Comparing Collocations Across Languages: An English-Czech Sample

Aleš Klégr[1]

1. Contrastive Comparison of Collocations

The importance of the concept of collocation, a recurrent combination of words, has long been recognized in theoretical linguistics. However, its relevance for the applied sphere, language teaching, lexicography and translation (including machine translation), *etc.*, is just as obvious. The concept has caught the imagination of professionals in applied linguistics and ELT and is currently getting more and more attention. This is documented by textbooks such as McCarthy and O'Dell's *English Collocations in Use* (2005) and others. The interest is understandable inasmuch as collocations are recognized as a source of naturalness in speech, and naturalness is one of the primary goals in language teaching. The inevitable limitation of English textbooks for foreign learners and even of the best learner's dictionaries is that without knowing how collocations in English relate to those of the learner's native language, they cannot alert the learner to the pitfalls of interference and asymmetries. Obviously contrastive analysis of collocations is called for, but research in this direction is still somewhat neglected.

There are several difficulties associated with collocations. For one thing, there is still a lack of consensus on their definition. Partington (1998) divides the definitions of collocations into textual (co-occurrence in a text), statistical (co-occurrence with greater than random probability) and psychological (co-occurrence due to a psychological link between words). Hoey (2005), discussing the merits of each type, opts for a statistical and psychological approach. It is not clear, however, whether a collocation should be identified with a syntagma or not. Collocation as a psychological association argues for a syntagmatic nature, yet from a statistical view this is an unnecessary restriction. In the following we shall take the position that a prototypical collocation is syntagmatic. Another difficulty with collocations is the absence of clear criteria that would delimit the range of a node's collocates and provide a firm basis for their (lexicographic) description. The problem is encountered in monolingual dictionaries, but in collocational dictionaries it becomes crucial. Nuccorini (2003), who analyzes several collocational dictionaries in terms of their features and congruence between the professed methodology and aims and the success of their implementation, concludes that these dictionaries are extremely useful for advanced users but finds that the definition and classification of collocations, together with terminological confusion, constitute a major stumbling-block and inevitably affect the criteria for the selection and treatment of both headwords and their collocates. The following is an attempt to address some of these issues.

[1] Department of English, Charles University, Prague
e-mail: ales.klegr@ff.cuni.cz

2. Comparing Collocations in Bilingual Corpora

Cross-language comparison of collocations drawing on bilingual corpora can make use of two types of corpus, comparable and parallel. Each type offers a somewhat different kind of information on collocations, provides different possibilities and involves different snags. With two independent comparable corpora, neither language is influenced by the other and the set of collocations found for a given node thus reflects a situation specific to either of the languages, whether it is the range or variety of collocations occurring in the environment of the node or their overall number. A parallel corpus, on the other hand, shows which and how many collocations of the SL translate compositionally in the TL, and which of them represent a complex lexical choice requiring a special translation solution. The main focus of this paper is a comparison based on comparable Czech and English corpora and describing two equivalent nodes by means of a simple, but as we hope effective strategy. In the final part, a few remarks on comparing collocations in a parallel corpus are made.

3. Comparing Collocations in Comparable Corpora: The Case of *sadness/smutek*

Experience from compiling a Czech-English noun-verb combinatory dictionary (Klégr *et al.*, 2005) shows that even with one type of syntagma a noun will typically combine with hundreds of verbs, most of which occur only once. This poses the problem of collocate selection and presentation. To give a full range of a word's collocates found in a corpus is too much space-consuming, and is likely to prove bewildering for the dictionary user. To include only habitual collocations such as *heavy smoker*, *rancid butter* or statistically significant collocations (*cf.* Sinclair *et al.*, 2004) would deprive the user of stylistic choices and fail to provide an adequate picture of the word's collocational potential. Similarly, to include collocates of a given node "down to a frequency cut-off point [the top 20 collocates], thereby automatically giving due weight to the most frequent cases" (Stubbs, 2002) has its drawbacks as the most frequent collocates prove to be the most general and least informative ones (*cf.* also Sinclair's upward collocates, 1991).

