English equivalents of the Czech preposition *v/ve* from the point of view of the 'open-choice principle' and the 'idiom principle'

Markéta Malá, Pavlína Šaldová, Aleš Klégr
Faculty of Arts, Charles University in Prague

Abstract

The paper is part of an on-going study of English equivalents of the 10 most frequent Czech prepositions. In this case it is limited to the most frequent of them, *v/ve* [in]. The general aim is to determine whether and to what extent the choice of translation equivalent is influenced by the fact that the sequence including *v/ve* is an open grammatical structure or a prefabricated lexical string. The first task is to identify which of the sequences including v/ve are grammatical and which are lexical, to find their representation and compare them with their translation equivalents. The assumption was that SL prefabricated strings tend to be translated by means other than a grammatical sequence (i.e. a prepositional equivalent) and that accordingly Czech lexical sequences (prefabricated, formulaic, etc.) with *v/ve* will have a higher proportion of non-prepositional equivalents than found in the whole sample. The results show, however, that the concept of prefabricated lexical strings and consequently their identification by purely statistical methods is not without problems. Still, the tests of correlation (using T-score) between presumably grammatical sequences and lexical sequences, and their corresponding translations, give some support to the hypothesis.

1. Introduction

The present paper is closely related to a previous study dealing with English equivalents of Czech prepositions[1]* in that it uses its sample and results. Its focus, however, is on a different aspect of cross-linguistic preposition equivalence and the scope was limited to only one preposition, *v/ve*. Still the reasons motivating this paper are very much the same as in the previous study. The subject of prepositions has been neglected for a long time, and that of English-Czech preposition equivalence even more so. The only attempts to correlate English and Czech prepositions are found in bilingual dictionaries, all of which pre-date the corpus

[1] Klégr A., M. Malá,2009, English Equivalents of the Most Frequent Czech Prepositions. A Contrastive Corpus-based Study. 5th Corpus Linguistics Conference, 2009, 20-23 July 2009, University of Liverpool, UK, http://www.liv.ac.uk/english/CL2009

era. The on-going compilation of the parallel Czech-English corpus under the *InterCorp* project organized by the Institute of the Czech National Corpus thus offers a unique opportunity to improve this state of affairs.

Apart from being easy to search for in a corpus even in an inflected language such as Czech, primary prepositions are also free of homography and their limited number offers the possibility, at least in theory, of their exhaustive survey. But amenability to search is certainly not the main reason for this study. What is far more important is the frequency of prepositions in both languages and the difficulty they present when it comes to their translation from one language to another.

It may not be surprising that the frequency list for English, based on *The British National Corpus*, gives 8 prepositions (*of, in, to, for, with, on, by, at*) among the 25 most frequent words. Dizier (2006, 3), drawing on the *WFWSE* web site, reports 9 prepositions (adding *from*) among the 30 most frequent words in English and points out that of is the second most frequently used word in English. What is surprising, though, is that the frequency dictionary for Czech, *Frekvenční slovník češtiny* (2004), based on a Czech corpus comparable to the BNC in size and composition, lists even more primary prepositions, i.e. 10, among the 25 most frequent words: *v/ve* (second most frequent word), *na* (5th), *s/se* (7th), *z/ze* (8th), *o/vo* (11th), *do* (13th), *k/ke/ku* (15th), *za* (17th), *pro* (19th), *po* (25th)!

Actually, it is not just the higher incidence of prepositions among the 25 most frequent words in a synthetic language compared to analytic English that is intriguing, but also the fact that most of the Czech and English prepositions are considered to be translation equivalents in standard Czech and English dictionaries. It just may be that regardless of the typological differences in both languages prepositions serve the same purpose of making explicit and extending the range of semantic and grammatical relations between clause elements in keeping with the growing demands on the stylistic diversification of the written language. The greater number of prepositions at the top of the ranking list in the highly-inflected Czech could also be influenced by the fact that Czech prepositions appear to be slightly less polysemous than the English prepositions and that among the top-scoring English words there appear more function words.

In addition to, and probably due to, frequency, prepositions are a potent source of errors. In fact, in her analysis of close to 1,200 errors appearing in translations into English by Czech learners, Klimšová (1999) found that prepositions were the third most frequent cause of errors (14 per cent), exceeded only by errors in the use of articles (24 per cent) and lexical errors (15 per cent). In other words, these three types of error accounted for more than 50 per cent of Czech speakers' errors in English. While errors in the use of articles are explained by the absence of in/definiteness as a grammatical category in Czech, errors in prepositions, if explained at all, are attributed to interference, or negative transfer, without specifying the mechanism whereby the negative transfer operates.

Inasmuch as in the previous study we found a remarkable distribution asymmetry between Czech and English prepositions and a relatively large proportion of non-prepositional equivalents, it seemed logical to ask about the cause of these phenomena. In addition to the extensive polysemy of the prepositions in either language and their semantic mismatch in varying degrees (in spite of similarities), there is another plausible explanation which the present study sets out to test. The fact that prepositions can be part of not only ad hoc constructions, but also of complex lexical structures (MWUs, idioms, etc.), may significantly influence the way they will be translated into another language.

