1. The Aims and Structure of the Thesis

The aim of the thesis is to introduce different types of multi-word combinations, to account for their principal aspects and describe the differences in their use by native and non-native speakers of English. The research part seeks to confirm the hypothesis that the language of non-native speakers is generally less idiom-principle based (to use Sinclair's terminology) than that of native speakers. There is one thing that should be stressed at the very beginning. The main purpose of this dissertation was not to conduct an exhaustive quantitative study, but to explore and test several ways in which the phraseological competence of Czech learners of English could be investigated. A decision to focus on four alternative ways of comparison meant that the extent of data examined had to be limited. Another reason was the difficulty of acquiring authentic suitable data. Although there does exist The International Corpus of Learner English (Granger et al. 2009), it proved to be inaccessible at the time of writing the dissertation and so I had to rely on my own modest resources.

After outlining the rapidly developing field of phraseology (Chapter 2) and describing the data, sample corpora and the applied methodology (Chapter 3), the thesis proceeds to the research part.

The research part of the thesis comprises three main chapters. Each of them attempts to capture a different type of multi-word combination: recurring non-idiomatic word-combinations, phrasal and prepositional verbs and collocations. All three chapters start with a short theoretical overview, which is followed by sample analysis and the presentation of the findings.

Chapter 4 tackles the issue of frequent non-idiomatic word-combinations. It draws heavily on Biber's investigation of lexical bundles. Despite the fact that the essays and reviews of which the samples are comprised aspire neither to academic writing nor to the language of conversation which are the registers that Biber examined, their analysis highlights the key differences, not only in the production of the recurring word-combinations between non-native and native speakers, but also in that it uncovers several divergences between the distribution of three- and four-word combinations. William Fletcher's BNC-based phraseological database Phrases in English (PIE; http://pie.usna.edu) is used as a control sample to see how many lexical bundles in the strict sense are present in the three samples.

Chapter 5 on multi-word verbs analyses and compares non-native speakers' phrasal verb and prepositional verb use with that of native speakers. Since phrasal verbs appear to be one of the most difficult multi-verb combinations for non-native speakers, the exploration of this area is expected to pinpoint the differences between non-native and native speakers.

Chapter 6 devoted to collocations contains two separate studies. The first one offers the comparison of selected nouns and their collocational behaviour in all three samples. The second is inspired by Granger's investigation of learners' collocational sense of salience, i.e. the ability to tell which collocates go best with a given node.

Chapter 7 reviews the major findings of the thesis and outlines perspectives for future research and ELT learning.

2. Introduction: developments in phraseology

Since the mid-1980s, the importance of chunk-based language has increasingly come to the fore. The grounds for such popularity in linguistic research as well as language teaching are manifold: the growing interest in the study of lexicon triggered the establishment of phraseology on distributional grounds as a field in its own right. The precursor, however, was the emergence of corpus linguistics, which allowed for the exploration of lexico-grammatical patterns to an extent that had not been previously feasible. Recent developments in phraseology as well as applied linguistics have heightened the need for research into multi-word combinations. Both disciplines offer ample evidence that language is strongly patterned and words hardly ever occur in isolation. This current view of language, however, is markedly different from what the previous approach used to be.

Since it is fully acknowledged that grammar and lexis are the principal aspects of any language, the original approach towards lexis was seriously undervalued, especially due to the influence of generative linguistics. Grammar and lexis were originally separated and it was grammar which was considered systemic while lexis was perceived as loosely organized (Hoey 2005, 9). Robins (1964, 18), for instance, expressly excludes lexis and claims that it is grammar that lies in the heart of all linguistics. He argues that grammatical categories are comprehensive whereas categories in lexis are merely particular. “Lexicon requires particular and different statements for each item” and is therefore described by Bloomfield (1933, 274) as “an appendix of grammar and a system of idiosyncrasies”. Such treatment of grammar and lexis resulted in the subjective significance of the paradigmatic axis at the expense of the syntagmatic one. However, the relationship between grammar and lexis in the language hierarchy has been largely reformulated since the times of Chomsky. The roles have been reversed throughout the recent decades and it is lexis that is perceived as being communicatively above grammar. One of the first to attach profound significance to lexis was Halliday (1978, 1). Halliday affirms that the process of language learning primarily includes the stage of getting familiarized with meanings. Even before Halliday, Firth (1957, 190) argued in favour of semantics as the basis and the most important part of linguistics - “indeed, the main aim of descriptive linguistics is to make statements of meaning”. With the gradual development and prevalence of corpus linguistics it has emerged that language is full of recurring patterns and the former idea of the word’s independence has been compromised (Sinclair 1991). Sinclair readily dismisses the traditional separation of grammar and lexis and advocates the mutual interdependence of both entities as is demonstrated through the recurring patterns offered by large corpora. He notes that grammar and lexis represent only different aspects of one and the same thing, and that sense and structure are interdependent (1991, 104). “The meaning of a text can be described by a model which reconciles both paradigmatic and syntagmatic dimension”. The most recent trends in linguistics point to a direct observation that “lexis is complexly and systematically structured and grammar is the outcome of this lexical structure” and that “grammar is the output of routines, collocational groupings, the repeated use of which results in the creation of a grammar for each individual” (Hoey 2005, 9).

The fact remains that the shifts towards a radical change in language description are consequently reflected in the views of language acquisition. The popularity of lexis forces language experts to reformulate traditional recommendations of how English is to be taught to foreign learners. It turns out that however much intimate knowledge of grammar a non-native speaker possesses, it does not guarantee successful native-like communication. Pawley and Syder (1983, 195) observe that “most of one’s productions are, to the native ear, unidiomatic. Each sentence may be strictly grammatical but the trouble is that native speakers just don’t say things that way”. Words acquire their meaning through combinations with other words and only a small number of words standing individually keep their independent meaning (Stubbs 2002, 1, 15).

Sinclair’s idea of language organization further illustrates the point. Sinclair (1991, 109-110) recognizes two ways of language processing: “the open-choice principle” and “the idiom principle”. The former corresponds with the terminological tendency whereby words have no possibility but bear a fixed meaning in reference to the world with the only restraint - grammaticalness. This “slot-and-filler model” is not applicable in the majority of cases since it is universally acknowledged that “each sense of phrase is coordinated with a pattern of choice that helps to distinguish it from other senses. Each is particular, each has its uses and specific environment” and it is not only grammatical restraints that have to be taken into account (Sinclair 1991, 78). The central aspects of the idiom principle are collocation and idiomaticity. This principle is in line with the phraseological tendency where collocation and larger patterns of language are encountered, where variation takes place commonly and the independence of words is dismissed. The possible constraints are not only grammatical, but also semantic, lexical, pragmatic and register-based. Words occur in the company of other “pre-selected” words, comprise particular grammatical structures, occur in a particular semantic environment and specific pragmatic associations are involved. The idiom principle, with its chunk-based nature, turns out to be crucial in language production. Several studies have confirmed (Granger in Cowie 2005), though, that while native speakers tend to operate largely on the idiom principle, non-native speakers prefer to convey meaning via the open-choice principle.

Indeed, it has emerged through translational-pedagogical practice that a great number of even very advanced learners and other non-native speakers still seem to lack something which makes their language production comparable with that of native speakers. What appears to be most problematic is how to select the “right” expression, which sounds both idiomatic and native-like, among those expressions which are constructed on the grammar basis, or represent a highly-marked usage. This is what Pawley and Syder (1983) term “the puzzle of native-like selection”. In this widely-cited paper, the authors argue that native-like selection is “the ability of the native speaker routinely to convey his meaning by an expression that is not only grammatical but also native-like” (1983, 191). Pawley and Syder point out that a native-speaker’s syntactic knowledge is not identical with the grammatical knowledge championed by grammarians. Native speakers do not fully make use of such grammatical rules and prefer to select a prefabricated expression from long-term memory, whereas non-native speakers are inclined to use grammar rules. They call these prefabricated expressions “memorized sequences” and “lexicalized sentence stems”. These are fixed, but minor alterations are permissible, and native speakers possess the knowledge to what extent these stems can be varied or extended. These extensions and alterations pose a significant obstacle for language learners since no explicit guidelines and regulations on them are available. No one is able to explain why the idiom *pass the buck (to blame someone or make them responsible for a problem that you should deal with)* allows a possible manipulation into the passive voice *the buck has been passed* or another alternative, such as *there was too much buck passing* while other idioms do not. Baker (1992, 64) notes that the matter of co-selection appears problematic not only for language learners but also for professional translators. “A person’s competence in actively using the idioms and fixed expressions of a foreign language hardly ever matches that of a native speaker“. It is also argued that it is often mother-tongue interference which triggers inappropriate co-selection of lexical items .

In the light of Pawley and Syder’s paper and many others touching upon this issue, it is obvious that at the core of the problem is simply the fact “whether we are familiar with the norms of co-occurrence in the language” (Stubbs 2002, 113). Stubbs’ term co-occurrence stands for a range of phenomena called variously, “multi-word units”, “extended lexical units”, “formulaic lexical combinations” etc. He stresses that our ability to use a particular expression appropriately depends on our familiarity and awareness of the cultural norms and customs: “The meaning of a particular word-combination relies on additional cultural knowledge which these combinations often encapsulate” - our ability to understand the meaning depends on “the inference from real world knowledge and conventions” (Stubbs 2002, 3, 13). The idea of examining the norms of co-occurrence is elaborated upon by Stubbs (2002, 110) in a more tangible form which can be seen as the fundamental methodological procedure: “ An entirely automatic method of discovering how many of such combinations in the text occur frequently in the language could take possible two-, three-, four-, five- or six-word combinations in the text, and check if the same combinations occur in a large corpus”.

This is exactly the method which Mason (2007) made use of in his study “Multi-word as a model of grammar”: “We start off by looking at a sentence taken from the call for papers of a conference: *The papers presented at the conference will be available in proceedings on the first day.* For each word in this sentence we retrieve the multi-word units from the BNC as described earlier. We then select those units which match the surrounding words in our sentence and display the results in tabular form”. His findings led him to the conclusion that “by tabulating the MWU as we have done here it becomes apparent that they overlap and link up to form a longer sequence, similar to what Hunston and Francis (2000) describe as “pattern flow”. That is to say, one pattern results in the selection of the following one with which it partly overlaps and thus we find that they “flow” into each other with no specific boundary”.

The term multi-word unit (MWU) is the primary concern of this thesis. Multi-word units have various names and sometimes there is confusion as to whether these names can be used interchangeably or whether slight differences exist between them. The most frequently used ones are “lexical phrases” (Nattinger and De Carrico 1989), “composites” (Cowie 1988), “gambits” (Keller 1988), “routine formulae” (Coulmas 1988), “phrasemes” (Melčuk 1988), “prefabricated routines and patterns” (Krashen 1981), “sentence stems” (Pawley and Syder, 1983), “formulaic sequences” (Wray, 2005), (Hunston and Francis 2000, 7). Several linguists have come up with taxonomies of recurrent expressions. The most influential and idiomatic taxonomy is proposed by Cowie (1988), who primarily distinguishes between composites (restricted collocations, figurative idioms, pure idioms) and formulae (which are classified into routine formulae and speech formulae). Melčuk (1988) distinguishes between semantic phrasemes (collocations, quasi-idioms, idioms) and pragmatic phrasemes. As opposed to idioms, which are regarded as rare and marginal phenomena, compositional multi-word units represent a key element in text. They are ready-made units and, according to Erman and Warren's research, form approximately 55 per cent of text (cf. Sinclair's claim of 55 per cent text being based on the idiom principle). The scope of their usage depends on a range of factors: the way units are classified, the method of calculation, text type etc. However divergent the individual taxonomies might be, though, several criteria must be fulfilled so that a multi-word sequence could be proclaimed as formulaic (Hickey in Wray 2005, 40). The sequence must contain at least two morphemes but it may stretch even further and comprise four, five or even more lexemes. The second important condition is for the sequence to be phonologically coherent, grammatically advanced, and idiosyncratic. Lexicalization and institutionalization are further important criteria. While lexicalization stands for "the process that a complex lexeme, once coined, tends to become a single complete lexical unit, a simple lexeme and through this process it loses the character of a syntagma to a greater or lesser degree" (Lipka 1990, 93), institutionalization involves "the process by which a string or formulation becomes recognized and accepted as a lexical item of language" (Moon 1998). The string must be situationally dependent and fixed. Pawley and Syder (1983) compare *first (and only attempt)* and **first and only aid* and thus indicate which string could qualify for membership in a formulaic (phraseological) group and which does not fulfil this criterion. Few formulaic strings are wholly fixed, most of them are variable and allow insertions which are, to some extent, permissible. The string must also be idiomatic - it must sound native-like. Formulaic sequences form a continuum with entirely transparent word-combinations at one end and completely opaque word-combinations at the opposite end of the scale with semi-fixed expressions lying in between (Wray 2005). Apart from the phraseological expressions mentioned previously, one more type of multi-word units has been introduced by Douglas Biber (1999). He speaks of "lexical bundles" which are significant in view of their high frequency in a specific register. Two interesting aspects are ascribed to them: "they are not usually idiomatic in meaning and they are not complete grammatical structures".

Drawing on Stubbs, Mason and others, one essential feature must be stressed: according to Stubbs (2002, 123) "text analysis must be always comparative: we can interpret patterns in an individual text only if we know what is to be expected in the language as a whole". The present analysis deals with the production of an English text by two groups of non-native speakers and one group of native speakers and hence provides a specific comparative element. The study conducted by Granger (2005), *Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae,* was a direct inspiration for one section in this thesis. In particular, Granger focuses on the comparison of native and non-native speakers in respect of similarities and differences in the use of collocations and sentence builders. She draws heavily on the corpus data and exclusively deals with the selected types of collocations (intensifying adverbs, i.e. amplifiers – *bitterly disappointed)* and the sentence builders (discourse frames: active and passive). The findings emerging from this investigation are partly in line with intuitive assumptions even though several unforeseen pitfalls await the researchers, especially as far as the application is concerned. Despite this, Granger puts forward other methodological paths on how to carry out a project from the contrastive view point.

Owing to the fact that the aim of this research was to identify, analyse and compare multi-word units in three sample corpora, it was necessary to take into account the fact that various types of multi-word units may be expected. In view of this, it is not very optimistic to remember Stubbs's (2002, 62) comment according to which one could witness that "it is an odd failing of linguistics that it has no convincing descriptive theory of units of meaning. It has, for instance, no widely accepted methods of segmenting spoken or written discourse into semantic units. Advances in a theory of units of meaning have come largely from practical activities (such as dictionary making and language teaching) rather than from theoretical linguistics".

As the focus of this study is on four specific types of phraseological phenomena, namely lexical bundles, multi-word verbs (phrasal and prepositional verbs) and collocations, each requiring an extensive description closely related to the sample analysis, it seemed appropriate to place the theoretical introduction to each of these phraseological categories at the beginning of the three chapters dealing with their sample analysis, rather than here.

3. Methodology

3.1 Preliminary issues

Language in a machine-readable form offers several advantages. Real language samples which amount to millions of words, replace the introspective approach when language examples were created, evaluated and based on linguists' intuition. A large collection of texts gathered from a variety of sources guarantees authenticity and objectivity. Computational tools allow for data processing in the most appropriate way for the purpose of researcher's analyses with the possibility of texts being analysed both qualitatively and quantitatively. PC software allows for extensive statistical computations which provide sophisticated interpretation of the data (Bowker 2001, 346).

It has been mentioned in the Introduction that language experts once believed that learners should first and foremost learn the target language grammar, whereas recurrent sequences of language were paid only scant attention. With the arrival of corpus linguistics, a great number of linguists started to prefer the corpus-driven "hypothesis-finding" approach (bottom-up approach) or the corpus-based "hypothesis-testing" approach (top-down approach) as described by Barlow (2000). In particular, the corpus-based approach draws on the data obtained from the corpus and additionally makes use of other sources (e.g. introspection, dictionaries), the corpus-driven approach concentrates on data obtained from the corpus only (Čermák 2006, 15).

The present study of three different types of multi-word units (lexical bundles, multi-word verbs, collocations) aims to prove that native speakers produce language in a more formulaic manner than non-native speakers. Each section is prefaced with several questions and hypotheses related to the particular type of multi-word unit, the obtained results subsequently help to modify the assumptions made at the beginning of the investigation. The investigation is performed in the light of the corpus-based approach, since different types of evidence, not only the British National Corpus are used.

3.2 Types of evidence, sources

Three different types of evidence are used including the British National Corpus (BNC) along with a British phraseological database – The Phrases in English (PIE). The Phrases in English is a large new interactive phraseological database created by William Fletcher and Isabel Barth, available at http://pie.usna.com. It offers quantitative information on very frequent phrases in the BNC, it deals with the distributions, functions and structure of recurrent sequences (Stubbs 2004). It mainly works with "n-grams", that is to say "recurrent strings of uninterrupted word-forms" (Stubbs 2004). The software is used to retrieve n-grams from the BNC, thus providing the user with information on frequency of a particular word-combination. The word combinations are sorted either according to the frequency or alphabetically and the minimal frequency is 3 occurrences in the database. For the purposes of this investigation, the database is used in Chapter 4 on lexical bundles to see whether any word-combinations produced in the three samples could be regarded as lexical bundles.