3.1 The Sample and Methodology

The study makes use of the fact that for each language there is a corpus namely the British National Corpus (2nd ed.) and Syn2000 compiled by the Czech National Corpus Institute. They were made available in the same year (2000) and are comparable in size (100 million words), methodology, and the range of texts. The choice of *sadness/smutek* as the node was motivated by the fact that an abstract noun tends to be monosemic, has simple morphology, and relatively straightforward translation. Still, the corpus data showed one striking difference between them: *sadness* had 751 (relevant) hit lines the BNC, *smutek* 2327 hit lines in Syn2000. The discrepancy is probably due to three factors: (i) polysemy of the Czech *smutek*; (ii) *sadness* having several widely used synonyms as competitiors, while the synonyms of *smutek* are formal and much more restricted in use; (iii) *sadness* being a complex word, unlike the synchronically simplex *smutek* (the difference in valency behaviour between derived and underived words is mentioned by Čermák, 2005). The analysis focused on the V-O type of collocation. Although the noun functions as a complement, the lexicographic experience and research in natural language generation suggest that the noun is indeed the starting point. As Heid (1994) notes, "for noun-verb-collocations (e.g. in the verb-object case: the object noun must be determined first, only then a collocationally adequate verb can be selected)". Despite the efforts to choose nodes with limited collocability and the restriction to the syntactically and semantically relatively clean-cut V-O syntagma (with the node as object and the collocate as finite verb), the number of collocates in the corpora and the task of identifying grammatically relevant collocations still proved daunting.

The first step in the analysis was to prepare a complete list of collocates for each node of a given syntactic type (V-O) from the total of all collocates obtaining in each corpus, taking care that in either language the same lexical unit (sense) is analyzed. The resultant sets of collocates were then subject to semantic analysis. The central premise was that while the number of collocates of the node may amount to hundreds, the semantic range of these collocates is much smaller. Accordingly the verbal collocates of *sadness/smutek* were divided into subsets of broadly substitutable items called synsets (i.e., forming sets composed of synonyms, hyponyms, hyperonyms; negative verbs forms – and their lexical parallels (antonyms) – are subsumed under the positive forms, the presence of an antonym in the subset is marked by (not)). Each synset is considered to represent one type of broadly conceived, but distinct node-collocate relationship. Their sum is expected to distinguish one node from another. This procedure reduced the total sets of verbal collocates of *sadness/smutek* to a relatively small number of synsets, 19 and 30 respectively, describing the collocational preferences of each node. The number and type of the resultant synsets (and their frequencies) then made it possible to compare the similarities and dissimilarities of the collocational sets of *sadness/smutek* both quantitatively and qualitatively. Each collocational synset can be described by two features: the number of distinct lexical items (types) it includes (lexical variety), and the total of occurrences of the verbs in the synset (token frequency). Needless to say, the semantic classification of the collocates is rather tentative. The collocate synsets, marked by square brackets, are usually designated by the most frequent (prototypical) verb within each subset (or its English translation in the case of Czech synsets). Semantically isolated collocates form separate synsets and their own headings.

3.2 The Verbal Collocates of *sadness*

The node *sadness* was found to be a finite-verb object in 205 collocations. It is not without interest that only 23, i.e. 26.4 percent of the verbal collocates have a frequency higher than 2. In other words, almost 3/4 of the verbs co-occurring with *sadness* appear just once. The respective instances (tokens) of the finite-verb collocates were assigned to 87 types (lemmas). The lemmas, in turn, were divided into 19 synsets (Table 1). We take it that these 19 synsets exhaustively summarize all activities performed with *sadness* (as object) in the BNC.