Accordingly, the present study focuses on two aspects: first, it will attempt to estimate the proportion of cases in the sample where the preposition is a component of a compositional free combination (e.g. an adverbial prepositional phrase) resulting from the "open-choice principle" (Sinclair 1991), and cases where the use of a preposition is determined in advance, i.e. the result of co-selection (the "idiom principle") due to the fact that it is part of a more or less fixed lexical string (whether idiomatic or merely collocationally/statistically prominent and compositional); second and more importantly, it will try to determine the correlation between the former and the latter cases and the type of equivalent used in their translation, i.e. the role of the "open-choice principle" and the "idiom principle" in the use of Czech prepositions and how this factor is reflected in the distribution of their English equivalents.

2. The results of the previous study

As the first stage in the investigation of the ten most frequent Czech prepositions and their English equivalents (Klégr, Malá 2009), the first and the last of these ten, the prepositions *v/ve* and *po*, were singled out and subjected to (a) *formal equivalent analysis* to find out how they are actually translated into English, i.e. to ascertain the types, frequency and diversity of their English equivalents; (b) *syntactic analysis* examining the syntactic function of the Czech prepositional phrase (PP) headed by the preposition and the correlation with the type of equivalent; and finally (c) *semantic analysis* focusing exclusively on the adverbial uses of the Czech PPs (assumed to include mostly compositional sequences in which the meanings of the preposition are best identified). However, only the formal equivalent analysis is crucial for this study and will be reviewed in some detail.

Formal equivalents of Czech prepositions, which are the focus of the research, were classified in the following way according to the manner of translation strategy:

(1) prepositional equivalents – the Czech preposition is directly translated by an English preposition, regardless of whether the function and meaning of the Czech preposition (-al phrase) matches exactly the function and meaning of the English preposition (-al phrase), e.g. *Myslí na ně s příchutí sody, když mění pohovku **v** postel.* (F) - *He thought of them with the taste of soda in his mouth, as he turned his sofa **into** a bed;*

(2) non-prepositional equivalents – these involve instances of indirect, implicit translation when the whole SL sentence is translated by an English sentence, its

purport is preserved, but the Czech preposition (and its meaning) contained in it is difficult or impossible to identify with any one form in the TL sentence. They were divided into two subtypes:

(a) lexical-structural equivalents, consisting of a lexical and/or structural transposition that compensates for the SL preposition; while the preposition becomes redundant in the TL text, the prepositional complement is preserved either as a free (noun, verb, etc.) or a bound morpheme; cases where the preposition was translated by a conjunction (homomorphous with a preposition) were assigned to this category as well, e.g. *Obrátila své hnědé oči **v sloup*** ... (V) – *She rolled her brown eyes **upwards** ...*;

(b) zero: the SL and TL sentences (and their meaning) correspond reasonably well, but neither the preposition nor its complement can be identified in TL. Textual equivalence is not impaired, only the message is possibly less detailed (or some kind of modulation appears), e.g. *Tomáš v sobě necítil žádný soucit.* (K) - *Tomas felt no compassion.*

Instances of textual non-correspondence (the Czech sentence has no correlate in the English text) were excluded from the samples.

Equivalent analysis of the 600 occurrences of the Czech *v* confirmed the expectation of a wide range of equivalents (see Table 1), more than two thirds of which were prepositional equivalents (409 occurrences; 68.2 per cent) and almost one third non-prepositional ones (191; 31.8 per cent). The total ratio of prepositional and non-prepositional equivalents is thus 2 : 1. A closer look at the equivalents shows that most of the non-prepositional equivalents are lexical-structural transpositions (82.7 per cent), which form 26.3 per cent of all types of equivalent. Prepositional equivalents (Table 2), somewhat surprisingly, include as many as 21 different English prepositions, of which only 7 occur at least four times. However, there is one preposition which accounts for almost 73 per cent of all English prepositions and almost a half of all English equivalents (49.7 per cent) – the preposition *in*. It is a dominant equivalent both among prepositional equivalents and among the whole group of English equivalents.

equivalent	**subtype**	**number**	**%**	**total**	**%**
prepositional	*in*	298	49.7	409	68.2
	others	111	18.5		
non-prepositional	lex.-struct.	158	26.3	191	31.8
	zero	33	5.5		
total		600	100.0	600	100.0

Table 1: A summary table of *v* equivalents

No	preposition	occurrence	%
1	in	298	72.9
2	at	34	8.3
3	on	27	6.6
4	into	13	3.2
5	about	5	1.2
6	to	4	1.0
7	with	4	1.0
8	during	3	
9	through	3	
10	under	3	
11	by	2	
12	from	2	
13	inside	2	
14	within	2	
15	among	1	
16	behind	1	
17	for	1	
18	in and out of	1	
19	in at	1	
20	out	1	
21	out of	1	
total		409	100.0

Table 2: The list of all English prepositional equivalents of the Czech *v*

Syntactic analysis of *v*-headed Czech PPs showed that 517 of them (86.2 per cent) were adverbials-modifiers (the two functions were not distinguished). Their equivalents were prepositional in 72.3 per cent of cases, while the ratio of prepositional and non-prepositional equivalents of the Czech PPs in the remaining functions was almost equal. (*Semantic analysis* showed that of the 454 Czech adverbial PPs 62.6 per cent were broadly spatial, 20.9 per cent temporal and 12.5 per cent adverbials of manner.) The results of syntactic analysis seem to confirm the assumed correlation between an adverbial function of the Czech PPs (tending to be realized by free combinations) and the high proportion of their English prepositional equivalents. However, the relation between a free combination PP in the SL and its translation by a prepositional equivalent is only indirectly implied. It is the task of this study to make this relationship more explicit.