Apart from the BNC and the PIE, the present study takes advantage of several dictionaries – The Oxford Advanced Learner's Dictionary (OALD), The Oxford Dictionary of Phrasal Verbs (ODPV), Oxford Collocations Dictionary for the Learners of English (OCDLE) and native speakers who have completed a degree in linguistics.

3.3 Data, materials, sample corpora

In order to investigate the similarities and differences in the non-native and native production of multi-word units, the data for the study consists of three electronic samples of written texts which comprise Czech students' essays, non-Czech students' essays and native speakers' reviews.

At the beginning of the investigation, it was necessary to consider appropriate topics for learners, topics that learners would be familiar with and find easy to cope with. Subsequently, relevant material was collected and three different topics were shortlisted: 1. *A book/film review* ; 2. *Money matters*; 3. *My home is where my heart is*.

The next step was to acquire texts with a similar number of words representing native-speaker data introducing the same topics, which was not entirely trouble-free. After protracted procedures to carry out this task it was necessary to narrow down the selection of topics - finally the topic *The book/film review* or *The book is my best friend* was opted for - it was the only topic which produced sufficient data.

A sample of 37 Czech learners were asked to write essays on the topic *The book is my best friend* or *The book/film review.* This collection of writing forms the sample corpus of 9 411 words. The essays were written by students at pre-intermediate level, intermediate level, and upper-intermediate/FCE level. All the students were between 17-18 years old at the time of writing, studying at a state secondary school in Prague. The pre-intermediate level group contains 11 students (seven girls and four boys). Despite being in their last year of secondary school, the students' skills were at pre-intermediate level and only some of them were going to take their school-leaving exam in English at that time. The upper-intermediate group consists of 16 students (eight boys and eight girls). Before writing the essays, some of them had already passed the First Certificate in English, a few of them had studied in the United States for one year. An important difference between the pre-intermediate and upper-intermediate/FCE group was that the former attended the secondary school for four years only (i.e. the students started to study at the age of 15; presumably, they had taken some English classes at primary school) while the latter group started to learn English at the secondary school at the age of 11-12, with four lessons of English a week. The final group group which was set up were 10 intermediate students.

The number of words required for each essay was between 200-250. However, in several (mostly) upper-intermediate level essays the limit was slightly exceeded, some students with lower proficiency levels did not write the requested number of words. The stories students described differed widely, however, ranging from *Harry Potter, The Adventures of Huckleberry Finn, The Lord of the Rings, The Da Vinci Code* to *Angels and Demons.*

The analyses of the individual types of multi-word units are carried out on the whole sample (disregarding a learner's level) with the exception of the phrasal verb section.

Apart from the data extraction from this sample corpus, another group of Czech learners were asked to sit a test on significant collocations (Chapter 6). The total of 15 secondary school students with intermediate to upper-intermediate level, and 6 adult learners at FCE level (first certificate) participated in this project. The test they underwent is a re-study and re-investigation of the collocation salience test by Sylviane Granger (2005).

The sample of the other group of non-native speakers comprises 19 texts, available at http://www.bookrags.com, a website offering a great amount of material provided by foreign writers. For the purposes of the present study, non-Czech students' essays were downloaded from this website. The final version of the sample contains 9 329 words and the number of words in each essay differs widely since no general requirement was imposed; some essays contain 300 words and some exceed 600 words. The majority of the topics correspond with the Czech learners' topics *(Harry Potter, The Adventures of Huckleberry Finn, The Lord of the Rings, The Da Vinci Code, The Pit and the Pendulum, The Battle of Agincourt etc.*).

The third sample contains 22 book reviews written by native speakers - mainly professional review writers and comprises 9 340 words. The reviews were downloaded from the website http://happypublishing.com/. Both non-native and native samples were cleaned so that no titles, captions, footers, headers or references would obscure the final results.

Although the size of the sample corpora is relatively small (about 28 100 words), it is arguably sufficient for the purposes of this study. Apart from methodological reasons (the focus on four different types of phrasemes), the size of the corpora was determined by practical considerations such as the limited access to authentic texts on suitable topics within the permitted span of time, etc. This study should be thus regarded as a pilot study aiming at a preliminary investigation of multi-word units and an attempt to develop a methodology that will accurately map the degree of idiomaticity in a learner's language production. The results are expected to show to what extent the methodology has been successful and can be applied to large samples to give a more detailed and representative picture.

3.4 The tools

Corpus data are analysed by using Collocate version 1.0 (Barlow 2004) and ConcGram (Greaves 2005-2009). The former is used for the extraction of non-idiomatic recurring word combinations (Chapter 4), the latter for the retrieval of concordance lines comprising phrasal verbs, prepositional verbs and for the collocation analysis (Chapter 5, 6). Both software packages consist of a set of tools, two of which are especially useful for the investigation: Wordlist and Concordancer (Waibel 2007). The Wordlist offers important statistical data of the sample corpora: it provides the overall number of types, tokens, type/token ratio. Words can be viewed according to frequency, in descending order, or organized in alphabetical order and in ascending order.

Concordancer provides the researcher with the surrounding context of the analysed words, phrases or distributed structures. This makes it possible to make statements about the collocational, colligation (provided the sample corpora are morphologically annotated), semantic and pragmatic behaviour of the linguistic items. There are several types of concordancers: "a corpus-based" concordancer where the entry is the user's word, phrase or a structure, the concordancer provides all occurrences of this word, phrase or structure in context. Other types of concordancers are either "text-based" or "story-based" concordancers. The present study makes use of the corpus-based concordancer.

The latter program, ConcGram, is especially useful for creating the phraseological profile of the linguistic items once it is morphologically annotated (tagged). According to the creator of the program, the program "is able to fully identify and describe the meaning shift units" (MWU called lexical items by Sinclair). It also enables to focus on the collocational frameworks, that is to say the co-selection of grammatical words ("the....of", "of ….the"). However, for the purposes of this investigation, the sample corpora are not morphologically annotated, the identification of individual word classes is carried out manually.

3.5 Research questions, hypotheses, procedures

Since the present study involves diverse types of multi-word units which are treated separately, the present methodology provides a brief outline of the central issues related to the research and individual chapters. Each chapter is prefaced with a list of questions relevant for the research, these include hypotheses which concern the particular type of multi-word units in relation to two groups of learners (Czech and non-Czech) and native speakers.

Studies which concentrate on the area of multi-word units are numerous. Established authorities in this field point out major differences between the strategies adopted by non-native and native speakers towards multi-word units. Granger (2005), for instance, argues that learners are inclined to what Sinclair (1991) calls "the open-choice principle" while native speakers operate predominantly on "the idiom principle". Learners tend to construct sequences of words by means of grammatical rules whereas native speakers make use of "prefabs" which are stored in the human mind. Even though it is a subconscious process as regards both non-native speakers and native speakers, the native and non-native language production gives rise to language which has different flavour with native speakers and non-native speakers.

The present study therefore assumes that learners will be more inclined to use the open-choice principle in text production, they will produce multi-word units through grammatical sequencing. Mother tongue interference, which is a source of errors of different nature, will inevitably play a certain role.

The three areas dealt with will be covered by three separate chapters. Chapter 4 (Recurrent non-idiomatic word-combinations) will presumably throw light on the use and variety of three-/four-word combinations in the learner samples and native speaker sample. The assumption is that non-native speakers are prone to greater repetitiveness in the production of non-formulaic word-combinations. Thus the ratio between the type/tokens is expected to be higher on the part of non-native speakers. Whether any of these word-combinations could receive the status of lexical bundles will be confirmed by the Phraseological Database of very frequent phrases in English (PIE). However, given that the register under scrutiny is neither academic English nor the language of conversation, a great number of lexical bundles in the strict sense may not be encountered. All three-/four-word combinations will be extracted by the application Collocate, the ordering of word-combinations will be frequency-based and the question of whether they can qualify into the lexical bundle group will be confirmed or denied by the PIE. Despite the creativity aspect, more lexical bundles in the strict sense are expected in the native speaker sample. The structural typology of three- and four-word combinations will be conducted so that it is possible to make statements about the structural richness of the texts. Greater structural richness is expected in the native speaker sample.

Chapter 5 on Phrasal and prepositional verbs aims to confirm the assumption that phrasal verbs are one of the major stumbling blocks for learners. The fact that phrasal verbs often elude learners could be explicable in view of the opaque nature of some phrasal verbs together with the fact that a great number of phrasal verbs often cover several meanings. On the other hand, studies which have confirmed learners' greater use of prepositional verbs are many and have proved that the main impediment for learners is the choice of the suitable preposition (Waibel 2007). The general assumption in this chapter relates to the low percentage of phrasal verbs on the part of the non-native speakers along with a less extended repertoire of lexical verbs, adverb particles and the range of phrasal verbs as such. On the other hand, native speakers are supposed to use a greater number of phrasal verbs than learners.

Concerning prepositional verbs, their number is expected to be double the phrasal verbs as regards both types and tokens (in both learner samples and the native-speaker sample). The question arises to what extent learners will be competent to use prepositions appropriately. The prepositional-verb investigation will also include semantic analysis of the prepositional verbs. The semantic taxonomy together with the corpus findings will draw on the data provided by LGSWE (1999). The semantic taxonomy will be carried out with a view to seeing whether the prepositional verbs used by non-native speakers represent frequently used prepositional verbs in the language. Given the above description, it is possible to infer that the analyses carried out on phrasal and prepositional verbs will partly differ in the aspects under scrutiny since phrasal and prepositional verbs are different in nature. One of the key methodological issues, to be taken into account, is how to extract multi-word verbs from ConcGram. Since neither of the sample corpora are morphologically annotated, all the verbs retrieved by means of ConcGram, will have to be manually sorted and arranged into three groups: phrasal, prepositional and phrasal prepositional verbs. Their status will be checked in Cowie's Oxford Dictionary of Phrasal Verbs (1993, 2010), the frequencies of phrasal verbs and prepositional verbs will be checked in the BNC and LGSWE (1999).

The last section devoted to Collocation (Chapter 6) is split into two parts. In the first part, the collocational behaviour of selected nodes related to *reading* will be investigated. In particular, the structural types *adjective + noun, noun + verb* will come under scrutiny. Concordance lines with the selected nodes will be extracted by ConcGram and the frequencies checked in the BNC. The present study will set an arbitrary limit of 5 and more occurrences. Accordingly, the percentage of frequent collocations will be focused on. The key issue here is to find to what extent individual groups produce collocations which occur in the BNC frequently. Also the range of collocates will be examined in detail. It is assumed that it will be greater on the native speakers' part, or will be formed by "more interesting" collocates.

In the second part of the collocational section, a group of 21 learners (15 secondary school students at intermediate level and 6 adult learners exhibiting FCE level) will be asked to undergo a test designed by Sylviane Granger (2005). This collocation salience test is a re-study and re-investigation of the salient collocations carried out by 112 participants in Granger's project. The test comprises *adverb + adjective type* of collocations and aims to find out to what extent learners are familiar with salient collocations, whether mother tongue interference plays a role in the collocational acquisition and focuses on the learners' sense of salience. It contains 10 adverbs and 15 adjectives. In this project, the learners will be asked to select, from a list of 15 adjectives in each case, the acceptable collocates of 10 amplifiers, by underlining all the adjectives which in their opinion can co-occur with an amplifiers. They will be asked to circle the adjective which they think is more frequently associated with the amplifier than the rest of the adjectives. The frequencies will be retrieved from the BNC and the arbitrary limit of at least 3 occurrences in the BNC will be set so that a combination can be labelled a collocation. The reason why different limits of occurrences will be set in these investigations is that in the latter case (Granger's collocation salience test), the combinations form, in the majority of cases, restricted collocations, which are not so frequent in the language, thus a greater tolerance as far as the minimum limit is shown. The time to complete the test will be set at 30 minutes, but more time will be provided if necessary. The learners will be assured that the test is intended for research purposes and incorrect responses cannot influence their final school achievement.

3.6 Questions related to terminology

Whenever a learner's performance in the process of learning a foreign language is assessed, some already established norm is important to consider. It is universally acknowledged that the English language is extremely wide-spread and serves as a kind of lingua franca. Therefore, the investigation requires the establishment of norms according to which learners' performance will be measured. Some language experts suggest that the performance of learners should be evaluated on the basis of so-called ”Nuclear English“, a recently established language norm, aimed purely at foreign learners who will be satisfied with achieving an intermediate level of English. On the other hand, a great proportion of learners would like to achieve almost native-like proficiency, with a view to becoming a translator, interpreter or an English language teacher (Nesselhauf 2005, 37-38).

If the concept of ”norm” is considered, a set of terms related to the concept of ”error“ emerges. Expressions such as “mistake”, standard” vs. “non-standard”, ”correct” vs. “incorrect”, “deviant”, “dubious”, ”acceptable” vs. “unacceptable” must be taken into account when the error analysis is considered. All these terms imply a deviation in terms of “a form or usage that is unlike the norm” (Nesselhauf 2005, 39). However, Nesselhauf (2005, 39) points out that acceptability or compliance with the norms should be considered a matter of degree.

The present study draws on the British and American standards such as embodied in widely recognized grammar books and dictionaries which represent the norms for this investigation. These norms will be used to measure the phraseological performance of non-native speakers and indicate to what extent learners' language performance “deviates” from the valid norm. The notion of “error”, which essentially comes up in relation to language learning will occur several times in the present study and the terms such as “mistake”, “incorrect”, “non-standard”, “deviant”, “dubious” or “unacceptable” will all be used interchangeably, that is to say all of them will refer to the notion of “error”.

4. Contrastive study of recurrent non-idiomatic word combinations

This section aims to reveal the major similarities and differences between recurring non-idiomatic word combinations which are produced by two learner groups - a Czech group of secondary school students and another group of non-Czech learners. These are subsequently contrasted with recurrent non-idiomatic word-combinations in book reviews written by native speakers. The objectives of this section are thus twofold: the first one is give a theoretical account of the term lexical bundles, the second is to compare the frequency, diversity, structural types of four-word and three-word combinations in all three samples. The frequency of these recurrent word-combinations is checked against the PIE in order to find out whether, which and how many lexical bundles in the strict sense occur in the samples. The adopted criterion in this analysis is that of Biber's et al. (1999), according to which a sequence has to occur at least ten times per million words to be considered a lexical bundle. Two terms are adopted throughout this investigation: "lexical bundles" and "word-combinations". Any three- or four-word sequences with a frequency of occurrence at least ten times per million words will be classified as "lexical bundles" while the term "word-combinations" is used for any sequence regardless of this frequency threshold. These two terms are thus not used interchangeably in this investigation.

4.1 Theoretical background

4.1.1 Multi-word sequences, lexical bundles

Phraseology and corpus linguistics are in their heyday and the subject of multi-word units has been the primary concern for many linguists as well as language teachers for several decades. Even though multi-word sequences is a general term for extended sequences of words, multi-word combinations have been given several other names. As noted in Hunston and Francis (2000, 7), the term "lexical phrases" is perhaps used most commonly (Nattinger and DeCarrico 1998, 1992), followed by a few others, including "routine formulae" (Coulmas in Cowie 1998), "gambits" (Keller in Cowie 1998), "composites" (Cowie 1988), "sentence stems" (Pawley and Syder 1983) or "prefabricated patterns" (Krashen 1981).

However, Wray (2005, 7) holds that the awareness of multi-word sequences goes back already to the 19th century. At this time the neurologist John Hughlings Jackson pointed out certain levels of fixedness in the language. In particular, Jackson noticed that even aphasic patients could be fluent in rhymes, prayers and routine greetings. At the beginning of the 20th century, de Saussure (1916/1966) observed that speakers subconsciously put two or more linguistic signs that co-occur together and form a whole unit. Similarly, multi-word sequences did not go unnoticed by Jespersen (1924) who maintained that to remember all words separately was almost unthinkable. A claim made by Bolinger (1976, 1) further illustrates the point that "our language does not expect us to build everything starting with lumber, nails, and blueprint, but provides us with an incredibly large number of prefabs". During the 1950s, the era of Chomsky, however, the attention from multi-word formulaic sequences was diverted to grammar and it was several decades later when the idea of recurrent patterns in language again emerged.

Linguists focusing on extended sequences of words – lexical phrases, chunks and idioms have developed two main approaches (see Chapter 6). Whereas the first group of linguists is concerned with those units which have idiomatic features or are pure idioms (e.g. *put all your eggs in one basket)*, others are more keen to look into the area of non-idiomatic expressions, which are characteristic for their perceptual salience (Biber and Barbieri 2007, 265). Lexical bundles occupy quite a different position among multi-word units since they are defined neither by the idiomatic aspect nor by perceptual salience but rather by their statistical occurrence.