The first six largest synsets among the verbs collocating with *sadness* as object are [feel], [express], [show (not)], [cause], [perceive] and [deal with]. Although they represent only 1/3 of the semantic synsets found with this noun, they account for 76.2 percent of its collocate tokens. In other words, speaking of *sadness* English speakers most frequently say that they feel, express, show/hide, speak of its cause, perceive or deal with it (by avoiding, controlling it, *etc.*). Among the six,

[feel] alone is responsible for 31.2 percent of tokens, which is in fact more than twice the percentage of the second largest synset [express]. Thus [feel], the richest synset in types and tokens alike, includes 13 verbs (*feel* 41x*, have* 6x*, sense* 4x, *etc.*) which account for 64 of the V-O collocations with *sadness*. These verbs express various kinds of feeling. While most of them can be regarded as (partial or near) synonyms (*feel, sense, have, experience*), the last four are somewhat different: *be drawn to, tend towards sadness* describe incipient feeling, *allow os* is a condensation of *allow os to feel*, and *love* describes how we feel about the feeling.

Each of the 19 synsets appears to have one principal verb realizing the synset which accounts for the type/token asymmetry. The poorer the lexical variety, the higher the correspondence between the number of types and tokens. While the token frequency tells us whether *sadness* is often or, conversely, hardly ever thought of in terms of the activity of the given semantic synset, the lexical variety presumably reflects the different possibilities of realization or stylistic variation (*feel/harbour sadness*).

	synset heading	**number of types**	**verbs (types) in the synset**	**number of tokens**	**%**
1.	feel	13	feel (41x) , have (6x), sense (4x) , experience (3x), endure (2x), bear, harbour, realize, suffer, love, be drawn to, tend towards, allow os	64	31.2
2.	express	11	express (14x), speak of (5x), convey, talk about, tell of , write of, portray, capture, sum up, measure, mean	28	13.7
3.	show (not)	9	show (6x), reveal, indicate, display, bring closer; hide (4x), conceal (2x), disguise (2x), deny	19	9.3
4.	cause	8	bring (7x), cause (5x), move to (2x), bring up, evoke, induce, provoke, work up	19	9.3
5.	perceive	4	see (12x), glimpse, hear of, watch	15	7.3
6.	deal with	11	avoid, control, counteract, dispel, heal (sb) of, outweigh, put aside, stop, ward off, fill up, sort out	11	5.4
7.	know	2	know (about, of)(6x), learn about	7	3.4
8.	include	4	include (2x), hold (2x), contain (2x), carry (with it)	7	3.4
9.	share	1	share (7x)	7	3.4
10.	relieve (not)	5	ease (2x), release, release from, relieve, deepen	6	2.9
11.	remember (not)	2	remember (3x), forget	4	1.9
12.	think of	2	think of (2x), reflect on	3	1.5
13.	associate with	3	associate (noise) with, mingle with, mix with	3	1.5
14.	respond to	3	respond to, acknowledge, accept	3	1.5
15.	change to	3	change (from) to s., cool into, give way to	3	1.5
16.	explain	2	explain, account for	2	1.0
17.	relate to	2	be (all) about, be concerned with	2	1.0
18.	attribute	1	attribute (to st)	1	0.5
19.	find	1	find	1	0.5
	Total:	**87**		**205**	**100**

Table 1: Semantic synsets of the verbal collocates of *sadness* (in V-O).

3.3 The Verbal Collocates of *smutek*

The number of the Czech verbal collocates participating in the V-O syntagma with *smutek* is, on the whole, proportionate to the higher incidence of the word *smutek* in the Czech corpus: the search yielded 235 distinct verbs (types) with 453 tokens. Counting verbs in Czech texts is somewhat more complicated than in English due to greater form variation (aspectual variants, negative and reflexive forms, *etc.*). Aspectual variants were generally disregarded where the variants differed only in terms of perfectiveness, repetition, *etc.* (*cítit/pocítit, vyjádřit/vyjadřovat, skrýt/skrývat*). Synthetic negative verb forms were subsumed under positive forms (just as the English (*not*) *remember/forget* were placed in the same synset). Reflexive forms with *se* are marked as such – *zbavit (se)* – but not counted separately. The case and number variability of the Czech node *smutek* (a plural form, for instance, appeared in almost a hundred cases) was disregarded.