3. Definition and identification of a collocational (formulaic, fixed) lexical sequence

In order to examine the influence that the nature of the structure with the Czech preposition *v/ve* may have on its translation into English, it is necessary to specify prefabricated "lexical strings" (in contrast to grammatical strings, ad hoc free combinations) and to decide on how they will be identified in the sample.

Needless to say, both tasks present formidable theoretical and practical problems which are beyond the scope of this study. An interesting review of the possibilities of detecting formulaic sequences[2], as she calls, them is found in Wray (2002, 19-43). There are two basic ways, she says, in which these sequences can be collected – first, to use an empirical method (an experiment, questionnaire) "to target the production of formulaic sequences ... as data" and, second, to search through linguistic material (in some more or less principled way) for potential strings according to some criteria. Focusing on the latter approach, she lists the following possible procedures:

native-speaker intuition as a basis for identification – Wray quotes Foster (2001), who used a panel of seven native speakers "to mark any language which they felt had not been constructed word by word", and goes on to enumerate the weaknesses (labour-intensive, inherently inconsistent, etc.);

shared knowledge, "the extent to which a word string, started by one person, can be reliably completed by others", used as a measure of formulaicity;

frequency counts using computer searches "which reveal which other words a given target word most often occurs with"; although relying on objective data, this approach is fraught with a number of problems of various kinds. If anything, it requires arbitrary decisions on frequency thresholds, i.e. how frequent an association has to be in order to count, decisions on relating frequency counts, i.e. ratio measures (what to measure and how) and many others depending on the circumstances. Wray concludes that "the frequency-based analyses conducted in corpus linguistics do not fully meet our needs when it comes to identifying formulaic sequences" and cautions that "just as there is evidence that a string generally agreed to be formulaic may or may not have a high frequency in even the largest of corpora, so it is also not possible to assert that all frequent strings are prefabricated".

She also reviews certain features associated with formulaic sequences that may be useful in identifying them as criteria:

structure as a form-based criterion, e.g., when based on the observation that "the first-occurring *invariable* word in a repeated sequence tends to be a function word or discourse marker";

[2] Her own working definition (Wray 2002, 9) of the formulaic sequence is: a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

compositionality (semantic transparency), or rather lack of it combined with grammatical irregularity; Wray regards this procedure as too conservative "because it excludes the formulaic sequences that are entirely regular in form and transparent in meaning";

fixedness disallowing, for instance, insertions of elements into formulaic sequences; the applicability of fixedness as a test is limited and only a small subset of formulaic sequences are entirely fixed;

phonological form (coherence), detection of formulaic sequences through phonological cues is restricted to the spoken language and again has limited applicability;

fluency as a criterion expects sequences retrieved whole from memory to be produced more fluently;

stress and articulation as an indication that a sequence is "felt and handled as a unit"; as with other phonological criteria, they are difficult to apply without having an independent way of determining the boundaries of formulaic sequences.

Wray lists three other specific procedures (liaison in French, identification criteria in specific data, such as children's language, and code-switching as a boundary indicator in formulaic sequences), but as might be expected she finds that "formulaicity seems to manifest too great a diversity of potential forms to submit to predictability beyond the most general and mundane level". Further on she points out four themes reappearing in various definitions of formulaic sequences - form, function, meaning and provenance - and subsequently proposes her own theoretical model (formulaicity as a dynamic solution involving three dimensions - processing, interaction and discourse marking - and responding to a unique situation and speaker). Although Wray's review does not offer any immediate solution to the problem at hand, it does point us in the most promising direction, i.e. the statistical approach (combined with linguistic knowledge) despite its limitations, such as too much noise due to functional words, low-frequency data resulting in unreliable scores and the need for manual evaluation.

4. Analyzing *v/ve* sequences: description of material and hypothesis

As implied above, the present study uses the same material and sample as the previous one. The sample was collected from three pairs of original fiction texts and their English translations (see *References*) and consisted of the first 200 occurrences of the preposition from each text and their translation counterparts (i.e. 600 pairs of parallel concordance lines).

The aim of the study being to determine whether and how much the fact that the sequence including *v/ve* is a grammatical or a prefabricated lexical string influences the way it is translated, it is necessary to identify the two kinds of sequence and compare them with their translation equivalents. The assumption is that SL prefabricated strings may be more prone to translation other than by grammatical sequence. Accordingly, the hypothesis is that in Czech lexical sequences (prefabricated, formulaic, collocational, etc.) with *v/ve* the proportion of prepositional and non-prepositional equivalents will be other than the 2 : 1 ratio manifest in the whole sample, i.e. the proportion of non-prepositional equivalents is expected to be somewhat higher, while in Czech grammatical sequences (free combinations) with *v/ve* the opposite tendency is assumed.

4.1. Identification of prefabricated sequences

It is clear that the issue of prefabricated sequences is bristling with unresolved theoretical questions such as whether grammatical and lexical sequences form a cline or whether there is a binary division (which, by some, Sinclair's two principles imply) or what the proportion of prefabricated sequences in the texts is like. Moreover, Wray believes that the prefabricated part of the lexis is not fixed, but unique to a situation (something which studies by Biber et al. 2004, Hyland 2008 and others suggest) and even to an individual. The crucial decision in our case was to choose between the two basic approaches to identification. For various reasons, objective statistical measures were preferred to subjective procedures (intuition, shared knowledge). However, the task at hand is specific in that statistical methods are not required to extract prefabricated sequences, but to find out whether a given set of sequences with *v/ve*, in the sample are prefabricated sequences or not. The preposition signals where to look for a sequence but not where the sequence begins or ends relative to the preposition.