The term lexical bundles came to be more widely known when introduced in the Longman Grammar of Spoken and Written English (Biber et al. 1999, 990) in which lexical bundles are defined as "sequences of word-forms that commonly go together in natural discourse“. Nevertheless, the concept of lexical bundles already emerged earlier in the investigation conducted by Salem (1987), whose research involved an exploration of a corpus of French government texts. A decade later, Butler (1997) and Altenberg (1998, 121) used the notion of lexical bundles when they carried out an investigation of Spanish and English Corpora. Namely, Altenberg analysed recurrent word combinations in spoken English and concluded that they could be found at all levels of linguistic organization. The result that emerged from his examination is clear: speakers make use of recurrent expressions which are typically used by native speakers, however, most of these are not idioms in the strict sense, i.e. frozen, semantically non-compositional sequences of words which allow no modification, extension or deletion of elements.

LGSWE (1999, 991) points out two major characteristics of lexical bundles: they are usually not idiomatic in meaning; rather, they are transparent (e.g. *do you want to; I don't know what you mean)*. Unlike idioms which are considered more or less rather marginal in everyday language production, the high frequency is the major factor when considering whether a particular extended sequence of words qualifies as a lexical bundle. Only if word-sequences appear at least in five different texts are they classified as lexical bundles. Nonetheless, some sequences of words might occur more frequently in a discourse, but it does not necessarily mean that they can be defined as lexical bundles (Biber 2004, 134). Recurrent sequences appearing frequently in a discourse may represent only a speaker's immediate needs or could be topic-bound. By contrast, true lexical bundles are building blocks that occur commonly in different situations.

According to Biber (2004, 135), the second significant aspect related to lexical bundles is that they usually do not represent a complete structural unit. Surprisingly, Biber's (2004, 135) investigation reveals that a mere 15 per cent of the lexical bundles in conversation are complete structural units and it is less than 5 per cent of lexical bundles in academic prose. Most lexical bundles stretch across structural units. To illustrate the point, *I am not sure what* presents two typical features related to lexical bundles: syntactic incompleteness and the fact that they cross phrase boundaries. The first clause *I am not sure* if followed by *what* which is the beginning of the second dependent clause.

Also, some lexical bundles typically occur in one specific register while others more commonly occur in a different register. Experts who regularly take part in a specific discourse are familiar with a respective set of bundles, specific for a given register. The use of these bundles then indicates to what extent speakers are communicatively competent in a particular discourse or whether their way of using the language is inappropriate (Hyland 2008, 5).

There are two approaches with regard to the statistical occurrence of bundles. One advocates the number of ten occurrences per million words for four-word bundles (LGSWE 1999). The other approach, taken by Biber and Barbieri (2007), is more conservative – a lexical bundle has to appear at least forty times per million words to be called a lexical bundle. Both of these limits are strictly speaking arbitrary and usually a less conservative frequency threshold is used with five-word or six-word bundles.

Shorter bundles often form a part of more than one longer bundle. The corpus findings in LGSWE (1999, 993) show that three-word bundles occur most commonly both in conversation and academic prose even though there are more lexical bundles in conversation.

Three-word bundles are almost ten times more frequent than four-word bundles; four-word bundles are ten times more frequent than five-word bundles. While three-word bundles occur over 80 000 times per million words in conversation and over 60 000 times in academic prose, four-word bundles appear 8 500 times in conversation and 5 000 in academic prose.

The lower frequency of longer bundles can be accounted for by the complexity of their production, which is especially the case of five- and six-word bundles. Speakers are required to make a greater effort to produce such long sequences. Most frequently occurring three-word lexical bundles in the BNC are *I don't know, I don't think, do you want, I don't want* in conversation. Lexical bundles such as *in order to, one of the* or *the fact that* are most common three-word bundles in academic prose.

4.1.2 Research into lexical bundles

Apart from Salem (1987), Butler (1997) and Altenberg (1998), who already used the concept of lexical bundles in their analyses as mentioned in the previous section, the investigation conducted by Biber et al. (LGSWE 1999) could be regarded as a pioneering study in the area of lexical bundles. Biber et al. (1999) in LGSWE deal with the distribution of lexical bundles across four major registers of language; this investigation thus serves as a guideline for further research into lexical bundles. Further analyses conducted by Biber, Conrad, Cortes (2004) and Biber (2006) provide more sophisticated functional and structural classifications of lexical bundles.

Subsequent research conducted by Stubbs and Barth (2003) was aimed at the frequency of pronouns. The research reveals that the frequency of certain pronouns allows to differentiate one text-type from another, such as fiction and academic articles. Several other studies related to bundles have been conducted. Biber, Conrad and Cortes (2004) focused on the native speaker production and contrasted lexical bundles in classroom teaching with lexical bundles in conversation and academic prose. Findings worth noting emerged from this investigation: Biber and his colleagues concluded that the majority of lexical bundles in academic prose were phrasal whereas lexical bundles recurring in the language of conversation showed signs of clausal structure. Further, Cortes (2004) performed a study comparing lexical bundles used by experts in biology and history with lexical bundles used by native speaker students. She explored to what extent native university students in respective fields produced bundles similar to those produced by native experts in these fields. She observed that even though the students did produce lexical bundles, they were not identical with those used by professional authors. Similarly, Hyland (2008) addressed the issues of lexical bundles: he focused on the differences in native speaker production. He carried out research into four-word bundles present in research articles, PhD dissertations and Master theses. Although the results of the investigation confirmed that bundles were essential building blocks, even more importantly it turned out that they helped to distinguish written texts as regards their focus of interest. By contrast, Chen and Baker (2010) were interested in comparing the non-native and native production of lexical bundles in academic writing: two native samples (expert and novice) together with a non-native sample were contrasted both from the structural and functional view point. They concluded that native novices and non-native Chinese students shared a good number of features: some traces of inexperienced writing could be observed in both samples, especially the overuse of verb phrase based bundles and discourse organizers, which are not typical of the academic register. Conversely, native professional writers kept the established academic norms.

4.1.3 Lexical bundles and collocations

The reason why there is a tendency for some authors to associate lexical bundles with collocations is that especially three-word lexical bundles are described as a kind of extended collocations (Cortes 2004). However, there are several differences between these two phenomena. Nesselhauf (2005), for instance, defines collocations as “some type of syntagmatic relations of words which is arbitrarily restricted“. In particular, the semantic aspect plays a considerable role as regards collocations (cf. Cowie’s “restricted collocation”) – it is in fact a defining feature - whereas lexical bundles are identified without reference to meaning. Both phenomena differ also in terms of the structure. While the structural element carries a lot of weight in terms of collocations (they are syntagmas), lexical bundles are usually not complete structural units. Rather than structural completeness, it is their statistical significance that is important (LGSWE 1999). Apart from that, Hoey (2005, 3) emphasizes the role of the psychological association in connection with collocations. Hoey argues that collocation is primarily a psychological concept (the feeling of semantic congruity between node and collocate), which is hardly essential with lexical bundles as speakers may not even be aware of their existence.

4.1.4 Lexical bundles and n-grams

The least complicated form of investigating recurrent non-idiomatic sequences is by means of a special computer program which is able to produce series of n-grams or word-clusters. The terms lexical bundle and n-gram can be used interchangeably, however, “n-gram“ is a rather a technical term which represents a sequence of variable characters that stands for a word or string of words in a corpus. The string can be either a fixed (continuous) sequence or a discontinuous one in a context of, at most, 11 words. Continuous n-grams are easy to identify while the retrieval of discontinuous strings requires special programs. The “n” carries the meaning of “any number of”, with the majority of bundles confined mainly to bi-, tri- or four-grams.

4.1.5 Structural taxonomy of lexical bundles

Biber, Conrad, Cortes (2004, 136) propose three main structural types of bundles. The first one contains verb phrase fragments. Prototypical structure in these bundles follows the pattern of a subject pronoun accompanied by a verb phrase (e.g. *this is one of*). The presence of a pronoun is not required and the structure can be introduced by a verb phrase (e.g. *is going to be*). The second major type includes verb phrase elements. In particular, verb phrase elements are again present, however, they mark the presence of dependent clause fragments (e.g. *if we consider the*). The third main structural type encompasses phrasal components with the presence of noun phrase components (e.g. *the beginning of the*).

LGSWE (1999) presents a detailed overview of structural types of the most typical lexical bundles occurring in academic prose and in conversation. Fourteen structural patterns of lexical bundles occur in the language of conversation; twelve structural types of lexical bundles are characteristic of academic English. Biber et al. (2004, 137) find that verb phrases form almost 90 per cent of lexical bundles (e.g. *I think that the*) in the language of conversation. On the other hand, 70 per cent of lexical bundles in academic prose contain noun phrase expressions (e.g. *the point of the; one of the main*).

4.1.5.1 Structural types of lexical bundles in conversation

LGSWE (1999, 1002-1012) provides a list of 14 structural types characteristic of the language of conversation (see Table 1, for abbreviations see the list at p.i). The corpus findings in LGSWE (1999, 996) provide evidence that the structural type with A personal pronoun and a lexical verb is the most frequent structural type in conversation; it is followed by the type An auxiliary with an active verb. In the majority of cases, most lexical bundles occurring in conversation are clausal rather than phrasal.

Table 1: Structural types of lexical bundles in conversation

Structural type	Description	Example
Type 1	A personal pronoun + a LVP	*I don' t know* *I didn't want to*
Type 2	A pronoun/noun + be	*it's got to be* *I thought I was*
Type 3	A VP + an active VP	*have a look at* *put them in the*
Type 4	Yes/no fragment	*can I have some* *do you know what*
Type 5	A *wh*-clause fragment	*what are you doing* *how do you know*
Type 6	A LV *to*-clause fragment	*to go to the* *would like to go*
Type 7	A lexical verb + *wh*-clause	*see what you mean* *know where it is*
Type 8	A verb + *that* clause	*said I don't know* *I don't think he*
Type 9	Adv clause fragment	*as long as you* *as soon as you* *if you want to*
Type 10	NP expressions	*the back of the* *the end of the*
Type 11	PP	*in the morning* *in the first place* *for the rest of*
Type 12	Quantifier expressions	*all the way round* *all of a sudden* *all over the place*
Type 13	Other expressions	*no no no no* *on and on and*
Type 14	Meaningless sound bundles	*la la la la* *mm mm mm mm*

4.1.5.2 Structural types of lexical bundles in academic prose

Biber’s taxonomy in LGSWE (1999, 996) provides 12 structural types which occur in academic prose (see Table 2). According to LGSWE, lexical bundles in academic prose tend to be phrasal and the type Prepositional phrases and Noun phrases with a post-modifier element are the most common structural types in academic prose. However, some structural types occur both in academic prose and conversation. These include Noun phrases, Prepositional phrases, Verbs followed by an adjective, Adverbial clause fragments.

Table 2: Structural types of lexical bundles in academic prose

Structural type	Description	Example
Type 1	A NP with an *of*-phrase	*the end of the* *the beginning of the*
Type 2	A NP with other postmodifier fragment	*that fact that is* *the extent to which* *the degree to which*
Type 3	A prep phrase with embedded *of*-phrase fragment	*in the course of* *in the development of* *as a consequence of*
Type 4	Other prep phrase fragments	*on the other hand* *at the same time* *in the next chapter*
Type 5	Anticipatory *it* + VP/Adj phrase	*it is possible to* *it is not clear* *it should be noted*
Type 6	Vpas + prep phrase fragment	*are shown in table* *can be found in* *is based on the*
Type 7	Copula *be* + NP/ADJ phrase	*is one of the* *is part of the* *be the result of*
Type 8	(VP) + *that*-clause fragment	*should be noted that* *does not mean that*
Type 9	(A verb + adj) + *to*-clause fragment	*are likely to be* *may be used to* *are more likely to*
Type 10	An adv clause fragment	*as shown in figure* *as we have seen* *as we shall see*
Type 11	A pronoun/NP + *be*	*this is not the* *there was no significant* *there is a number*
Type 12	Other expressions	*as well as the* *may or may not*

4.1.6 Functional classification of lexical bundles

Lexical bundles can be classified according to their functional role in texts. The typology proposed by Biber, Conrad and Cortes (2004) as well as a more recent typology developed by Hyland offer three main functional types of bundles with little difference in terminology: 1) Biber' s ”stance bundles” correspond to Hyland's (2008, 13) “participant-oriented bundles” *(*e.g. *it is essential to; it should be emphasized)*. The role of this functional type is to convey the speaker's opinion of probability or certainty of the expressed proposition or attitude towards something (epistemic *it is possible* x attitudinal bundles *I don*'*t want to*). The second category suggested by Biber is that of “discourse organizers” corresponding to Hyland's “text-oriented bundles”. The primary focus of such bundles is the organization of text and its meaning (e.g. *in contrast to the; it was found that).* It includes the subtypes of topic introduction/focus (I would like to touch upon) and topic elaboration/identification *(*e.g. *on the one hand...one the other hand)*. The last of Biber's types is that of “referential expressions” whose subtypes are a) imprecision bundles (e.g. *and the like)*; b) bundles specifying attributes (e.g. *the core of the problem)*; c) bundles referring to time, place or text (e.g. *the top of the; the beginning of the).* Biber's ”referential bundles” correspond to Hyland's “research-oriented bundles” (e.g. *at the top of; in the end),* Biber (2006).

The functional taxonomy will not be included in this research. For the sake of completeness, positional classification of lexical bundles can be mentioned in connection with functional taxonomy of lexical bundles. They are divided into text initial, medial and final.

4.2 Sample analysis - research into recurrent non-idiomatic word-combinations

The key concerns of the subsequent analyses of three- and four-word combinations will be outlined in the following sections. Issues related to the investigation of the recurring non-idiomatic combinations will be introduced and discussed.

4.2.1 General overview

This examination of the recurrent non-idiomatic three- and four-word combinations is based on three samples, namely two learner samples and one native speaker sample. The Czech sample includes 37 essays written by Czech secondary school students, another sample is that of non-Czech learners and contains 19 essays. The native sample consists of 22 reviews. All three samples contain approximately 9 400 words. Czech learners are students from a grammar school in Prague; they are sixteen and seventeen year old students with pre-intermediate, intermediate and upper-intermediate to FCE level. The other group of non-native speakers are students from various linguistic backgrounds; their essays were downloaded from the website http://bookrags.com/. The total of 22 book reviews written by native speakers, mainly professional review writers, were downloaded from the website available at http://happypublishing.com/ (for a detailed description see Section 3.3).

It is taken for granted that every analysis is meaningful only if it is compared with some previous findings and if the new findings can be related to the previous ones. In this investigation, all three- and four-word combinations with the minimum frequency of two occurrences in the samples will be identified using the application Collocate. Subsequently, the samples of Czech, non-Czech and native speakers will be checked against the data provided by the large new interactive phraseological database Phrases in English (PIE), which can be visited at http://pie.usna.com/ (see Section 3.2). In this investigation, the database will be used as a reliable source of n-grams (lexical bundles) to see whether, which and how many true lexical bundles are used in all three samples. The frequency threshold for this analysis is at least 10 occurrences per million words, i.e. the condition stated by Biber et al. (1999). Two terms are adopted throughout this examination: “word combinations“ and “lexical bundles“. The former is used for any three- and four-word combination retrieved by Collocate regardless of its frequencies in the PIE. The latter is used only for the word combinations which occur at least ten times in the PIE.

4.2.2 Issues related to the investigation of three- and four-word combinations

Several assumptions need to be considered before the investigation is launched. First of all, native speakers are expected to produce more creative language, non-native speakers' word-combinations are likely to be repetitive. In other words, it is expected that recurrent word-combinations produced by native speakers will be less frequent in contrast to non-native speakers. With the above proviso in mind, the focus of attention in this investigation is to compare the similarities and distinct features in the recurrent word-combinations, their frequency and diversity and the number of lexical bundles in the strict sense. Despite the expectations of greater diversity and creativity in the native sample, it is possible to assume that native speakers will create more lexical bundles in the strict sense in comparison with non-native speakers. However, the sample corpora do not contain pieces of academic English or transcripts of conversations - an extensive list of lexical bundles cannot be expected in them. Three groups of word-combinations emerging from the samples are expected after we check their frequency against the PIE: the first group will include lexical bundles in the strict sense, the second group will subsume the word-combinations which occur in the PIE but fall below the frequency threshold, i.e. ten occurrences per million words. The last group will be represented by word-combinations with zero occurrence in the PIE. It is thought that non-native word-combinations will mainly include sequences which have some matches in the PIE, however, not enough to call these word-combinations lexical bundles.

4.3 Study of four-word combinations in the Czech learner sample

The following sections will focus on the number of types and tokens and structural types of four-word combinations in the Czech sample. Subsequently, three groups of four-word combinations will be yielded from the search in the PIE. The differences among these groups will be accounted for.