As might be expected, the larger number of collocates in Czech produced a larger number of semantic synsets: 30 in all instead of the 19 in English. Perhaps surprisingly, the first six largest synsets are the same ones as for *sadness*, [feel], [deal with], [cause], [express], [show (not)], and [perceive]. However, with the exception of [feel], which is the most frequent in either language, the order of the categories is markedly different. The most conspicuous synset in Czech is [deal with]. Although second in frequency after [feel], the difference is quite small: [feel] 23.8 percent, [deal with] 19.9 percent. It certainly surpasses the [feel] synset in lexical diversity: while [feel] is realized by 31 verbs (13.2 percent), including such special 'feel' words as *vychutnávat*-relish, *libovat si v*-delight in, *etc.*, [deal with] includes 63 distinct verbs (26.8 percent of the total), which is more than twice as many. The most frequent [deal with] verb, *rozmlouvat*-talk out of (11x) has no counterpart in English. By contrast, there is no counterpart in the Czech synset to the English *control* or *avoid sadness*. It is difficult to say how much can be read into finings. It seems that Czech prefers to 'deal with sadness' in a more varied way and that, whatever the manner, 'dealing with' *sadness* is almost as popular, if not more, as 'feeling' *sadness*. The question remains which of the two, token frequency or lexical variety, is to be given more weight.

	synset heading	**no. of types**	**verbs (types) in the synset**	**number of tokens %**
1.	feel	31	cítit (27x), mít (24x including 8x: na/v očích, v sobě, v tváři, na duši), pocítit (11x), prožívat (7x), pociťovat (6x), prožít (3x), necítit (3x), zažít (3x), trpět 7 (2x), dostat, nasávat, nezažívat, nosit v sobě, unést, vléci s sebou, procítit, rozmlouvat se 7, soucítit se 7, utápět se v 6, vydržet, zakoušet, zakusit, zažívat // bát se 2, libovat si ve 6, postrádat, radovat se ze 2, stydět se za, tíhnout ke 3, velebit, vychutnávat (něčí)	108 23.8
2.	deal with	63	rozmlouvat (11x), zahánět (4x), zaplašit (4x), přehlušit (3x), vyrovnat se se 7 (3x), odložit (2x), odnášet (2x), překonat (2x), řešit pl (jídlem) (2x), utápět (v sobě) (2x), zapíjet (2x), zpracovávat (2x), odpustit, odvrátit, prominout, uzamknout, zapít, bojovat proti 3 (prací), čelit, dávkovat, chránit proti 3, nebránit se 3, nedbat 2, nechat doma, nechávat za sebou, nepřemáhat, obelstít, odbýt si, odehnat, odnést, odplavit, pohřbít, pomoci od 2, popřít, probudit ze 2, překonávat, překousnout, přemoci, přepíjet, rozkoukat se ze 2, setřást ze sebe, shodit, sloužit proti 3, spláchnout, stírat, ubránit se 3, utopit 4pl, vydýchávat 4pl, vykoupit, vymanit se ze 2, vymazat, vyplakat 4pl, vyrovnávat se se 7, vysvobodit z 2, vyzrát na, zabránit 3, zahnat, zahodit 4pl, zapudit, zaříkávat , zbavit 2, zbavit se 2, odstranit	90 19.9
3.	cause	21	vyvolávat (10x), přinést (7x), způsobit (5x), vyvolat (4x), přinášet (3x), uvrhnout do 2 (3x), budit (2x), evokovat (2x), (vrána) nosit (2x) , vzbudit (2x), rozprostírat, vdechnout (pohledu), vystavit 3, kouzlit (= vykouzlit), nechat to v kom, probouzet, probudit, vhánět, vnášet, vykouzlit, vzbuzovat	51 11.3