Hence the first step was to determine the scope of the *v/ve* sequences identified in the Czech texts of our parallel corpus. The situation in Czech as an inflected language with relatively great mutual positional variability of clause elements (free word order) which allows for extensive discontinuity in some types of sequences does not simplify the task. After preliminary tests it was decided to use two types of queries, depending on whether the construction is headed by the preposition (a prepositional phrase - Type B, where the word-form of the prepositional complement is governed by the preposition and the preposition and the complement are sequentially ordered) or whether the prepositional phrase itself complements or modifies a superordinate clause element (Type A with sequential and formal variability):

Type A:preposition (PP) complements a superordinate element,
e.g. [lemma="pokračovat"][lemma="v"] within the search scope of [-3, 3]
(*pokračoval v prohlídce, v prohlídce budeme pokračovat*)
Type B: preposition (PP) is not a complement of a superordinate element,
e.g. [lemma="v"][word="týdnu"] within the search scope of [0, 3]
(*v týdnu, v každém týdnu*)

The scope of co-occurrence was set to three tokens to allow for variation in word order (Type A) and modification. Where the prepositional construction was found to be indeterminate between Type A and Type B both types of queries were performed hoping that either type of query would have a desirable outcome.

Some of the Type B sequences were excluded as they did not qualify for the status of prefabricated expressions, e.g. sequences in which the complement of the preposition is a pronoun or proper noun.

As the texts of the three novels from which the samples were gathered are relatively short, we decided to measure the collocation strength of the sequences using not the texts of the novels but the 100-million-word *SYN2005* corpus in an attempt to avoid problems with low-frequency data reported in the literature. The next step was to choose a statistical measure of the strength of a collocation between the two words of the sequences. Two of the most common measures, MI-score and T-score, were opted for.

When the results produced by measuring MI-score and T-score were compared and manually evaluated, the results of T-score proved on the whole to be more meaningful than the MI-score[3] ones in that the sequences with the highest T-score values included sequences that can be classified as multiword prepositions, phrasal verbs, intensifiers, etc. On the other hand, some low-frequency sequences traditionally regarded as idioms (*obrátit oči v sloup*) score very low, which indicates that even T-score is not fully reliable in such cases. It has to be admitted though that it is not always clear what the available statistical measures actually do and show.

In short, our experience with these statistical measures is very much in keeping with what McEnery et al. (2006, 57) have to say about them: "While the MI test measures the strength of collocations, the *t* test measures the confidence with which we can claim that there is some association (Church and Hanks 1990). Collocations with high MI scores tend to include low-frequency words whereas those with high *t*-scores tend to show high-frequency pairs. As such, Church, Hanks and Moon (1994) suggest intersecting the two measures and look at pairs that have high scores in both measures."

Certainly T-score (and MI-score even less so) can hardly be expected to provide guidelines as to which of the sequences marked as statistically significant are idioms (non-compositional, fixed, lexically and/or grammatically irregular), compositional multiword units (such as phrasal verbs) or simply collocations (two-word or extended, i.e. lexical bundles). Moreover, contrary to Church, Hanks and Moon's suggestion, sometimes the position of some sequences in the list and the values of both types of score are somewhat puzzling and look suspiciously like statistical flukes. Likewise the cut-off point between such presumably prefabricated sequences and grammatical sequences (free combinations) is obviously arbitrary.

4.2 Selection of data for correlation with equivalents

In view of the difficulties and uncertainties with identifying prefabricated sequences in the sample, a strategy was devised to meet these problems. In order to obtain results as representative as possible, the following two procedures of

[3] MI-scores appear to be affected negatively by the high frequency of the preposition v/ve (f (v) = 2348446).

data selection were used: (a) *top and bottom selection* – based on T-score values, the 50 highest scoring and 50 lowest scoring sequences were chosen for the correlation in the hope that the influence of "prefabrication", if any, would thus be maximally highlighted; (b) *reverse selection* – starting from the English correspondences of the preposition *v/ve*, 50, or rather 48, non-prepositional equivalents (other than in the previous two lists) were randomly chosen on the assumption that due to the prefabrication influence their corresponding Czech sequences would include a higher proportion of top-scoring items.

Accordingly, the following two lists of sequences based on T-score results were compiled (fourteen of the T-score top sequences also appear in the top-100 MI-score list):