4.3.1 Four-word combination frequencies in the Czech learner sample

For the retrieval of the four-word combinations, the program Collocate was used. The search found 127 four-word combination types with a minimum frequency of 2 occurrences in the Czech sample (see Appendix 2a). The results show that *I would like to* is by far the most frequent of these, it occurs 12 times. This sequence is obviously topic-bound, it occurs in such contexts where students recommend a book, a film, describe a story line. The following three four-word combination types *one of the most, would like to write, is one of the* occur 5 times. The sequences *one of the most* or *is one of the* are used in the contexts in which students write about *one of the most popular* or *famous authors* or *books.* Two four-word combination types show the frequency of four *(the Lord of the; about a book which).* Nineteen four-word combination types appear three times. Several of these owe its high frequency to the topic (*of the rings is; the Adventures of Huckleberry; girl who wants to; the story takes place; from cover to cover; like to write about; the main character is; is based on a; book Harry Potter and; in the middle of; my cup of tea; Adventures of Huckleberry Finn; the Da Vinci code; it does not matter; J R R Tolkien; my point of view).* The remaining word-combination types occur twice. The frequency of types is relatively stable; there is only one type occurring twelve times, very few word-combinations have five or four occurrences, the majority of word-combinations occur three or four times (see Table 3).

Table 3: Distribution of 4-word combination tokens and types in the Czech learner sample

Order	**Tokens per type**	**Types**	**Example**
1	12	1	*I would like to*
2	5	3	*one of the most*
3	4	2	*about a book which*
4	3	19	*the story takes place*
5	2	102	*the book is a*
Total	**296**	**127**	

4.3.2 Structural types of 4-word combinations in the Czech learner sample

The data obtained is classified structurally. Accordingly, the total of 127 four-word combination types can be arranged into 10 structural types (see Table 4). The taxonomy adopted for this structural analysis is that of Biber et al. (1999). The arbitrary limit for a self-standing structural group was set at 4 and the structures with fewer than four word-combinations are included in the type Others. The number in brackets refers to the number of word-combination types. The four-word combinations in bold letters imply that the combination does not occur in the PIE which used is as a control sample.

Type 1: A noun phrase with an of-fragment (7)

Seven four-word combinations in this grammatical pattern were identified. This structure consists of a noun phrase followed by a fragment of an *of*-phrase. The examples are as follows: *The Lord of the; the middle of the;* ***a movie out of****;* ***movie out of the****; the magic of the; the rest of the;* ***the cat out of****.* Some of them overlap: *a movie out of* as well as *movie out of the.* As noted above, the word-combinations in bold letters are not attested in the PIE.

Type 2: Noun phrases with post-modifier fragments or other noun phrases (25)

Even though LGSWE (1999) distinguishes between noun-phrases with an *of*-fragment and noun phrases with other post-modifier fragments (where the noun-phrases with a post-nominal clause fragment and a prepositional phrase fragment are distinguished), this structural type encompasses any type of a noun phrase excluding only those which fall into Type 1. The Czech sample includes the following word-combinations: *girl who wants to; a book which is;* ***book to anyone who****;* ***to anyone who likes****; it to anyone who;* ***book Harry Potter and****;* ***J RRR Tolkien****;* ***this book to anyone; The Lost symbol I****;* ***Da Vinci Code and; the Adventures of Huckleberry;*** *Lord of the Rings; my cup of tea;* ***Adventures of Huckleberry Finn; book Harry Potter and; The Da Vinci Code****; my point of view;* ***the first book Harry****;* ***The Battle of Agincourt; book by Dan Brown****; one of the most; rest of the film; point of view I; few years ago I;* ***Harry Potter and the****.*

Several phrases contain books titles, names of authors, and fiction characters. As will be discussed in detail below, a good number of them have no matches in the PIE database.

Type 3: Prepositional phrases with of-fragment/other prepositional phrases (14)

As opposed to Biber's classification, a single category of prepositional phrases was opted for. The prepositional phrases with an embedded *of*-phrase fragment are not treated separately.

Unlike prepositional phrases often found in academic style, the collection of examples identified in the Czech sample seems somewhat randomly gathered strings containing a preposition with some classic prepositional phrases (*in the middle of, from my point of* etc.) This structural type provides the following examples in the Czech sample: *in the middle of; from my point of; from cover to cover; in the first book; in a nutshell the;* ***by J RRR; for the best picture****;* ***on a real story****;* ***out of the Da Vinci; of the rings is; of Harry Potter and****;* ***about a book which****;* ***of the Da Vinci****;* ***during the WWII the****.* Half of the combinations in bold letters (not occurring in the PIE) contain the name of a character, author etc. It is clear that they have become prominent only because the sample is very small and topic-bound. Small wonder that these combinations are not attested in the PIE.

Type 4: (A noun/pronoun) passive verb + preposition (9)

Passive constructions are extremely common in the English language. A few structures using the passive voice were found in the Czech sample. All of them describe some aspects related to books or films (location, the name of the author or film director). This could be the reason why this structure occurs occasionally in the Czech sample. The word-combination types include: *is based on a; was written by a; it was written by;* ***filmed in New Zealand; was filmed in New****; based on a real;* ***book was written by; novel was written by; it was directed by****.* Several of these word-combinations overlap.

Type 5: Copula *be* + a noun phrase/adjectival phrase (7)

Two groups are included in this type. The former is created from strings where the copula is followed by a noun phrase, the latter type contains those where the copula is followed by an adjective: *is one of the;* ***is very funny and****; is also full of; was one of the;* ***am not a bookworm****;* ***is my cup of****; is the fact that.*

Type 6: A personal pronoun/noun + lexical verb phrase (+ VP fragment of a complement) (17)

This grammatical pattern of structures is extremely common in the language of conversation. According to LGSWE (1999), these phrases often contain a fragment of the following phrase. Unlike in LGSWE, the type wherein a noun occurs was included. Most of the phrases express a learner's recommendation or a personal attitude. The examples are as follows: *I would like to; I would recommend it; I am not going ; I read the book;* ***I totally forgot the****;* ***I recently read the****; I can say I; I do not have;* ***I read two books****; I must admit that;* ***the story takes place; story takes place in;*** *now I do not; but I think that;* ***but now I do****; and I have read; and I would recommend.* As will be seen later, the native sample is almost devoid of the four-word combinations with a personal pronoun, especially the first person singular.

Type 7: A noun phrase/pronoun + be (16)

Several of these strings refer to the description of a book aspect (the plot, characters); sometimes they refer to a recommendation, an expression of likeness or indifference. The following examples were found: *this book is very; there is a large; the main character is; I* *am not a; the book is a; story is about a; the book is full; the story is about; it is the most;* ***the plot is very****; it is about a; book is full of;* ***book is very readable****; I am not interested;* ***the plot is simple****; it is crucial to.*

Type 8: A verb phrase with an active verb (12)

Although LGSWE points out that many word-combinations of this structural type are often idiomatic, the occurrences found in the Czech sample do not confirm this. The majority of these examples are simple verb phrases and, as will be seen later, many of them have no matches in the PIE database. The examples are as follows: ***write about a book; read a lot but****;* *falls in love with;* ***read the Da Vinci; has a surprising end****; takes place in the;* ***realizes there is*** *a; recommend it to anyone;* ***recommend this book to; strongly recommend this film; would recommend this book****; would recommend it to.*

Type 9: Lexical bundles with *to*-infinitives (7)

This structural type consists of two groups. A verb phrase followed by a *to*-clause or phrases which begin with *to* are found. This type very often expresses intention, desire to describe something, recommend something or write about something (e.g. *would like to write; like to write about; to write about a; to go to the;* ***to recommend this book****; to sum it up;* ***would like to explain)***. Even though word-combinations with *recommend* such as *I would recommend* or *would like to etc.* proliferate in the PIE, the word-combinations *to recommend this book* and *would like to explain* are not attested in the PIE. Czech students produced nine word-combinations with *recommend* in which *recommend* is followed by *a book* or *film.*

Type 10: Others (13)

Structures which occur less than four times were automatically placed in this last type. This type thus includes various grammatical structures. Again the combinations include structures both attested and not attested in the PIE *(*e.g. *it does not matter;* ***that a book is****; that he is a; who is one of; because it is very;* ***before the Da Vinci****;* ***a lot but now****; but it does not;* ***lot but now I****; not only because of;* ***this book was written****;* ***book which is called;*** *am not interested in).*

Table 4 presents the structural types obtained from the Czech sample. The first and second column give the list of the structural types, the third column shows the number of word-combination types appearing in this structural type, the fourth column contains examples of such structural types. The structural types are arranged in descending order.

Table 4: A survey of the structural types of 4-word combinations obtained from the Czech learner sample

Structural type	Description	Types	Example
Type 2	A NP s with post-modifier fragments or other NP	25	*adventures of Huckleberry Finn*
Type 6	A personal pronoun/noun + LVP	17	*I would like to*
Type 7	A NP/pronoun + *be*	16	*the plot is very*
Type 3	PP with *of*-fragments or other PP	14	*in the middle of*
Type 10	Others	13	*it does not matter*
Type 8	A VP with a Vact	12	*falls in love with*
Type 4	(A noun/pronoun) Vpas + prep	9	*was written by a*
Type 5	Copula *be* + a NP /adj phrase	7	*is also full of*
Type 1	A NP with an *of*-fragment	7	*the middle of the*
Type 9	LB with *to*-fragments	7	*to recommend his book*
Total:		**127**	

4.3.3 Czech learner four-word combinations and the PIE

After the structural type categorization, the next step was to check the frequency of the word-combinations against the phraseological database (see Appendix 2a). For this investigation, a less conservative approach for the frequency cut-off, i.e. the one advocated by Biber et al. (1999), was adopted. According to this approach, a word-combination has to occur at least 10 times per million words in order that it could be qualified as a lexical bundle. The reason why we opted for this cut-off limit was that in preliminary probes the results yielded made sense. The following results were obtained:

Overall, only 9 four-word combinations (7.1 per cent) out of 127 in the Czech sample occur more than 10 times per million words and qualify as true lexical bundles: *I would like to, one of the most, is one of the, in the middle of the, was one of the, to go to the, the rest of the, the middle of the, rest of the film.* The sample contains 58 word combination types (45.6 per cent) which do occur in the PIE but fall below the minimum frequency of ten occurrences. Next, 60 word-combination types (47.3 per cent) have no matches in the PIE (see Appendix

2a).

To comment briefly, *I would like to* is the most frequent in the Czech sample (12 occurrences). As mentioned previously, the high frequency of *I would like to* in the Czech sample occurs in the contexts where Czech learners recommend a book, describe a story line, write about a book. Its high frequency is explicable due to the fact that students write a review. O*ne of the most* and *would like to write* occur five times in the Czech sample. Whereas *one of the most* qualifies as a lexical bundle, *would like to write* falls well below the frequency cut-off. Two frequently occurring strings in the Czech sample occur very scarcely in the PIE: *the lord of the*, which is obviously topic-bound and *about a book which*. Out of 19 four-word combinations which occur three times in the Czech sample only 10 have their matches in the PIE and only *in the middle of* qualifies as a lexical bundle.

A detailed comparison of the three groups of four-word combinations reveals that 60 four-word combination types with zero occurrence in the PIE represent almost one third of all four-word combinations. These contain names of authors, book characters or book titles (*book Harry Potter and, J R R Tolkien, The Adventures of Huckleberry, Da Vinci Code and, book by Dan Brown)* and are clearly not a part of native speakers' language repertoire. The rest of the four-word combinations within this group mainly relate to books or films in general (*this book to anyone, novel was written by, it was directed by* etc.). The group totalling 58 four-word combination types attested in the PIE falls below the required limit of at least 10 occurrences per million words. These are sequences created on an "ad hoc" basis and include the following examples *I do not have, this book is very, and I have read, the story is about, book is full of.* As opposed to the group mentioned with zero PIE occurrences, a good number of these sequences concern a learner's opinion of a book, a novel or an attempt to describe the plot of the story, some of them are used in order to give a recommendation *(and I would recommend, I must admit that, would like to write).* The native sample is almost devoid of these combinations.

4.4 Study of four-word combinations in the non-Czech learner sample

The following section provides the list of types and tokens followed by a structural analysis of four-word combination types in the non-Czech learner sample. The search in the PIE yields three groups of four-word strings found in the non-Czech learner sample.

4.4.1 Four-word combination frequencies in the non-Czech learner sample

Search in the sample of non-Czech learners retrieved 119 four-word combination types by means of the software application Collocate (see Appendix 2b). The word-combination *please, please, please, please* comes first (7), it is followed by *the end of the,* which occurs 6 times in the non-Czech learner sample. The other six examples of four-word combination types occur only 3 times: *to the old man, Da Vinci code is, I think that the, to go through with, The Priori of Sion, were discriminated for their.* The word-combination *Da Vinci code is* and *The Priori of Sion* are clearly local repetitions due to the selected topic. The rest of the four-word combinations occur twice. The sample of non-Czech learners confirms that the frequency of types is relatively stable. Table 5 shows the distribution of the four-word combination types in the non-Czech sample.

Table 5: Distribution of 4-word combination tokens and types in the non-Czech learner sample

Order	Tokens per type	Types	Example
1	7	1	*please, please, please, please*
2	6	1	*the end of the*
3	3	6	*I think that the*
4	2	111	*one of the greatest*
Total	**253**	**119**	

4.4.2 Structural types of four-word combinations in the non-Czech learner sample

In comparison to the Czech learner structural types, there are 8 main structural types in the sample of non-Czech learners. The results do not produce a group of the passive voice type or the type consisting of copula *be* in combination with a noun phrase or an adjectival phrase. A great number of instances are noun phrases (37 examples) and prepositional phrases (28). The especially higher number of noun-phrases is due to the selected topic since the majority of these noun phrases contain the name of a character and are not lexical bundles in the strict sense.

The four-word combinations which do not qualify for any specific structural pattern, since they do not occur at least four times, are placed into the structural type Type 8 (Others). The four-word combinations in bold letters have no matches in the PIE. The number in brackets refers to the number of four-word combination types (see Table 6).

Type 1: A noun phrase with an *of*-fragment (5)

The search yielded the following extended word-combinations: *the end of the; the faces of the; the death of his; the story of his; the names of his.*

Type 2: Other noun phrases (37)

This type is most prevalent in the non-Czech learner sample. Some of the phrases include postmodification and the rest are formed by miscellaneous noun phrases. As noted above, the word-combinations in bold letters have no matches in the PIE.

The examples include: ***another key incident that; such instances as the***; *one of the greatest;* ***The Priori of Sion; faces of the death; death of his love;*** *end of the story; end of the novel;, masque of the red;* ***his philosophy about slavery***; ***murder of Jacques Saunier***; *the main character in; main character in the; character in the story;* ***The Adventures of Huckleberry; Adventures of Huckleberry Finn***; ***the tell-tale heart Poe***; ***parts I and part II; one final march down;*** *change in his attitude;* ***his attitude towards Jim; constant fight with life; Henry IV Part I; philosophy about slavery as***; ***his love Sybil Vane***; ***the old man because***; ***the Holy Grail the***; *a change in his;* ***attitude towards Jim by; property by changing his***; ***the Mississippi river together***; ***the world the story; world the story of; Jim and by reevaluating; Jim by beginning to; slavery as they go; 4 Parts 1 and****).*

Type 3: Prepositional phrases with *of* and other prepositional phrases (28)

Similar to the Czech sample, prepositional phrases from the non-Czech learners form a small group of fixed propositional phrases, describing either a locative or temporal aspect: *near the end of; at the battle of;* ***in the end Frodo***; *in the story is;* ***in the novel the***; *at the same time.* The rest encapsulate miscellaneous phrases: *according to the book;* ***for anyone but himself***; *in his attitude towards;* ***with the simple operation***; ***towards Jim and by; by beginning to view; by reevaluating his philosophy;*** *to the old man;* ***of his love Sybil***; *of the story the;* ***of the novel Tom***; *of the red death;* ***through with the simple***; ***to the book the; of the priori of;***

with the world the*;* ***to the caf 65 533; about slavery as they; down the Mississippi river; towards Jim by beginning;*** *of the story is; of the book are.*

Type 4: Personal pronoun or noun + lexical verb phrase (7)

Non-Czech learner instances within this structural type are more or less concerned with a pure description of the event in the story (*I think that the; he had ever had;* ***Tom doesn't like; Huck reveals a change****; he came back from;* **this story takes place; they go down the)***.* As opposed to the Czech sample, non-Czech learners almost avoid using word-combinations with the first person singular pronoun.

Type 5: A noun/pronoun phrase + be (5)

There are very few examples compared to the Czech group. However, most of the non-Czech examples give the impression of very simple observations, with the possible intention to fulfil the limit of obligatory words or, if that is not the case, most of them describe the plot/book. The examples are as follows: ***Da Vinci Code is; the American dream is****;* ***old man who is****; the story is of;* ***E. A. Poe was****.*

Type 6: A Verb phrase with an active verb (9)

Non-Czech learners produced the following word-combinations within this type: *have to deal with;* ***keep himself from being*** *;* ***go through with the****;* ***reveals a change in****; go through with it; came back from the;* ***go down the Mississippi; share with the world****;* ***distinguished himself at the.*** A mere glance at the list reveals that the sequences mostly refer to story lines.