4.	express	17	vyjadřovat (11x), vyjádřit (10x), znamenat (7x), sdělovat (3x), dát průchod 3, hučet si (svůj), uvádět, hovořit o 6, malovat, mluvit o 6, nevypovídat o 6, přiblížit, říci, tématizovat, vykřičet, vyprávět o 6, neznamenat	44 9.8
5.	show (not)	22	neskrývat (6x), skrýt (4x), skrývat (4x), vyzařovat (3x), ukázat (2x), tajit (2x), zakrývat (2x), dát najevo (2x), odhalit (2x), projevovat (2x), dát na sobě znát , nedávat na sobě znát, nejevit, neprojevit, netajit, netajit se 7, nezastírat, ukazovat, vyjevit, vyjevovat, ukrývat, zastírat	41 9.1
6.	perceive	15	vidět (10x), spatřit (na tváři, v očích) (3x), ohlédnouti se po 6 (2x), setkat se se 7 (2x), číst s. (v obličeji, z očí) (2x), uvidět (2x), vycítit (2x), představit si, nevidět, přihlížet 3, uslyšet, všimnout si 2, vyčíst (z pozdravení), zahlédnout, zaznamenávat	31 6.9
7.	know	11	pochopit (3x), poznat (3x), znát (3x), chápat (2x), nechápat, neznat, porozumět 3, reflektovat, rozeznat, rozpoznat, vědět o 6	18 4.0
8.	change to	9	propadnout 3 (3x), střídat (radost) 4 (2x), propadat 3, dojít k 3, neupadnout do 2, propadnout se do 2, upadnout do 2, změnit se v, zvrhnout se do 2	12 2.7
9.	relieve (not)	8	násobit (2x), přidat, neakcentovat, podtrhovat, tišit, umocnit, uvolnit , ztlumit	9 2.0
10.	respond to	4	uklonit se 3 (3x), přijmout (2x), respektovat (2x), reagovat na	8 1.8
11.	pretend	4	nehrát (2x), hrát, předstírat, přehrávat	5 1.1
12.	associate with	4	vměšovat s. (do čeho), mísit se se 7, přidružit se k 3, spojovat (X,Y) s.	4 0.9
13.	remember (not)	2	zapomenout na, zapomínat na	4 0.9
14.	share	1	sdílet	4 0.9
15.	include	4	zahrnovat, nepočítat (mezi co), vložit (do hudby), vzít (s sebou)	4 0.9
16.	disturb	3	poskvrnit, rušit v 6, vyrušit ve 6	3 0.7
17.	compensate	2	vynahradit, vyvažovat (radostí)	2 0.4
18.	relate to	1	týkat se 2	2 0.4
19.	use	2	využít 2, přisoudit 3 (roli)	2 0.4
20.	resemble	1	podobat se 3	1 0.2
21.	caress	1	pohladit	1 0.2
22.	expect	1	očekávat	1 0.2
23.	get used to	1	zvyknout si na	1 0.2
24.	return to	1	vrátit se do 2	1 0.2
25.	participate in	1	podílet se na 6	1 0.2
26.	continue	1	setrvávat ve 6	1 0.2
27.	lose	1	ztrácet	1 0.2
28.	mistake for	1	zaměňovat se 7	1 0.2
29.	need	1	potřebovat	1 0.2
30.	originate from	1	vzniknout ze 2	1 0.2
	Total:	**235**		**453 100**

Table 2: Semantic synsets of the verbal collocates of *smutek* (in V-O).
Note on column 4: numbers in round brackets give frequency higher than 1; unbracketed numbers show case of the object

As with *sadness*, there seem to be few, if any collocations of *smutek* that would display mutual or at least unidirectional expectation on the part of the noun or the verb. A possible candidate is *zapít* or *utopit* ('drown'), though in Czech as in English *žal sorrow(s)* rather than *smutek/sadness* is the first choice. In common with English, *cítit*-feel is the most frequent verb, and obviously the number-one collocate, though it can hardly be called *sadness*-specific. On the whole, there is an even greater proportion of one-off instances than in English. Although they sound perfectly natural, they are highly context-specific. Only few of them are distinctly metaphoric (*pohladit*-caress, *poskvrnit*-tarnish/sully).