	Sequence	T-score	MI-score
1	být v	513.970	2.407
2	v roce	247.719	5.627
3	v době	196.171	5.595
4	v případě	188.372	5.698
5	v letech	164.163	5.267
6	v rámci	125.786	5.709
7	v oblasti	124.707	5.037
8	v chvíli	118.351	4.384
9	místo v	115.400	3.067
10	v noci	114.008	5.111
11	žít v	113.923	4.175
12	v životě	102.024	5.321
13	v městě	97.994	5.502
14	v případech	94.365	5.644
15	v smyslu	93.935	5.319
16	v skutečnosti	93.069	5.069
17	v domě	90.741	5.550
18	vidět (se) v	89.840	2.650
19	v sobotu	88.868	5.498
20	v světě	88.642	4.242
21	v podobě	86.924	5.525
22	pokračovat v	86.196	4.114
23	zůstat v	85.935	3.407
24	v ulici	85.493	4.690
25	v dnech	85.459	5.346
26	v ruce	84.715	4.052
27	v pořádku	84.308	5.285
28	v řadě	82.177	5.333
29	objevit (se) v	82.134	3.653
30	v okamžiku	81.317	5.338
31	v duchu	81.298	5.578
32	v neděli	80.910	5.365
33	změna v	80.421	3.424
34	sedět v	80.400	3.712
35	v části	79.635	4.001
36	v výši	78.185	5.298
37	spočívat v	76.929	5.360
38	tvář v	74.308	3.391
39	v pátek	73.060	5.312
40	v pondělí	72.924	5.216
41	ležet v	71.185	3.670
42	v pokoji	69.986	5.214
43	v okolí	69.952	4.447
44	v století	69.880	3.383
45	v očích	69.617	5.409
46	v dvou	68.963	3.373
47	bydlet v	68.076	4.454
48	držet v	68.034	3.631
49	v prostředí	66.214	3.472
50	v škole	65.813	4.825

Table 3: The first 50 Czech sequences at the top of the T-score list

	Sequence	T-score	MI-score
1	v ředitelně	6.784	5.589
2	v poklusu	6.671	5.214
3	v vzrušení	6.663	1.555
4	v kabinetě	6.580	5.704
5	v křesílku	6.474	5.382
6	nevíra v	6.260	3.124
7	v podprsence	6.161	5.271
8	v polovici	6.139	4.602
9	v předsíňce	6.125	5.704
10	zasutý v	6.037	3.867
11	v čepci	5.950	4.843
12	poskočit v	5.929	2.318
13	v policích	5.651	3.586
14	v keřích	5.608	4.263
15	odplout v	5.499	2.278
16	v úschovně	5.190	5.704
17	bělat (se) v	5.167	3.314
18	v baráčku	5.136	4.433
19	v záhonu	4.903	2.652
20	v koncentrácích	4.782	5.382
21	v internátu	4.631	3.201
22	v úprku	4.620	3.718
23	v variaci	4.562	3.863
24	v ozvěnách	4.486	5.573
25	v intonaci	4.339	3.737
26	v halence	4.288	4.600
27	v ponožce	4.271	5.630
28	v návratu	4.224	0.618
29	úctu v	4.095	1.269
30	snídaně ve	4.095	1.748
31	v pěstích	4.044	5.704
32	v družnosti	3.664	3.575
33	lechtat v	3.624	2.568
34	v pankreatu	3.503	2.206
35	v argotu	2.859	4.414
36	v planetáriu	2.767	5.534
37	v gymnasiu	2.605	3.660
38	v kožeňácích	2.595	5.704
39	v korekturách	2.479	3.988
40	hroužit (se) v	2.259	2.312
41	v kastlíku	2.133	4.441
42	vzdout (se) v	1.833	2.471
43	v kvízu	1.799	3.312
44	v jídelničce	1.699	5.704
45	v cítění	1.463	0.499
46	v dóziččce	1.387	5.704
47	v ataku	1.101	1.456
48	sklenout v	1.034	1.897
49	v ukazovátku	0.981	5.704
50	v matce	0.733	0.141

Table 4: The last 50 Czech sequences at the bottom.of the T-score list

The reverse selection consists of 48 non-prepositional equivalents, i.e. lexical-grammatical transpositions (zero equivalents were omitted as they are mostly higher-level translational modulations), chosen from all three texts in roughly the same proportion. Only those sequences which are neither in the top nor in the bottom T-score list were included. Considering that the total of lexical-grammatical transpositions is 158, the non-prepositional equivalent list consists of all the available remaining items not included in the T-score lists (F, K, V indicates the respective authors of the texts):

cítit v (F), *číst v* (K), *růst v* (V), *slít se (v kus)* (F), *v budoucnu* (K), *v bytě* (K), *v cizině* (K), *v dopisu* (V), *v hlase* (V), *v hloubi* (K), *v jícnu* (F), *v klidu* (V), *v kombinaci* (V), *v kompetenci* (V), *v kufru* (K), *v míru* (F), *v moci* (K), *v náladě* (K), *v ničem* (K), *v normě* (V), *v nouzi* (F), *v obálce* (V), *v olově* (F), *v podání* (V), *v pohledu* (V), *v poschodí* (F), *v postavení* (K), *v prosinci* (V), *v předklonu* (V), *v předsíni* (K), *v předtuše* (F), *v rozporu* (K), *v rychlosti* (K), *v spěchu* (V), *v spojení* (K), *v střehu* (V), *v týdnu* (V), *v umění* (V), *v úprku* (F), *v úvahu* (F), *v válce* (F), *v vztazích* (K), *v zázemí* (F), *v žaludku* (K), *v čtvrtek* (V), *vězet v* (F), *viset v* (F), *zahrnovat v* (K)

The range of the possible T-score values of these sequences is between those of the T-score lists, i.e. between 65.813 and 6.784, and the assumed "prefabrication influence" will be signalled by values in the upper zone of this range.