Type 7: Lexical bundles with a to-clause fragment (10)

The majority of examples represent strings which begin with a *to*-clause. The rest are verb phrases followed by a *to*-clause: *to go through with;* ***relate to the old****; can relate to the; to share with the; to keep himself from;* ***decide to have the****;* ***to be happy afterwards****; to take responsibility for; to carry out the;* ***to face his destiny.***

Type 8: Others (18)

Diverse structural types were grouped under this heading. Some of the structures follow the pattern of adverbial clause fragments, two phrases are in the passive, a few are examples of non-finite constructions. In each case, there is only a small number of them. The examples include: *is set in a;* ***were discriminated for their****;* *was one of the; was the death of ;that it was a;* ***when he is drunk****;* *when he came back; as they go down;* ***please, please, please, please****;* ***and by reevaluating his****;* ***himself from being punished****;* ***reevaluating his philosophy about****;* ***himself at the battle****;* ***seeing him as a; is a black eye; as a person instead****;* *while at the same time;* ***back from the play****).*

Similarly, as was evidenced in the Czech learner sample, type 2 with A Noun phrase predominates. Another aspect in which both non-native samples resemble one another is the greater number of structures with no matches in the PIE.

Table 6 presents the structural types of word-combinations found in the non-Czech sample. Column 2 lists the structural types, column 3 shows the number of word-combination types, the fourth column provides an example of the particular structural type. The structural types are arranged in descending order.

Table 6: A survey of the structural types of 4-word combinations obtained from the non-Czech learner sample

Structural type	Description	Types	Example
Type 2	Other NP	37	*death of his love*
Type 3	PP with *of* and other PP	28	*at the battle of*
Type 8	Others	18	*please, please, please, please*
Type 7	LB with a *to*-clause fragment	10	*to take responsibility for*
Type 6	A VP with a Vact	9	*have to deal with*
Type 4	A personal pronoun or noun + LVP	7	*Huck reveals a change*
Type 5	A NP/pronoun phrase + *be*	5	*the American dream is*
Type 1	A NP with an *of*-fragment	5	*the end of the*
Total		**119**	

4.4.3 Non-Czech learner four-word combinations and the PIE

After sorting out the four-word combinations into corresponding structural types, the next step was to obtain the data from the PIE. The PIE is used as a control sample to find out whether, which and how many word-combinations are true lexical bundles. The criterion of Biber et al. (1999) is adopted, i.e. the sequence must occur at least 10 times per million words to be considered a lexical bundle in the strict sense. The results obtained are as follows: only 3 of the 119 four-word combination types (2.5 per cent) occur more than 10 times per million words (*the end of the,at the same time, was one of the*); 42 four-word combination types (35.5 per cent) occur in the PIE but less than 10 times per million words; the last 74 four-word combination types (62 per cent) have no matches in the PIE (see Appendix 2b).

Close observation shows that the group of four-word combinations with zero occurrence in the PIE forms approximately one third of the four-word combinations in the non-Czech sample. These word-combinations mainly contain book titles, names of authors, film characters or a geographical location such as *down the Mississippi river, his love Sibyl Vane, attitude towards Jim, of the novel Tom, in the end Frodo.* These sequences cannot qualify as true lexical bundles, they do no form a standard part of a native speaker's repertoire. The four-word combinations which fall below the required limit of 10 occurrences per million words are represented by the sequences localized to a specific context, such as the story line, for instance (*to the old man, at the battle of of, was the death of, is set in).*

4.5 Study of four-word combinations in the native speaker sample

The following sections provide the list of types and tokens followed by a structural analysis of four-word combination types in the native speaker sample. The search in the PIE yields three groups of four-word strings found in the native speaker sample.

4.5.1 Four word-combination frequencies in the native speaker sample

The native speaker sample shows one striking difference in comparison with both non-native learner samples. The number of four-word combinations is remarkably lower – only 54 four-word combination types with a minimum frequency of 2 occurrences (see Appendix 2c). This result contributes to the confirmation of the initial assumptions: lower repetitiveness on the part of the native speakers is evident. Furthermore, Table 7 shows the distribution of four-word combination types and tokens. The situation is somewhat similar to the non-Czech learner sample in that almost all the word combinations occur twice. The most frequent word-combinations, i.e. *fear of speaking in, of speaking in public, father of the rain* occurring four times in the native speaker sample are not attested in the PIE. These sequences relate to the book titles and appear only in one or two native speakers' reviews. They can be regarded only as local repetitions which do not commonly occur in the language. The frequency of types is relatively stable.

Table 7: Distribution of 4-word combination tokens and types in the native speaker sample

Order	Tokens per type	Types	Example
1	5	2	*fear of speaking in*
2	4	2	*history of the world*
3	3	2	*is the one that*
4	2	48	*the rest of the*
Total	**120**	**54**	

4.5.2 Structural types of four-word combinations in the native speaker sample

Since only 54 four-word combination types were obtained from the native speaker sample, the classification into structural types turned out less complicated (see Table 8). The analysis yielded 7 structural types of four-word combinations. Due to the relatively low incidence of word-combinations, a few changes were made: the word combinations which exhibited less than two occurrences were placed into the type Others. Again, the word-combinations with zero occurrence in the PIE are in bold letters. The number in brackets represents the number of four-word combination types.

Type 1: A noun phrase with an *of*- fragment (3)

Only 3 four-word combination types found in the native speaker sample fall into this structural type: *the murder of the, the rest of the, a history of the.*

Type 2: A noun phrase with post-modifier fragments or other noun-phrases (12)

In comparison with the previous type, this type includes diverse examples with a noun phrase accompanied by a post-modifier element or another noun phrase. These include the following 12 four-word combination types: ***Father of the rain*, *fear of speaking in, a fear of speaking, the coincidences of our, coincidences in which, his best friend's dad, Rich dead poor dad, a***

work of fiction, the rules of engagement, the Song of Kahunsha, such a good book, *history of the world.* This type contains most examples of all the structural types in native speaker sample. Most of these occurrences refer to the titles of books. Only one example from this group was found in the PIE (those in bold letters have no matches in the PIE).

Type 3: Prepositional phrase with of-fragment/other prepositional phrase (10)

The type of prepositional phrases appears to be comparably rich in examples. This type comprises 10 four-word sequences (***of speaking in public***; *in the form of; in search of a;* ***of her mother and***;*in search of the; with that in mind; from the view point of ;* ***for any writer there***; *on all aspects of;* ***from foster home to***).

Type 4: (A noun/personal pronoun) + a passive verb + (a preposition) (4)

Even though the passive voice is used relatively frequently in English, especially in certain registers, only 4 four word-combination types of this structural type of lexical bundles were marked in the native speaker sample. Two of them overlap - the latter is part of the longer sequence *it should be required reading for* (***girls raised by wolves; should be required reading; be required reading for***; ***the story is told***.

Type 5: A personal pronoun/noun + lexical verb phrase (7)

This type includes mainly word-combinations where a personal pronoun occurs. Some of them indicate the author's subjective view point. The sequences are as follows: ***any idiot can argue, and I don't think***, ***the rich invest in***, ***I have told so***, ***you'll feel more***, ***you read this book***, ***those we love are.***

Type 6: A noun phrase/pronoun + be (10)

The occurrences which fall into this type are comparatively frequent. A lot of them refer to the book in question, especially the title of the book (***kiss is the one***; *that there is no;* ***poor dad is the; first kiss is the;*** *the book is a;* ***dad is the story***; ***any writer there ar***e; ***and poor dad is; The pocket muse*** is; ***other person is wrong)***.

Type 7: Others (8)

The type Others exhibits miscellaneous structures. Even though some of these would fit neatly into Biber's structural types, the number of occurrences matching a particular type would be extremely low (one or two examples). The following occurrences were identified: *is the one that;* ***when you're not there; Who Moved my Cheese; Picturing your Audience Naked; Stop Picturing your Audience****;* ***heard of this book****; to find out what;* ***have told so many).***

Table 8 provides the list of structural types obtained from the native speaker sample. Column 2 presents the list of the structural types, column 3 shows the number of the structural types, column 4 provides an example of the corresponding structural type. The word-combinations are arranged in descending order.

Table 8: A survey of the structural types of 4-word combinations obtained from the native speaker sample

Structural type	**Description**	**Types**	**Example**
Type 2	A NP with a post-modifier fragment/other NP	12	*father of the rain*
Type 6	A NP/pron + *be*	10	*kiss is the one*
Type 3	PP with *of*-fragment/ other NP	10	*in the form of*
Type 7	Others	8	*picturing your audience naked*
Type 5	A personal pron/noun + a LVP	7	*I have told so*
Type 4	(A noun/pron) Vpas + prep	4	*should be required reading*
Type 1	A NP with an *of*- fragment	3	*the murder of the*
Total		**54**	

4.5.3 Native speaker four-word combinations and the PIE

The following data was yielded in the PIE. Only 3 four-word combination types (5.6 per cent) from the native speaker sample (*that there is no; in the form of; the rest of the)* exceed the number of more than 10 occurrences per million words. Furthermore, 40 four-word combination types (74 per cent) produced by native speakers have no matches in the PIE, 11 four-word combination types (20.4 per cent) occur scarcely in the PIE (see Appendix 2c).

Almost one half of the four-word combinations in the group with zero occurrence in the PIE include book titles, names of authors or book characters (*The Song of Kahunsha, Who moved my cheese, Rich dad poor dad, Father of the Rain).* The group with the four-word combinations with low frequencies in the PIE is represented mainly by noun phrases or prepositional phrases (*in search of, the murder of the, the book is a*). Most of these word-combinations refer to the plot of the story but do not occur in the language frequently enough to be considered lexical bundles.

4.6 Comparison of four-word combinations: Czech, non-Czech learner and native speaker sample

A closer look at Table 9 reveals that lexical bundles (LB) in the strict sense occur rarely in all three samples: only 9 four-word lexical bundle types were found in the Czech sample, 3 lexical bundles were identified both in the non-Czech and native speaker sample. Approximately half of the four-word combinations in the Czech sample are attested in the PIE, however their frequency is less than ten occurrences per million words. The other learner group produced approximately one third of such four-word combinations; the native speaker sample contains approximately 20 percent of such four-word combinations. The last group of four-word combinations not found in the PIE is not small, especially in the non-Czech learner sample (62.3 per cent) and native sample (74 per cent). The second column shows lexical bundles in the strict sense. The third column lists types which are attested in PIE but are not true lexical bundles. The fourth column contains word-combination (WdC) types which are not attested in PIE. The last column provides information about the total of four-word combination types in all three samples. For further discussion see the paragraphs below Table 9.

Table 9: A comparison of 4-word combination types from all three samples and the PIE

Sample	**LB (10+)**	**%**	**Matches in the PIE (10-)**	**%**	**No matches in the PIE**	**%**	**WdC types**
Czech learner	9	7.1	58	45.6	60	47.3	**127**
Non-Czech learner	3	2.4	42	35.3	74	62.3	**119**
Native speaker	3	5.6	11	20.4.	40	74	**54**

The primary concern of the previous analyses focused on the number of types and tokens of four-word combinations. The investigation also aimed to find out whether, which and how many lexical bundles in the strict sense occur in the samples. The adopted approach was that of Biber et al. (1999), i.e. a sequence has to occur at least 10 times per million words to be considered a lexical bundle. The PIE was used as a reliable source of lexical bundles.

The results indicate that four-word lexical bundles are extremely rare in essay writing or reviews. The low number of lexical bundles seems plausible since the samples represent neither academic prose nor the language of conversation. Namely, Czech learners produced 9 lexical bundles in the strict sense, 3 lexical bundles were identified in the non-Czech learner and the native speaker sample. The outcome of these investigations confirms the initial assumption that the native speaker sample will contain fewer word-combinations that non-native samples. Indeed, only 54 four-word combination types were found in the native speaker sample whereas both non-native learner samples contain twice as many four-word combination types: Czech learner sample contains 127 types, non-Czech 119 types. The data points to obvious conclusions: greater repetitiveness in the non-native learner samples on the one hand and greater diversity in the native speaker sample on the other one.

Some distinct features are to be commented upon in connection with the structural taxonomy. Both non-native learner samples seemingly provide more structural types than the sample of native speakers. The Czech learner sample consists of 10 structural types, the non-Czech learner sample encompasses 8 structural types, the native speaker sample contains 7 structural types. However, given that the number of both types and tokens is less than half in the native speaker sample, it emerges that the 7 structural types produced by the native speakers could suggest greater structural richness in comparison with both of the non-native speaker groups. Native speakers could theoretically produce twice as many structural types as non-native speakers provided the number of word-combinations was higher. On the other hand, this assumption is purely hypothetical and a greater number of word-combinations in the native speaker sample would not guarantee the increase in the number of structural types.

Another observation from the structural analysis suggests that the most frequent word-combinations in all three samples are represented by the structural types of Noun phrases and Prepositional phrases. This is hardly surprising since a large number of noun phrases are topic-bound in all three samples. The titles of books or names of characters are used repeatedly. The in-depth analysis shows that other subtle differences exist between the individual samples. In spite of the frequent use of the passive voice in English, there are very few four-word combinations in the native speaker sample. The Czech sample, on the other hand, provides several four-word combinations in the passive voice. All the word-combination types using the passive voice in the Czech sample are apparently topic-bound (*is based on a, was written by, filmed in New Zealand, it was directed by)*. The reason why several structures with the passive voice occur in the Czech sample could be explained by the fact that passive structures often appear in textbook sections devoted to review writing. The missing types in all three samples are for instance The adverbial clause fragment type or the type with Anticipatory *it*.

The results obtained from the PIE also are worth commenting upon. As mentioned above, the number of true four-word lexical bundles used in all three samples was extremely low. Czech learners produced most true lexical bundles (7.1 per cent) in comparison with the other learner group (2.5 per cent) and the native-speaker group (5.6 per cent). However, the count is so small that it is not possible to make broad generalizations about the highest number of lexical bundles in the Czech sample. The data obtained from the two remaining groups of word-combinations, namely the four-word combinations which are not attested in the PIE and the word-combinations with a low frequency in the PIE, needs a few comments. Approximately one third of all four-word combination types out of the former group include sequences with names of authors, book titles in both learner samples (*The Adventures of Huckleberry, book by Dan Brown, book Harry Potter and, J R R Tolkien*). In the native sample, the number of such sequences is almost half. These sequences are not lexical bundles in the strict sense, they reoccur only due to the selected topic. The latter group with less than 10 occurrences per million words are sequences such as *the story is about, the book is full, of the book are, the murder of the, to find out what*. Again, these are obviously topic-bound sentences related to the semantic field of reading.

Table 10 provides a survey of 4-word combinations, their structural types and the percentage in which they occur in all three samples. The second column relates to the Czech learner sample (CZL), the third column refers to the non-Czech learner sample (NCZL), the fourth column provides the data obtained from the native speaker sample (NS).

Table 10: A survey of the 4-word combination structural types in the Czech, non-Czech and native speaker sample

Structural type	**CZL** **types**	**%**	**NCZL** **types**	**%**	**NS** **types**	**%**
Type 1: A NP with an *of*-fragment	7	5.4	5	4.2	3	5.5
Type 2: A NP with a postmodifying fragment/other NP	25	19.6	37	31.1	12	22.2
Type 3:PP	14	11.1	28	23.5	10	18.5
Type 4: (A noun/ pron) Vpas+ (prep)	9	7.1	-	-	4	7.4
Type 5: Copula *be* + NP/ adj phrase	8	6.3	-	-	-	-
Type 6: A personal pron/N + LVP	17	13.4	7	5.9	7	13.0
Type 7:A noun/pron + *be*	16	12.6	5	4.2	10	18.5
Type 8: A VP with an Vact	12	9.5	9	7.6	-	-
Type 9: LB with *to*-fragment	7	5.5	10	8.4	-	-
Type 10: Others	12	9.5	18	15.1	8	14.9
Total	**127**	**100.00**	**119**	**100.00**	**54**	**100.00**

4.7 Study of three-word combinations in the Czech learner sample

The same procedure is used with the three-word combinations: the identification of the three-word combinations using Collocate, their type and token ratio, organizing the word-combinations into structural types, identifying true lexical bundles in the PIE.

4.7.1 Three-word combination frequencies in the Czech learner sample

The search for three-word combinations with a minimum frequency of 2 occurrences retrieved 370 strings from the Czech learner sample (see Appendix 2d). The results show that the number of three-word combinations is almost double the four-word combinations. This result confirms the data obtained from LGSWE (1999) in that three-word combinations occur more frequently than four-word combinations in the language. The top positions with the frequency of 12 occurrences are occupied by 3 three-word combinations in the Czech sample *(would like to, I would like, one of the);* 1 three-word combination occurs 10 times *(the main character)*; 4 three-word combinations occur 9 times *(the book is, it is a, this book is a, a lot of).* Apparently, the frequency of types is relatively unstable – it ranges from 12 to 2. Apart from the top positions, however, a great number of types occur four or three times. The majority of three-word combinations occur two times (see Table 11).