3.4 Contrastive Analysis of the *sadnes/smutek* Collocates

Some of the differences have already been mentioned above: the V-O collocations differ in absolute figures, which is due to different representations of *sadness* and *smutek* in the respective corpora. They differ in the number and type of semantic synsets they have. In the total of 34 synsets identified in the verbal collocates of *sadness/smutek*, 15 are common to both *sadness/smutek*, 19 are specific:

synsets shared	synsets only in English	synsets only in Czech
1. **cause**	1. attribute	1. resemble
2. **deal with**	2. explain	2. caress
3. **express**	3. find	3. compensate
4. **feel**	4. think of	4. continue
5. **perceive**		5. disturb
6. **show (not)**		6. expect
7. *change to*		7. get used to
8. *include*		8. lose
9. *know*		9. mistake for
10. *relieve (not)*		10. need
11. *remember (not)*		11. originate from
12. *respond to*		12. participate in
13. *share*		13. pretend
14. associate with		14. return to
15. relate to		15. use

Table 3: Comparison of the distribution of *sadness/smutek* synsets in English and Czech.

In terms of proportional representation the synsets can be conveniently, though somewhat arbitrarily, divided into three zones: the core group with token frequencies of 5 percent and more (in bold type in Table 3), a mid-level group of 5–2 percent (in italics), and a marginal group below 2 percent (normal type). The fact that *sadness/smutek* share fifteen, i.e. almost half, of the synsets distinguished, especially the core and mid-level ones, is a highly significant finding as it suggests that the collocational range of the two nodes is very similar. This is accentuated by another, probably even more important finding that the core group in either language consists of the same six synsets which account for 76.2 percent of collocations in the English sample and 80.8 percent in the Czech one. The six synsets can be seen very much as defining the collocational preferences of *sadness/smutek* as regards their transitive verb collocates.

The 21 marginal synsets with frequencies below 2 percent, on the other hand, signal the potential, rather than normal collocations of *sadness/smutek*. Two things are worth noticing. Firstly, only two of them are shared, [associate with] and [relate to]. The lack of overlap between the English and Czech marginal synsets may partly be a function of their low incidence; partly it may signify different frames of thought in either language. The English speakers apparently do not think of *sadness* in the same terms or contexts as the Czech speakers, although it would be easy to match the Czech verbs with English ones and vice versa. The other remarkable thing is the semantic diversity of the Czech marginal synsets. In fact, they are as numerous as the synsets shared by both nodes. This could be attributed to the higher incidence of *smutek* in the Czech corpus, but it may also be the case that the positions of *sadness* and *smutek* in their respective lexical fields are simply different. We may hypothesize that *sadness* is less important compared to *smutek*, being frequently replaced by synonyms such as *unhappiness, sorrow, grief, etc.* As a consequence, it appears in fewer contexts than its Czech counterpart.

The core groups of the synsets (see Table 4) deserve closer attention. The fact that the first six are the same for both languages suggests similar collocational preferences of *sadness/smutek* and, by the same token, similar communicative preferences in the two speech communities. Evidently, for speakers of English and Czech conveying that they ‘feel sadness’ has top priority (31.2 percent and 23.8 percent respectively). Conversely, the finding that these synsets are differently distributed and show different lexical varieties suggests subtle dissimilarities between the communicative preferences in either speech community. While in English the second most frequent synset [express] has less than half the percentage of [feel] and the last of the six, [deal with], is actually very near the cut-off point of 5 percent, in Czech [deal with] comes second with a proportion quite close to [feel] and [express] is sandwiched between [cause] and [show (not)] in the middle. The position of [cause] and [show (not)] is switched in English, though again roughly in the middle. The representation of [perceive] is about the same in both languages. We may conclude that, quite interestingly, speakers of English primarily seem to feel and express sadness, whereas speakers of Czech feel and deal with it.

sadness collocates				
semantic synset	types	%	tokens	%
feel	13	14.9	64	31.2
express	11	12.6	28	13.7
show (not)	9	10.3	19	9.3
cause	8	9.2	19	9.3
perceive	4	4.6	15	7.3
deal with	11	12.6	11	5.4

smutek collocates				
semantic synset	types	%	tokens	%
feel	31	13.2	108	23.8
deal with	63	26.8	90	19.9
cause	21	8.9	51	11.3
express	17	7.2	44	9.8
show (not)	22	9.4	41	9.1
perceive	15	6.4	31	6.9

Table 4: Comparison of the six core synsets of *sadness/smutek*.