5. Correlation between selected sequences and their equivalents/T-score values

The sequences which appear in the top T-score list typically crop up several times in the parallel texts, and each time the equivalent may be different. As we are interested in the types of equivalent rather than their frequency, whenever this happens each type is noted regardless of whether it appears just once or more times. Thus the sequence *v* + *chvíli* was translated by all three formal equivalents, prepositional: ***v poslední chvíli*** *ho napadne varovat ji i jinak* (F) - ***at the last moment*** *it occurred to him to warn her about something else*, lexical-grammatical transposition: ***v té chvíli*** *poznal, že mu nezbývá, než aby vyšel s tím svým pomyslem o rozvratu* (F) - *He realized* ***then*** *there was nothing for it but to bring out the breaking down of morale he had thought up*, and zero: *Ještě že to nedal slepici, řekne Mon a* ***v té chvíli*** *zmizí.* (F) – *"A good job he didn't give it to that hen to eat," said Mon and disappeared.* The proportion of the types of equivalent is computed from the total of different types occurring with each sequence whose number then exceeds the number of sequences in the list. As Tables 5 and 7 show, the 50 sequences in the top T-score list were translated by 76 equivalents, of which 42 were prepositional, 28 were transpositions and 6 zeros. The ratio of prepositional and non-prepositional equivalents is 42 to 34 or 55.3 per cent to 44.7 per cent.

	Sequence	T-score	Equivalent		
			prep. PE	non-prep.	
				NPT	NPZ
1	být v	513.970	•	•	
2	v roce	247.719	•		
3	v době	196.171	•	•	
4	v případě	188.372	•	•	•
5	v letech	164.163	•	•	
6	v rámci	125.786		•	
7	v oblasti	124.707	•	•	
8	v chvíli	118.351	•	•	•
9	místo v	115.400	•		
10	v noci	114.008	•	•	•
11	žít v	113.923		•	
12	v životě	102.024	•	•	
13	v městě	97.994	•	•	
14	v případech	94.365	•		
15	v smyslu	93.935	•		
16	v skutečnosti	93.069		•	•
17	v domě	90.741	•	•	
18	vidět (se) v	89.840	•	•	
19	v sobotu	88.868	•		
20	v světě	88.642	•		
21	v podobě	86.924	•		
22	pokračovat v	86.196	•	•	
23	zůstat v	85.935	•		
24	v ulici	85.493		•	
25	v dnech	85.459	•		
26	v ruce	84.715	•	•	•
27	v pořádku	84.308	•	•	
28	v řadě	82.177	•		
29	objevit (se) v	82.134	•		
30	v okamžiku	81.317	•	•	
31	v duchu	81.298	•	•	
32	v neděli	80.910	•	•	

	Sequence	T-score	Equivalent		
			prep. PE	non-prep.	
				NPT	NPZ
33	změna v	80.421	•		
34	sedět v	80.400	•		
35	v části	79.635	•		
36	v výši	78.185		•	
37	spočívat v	76.929	•	•	
38	tvář v	74.308	•		
39	v pátek	73.060	•		
40	v pondělí	72.924	•	•	
41	ležet v	71.185		•	
42	v pokoji	69.986	•	•	
43	v okolí	69.952			
44	v století	69.880	•		•
45	v očích	69.617	•	•	
46	v dvou	68.963		•	
47	bydlet v	68.076	•		
48	držet v	68.034	•		
49	v prostředí	66.214	•		
50	v škole	65.813	•		
Total			**42**	**28**	**6**

(PE – prepositional equivalent; NPT – non-prepositional equivalent/transposition;
NPZ - non-prepositional equivalent/ zero)

Table 5: The correlation between the top T-score sequences and the type of equivalent

By contrast, sequences in the bottom T-score list occur mostly once and accordingly there is only one sequence (*v kabinetě*) which was translated by more than one type of equivalent. As Tables 6 and 7 show, the 50 sequences in the bottom T-score list were translated by 51 equivalents, of which 32 were prepositional, 15 were transpositions and 4 zeros. The ratio of prepositional and non-prepositional equivalents is thus 32 to 19 or 62.7 per cent to 37.3 per cent.

	Sequence	T-score	Equivalent		
			prep.PE	non-prep.	
				NPT	NPZ
1	v ředitelně	6.784	•		
2	v poklusu	6.671	•		
3	v vzrušení	6.663		•	
4	v kabinetě	6.580	•		•
5	v křesílku	6.474	•		
6	nevíra v	6.260	•		
7	v podprsence	6.161		•	
8	v polovici	6.139		•	
9	v předsíňce	6.125	•		
10	zasutý v	6.037		•	
11	v čepci	5.950	•		
12	poskočit v	5.929			•
13	v policích	5.651	•		
14	v keřích	5.608	•		
15	odplout v	5.499	•		
16	v úschovně	5.190			•
17	bělat (se) v	5.167	•		
18	v baráčku	5.136	•		
19	v záhonu	4.903	•		
20	v koncentrácích	4.782	•		
21	v internátu	4.631		•	
22	v úprku	4.620		•	
23	v variaci	4.562	•		
24	v ozvěnách	4.486	•		
25	v intonaci	4.339		•	
26	v halence	4.288	•		
27	v ponožce	4.271		•	
28	v návratu	4.224	•		
29	úctu v	4.095		•	
30	snídaně ve	4.095		•	
31	v pěstích	4.044	•		
32	v družnosti	3.664	•		
33	lehtat v	3.624		•	