Table 11: Distribution of 3-word combination tokens and types in the Czech learner sample

Order	Tokens per type	Types	Example
1	12	3	*I would like*
2	10	1	*the main character*
3	9	4	*a lot of*
4	8	1	*the plot is*
5	7	1	*the story is*
6	6	6	*is one of*
7	5	11	*I am not*
8	4	24	*this book was*
9	3	78	*this is the*
10	2	241	*I have to*
Total	**1000**	**370**	

4.7.2 Czech learner three-word combinations and the PIE

After making the frequency list, the next step was to check the frequency of occurrences against the PIE to see whether any of the three-word combinations are true lexical bundles. In comparison with the four-word bundles, the results differ. Out of the 370 three-word combination types, 84 three-word combination types (22.7 per cent) occur more than 10 times per million words; 219 three-word combination types (59.2 per cent) occur in the PIE less than 10 times per million words; 67 three-word combination types (18.1 per cent) do not have matches in the PIE (18.1 per cent); see Appendix 2d.

A closer look at the three-word combinations with zero occurrence in the PIE reveals that one third of these sequences contain names of authors, characters or book titles (*The Lost Symbol, Harry Potter and, the Da Vinci, of Huckleberry Finn)*. The three-word combinations with low frequencies in the PIE are represented mainly by phrases which occasionally occur in the language but not frequently enough to qualify as lexical bundles. They mostly do not contain names of characters or book titles. Still, they are topic-bound and relate to the semantic field of reading (*to write about, I would recommend, the main character, the first book, this film is).*

4.7.3 Structural types of three-word combinations in the Czech learner sample

The three-word combination types were sorted manually. The analysis of the Czech sample produced 11 structural types and basically the same structural types as in four-word combinations were obtained. The most numerous groups contain structures such as noun phrases, prepositional phrases, a noun or pronoun followed by a verb phrase.

Table 12 presents the findings obtained from the structural analysis and provides the list of the three-word combination structural types with one other structural type Adverbial clause fragments, not used in the four-word combinations.

Table 12: A survey of the structural types of 3-word combinations obtained from the Czech learner sample

Structural type	Description	Types	%	Example
Type 4	Other NPs	59	16.0	*cover to cover*
Type 11	Others	54	14.6	*and it also*
Type 5	PP expressions	49	13.2	*in the end*
Type 8	A noun/pron + VP	48	13.0	*I would like*
Type 1	A noun/pron + *be*	46	12.4	*the book is*
Type 2	VP + active VP	34	9.2	*doesn't like her*
Type 3	VP + active VP	28	7.6	*the lord of*
Type 6	be + NP/adj phrase	17	4.6	*is one of , is based on*
Type 10	Adv clause fragments	12	3.2	*if you are*
Type 9	*To*- inf. cl. fragment	12	3.2	*to write about*
Type 7	(A pron/N) + *be* +Vpas	11	3.0	*book was written*
Total		**370**	**100.00**	

4.8 Study of three-word combinations in the non-Czech learner sample

The following section focuses on the number of types and tokens in the non-Czech sample. Next, the structural type analysis is carried out and the PIE is used as a reliable source of true lexical bundles.

4.8.1 Three-word combination frequencies in the non-Czech learner sample

The list of the most frequent three-word combinations retrieved from the non-Czech learner sample provides the following results: the overall number of three-word combination types is 320 (see Appendix 2e). The top position is occupied by *please, please, please* (8)*;* there are two word-combination types which occur 7 times (*the old man, I think that);* 4 three-word combination types which occur 6 times (*the end of, one of the, end of the, in the story*); 7 three-word combination types which occur five times (see Table 13). In comparison with the Czech sample, the frequency of types in the non-Czech sample is relatively stable, it ranges from 8 to 2, however, the types with 8 or 7 occurrences appear only once or twice. Few three-word combinations occur 6, 5 or 4 times and the majority of the word-combinations occur two times. Table 13 shows the distribution of three-word combinations types. The occurrences are arranged in descending order.

Table 13: Distribution of 3-word combination tokens and types in the non-Czech learner sample

Order	**Tokens per type**	**Types**	**Example**
1	8	1	*please, please, please*
2	7	2	*I think that*
3	6	4	*one of the*
4	5	7	*the fact that*
5	4	10	*in the end*
6	3	34	*it was a*
7	2	244	*it is not*
Total	**711**	**320**	

4.8.2 Non-Czech learner three-word combinations and the PIE

The frequencies obtained from the PIE are as follows (see Appendix 2e): 55 three-word combination types (17.2 per cent) occur more than ten times per million words; 165 three-word combination types (51.5 per cent) occur in the PIE less than 10 times per million words, 100 word-combination types (31.3 per cent) created by non-Czech learners were not attested in the PIE. Almost one half of the three-word combinations in the group with zero occurrence in the PIE contain examples with names of authors, characters, book titles or geographical locations (*Da Vinci Code, Finn is a, murder of Jacques, the Mississippi river).* The remaining examples are the three-word combinations which relate to the story line or the semantic field of books and reading (*story the narrator, have the abortion, river and mountains* etc). The group with few occurrences in the PIE is formed by the three-word combinations which again relate to the semantic field of reading. These three-word combinations are localized to particular contexts and occur in the language only occasionally.

4.8.3 Structural types of three-word combinations in the non-Czech learner sample

The non-Czech learners produced 10 structural types which slightly differed from the structural types produced by the Czech learners. Since the frequency cut-off was set at minimally 4 occurrences for a self-standing group, the structural type with the passive voice was not included (only 3 three-word combination types were found in the non-Czech learner sample). By contrast, non-Czech learners produced a structural type of Lexical verb phrase + infinitive*,* which was rare in the Czech sample. The structural types produced by the analysis are listed in Table 14. The structural types Other noun phrases and Prepositional phrases form the biggest groups. A relatively big groups are that of Other expressions and Verb phrases. The structural taxonomy does not include the types Adverbial fragments or Anticipatory *it* (see Table 14).

Table 14: A survey of the structural types of 3-word combinations obtained from the non-Czech learner sample

Structural type	Description	Types	%	Example
Type 5	Other NP	95	29.7	*the main character*
Type 6	PP	49	15.3	*in the story*
Type 10	Others	42	13.1	*him as a*
Type 3	(A VP) + Vact	35	10.9	*go through with*
Type 2	A pron/N + *be*	28	8.7	*this book was*
Type 4	A NP with an *of*-fragment	21	6.6	*the end of*
Type 8	A noun/pron + a LVP	19	6.0	*I think that*
Type 9	*To*- infinitive + clause fragment	15	4.7	*to go through*
Type 1	A LV + infinitive *to*	9	2.8	*has to face*
Type 7	*be* + a NP/adj phrase	7	2.2	*was one of*
Total		**320**	**100.00**	

4.9 Study of three-word combinations in the native speaker sample

The following section focuses on the number of types and tokens in the native sample. As the next step, the structural type analysis is carried out and the PIE is employed as a reliable source of true lexical bundles.

4.9.1 Three-word combination frequencies in the native speaker sample

The number of three-word combinations in the native speaker sample is much lower than in the non-native learner samples. Whereas the learner samples contain 370 and 320 three-word combination types, only 220 three-word combination types were identified in the native speaker sample (see Appendix 2f). The majority of three-word combination types in the native speaker sample occur twice. Table 15 provides evidence that the frequency of the three-word combination types is relatively stable (6,5,4) in the native speaker sample in comparison with the Czech sample which ranges from 12 to 2 occurrences.

Table 15: Distribution of 3-word combination tokens and types in the native speaker sample

Order	Tokens per type	Types	Example
1	6	1	*in your life*
2	5	7	*the story of*
3	4	11	*in search of*
4	3	25	*there is no*
5	2	176	*the rest of*
Total	**512**	**220**	

4.9.2 Three-word combination types in the native speaker sample and the PIE

In order to find out how many three-word combinations could be considered lexical bundles, the PIE was used as a reliable control sample. Again, the approach by Biber et al. (1999) was adopted for this analysis. That is to say, the frequency cut-off was set at minimally ten occurrences per million words. Appendix 2f provides evidence that the PIE identified approximately 42 lexical bundles types (19.1 per cent) in the strict sense. Another group is formed by 108 three-word combination types (49.1 per cent). These three-word combinations occur in the PIE less than 10 times per million words, they are not prevalent in the language. However, they do occur in particular contexts. Most of these three-word combinations in the native sample relate mainly to the story line or reading as such (*you read this, the reader and, the murder of, of her mother*). The last group with no matches in the PIE contains 70 three-word combination types (31.8 per cent). Almost in one third of the cases, these three-word combinations include the book titles, such as *The pocket muse, moved my cheese, Stop picturing your etc.* It is worth noting that the percentage of lexical bundles in the native speaker sample is basically the same as in both learner samples: approximately 20 per cent of lexical bundles in the strict sense occur in all three samples (see Table 17 in Section 4.10).

4.9.3 Structural types of three-word combinations in the native speaker sample

The total of 220 three-word combination types occurring in the native speaker sample were manually sorted and this sorting produced 10 structural types. Similar to the previous structural analysis in the Czech and non-Czech learner samples, the structural types of Noun phrases and Prepositional phrases are the most numerous ones. With the exception of the structural type Adverbial clause fragment (which is present only in the Czech learner sample), the same structural types were produced by the native speakers (see Table 16). The types are arranged in descending order.

Table 16: A survey of the structural types of 3-word combinations obtained from the native speaker sample

Structural type	Description	Types	%	Example
Type 5	PPs	38	17.3	*in search of*
Type 4	Other NPs	37	16.8	*the pocket muse*
Type 10	Others	33	15.0	*a lot of*
Type 3	A NP with an *of*-fragment	30	13.7	*the story of*
Type 2	A VP + an active VP	22	10.0	*buy this book*
Type 1	A pron/N + *be*	20	9.1	*this book is*
Type 7	A N/pron + verb	19	8.6	*you will feel*
Type 6	A pron/N + be + Vpas	8	3.6	*can be said*
Type 8	*To*-infinitive clause fragment	7	3.2	*to find out*
Type 9	A NPs with other postmodifying fragment	6	2.7	*assets that produce*
Total		**220**	**100.00**	

4.10 Three-word lexical bundles found in all three samples

Table 17 shows three groups of word-combinations in all three samples. The second column identifies lexical bundles in the strict sense, i.e. three-word combinations which occur more than 10 times per million words. This group forms approximately 20 per cent of lexical bundles in all three samples. The third column shows the number and percentage of three-word combinations which are attested in the PIE but fail the minimum frequency cut-off, i.e.10 occurrences per million words. Such three-word combinations form approximately one half of all three-word combinations in all three samples. The fourth column contains the number of three-word combinations with no matches in the PIE. They form only 18 per cent in the Czech sample whereas almost one third in the non-Czech and the native speaker sample.

Table 17 : Distribution of 3-word combinations from all three samples checked with the PIE

Sample	**LB (10+)**	**%**	**Matches in the PIE (10-)**	**%**	**No matches in the PIE**	**%**	**Total**
Czech	84	22.7	219	59.2	67	18.1	**370**
Non-Czech	55	17.2	165	51.5	100	31.3	**320**
Native	42	19.1	108	49.1	70	31.8	**220**

Table 18 shows the most frequent three-word lexical bundles in the PIE found in both learner samples and the native speaker sample. The first, third and fifth column contain the most frequent lexical bundles in the PIE, the second, fourth and sixth column show the number of tokens of these most frequent lexical bundles found in the Czech learner, non-Czech learner and native speaker sample. The word-combinations in bold italics are shared among all three samples; the word-combinations in bold lower cases refer to those found in both learner samples; the word-combinations in capital letters contain both the non-Czech learner and the native speaker sample.

Table 18: PIE most frequent 3-word lexical bundles found in the Czech, non-Czech and native speaker sample

PIE most frequent LB	**Tokens per type CZL**	**PIE most frequent LB**	**Tokens per type NCZL**	**PIE most frequent LB**	**Tokens per type NS**
1. ***one of the***	12	*1.* ***one of the***	6	1.o***ne of the***	2
2. **as well as**	2	*2. the end of*	6	2.*out of the*	3
3. out of the	3	*3.* **as well as**	2	3. SOME OF THE	2
4. there is a	4	*4.* SOME OF THE	2	4. ***in order to***	2
5. it was a	3	*5. end of the*	6	5. THERE IS NO	3
6. the fact that	3	*6. the fact that*	5		
7. to be a	3	*7.* ***in order to***	2		
8. ***in order to***	2	*8.* THERE IS NO	3		
9. **it is not**	3	*9.* **it is not**	2		

Table 19 shows the frequency (obtained from the PIE) of the lexical bundles which occur in all three samples.

Table 19: Three-word lexical bundles found in all three samples and their frequency in the PIE

No.	LB	Freq per mil. words
1	*one of the*	350
2	*as well as*	171
3	*out of the*	154
4	*some of the*	151
5	*there is a*	149
6	*end of the*	134
7	*it was a*	132
8	*the fact that*	129
9	*in order to*	118
10	*to be a*	115
11	*there is no*	111
12	*it is not*	103

4.11 Comparison of three-word combinations in all three samples

All three samples show that there are great differences between three- and four-word combinations. There are almost twice as many three-word combinations than four-word combinations (both in types and tokens) in all three samples. This result is not surprising given that the corpus findings in LGSWE (1999) indicate that three-word bundles occur more frequently in the language than four-word bundles. Again, the number of three-word combinations in the native speaker sample is much smaller compared to both learner samples. Apart from that, the distribution of types in the native speaker sample is relatively stable whereas this cannot be said about the Czech learner sample, in which the frequency of types ranges from 12 to 2. The PIE also provides useful data concerning true lexical bundles. Approximately 20 per cent of the three-word combinations (in all three samples) qualify as lexical bundles - they occur more than ten times per million words. A further qualitative analysis reveals that the group with zero occurrence in the PIE contains sequences which are meaning-specific. They include sequences with names of authors, characters or book titles, such as *The Adventures of Huckleberry, book by Dan Brown, Who moved my, Picturing your audience*. Such sequences form in all three samples approximately one third of three-word combination types. The other group formed by word-combinations with only few occurrences in the PIE is represented by sequences created on "an ad hoc basis". Most of the sequences are related to the semantic field of books and reading such as *the book I , main character is, like to write.*

The in-depth analysis reveals that only 2 lexical bundles in the strict sense are shared among all three samples: *one of the, in order to.* Both of them are the most frequent lexical bundles in academic prose (LGSWE 1999, 993). There is a small number of other three-word combinations shared between all three samples; these do not belong to the bundle group, though and all of them relate to the essay topics *(of the book, the story is, the book is, this book is).* Apart from the 2 lexical bundles in the strict sense shared between Czech and non-Czech learners, 24 three-word combinations were identified in both non-native learner samples: *I think that, of the book, Da Vinci Code, the fact that, the story is, in the end, him in the, based on the, in the beginning, the book is, the adventures of, this book is, it is not, is set in, was one of, the time of, it is a, the battle of, because of their, Adventures of Huckleberry, as well as, Harry Potter and, as a result, of Huckleberry Finn.*

Greater structural richness in the native speaker sample is debatable. The number of three-word combination structural types is approximately the same in all three samples. In particular, Czech learner sample provides 11 structural types, non-Czech learner sample contains 10 structural types, 10 structural types of three-word combinations were found in the native speaker sample. Providing that the native speakers produced more three-word combination types, a greater number of structural groups in the native sample could be formed. Nevertheless, this is just a hypothetical assumption. The sample does not provide sufficient evidence for this statement and greater structural richness cannot be relied upon with the increasing number of word-combinations.

Few similarities exist: the biggest structural groups are in all three samples the same, with only little differences in counts (see Table 12, 14, 16). The structural type of Other noun phrases occupies the top position in the Czech and non-Czech learner sample. The high frequency of noun phrases is due to the fact that a great number of them contain names of characters, authors etc. In the native speaker sample, noun phrases are the second biggest structural type, preceded only by Prepositional phrases. The structural type of prepositional phrases in the Czech learner sample forms the third biggest structural group, following the structural type Others. The structural types Prepositional phrases represents the second biggest group also in the non-Czech sample, followed by the structural type Others. A great number of the noun phrases, prepositional phrases and sequences from the type Others concern the topic, i.e. either book titles, names of characters, authors and generally relate to the semantic field of reading. As opposed to the four-word combinations, structures with Adverbial fragments appear to some extent in all three samples.

4. 12 Discussion, comparison and summary of findings

The analyses in the previous sections compare three- and four-word combinations in terms of their frequency, diversity and syntactic structure in two non-native learner samples and one native speaker sample. Another important part of the investigation was to find out whether, which and how many lexical bundles in the strict sense occur in all three samples. The PIE database was used as a control sample. This investigation adopted a less conservative approach, i.e. the one proposed by Biber et al. (1999), i.e. the sequence must occur at least ten times per million words to qualify as a lexical bundle. The terms "word combinations" and "lexical bundles" were not used interchangeably in this investigation. While the term "word-combinations" refers to any three-word and four-word non-idiomatic sequences present in the samples regardless of the frequency in the PIE, the term "lexical bundles" concerns such word-combinations which occur at least 10 times per million words.