When lexical diversity is taken into account, the order of synsets in English remains the same, except for [deal with] which draws level with [express]. In Czech there are two marked changes compared to token frequency: [deal with] is by far the most varied lexically, in fact twice as much as [feel], and [show (not)] comes third before [cause]. In sum, lexical variety somewhat raises the role of [deal with] in English but generally is in agreement with token frequency, whereas in Czech it underscores the importance of [deal with] quite dramatically compared to [feel] and accentuates the role of [show (not)]. The overall picture of *sadness/smutek* collocations thus speaks of similar speech habits but distinct specific tendencies and preferences.

The strategy of describing collocations through their synsets appears capable of handling the collocability of a node in an economical way by reducing the hundreds of collocates to a few groups. Fears that the large number of one-off collocates might result in too many synsets proved unnecessary. The method works equally well in one language or in two, making possible a meaningful contrastive analysis of equivalent nodes. By assigning their numerous collocates to synsets it manages to arrive at a coherent picture of the collocational preferences of the contrasted nodes, highlighting their similarities and dissimilarities. To some extent, it may also be predictive of potential collocates (synonyms, co/hyponyms, *etc.*, of the actual collocates).

4. Comparing collocations in a parallel corpus

Collocations in a parallel corpus involve SL collocations and their translation equivalents in the TL which may, but need not be structurally parallel (*cf.* McDonald *et al.*, 2004). While SL collocations reflect authentic collocational patterns of the source language, their TL equivalents depend on the degree of isomorphism between the two languages. In principle, the translation will be a continuum running from a node-collocate match (direct or compositional translation) to the complete absence of an equivalent (non-match or lexical gap). In between are various forms of indirect translation whereby equivalence is achieved at the expense of formal correspondence. The relevant techniques of indirect translation include transposition, modulation and adaptation or their combination (*cf.* Vinay-Darbelent, 1958/1995). Given the complexity of translation and the differences between languages, there is naturally a plenty of scope for interference or negative transfer resulting in instances of deviation from the collocational preferences of the target language. By analogy with a phenomenon well-known in language teaching as faux amis we may call these instances ‘collocational faux amis’.

Work on Czech-English combinatory dictionaries (Klégr *et al.*, 1991, 1994, 2005) showed three principal types of collocational faux amis (CFA) depending on which of the three main techniques of indirect translation is required instead of compositional translation to render the SL collocation naturally. Transpositional CFA require a grammatical shift between the SL collocation and its TL counterpart (without change in denotative meaning), such as word-class shift (*třít bídu* “be poor/destitute”: n>adj and lexical verb>copula), unit shift (phrase for word, clause for word, *etc.*: *srdce vynechává* “heart beats irregularly”: V>V-Adv); structure shift (change in word order, syntactic function, *etc. – inspect a school* “provest inspekci ve škole”: v > nv; n> PrP; O > Adv), *etc.* Modulative CFA necessitate lexical variation: one or more components of the SL collocation have to be translated by a word which is not their usual, dictionary equivalent (such as a synonym, (co-)hyponym or meronym), *cf. protancovat boty* (= shoes) “dance the soles off”; *vraštit čelo* (= forehead) “knit (one’s) brows, frown”. Adaptational CFA are instances where – in the absence of “structural and conceptual parallels between SL and TL” and “when the situation referred to in ST does not exist in the target culture, or does not have the same relevance or connotations” (Shuttleworth-Cowie, 1997) – the translation of a collocation is possible only by adaptation. They are not easy to find and, moreover, the distinction between modulation and adaptation is often very difficult to draw: *reconstitute dried milk* “rozmíchat/rozpustit (sušené) mléko ve vodě”. It was found, however, that contrary to claims that collocations are “completely lexically determined and thus need to be memorized” (see Heid, 1994, referring to Mel’chuk, Polguere, 1987), at least collocations of the vb-n or adj-n type seem to be largely semantically determined and have straightforward correlates, while unpredictable collocations form a distinct minority.