	Sequence	T-score	Equivalent		
			prep.PE	non-prep.	
				NPT	NPZ
34	v pankreatu	3.503		•	
35	v argotu	2.859	•		
36	v planetáriu	2.767	•		
37	v gymnasiu	2.605	•		
38	v koželuhách	2.595	•		
39	v korekturách	2.479	•		
40	hroužit (se) v	2.259	•		
41	v kastlíku	2.133	•		
42	vzdout (se) v	1.833	•		
43	v kvízu	1.799		•	
44	v jídelničce	1.699	•		
45	v cítění	1.463		•	
46	v dózičce	1.387	•		
47	v ataku	1.101	•		
48	sklenout v	1.034		•	
49	v ukazovátku	0.981	•		
50	v matce	0.733			•
Total			**32**	**15**	**4**

(PE – prepositional equivalent; NPT – non-prepositional equivalent/transposition; NPZ - non-prepositional equivalent/ zero)

Table 6: The correlation between the bottom T-score sequences and the type of equivalent

The results of the correlation are summed up in Table 7. It shows that compared to the proportion of prepositional and non-prepositional equivalents in the whole sample of *v/ve* sequences, the sequences in the top and the bottom T-score list display an opposite tendency. The items in the top T-score list, which presumably represent (mostly) prefabricated sequences, display a markedly higher tendency to have non-prepositional and so (presumably) prefabricated English equivalents. While in the whole sample the ratio of prepositional and non-prepositional equivalents is 2 : 1 in the top T-score sequences it approaches the ratio of 1 : 1 as the hypothesis predicted. The reason why more than half of the presumably prefabricated sequences have prepositional (i.e. mostly compositional) equivalents is that prefabricated sequences include a number of strings which are both compositional and (lexically and grammatically) regular and as such will have a relatively large number of compositional and regular, i.e. prepositional, equivalents.

On the other hand, Table 7 shows that in the sequences in the bottom T-score list the proportion of prepositional and non-prepositional equivalents is very close to that of the whole sample of *v/ve* sequences. This fact, which somewhat weakens the initial hypothesis which predicted a higher proportion of prepositional equivalents in them, can be in part explained by the low frequency of these sequences. As mentioned above, both T-score and MI-score are reported to be adversely affected by low frequency and so the list includes some items which may be easily regarded as prefabricated, e.g. *v úprku, (budit) úctu v*. Translation of other items by a transposition can be attributed to translator's licence as they could easily have a prepositional equivalent, e.g. *jako chovanka v internátu > like a boarding-school girl* (*like a girl in a boarding-school*).

Sample	**PE**	**NPT**	**NPZ**	**PE proportion**	**NP proportion**	**total %**
***v/ve* sample**	409	158	33	68.2 %	31.8 %	600 100 %
top T-score	42	28	6	55.3 %	44.7 %	76 100 %
bottom T-score	32	15	4	62.7 %	37.3 %	51 100 %

Table 7: Comparison of the distribution in the total sample and the T-score subsets

Table 8 below lists the *v*-sequences which were translated by non-prepositional equivalents – transpositions and their respective T-score values (in the *SYN2005* corpus). As they include items not included in the top and bottom T-score items, their values can range between the values of the T-score lists: 65.813 to 6.784. Transpositions are assumed to correlate with Czech prefabricated sequences which are expected to have high T-score values. Hence T-score values above the mid-point of the 65.813 to 6.784 range, i.e. T-score 29.5, in the Table 6 sequences may be cautiously considered indicative of the influence of prefabricated sequences on the choice of translation equivalence in favour of the non-prepositional ones (and of their correlation).

However, only 20 of the sequences in Table 8, less than a half, exceed the T-score of 29.5. The reason for this can again be attributed to T-score unreliability due to the low frequency of the data. Below the midpoint T-score value we find items such as *v hloubi, vězet v, ve střehu, v míru, v předtuše* and *v úprku*, which one intuitively associates with prefabricated structures. So the result of the correlation is inconclusive and does not offer further corroboration to the findings in the T-score lists.

	Sequence	T-score
1	v úvahu	63.054
2	v bytě	60.578
3	v čtvrtek	60.355
4	růst v	55.996
5	v klidu	52.869
6	v budoucnu	50.658
7	v prosinci	49.549
8	cítit v	49.354
9	v rozporu	46.700
10	v týdnu	46.040
11	v válce	42.181
12	číst v	38.508
13	v dopise	38.044
14	v cizině	36.876
15	v umění	35.926
16	v hlase	35.270
17	v spojení	34.331
18	v podání	33.587
19	v kombinaci	32.406
20	v vztazích	32.295
21	viset v	28.868
22	v náladě	28.276
23	v předsíni	25.278
24	v hloubi	23.994
25	v žaludku	23.177
26	v rychlosti	23.120
27	v postavení	22.720
28	v poschodí	21.823
29	vězet v	21.689
30	v nouzi	21.605
31	v ničem	21.499
32	v pohledu	20.064
33	zahrnovat v	19.986
34	v střehu	19.916
35	v spěchu	19.792
36	v míru	19.592
37	v kufru	18.564
38	v moci	18.512
39	v kompetenci	18.335
40	v normě	14.963
41	v obálce	13.895
42	v předklonu	12.456
43	v předtuše	11.255
44	v zázemí	9.084
45	slít se (v kus)	8.161
46	v jícnu	6.667
47	v úprku	4.620
48	v olovu	1.279