The results from the analyses indicate that recurrent non-idiomatic word-combinations produced by two learner groups and one native speaker group do show some differences in spite of the same size of the samples. When the four-word combinations and three-word combinations were retrieved by the application Collocate (see Appendix 2), it was found that non-native speakers produced twice as many three- and four-word combinations than native speakers. Since essays and reviews represent a creative form of the language, it was assumed that non-native learners' output would be more repetitive than the writing of native speakers. Indeed, both non-native samples do show signs of rather repetitive language compared to the native speakers. Namely, the Czech learners produced 127 four-word combination types, non-Czech learners 119 and native speakers twice less – 54 four-word combination types. The situation looked similar with the three-word combinations: 370 three-word combinations were found in the Czech sample, 320 in the non-Czech sample, 220 in the native speaker sample (see Figure 1). When a detailed analysis was carried out, the aspect of greater repetitiveness became even more obvious in that the most frequent three-word combinations in both groups of non-native writing occurred more frequently than in the native speaker sample. The frequency of types was stable in the native speaker sample (the majority of three-word combinations occurred twice) whereas the frequency of types in the Czech learner sample ranged from 12 to 2. A great number of the three-word combinations in the Czech learner sample occurred five, four, three and two times. These were mainly sequences created on "an ad hoc basis" and thus cannot be regarded as true lexical bundles. Some word-combinations were particularly popular among Czech learners (e.g. *I would like to, would like to write, I am going to, my point of view, I would recommend it)* whereas they were completely absent in the native speaker sample. Czech learners apparently used such word-combinations as fillers in order to make the essay longer.

The structural analysis of word-combinations was conducted following the taxonomies proposed by Biber et al. in LGSWE (1999). The analysis yielded similar structural groups in all three samples. The four-word combinations include the biggest structural group of A noun phrase with postmodifying clause/fragment in all three samples. These structures were mainly used to identify or specify book characters (*the lord of the, book Harry Potter),* a place (*the Mississippi river),*some type of quantity (*the rest of the)* or to highlight qualities (*the magic of the).* The type with the passive voice yielded several sequences in the Czech sample, however, very few in the native speaker sample even though the passive voice is relatively frequently used in written English. The reason why Czech learners used the passive voice is influenced by the topic, that is to say the sections in text-books providing a recommendation on how to write a review often emphasize the use of the passive voice. Since the reviews in the text-books often contain a book or film review, no wonder learners are familiar with such structures as *the film is set, the book was written* etc. The type with An adverbial clause was not identified in any sample, also the structural type with Anticipatory *it* was missing as far as the four-word combinations are concerned. The situation looked similar with the three-word combinations. The biggest structural groups of three-word combinations are also Other noun phrases, Prepositional phrases and Others and again they mainly contain book titles such as *lord of the, book Harry Potter* and the like.

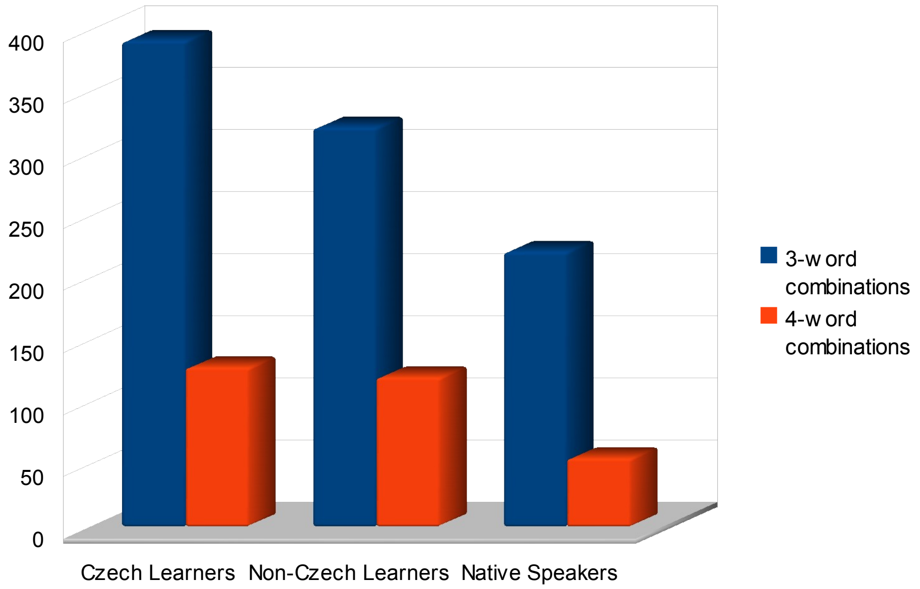
A further comparison with the PIE revealed the number of lexical bundles in the strict sense in the samples. From the start, it was emphasized that a great number of true lexical bundles were not expected owing to either non-academic register or the language of conversation. This assumption was confirmed with the four-word combinations. It turned out that the majority of the four-word combinations in the non-native and native writing could hardly be regarded as lexical bundles. A mere 7.1 per cent of four-word combination types in the Czech sample; 2.4 per cent in non-Czech learner sample and 5.6 per cent in the native speaker sample are lexical bundles in the strict sense. Almost one half of the combinations in the Czech learner sample were not attested in the PIE, in the native speaker sample it was almost two thirds of the four-word combinations.

The comparison with the PIE yielded interesting results as regards the three-word combinations in the samples. Approximately 20 per cent of the three-word combinations in all three samples are lexical bundles in the strict sense. Despite the assumption of greater creativity in the native speaker sample at the expense of formulaic language, lexical bundles in the strict sense were expected to be well represented in the native speaker sample. The results show, however, that it was the Czech learners who produced slightly more lexical bundles in the strict sense than the native speakers. In view of a small size of the corpora, however, the conclusions are somewhat counter-intuitive, even though an explanation can be offered: the learners use "safe" sequences of words, the phrases they are familiar with while native speakers are not afraid to use more creativity in their writing. A great number of word-combinations with a relatively high frequency in the learner samples but with zero occurrence in the PIE suggest that these word-combinations have become prominent only because the sample is very small and most of these sequences are topic-bound. It is hardly surprising that such combinations are not attested in the PIE. The search in the PIE also indicates that apart from the true lexical bundles, both learners and native speakers produced a great number of word-combinations with either zero occurrence in the PIE or sequences which occur in the PIE not frequently enough to be called lexical bundles. The sequences with zero occurrence in the PIE mainly contain book titles, book characters and so on. Those with low occurrence in the PIE mainly involve word-combinations created on an ad-hoc basis and topic bound sequences, which mostly refer to the semantic field of reading.

To sum up, analysis of lexical bundles shows that the presence of distributional multi-word sequences in text need not be an unequivocal test of native-like competence. Somewhat paradoxically, their abundance may indicate lack of effective writing skills and linguistic confidence. Obviously, other factors than merely quantitative must be taken into account when assessing the use of lexical bundles.

Figure 1 shows the distribution of three- and four-word combination types and tokens in all three samples.

Figure 1: Three- and four-word combination types in all three samples



5. Phrasal and prepositional verbs

While the previous chapter focused on distributional phraseology, this chapter reports on the findings which emerged from the investigation of multi-word verbs. The research is introduced by a short theoretical overview of multi-word verbs - phrasal verbs and prepositional verbs. The analysis of the data is again based on the comparison of two non-native samples with a native speaker sample, focusing first on the overall frequency of phrasal verbs, the variety of their meanings, then on the range of adverb particles, the frequency and range of prepositional verbs. Dubious cases of phrasal verbs and prepositional verbs created by non-native speakers are also discussed.

5.1 Theoretical background

The chapter on multi-part verbs is divided into two parts: the theoretical and the research part. The theoretical part addresses the key issues relevant for the comparison of the samples.

5.2 Multi-word verbs

The following sections will provide a brief theoretical outline of all types all multi-word verbs, that is to say phrasal, prepositional verbs and phrasal-prepositional verbs.

5.2.1 General overview

The literature concerning multi-word verbs is quite extensive with authors adopting different approaches towards the notion of multi-word verbs. In general terms, multi-word verbs are defined as constructions which contain a lexical verb followed by a free morpheme. The free morpheme can be: 1. an adverb particle; 2. a preposition; 3. an adverb particle accompanied by a preposition. Thus the notion of multi-word verbs encompasses a set of constructions which are commonly described as: 1. phrasal verbs *if you* ***go out*** *drinking every night you will never pass your exams or save any money;* 2. prepositional verbs *a lot of small grocers have* ***gone out of*** *business since the advent of the supermarket.*; 3. phrasal prepositional verbs *our heart* ***go out to*** *all the victims of the earthquake in Yugoslavia* (ODPV 2010 140-2).

Quirk et al. (1985) define multi-word verbs as “units which behave to some extent either lexically or syntactically as single verbs”. Their classification of multi-word verbs follows the traditional categorization mentioned above: 1. phrasal verbs (*find out; carry out)*; 2. prepositional verbs (*cope with; depend on)*; 3. phrasal prepositional verbs (*put up with).*

Additionally, Quirk et. al (1985) propose two other types of multi-word verbs, without any particle (*put paid to, cut short).*

Cowie and Mackin (1993, 2010), Dušková (1988), LGSWE (1999) and Claridge (2002) adopt a similar classification. However, Cowie and Mackin (1993, 2010) include other combinations, for example the verb-adjective or the verb-pronoun type. LGSWE (1999, 403) proposes the categories of verb + noun phrase + preposition (*take a look at*); verb + prepositional phrase (*take into account);* verb + noun. Claridge (2002) extends the classification a little further by including a verb-adjective type (intransitive *-hold good x* transitive *break open);* a verbo-nominal type (with the noun regarded as an obligatory part); a verb-verb type (either in the form of infinitive *let go* or a present participle *send N packing).*

A different approach towards the notion of the adverb particle is taken by Huddleston and Pullum (2002, 273). They do not use the term "adverb particle" but replace it by the term "intransitive preposition", since it functions as the complement to the verb *(I have to carry out this task)*. The omission of the intransitive preposition would result in a sentence which would be agrammatical. On the other hand, constructions such as *the book belongs to me* or *he came to the office late,* contain a preposition and yet they would not categorize it similarly. While the first example is classified as a prepositional verb and the preposition *to* in *belong to* cannot be replaced by a different preposition, the other example *he came to the office late* does not contain a prepositional verb. Huddleston and Pullum (2002) argue that the preposition in the latter example functions as an unspecified preposition since it can accompany other verbs suggesting some kind of motion *walk, go etc.*. The same applies to the preposition which can be replaced by a different preposition - *from.*

5.2.2 Multi-word verbs versus free combinations

Multi-word combinations must be distinguished from free combinations. Free combinations are defined as constructions where each element is not grammatically seen dependent on another element. Semantically speaking, each element included in a free combination carries its own independent meaning whereas the meaning of a multi-word verb cannot be deduced from the single elements. There are several, mostly syntactic, tests which are used to differentiate between multi-word verbs and free combinations. The following are proposed in LGSWE (1999, 404): 1. adverb insertion; 2. stress; 3. the possibility to transfer the combination into passive voice; 4. creation of a relative clause; 5. formation of *wh*-questions;

6. fronting of the preposition; 7. particle movement. The majority of transitive phrasal verbs allow for a different position of the adverb particle (if the object is noun). It can stand either before or after the direct object such as in *look up the word in the dictionary* or *look the word up in the dictionary.* This fact does not apply to prepositional verbs and free combinations *I`m waiting for the taxi to come; *I`m waiting the taxi for.*

Quirk et al. (1985) claim that a clear-cut categorization is impossible. They suggest instead that multi-word verbs and free combinations form a cline involving units which range from the idiomatic, fixed and syntactically cohesive ones to those which are connected loosely.

5.2.3 Multi-word verbs from the historical perspective

Claridge (2002, 41) stresses the importance of multi-word verbs from a historical point of view. She notes that multi-word verbs are analytic constructions and represent one of the phenomena indicating the transition of English towards analyticity.“Their most obvious analytic characteristics are of course the fact that one meaning is expressed by a combination of separate words (free morphemes). The alternatives to this procedure are, or would have been, compounding and affixation, and in this respect the decline in the productivity of prefix verbs (*overtake)* is noteworthy when seen against the rise of phrasal verbs” (2002, 41).

Claridge (2002, 41), viewing multi-word verbs from the historical point of view, claims that prepositions and zero-derivation started to be topical when English became analytic. Since inflectional endings have been reduced to a minimum, the use of prepositions has increased, especially of those following verbs. The process of shifting from one word class to another by means of zero-derivation has become very common and so has the use of nouns as verbs which then combine with free morphemes to become multi-word verbs.

5.3 Phrasal verbs

The following sections focus on phrasal verbs from the theoretical view point. The sections provide the outline of phrasal verb characteristics and classification. Further, the following sections concentrate on the differences between phrasal verbs and free combinations, the divergences between adverb particles and prepositions, and corpus findings related to phrasal verbs are presented. Finally, Sinclair´s model of extended lexico-grammatical units is touched upon.

5.3.1 General characteristics and phrasal verb classification

Apart from the term “phrasal verb”, other names have been used: “separable verbs” (Francis, 1958); “two-word verbs” (Taha, 1960); “discontinuous verbs” (Live, 1965) or “verb-particle combinations” (Fraser, 1976).

To define a phrasal verb, it is a multi-word verb which contains a lexical verb, usually a monosyllabic one (*take, put, get, set*), and which combines with an adverb particle. Quirk et al. (1985) define a phrasal verb as “a verb followed by a morphologically invariable particle, which functions with the verb as a single grammatical unit”. Hence phrasal verbs are regarded as single units, where the intended meaning is expressed only thanks to the cooperation of both elements. If the adverb particle is removed, the meaning of the lexical verb changes (cf. *The plane takes off x The plane takes).*

5.3.2 Phrasal verbs versus free combinations

As mentioned above, several syntactic tests have been proposed to distinguish between multi-word verbs and free combinations. Quirk et al. (1985) describe in detail the differences between phrasal verbs and free combinations and focus on the elements which allow for the differentiation:

1.The idiomatic meaning of the phrasal verb cannot be deduced from the single components while the meaning of a free combination is quite transparent. For example, the meaning of s*he took in the box* is *she brought the box inside,* while *she took in her teacher* expresses the idiomatic meaning of *deception.*.

2. Both elements of free combinations can be separated and replaced by another one from the same word class (*she walked past).* The empty slot for *walk*, could be filled with *run, rush, swim ; etc.; past c*ould be replaced by *in, through, over.*

3. Syntactically speaking, free combinations allow other elements to be inserted *(go straight on)* whereas this is not the case in phrasal verbs (*s*he turned right up*).

4. The adverb can occupy the initial position in free combinations (o*ut came the sun*) but never in the case of phrasal verbs (**up blew the tank).*

When a a semantic perspective is taken into account, Dušková (1994, 204) holds that the meanings of the individual components of the phrasal verbs are different from the meaning of the unit (*look up* vs. *look – podívat se, up – nahoru).* Apart from this*,* a great number of phrasal verbs have their one word equivalents, which are usually of a Latinate origin (*put up – accommodate; track down – discover, find; bog up – confuse)*.

Dušková also (1994, 205) distinguishes between: a) idiomatic expressions (verbs with the adverb particle create a new semantic unit, *(see off; size up; bring about; bear up)*; b) non-idiomatic expressions, in which verbs and adverb particles retain their meaning (*blow up; break off etc.)*; c) intensifying expressions or phrasal verbs wherein the particle has intensifying or perfectivizing function (*fasten up; eat up; break up).* A great number of phrasal verbs have both a literal and idiomatic meaning (*take in a journal – odebírat časopis, take in a skirt – zabrat sukni).*

5.3.3 Adverb particles versus prepositions

There are several differences between prepositions and adverb particles. The adverb accentuation is the most crucial factor according to Lipka (1972). Adverb particles can be stressed, whereas prepositions do not have this capacity. Quirk et al. (1985) list other factors. One of these is the adverb particle placement: whereas the adverb particle can precede or follow the direct object, the preposition must precede the direct object *(she called up her friends or she called her friends up x she called on her friends x *she called her friends on)*. Also the position of a personal pronoun with regard to adverb and the preposition is fixed and differs in both constructions. The adverb particle follows the pronoun *(she put it off),* the preposition precedes it (*look at it.).* The verb and the preposition can be separated by an adverb *(she called angrily on her friends x She called angrily up her friends*), but not the verb and the adverb particle. Another difference concerns the position of the adverb particle and the preposition when there is a relative pronoun or *wh*-interrogative in the structure. While the adverb particle cannot stand between a relative pronoun or a *wh*-interrogative, this position is possible with a preposition (*the friends on whom she called x the friends up who she called).*

Claridge (2002, 50) focuses on the historical perspective of adverb particles and notes that adverb particles are regarded as items expressing *location* and/or *direction in space*. Concerning the non-semantic aspect of adverb particles, Claridge holds that if the adverb particle joins a verb it results in the change of transitivity – the intransitive verb becomes transitive (*he was just staring x each boxer tried to stare the other down)* or the transitive verb becomes intransitive *(take a book x the plane takes off).* Interestingly, some words start to function as verbs only when the adverb particle is added, otherwise they function only as nouns, for instance (*zip x zip up).*

5.3.4 Corpus findings related to phrasal verbs

A detailed overview of corpus findings concerning phrasal verbs is presented in LGSWE (1999, 409-412). The findings reveal that phrasal verbs are mainly pervasive in fiction and the language of conversation but they also occur, to a certain extent, in journalistic English or academic prose. Several phrasal verbs have a core meaning but can take on a different meaning if used in a different register.

