idioms having its major concentration in verbal types.

Linguistically, it would be very important to compare statistical data here from more languages, to see if there are some general tendencies behind this single-language picture.

Traditionally, a vast part of what is usually covered by the term of **collocation** (whatever that might be) consists of **verb-noun combinations**, made up of abstract nouns, mostly. If one tries to look up relevant combinations offered in corpus (Word Bank 1995, Collins, 5 million words) of such a typical **abstract noun** as *ATTENTION*, what one gets is an amazing and perhaps confusing multitude of combinations based on such verbs, as these:

arrest, attract, bring to, call a. to, catch, capture, centre, claim, come to, command, concentrate, crave for, demand, detract, develop, devote, direct, distract, divert, draw, escape, focus,

gain, get, grab, have, hold, keep, lavish, love, need, pay, pull, receive, require, return a. to, share, shift, slip, take a. away, turn, vary, vie for, want, win (ATTENTION).

Singling out as special idioms *stand at attention* and, perhaps, *call somebody to attention*, what one can use for classification are only frequency data, which are not really much help. Using them in descending order, they read here roughly like this: *pay, draw, focus, get, attract, turn*; oddly enough, they rather correspond to the choice made by many dictionaries too, but that is hardly a way how to deal with the rest. One way how to handle these bases is to treat them on the basis of restrictions, in the various transformations they may or may not undergo (such as insertion of an adjective or *your*, passivization etc.).

In the Czech theory of **idioms**, it has been shown that their substance should not be sought in deviant, non-compositional semantics, primarily, but in the **anomaly of the function** of their components (see, for example, Čermák 1994a). In this approach, based on a commutation test, the above-mentioned combinatory continuum of words can, indeed, be further split and many cases of what has been traditionally passed for collocations, may rather safely be classed as semi-idioms. It is purely on the basis of a limited inventory of stable combinations of such typical abstract nouns as *attention* that these combinations begin to appear as a limited and small group of **semi-idioms** most of which a part of one or more larger structures.

Compare the simple **IN**choative, **DUR**ative and **TERM**inative **Phases** (which may not always have a verb representative for every noun) and its parallel **C**ausative series of expressions in three phases, where the identity of the meaning of the noun and agents involved has to be retained throughout:

ATTENTION^1: (two human agents)

IN	pay, focus, turn, devote	**DUR** pay, devote	**TERM** ? [withdraw]
C-IN	draw, attract, catch, bring to, call to, capture	**C-DUR** command, hold, have	**C-TERM** distract, divert, take away

ATTENTION^2: (non-human agent)

IN	come to	**DUR** command, have, require?	**TERM** ? [lose]

However provisional the inclusion of only some verbs may be, the intention being to include the more stable and frequent ones, the picture shows the patterning rather clearly. Of course, a further corroboration (such as for the inclusion of *withdraw*) is necessary on a much larger corpus.

A successful attempt has been made to describe in this way Czech verbal semi-idioms of this type in a rather exhaustive way; it has entered the last volume of the Czech Idiom Dictionary (SČFI 3, 1994). At the same time, this approach pointed to the existence of both larger structures in the human mind and in human lexicon.

One might consider also other stable combinations, which might be passed for multi-word units, or rather for semi-idioms; even though this is usually not the case with many dictionaries.

Briefly, there are **valency** prepositional combinations, such as *insist on/that/0*. On a closer inspection, it is apparent that *insist* does not enter, as a rule, into any other combination. Why, then, is this information usually submerged inside the dictionary entry, if these two combinations could be viewed as obvious multi-word units, namely *insist on*, and *insist that*?

To conclude this part, let me stress that in the logic of this approach **sentence-type** multi-word entities should be given the same treatment in dictionaries as single words, whether they are idioms or belong elsewhere.

Most of the latter cases are subject to different treatments and represent very much open problems. It is, however, necessary to view them as separate entities, no matter what labels they might have taken, and treat them both systematically and in due detail. They are usually much more complicated than single words (lexemes) which make them up.

5. A Survey and Summary

To sum up, it seems that word combinations (C), given the present state of our knowledge, fall into several classes.

		SYSTEM -	**1** regular:	terms (*nitric acid*)
	/	(stable)	**2** irregular:	idioms (*give a hand to*),
				semi-idioms (*turn attention to*)
C-				
	\			
		TEXT -	**3** regular:	a-grammatico-semantic expressions (*read a book*)
				including some of traditional collocations
				b-grammatical (*would read, has been*)
				c-terms (*nitric acid*)
			4 irregular:	a-idioms (*give a hand to*),
				semi-idioms (*turn attention to*)
				b-author metaphors
				c-accidental combinations representing no entities
				(*[the general consensus in Russian] elite was that [the major threats]*)
				d-gibberish, nonsense

Not all of these combinations are, however, language entities (as in 4b-d), some of which do not make much sense (4c-d). Some (as in 3b) are text entities only and belong, in the system, to a single-word lexeme. Since, however, there is no way yet how to incorporate in the dictionaries the combinatorial **potentiality** of lexemes, this picture may seem to consist of more clear-cut classes than the language reality suggests.

6. References

Benson, M. (1990). Collocations and General-Purpose Dictionary, in: *International Journal of Lexicography*, Vol. 3, 1, pp. 23-34.

Cobuild on Compact Disc, Ver. 1.2. (1995) Harper Collins Publishers: Birmingham (Word Bank Corpus based on 200 million Bank of English Corpus).

Cruse, D.A. (1986). *Lexical Semantics.* Cambridge U.P.: Cambridge.

Čermák, F. (1988). On the Substance of Idioms. *Folia Linguistica* XXII/3-4, pp. 413-438.

Čermák, F. 1994a. Czech Idiom Dictionary, in *Euralex 1994-Proceedings*, (eds.) W. Martin, W. Meijs et al., Euralex: Amsterdam, pp. 426-431.

Čermák, F. 1994b. Idiomatics, in: *The Prague School of Structural and Functional Linguistics*, ed. P. A. Luelsdorff, Amsterdam/Philadelphia: J. Benjamins pp. 185-195.

Čermák, F.-Hronek, J.-Machač, J. (eds.). *Slovník české frazeologie a idiomatiky* (=SČFI) *Přirovnání* (1983), *Výrazy neslovesné* (1988), *Výrazy slovesné A-P, R-Ž* (1994) Praha: Academia (= Dictionary of Czech Phraseology and Idiomatics. Comparisons. Non-Verbal Expressions. Verbal Expressions).

Engler, R. (ed.). *F. de Saussure, Cours de linguistique générale. Édition critique. Tome I: fasc. 1, 2,* (1967), *fasc. 3* (1968). *Tome II: fasc. 4* (1974). Wiesbaden: Harrassowitz.

Filipec, J.-Čermák, F. (1985). *Česká lexikologie*. Praha: Academia (= Czech Lexicology).

Fontenelle, T. (1992). Collocation Acquisition from a Corpus or from a Dictionary: a Comparison, in: *Euralex '92 Proceedings*, pp. 221-228.

Heid, U. (1994). On Ways Words Work Together - Topics in Lexical Combinatorics, *Euralex '94 Proceedings*. Amsterdam, pp. 226-257.

Mathesius, V. (1975). *A Functional Analysis of Present-Day English on a General Linguistic Basis*. Prague: Academia.

Mel'čuk, I. et al. (1995). *Introduction à la lexicologie explicative et combinatoire.* Aupelf-Uref, Duculot Louvain-La-Neuve.

de Saussure, F. (1982). *Cours de linguistique générale*. Paris: Payot.

Sinclair, J. (1991). *Corpus Concordance Collocation.* Oxford: Oxford U.P.