5. Conclusions

Believing that contrastive study of collocations is a long-overdue task, we have tried to suggest how comparable and parallel corpora can be made use of. The proposed synset procedure, applicable to parallel corpora, gives a comprehensive, yet succinct account of the collocation patterns of a node and allows comparison with an equivalent node in another language. The procedure could presumably be simplified by using Kilgariff's Word Sketch. It could do away with the laborious gathering of collocates (and reduce their number), parsing and assignment of syntactic functions and make it immediately possible to start with the key stages: division of collocates into synsets and the actual comparison of the situation in each language. Analysis of collocations in a parallel corpus introduces another aspect, degrees of correspondence between collocations across languages, and raises the issue of mismatches, or collocational faux amis.

References

Čermák, F. (2005) Abstract noun collocations: their nature in a parallel English-Czech corpus, in G. Barnbrook, P. Danielsson and M. Mahlberg (eds), Meaningful Texts, pp. 143–51. London-New York: Continuum.

Heid U. (1994) On Ways Words Work Together – Topics in Lexical Combinatorics, in Euralex 1994 Proceedings, pp. 226–57. Amsterdam: Vrije Universiteit.

Hoey, M. (2005) Lexical priming. A new theory of words and language. London / New York: Routledge.

Kilgarriff A., Sketch Engine. Available on-line from http://www.sketchengine.co.uk/ (last modified: Adam Kilgarriff, 25 May 2007).

Klégr A., N. Hronková, Z. Hron (1991) Znáte anglická slovesa? Česko-anglický slovník nejužívanějších spojení podstatných jmen se slovesy [Do you know your English verbs? A Czech-English dictionary of noun-verb combinations]. Praha: SPN.

Klégr A., N. Hronková (1994) Znáte anglická přídavná jména? Česko-anglický slovník spojení podstatných jmen s přídavnými jmény [Do you know your English adjectives? A Czech-English dictionary of noun-adjective combinations]. Praha: Leda.

Klégr A., P. Key, N. Hronková (2005) Česko-anglický slovník spojení: podstatné jméno a sloveso / Czech-English combinatory dictionary: noun and verb. Praha: Karolinum.

Klégr A., P. Šaldová (2006) Kolokační faux amis, in Čermák, F., M. Šulc (eds) Kolokace, pp. 168–77. Praha: NLN.

McCarthy, M., F. O'Dell (2005) English Collocations in Use. Cambridge: Cambridge University Press.

McDonald, S., D. Turcato, P. McFetridge, F. Popowich, J. Toole (2004) Collocation Discovery for Optimal Bilingual Lexicon Development. Berlin / Heidelberg: Springer.

Nuccorini, S. (2003) Towards an 'ideal' Dictionary of English Collocations, in P. van Sterkenburg (ed.) A Practical Guide to Lexicography. Amsterdam / Philadephia: John Benjamins.

Partington, A. (1998), Patterns and Meanings: Using Corpora for English Language Research and Teaching. Amsterdam / Philadelphia: John Benjamins.

Shuttleworth M., M. Cowie (1997) Dictionary of Translation Studies. Manchester: St. Jerome.

Sinclair, J. (1991) Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sinclair, J., S. Jones, R. Daley (2004) English Collocation Studies: The OSTI Report. London / New York: Continuum.

Stubbs, M. (2002) Words and Phrases. Corpus Studies of Lexical Semantics. Oxford UK / Cambridge USA: Blackwell Publishing.

Vinay J.-P., J. Darbelnet (1958/1995) Comparative Stylistics of French and English: A Methodology for Translation (transl. J. C. Sager, M.-J. Hamel). Amsterdam / Philadephia: John Benjamins.