Table 8: Correlation between *v*-sequences which were translated by non-prepositional equivalents (transpositions) and their T-score values (in the *SYN2005* corpus)

6. Conclusions

The study has brought home the realization that the concept of prefabricated strings, however attractive it may be, is far more complex than the general descriptions in the literature often suggest. It subsumes a diversity of different types of sequences so huge that it complicates their definition and detection, which apparently presents serious problems to standard statistical measures. The problems seem to derive from difficulties with formulating queries due to fuzzy boundaries of these sequences, especially in Czech, and in the case of this sample from the suspected unreliability of the statistical measures used, T-score and MI-score, caused by the low frequency of the data in the texts (despite the attempt to compensate for this by having recourse to a large corpus). The goal of estimating the proportion of free combinations and prefabricated strings in the *v/ve* sequences by purely statistical methods was therefore abandoned.

Nevertheless, the tests of correlation between two sets of sequences, those presumed to be prefabricated (the 'idiom principle') and those considered grammatical (the 'open-choice principle') on the basis of T-score, and their corresponding translations give a reasonable support to the hypothesis that the prefabricated sequence shows some preference for a non-prepositional (prefabricated?) equivalent. A third test, reversing the approach by trying to find out to what extent non-prepositional equivalents tend to translate sequences presumed to be prefabricated on the basis of their T-score values, proved inconclusive.

It may be that the underlying assumption 'non-prepositional equivalent equals prefabricated equivalent and so automatically correlates with prefabricated SL sequence' is too simplistic. There seem to operate other factors which lead to the choice of a divergent (non-prepositional) counterpart. Some of the non-prepositional correspondences appear to be tied to the typological differences between the two languages (as far as the correlation between a clause element and its semantic role is concerned), to the preferences in the area of anaphoric devices (prepositional phrases in Czech as opposed to pronouns in English), or to the use of postmodifying clauses with general temporal antecedents in Czech whose English counterparts are adverbial clauses of time (*Už ve chvíli, kdy odcházel za některou z milenek ... Even as he set out to visit another woman*), etc. This, however, remains an area to be further explored in future.

References

Bennett D.C., 1975, *Spatial and Temporal Uses of English Prepositions*. Longman.

Biber D., 2006, *University Language: A corpus-based study of spoken and written registers*. John Benjamins, Amsterdam/Philadelphia.

Biber D., S. Conrad, V. Cortes, 2004, If you look at... : Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25:3, 371–405.

Biber D., S. Johansson, G. Leech, S. Conrad, E. Finegan, 1999, *Longman Grammar of Spoken and Written English*. Pearson, Harlow.

Church K. W., P. Hanks, 1990, Word association norms, mutual information & lexicography. *Computational Linguistics*, 16(1): 22–29.

Church K. W., P. Hanks, R. Moon, 1994, Lexical substitutability. In *Computational Approaches to the Lexicon*, eds. B. Atkins, A. Zampolli, Oxford University Press, Oxford, 153–177.

Cortes V., 2004, Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23, 397–423.

Čermák F., 2007, *Frazeologie a idiomatika česká a obecná* [Czech and General Phraseology]. Karolinum, Praha.

Čermák F., M. Křen et al., 2004, *Frekvenční slovník češtiny* [Czech Frequency List], Nakladatelství Lidové noviny, Praha.

Foster P., 2001, Rules & Routines: a consideration of their role in the task-based language production of native and non-native speakers. In *Researching pedagogical tasks: second language learning, teaching and testing*, eds. M. Bygate, P. Skehan, M. Swain, Longman, London, New York ,75-94.

Hyland K., 2008, As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27, 4–21.

Klimšová K., 1999, *Zdroje chyb českých mluvčích v angličtině* [Sources of Errors in Czech speakers of English], MA thesis, FF UK, Praha.

McEnery T., R. Xiao, Y. Tono, 2006, *Corpus-based language studies. An advanced resource book*. Routledge, London and New York.

Osborne G., 1993, *Computer Based Analysis of Idioms and Idiom-like Phrases in English*. M. Phil. thesis, University of Birmingham.

Saint-Dizier P. (ed.), 2006, *Syntax and Semantics of Prepositions*. Dordrecht: Springer.

Scott M., C. Tribble, 2006, *Textual Patterns: Key words and corpus analysis in language education*. John Benjamins, Amsterdam/Philadelphia.

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

Tyler A., V. Evans, 2003, *The Semantics of English Prepositions*, Cambridge University Press, Cambridge.

Wray A., 2002, *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge.

Sample sources

Czech National Corpus - InterCorp. Institute of the Czech National Corpus FF UK, Praha. Accessible at WWW: <http://www.korpus.cz/>.

Kundera M., 2006, *Nesnesitelná lehkost bytí*. Atlantis, Brno.

Kundera M., 1984, *The Unbearable Lightness of Being*. Penguin – transl. Michael Henry Heim.

Viewegh M., 1994, *Výchova dívek v Čechách*. Český spisovatel, Praha.

Viewegh M., 1997, *Bringing up Girls in Bohemia*. USA: Readers International – transl. A.G. Brain.

Fuks L., 1963, *Pan Theodor Mundstock*. Odeon, Praha (4th ed. 2005).

Fuks L., 1968, *Mr Theodore Mundstock*. Orion Press, New York – transl. Iris Urwin.