The most common phrasal verbs in English form the following subtypes: activity intransitive phrasal verbs (*come on, get up, sit down, get out, come over, stand up, go off, shut up, come along, sit up, go ahead)*, activity transitive verbs (*get in*, *pick up, put on, make up, carry out, take up, take on, get back, get off, look up, set up, take off, take over*) mental transitive verbs*(find out, give up)*, communication transitive verbs *(point out*), occurrence transitive verbs *(come off, run out),* copular verbs (*turn out)* and aspectual transitive phrasal verbs *(go on)*. LGSWE (1999) provides a list of the most productive lexical verbs which form a phrasal verb together with a particle includes: 1. *take*; 2. *get*; 3. *put*; 4. *come*; 5. *go*; 6. *set*; 7. *turn*; 8. *bring;* the six commonest adverbial particles are as follows: 1. *up*; 2. *out*; 3. *on*; 4. *in*; 5. *off*; 6. *down*.

5.3.5 Phrasal verbs and The model of extended lexico-grammatical units

Sinclair's (2004) proposed model of extended lexical units clearly demonstrates how intricately language is patterned. Lexical units can be approached from several angles which results in a comprehensive and detailed description of the lexical unit. Sinclair puts forward a model of extended lexical units, which comprises 1. collocation (words that keep a particular expression company); 2. colligation (grammatical structure); 3. semantic preference (associations which words provoke in our mind); 4. discourse prosody of lexical units (reflects what is not explicitly worded by the speaker but yet understood). Thus every lexical item includes lexical, semantic, syntactic and pragmatic level of description.

This model also applies to phrasal verbs and through this model Sinclair (2004) proves that the environment of phrasal verbs often indicates that the grammatical structure, semantics as well as the words in the neighbourhood are predictable to some extent. There is an implied meaning which can be inferred without being explicitly worded. Sinclair claims (1991), that "the disposition of the words involved and their syntax are governed by complex and predictable rules and the semantics of phrasal verbs are not as arbitrary as it was often held to be“. Sinclair’s theory is illustrated by a variety of examples related to the verb *set* followed by numerous adverb particles. For instance, the phrasal verb *set about* is mostly followed by the *-ing* form of another verb, which is usually transitive and preceded by either modals, negatives, interrogative words; structures which give the associations of *uncertainty or problem solving* often come first *(she had not the faintest idea of how to set about earning any money)*.

5.4 Prepositional verbs

The sections related to prepositional verbs offer the general overview of prepositional verbs, their semantic domains, the discussion of the differences between prepositional verbs and free combinations will be presented together with the corpus findings related to prepositional verbs. Finally, brief mention is made of phrasal-prepositional verbs.

5.4.1 General overview

Prepositional verbs are defined as verbs which take a prepositional object, i.e. the noun phrase coming after a preposition (LGSWE 1999, 413). LGSWE distinguishes two main structural patterns: Pattern 1 contains a noun phrase which is followed by a verb a preposition and another noun phrase *(it just looks like the barrel).* Pattern 2 contains a noun phrase which is followed by a verb, noun phrase, preposition, noun phrase *(they like to accuse women of being mechanically inept*). The passive voice usually occurs in Pattern 2, in this case the noun phrase corresponds to the direct object and occupies the subject position (LGSWE 1999, 413).

LGSWE (1999, 413) says that it is possible to come across an adverbial element between the verb and the prepositional phrase, as it can be seen in the following example *I have never thought much about it*. On the other hand, the structure comprising a verb and a preposition in Pattern 1 can be regarded as a single entity, a prepositional verb. Pattern 1 has the capacity to function semantically, as a single unit, whose the meaning does not follow from the meanings of the two parts (LGSWE 1999, 413). Similarly as with phrasal verbs, which often have one-word equivalents, prepositional verbs can be substituted by a single lexical verb (*think about* can be replaced by *consider, ask for* by *request*).

5.4.2 Semantic domains of prepositional verbs

LGSWE (1999, 414) notes that a great number of prepositional verbs bear more than one meaning which is mainly true of so called activity verbs *(deal with, get into, go through, look at, return to, arrive at, engage in, get at, get through, look into, derive NP from, reduce NP to)*. Apart from this group of verbs, prepositional verbs also form the semantic group of communication verbs (*talk to, talk about, speak to, ask for, refer to* etc.), mental verbs (*think of, think about, listen to, worry about, know about, be known as, be seen in, be regarded as, be seen as*), causative verbs (*lead to, come from, result in, be required for*), occurrence verbs (*look like, happen to, occur in*), existence and relationship verbs (*depend on, belong to, account for, consist of, be based on, be involved in*).

5.4.3 Prepositional verbs and free combinations

Quirk et al. (1985, 1152) introduce several criteria to set apart prepositional verbs from free combinations. The possibility to make the prepositional object the subject of a corresponding passive clause points to prepositional verbs; the preposition stands in its post-verbal passive (*This matter will be dealt with immediately).* Additionally the *wh*-questions should be mentioned in this respect: those which elicit prepositional object take the form of *who(m), what* (the same applies to direct objects), for example *John called on her - who(m) did John call on?* The situation looks different with free combinations *(John called from the office x where/ did John call from?*).

5.4.4 Prepositional verbs and corpus findings

According to LGSWE (1999, 415), prepositional verbs occur frequently in all four registers, they are especially popular in fiction. In particular, approximately 4 800 prepositional verbs per million words can be found in the language of conversation, more than 6 000 in fiction; journalistic English or academic prose mark slightly above 4000 words per million words. Since they lack the informal tone (as opposed to phrasal verbs), they are comparably common also in academic prose. They occur more frequently than phrasal verbs and their set of prepositions includes also the non-spatial relations -*as, with, for, of.* Phrasal verbs confine themselves only to the meaning of location and direction since the range of adverbial particles is relatively narrow .

Corpus findings in LGSWE (1999, 416-418) provide evidence that prepositional verbs differ substantially as far as different registers are concerned. The verb *look at* is by far the most common in all registers even though in the language of fiction and the language of conversation occupies the top position (in both cases 400 occurrences per million words). Similarly, *look for* is widely distributed across all four registers, it is particularly common in fiction, though (100 occurrences per million words). Also the prepositional verbs *think of* (per million words 300 occurrences were found in fiction, 100 occurrences in conversation, 40 occurrences in journalistic English as well as academic prose) and *depend on* (per million words, 200 occurrences in academic prose, 40 occurrences in journalistic English and 20 occurrences in the language of conversation and fiction) receive immense popularity in all four registers.

According to LGSWE (1999,419), semantic domains of prepositional verbs are distributed in the following way: activity and mental verbs occur equally frequently in all registers (activity verb: 38 per cent conversation, 41 per cent fiction and news, 33 per cent academic prose); relatively frequent in all registers with the exception of academic prose are communicative verbs (around 20 per cent in each; only 5 per cent in academic prose). Causative prepositional verbs as well as existence verbs occur in abundance in academic prose whereas other registers somewhat fall behind. As far as the syntactic pattern is concerned, conversation and fiction tend to favour Pattern 1*,* academic prose is more inclined towards Pattern 2.

5.5 Phrasal-prepositional verbs

Phrasal-prepositional verbs resemble both phrasal verbs and prepositional verbs in that they contain a lexical verb followed by both adverb particle and preposition. The complement of the preposition fulfills the function of a direct object of the phrasal-prepositional verb. In the previous section relating to prepositional verbs, it was said that two structural patterns are distinguished here, and the same applies to the phrasal-prepositional verbs: structural Pattern 1 comprises a noun phrase followed by a verb,particle, preposition and a noun phrase such as in *I shall look forward to this now.* Pattern 2 comprises *a noun phrase* which is followed by a verb, noun phrase, adverb particle and preposition (*I could hand him over to Sadia),*.

5.5.1 Phrasal-prepositional verbs and corpus findings

Although it was pointed out that phrasal verbs are confined more or less to conversation and fiction, phrasal-prepositional verbs are extremely rare in comparison with phrasal verbs. The findings presented in LGSWE (424) show that phrasal-prepositional verbs tend to be connected with the informal spectrum of the language; there are approximately 1400 phrasal verb occurrences per million words, 2400 prepositional verb occurrences and mere 300 phrasal-prepositional verb occurrences in the corpus. However, a certain degree of similarity exists between phrasal and phrasal-prepositional verbs in that both groups of verbs usually associated with physical activities. On the contrary, the repertoire of prepositional verb semantic meanings is much more extensive, reaching far beyond the physical activities only. The most frequent phrasal-prepositional verbs are activity verbs, *get out of* occupies the first position and is followed by *come out of, get back to.* However, compared to prepositional verbs, their frequency is comparably low – approximately 10-30 occurrences per million words for the most common phrasal-prepositional verbs. Also the semantic group of mental verbs (*look forward to)* does occur quite commonly in comparison with the occurrence, existence and causative semantic groups.

5.5.2 Semantic domains of phrasal-prepositional verbs

As noted above, corpus findings (LGSWE 1999,424-5) show that the most common phrasal-prepositional verbs in the language are verbs linked to activity and mental semantic domains. The most common activity verbs are *.get out of*, *come out of, get back to, go up to, get on with, get away with, get off at, get off with, go out for, catch up with, get away from, go over with, hold on to, turn away from, turn back to, be set up in;* in the mental domain it is *look forward to, come up with, put up with;* occurrence *come down to,* existence *set out in, be made up for, be cut off from* ; causative *end up with;* aspectual *go on to, move on to* .

5.6 Sample analysis – research into the multi-word verbs

The following sections present analyses of multi-word verbs in two learner samples and one native speaker sample.

5.6.1 Questions related to phrasal verb investigation

As has been established, phrasal verbs present a potential pitfall for learners. It is either due to the opacity or polysemy of some phrasal verbs. With increasing learner proficiency, learners are expected to have a better command of phrasal verbs than at the initial stages of language learning. This part presents data obtained from two non-native sample corpora, 37 essays from Czech learners, 19 essays written by non-Czech learners, and one native sample comprising 22 book reviews. All three samples contain approximately 9 400 words. Czech learners are students from a grammar school in Prague; they are sixteen and seventeen year old students with pre-intermediate, intermediate and upper-intermediate to FCE level. The other group of non-native speakers are students from various linguistic backgrounds; their essays were downloaded from the website http://bookrags.com/. The total of 22 book reviews written by native speakers, mainly professional review writers, were downloaded from the website available at http://happypublishing.com/ (for a detailed description see Section 3.3).

It is possible that it will be necessary to tackle the following issues:

It is expected that native speakers will use more phrasal verbs than non-native speakers; the variety of meanings and the range of adverb particles will differ; the native speakers' range of lexical verbs which form phrasal verbs will presumably be wider. Even the most common phrasal verbs with very frequent lexical verbs pose a problem for learners and thus it can be expected that if learners use some phrasal verbs, they will belong to the most frequent ones. It is necessary to reckon with errors of a different origin in the non-native writing which are largely due to mother tongue interference. It often subsumes the following factors (Nesselhauf 2005, Waibel 2007):

7. The inappropriate extension of the collocational range;
8. The use of a wrong lexical verb or an adverb particle, whose combination results in the inappropriate phrasal verb;
9. The use of rather a vague verb due to the learner' s insufficient vocabulary skills ;
10. The grammatical structure in which a phrasal verbs prototypically occurs could be distorted (colligation);

11. Omission of the adverb particle;
12. The use of one word equivalent instead of the phrasal verb more appropriate in the particular context.

These aspects require further clarification: collocational deviations are a common occurrence among non-native speakers. They lack the sensitivity of a native-speaker to judge objectively how a particular collocational range can be extended (Baker 1992, 51). Learners often erroneously assume that words sharing the same semantic field have the same collocational range, which is not always the case (e.g. *carry out a task* but not *carry out homework).*

A further problem closely linked to the use of phrasal verbs by non-native speakers is the appropriate choice of a phrasal verb. Here two possibilities arise: either the correct lexical verb is selected given the particular context while the adverb particle proves inappropriate, or vice versa.

Another difficulty refers to the use of a vague expression instead of a phrasal verb (learners do not have the knowledge and try to find other means how to express meanings).

Omitting the adverb particle where it is appropriate may be encountered (e.g. *drink x drink up; pay x pay off*).

Different types of evidence to judge the appropriate use of a particular unit are recommended by Sinclair (1991): a native-speaker introspection, the corpus and dictionary consultation.

5.6.2 Initial procedures related to phrasal verbs

Before the investigation of phrasal and prepositional verbs had been launched, the question of how to extract the different types of multi-word verbs presented a major methodological problem. Since none of the sample corpora are morphologically annotated (tagged), it was necessary to consider how to distinguish phrasal verbs from prepositional verbs. The theoretical introduction, which outlines the possible pitfalls in distinguishing between phrasal and prepositional verbs, is provided.

The first step was to identify all candidates in the samples and then sort all the verbs manually. Cowie's Oxford Dictionary of Phrasal Verbs (1993, 2010) was used to verify the status of the verbs. The verbs were divided into three groups - phrasal, prepositional and phrasal-prepositional. To some extent different analyses were carried out on phrasal and prepositional verbs.

As stated above, the analysis concerns the data obtained from three samples: the Czech learner sample (37 essays), the non-Czech learner sample (19 essays), and the native speaker sample (22 texts). For the purposes of the investigation, the program ConcGram was used. All multi-word verbs were sorted out manually, the phrasal verbs being selected first. Three types of evidence are used: 1. ODPV (2010) is used for the verb identification; 2. the BNC is used as a control sample if a phrasal verb is not listed in the dictionary; 3. native speakers as informants.

5.7 General overview of the phrasal verbs obtained from the Czech learner sample

It has been pointed out in several studies that the use of phrasal verbs is closely linked to a learner's proficiency. That is to say learners exhibiting more advanced levels of English tend to use more phrasal verbs in their language production; the investigation of the Czech learner writing was carried out at each level separately – pre-intermediate, intermediate and upper-intermediate. In view of this, a low incidence of phrasal verbs was expected, especially in the pre-intermediate and intermediate writing

Table 20 presents all the phrasal verbs found in the Czech learner sample. The first column corresponds to phrasal verb types. The second column shows the list of phrasal verbs sorted alphabetically; doubtful cases of phrasal verbs are marked in bold italics. The third column relates to the phrasal verb meaning. The last column concerns the frequency of the phrasal verbs. The phrasal verbs are arranged in alphabetical order.

Table 20: Phrasal verbs in the Czech learner sample

No.	**PV types**	**Meaning**	**Tokens per type**
1	*bring up*	raise a child	1
2	*close in*	approach	1
3	*come back*	return	2
4	*come out*	be released from prison	2
5	*cut out*	remove	1
6	*end up*	finish	1
7	*fall down*	collapse	1
8	*find out*	discover	4
9	***get up***	help sb to climb the career ladder	1
10	*get back*	return	1
11	*go back*	return	1
12	*go on*	continue	2
13	*grow up*	be raised	3
14	*knock out*	eliminate in competition	1
15	*look up*	find a word in a dictionary	1
16	*pass out*	faint	1
17	*point out*	indicate	1
18	*run away*	escape	1
19	*run down*	criticize unkindly	1
20	*set up* ***set up (a journey)***	1. establish a company 2. start a journey	2
21	*slow down*	drive less quickly	1
22	*sum up*	summarize	2
23	*take one's breath away*	surprise	1
24	*turn out*	show	1
25	*wake up*	stop sleeping	2
Total			**36**

Table 21 presents the findings related to the productivity of lexical verbs in the Czech learner writing. In particular, it outlines the list of lexical verbs with the number of adverb particles it combines with. It is evident that learners were not highly inventive - only four verbs occur with two particles, the rest of lexical verbs (not presented in the table) combine only with one adverb particle.

Table 21: Adverb particle productivity (combination with different verbs) in the Czech learner sample

Lexical verb	Adverb particle	No. of particles
come	*out, back*	2
run	*away, down*	2
get	*up, back*	2
go	*on, back*	2

The data collected from the pre-intermediate learners support the initial assumptions: a very low incidence of phrasal verbs was found in the pre-intermediate and intermediate samples. Perhaps because of the low number of occurrences, errors occur only scarcely. There are 25 phrasal verb types, 36 tokens in the Czech learner writing. The pre-intermediate students produced 8 tokens which include 2 dubious cases; intermediate learners produced 8 tokens, all are used appropriately; the upper-intermediate learners produced 20 tokens, all of them are used appropriately.

The most frequent phrasal verbs in the Czech learner sample are the following: 1. *find out* (4 occurrences); 2. *grow up* (3 occurrences); 3. six phrasal verbs occur twice (*go on, come out, wake up, sum up, set up, come back);* the rest of the phrasal verbs occur only once.

The frequency and the selection of the phrasal verbs suggest what style the learners adopted. While some phrasal verbs are common in academic prose (*sum up, point out*), others represent a rather colloquial style (*go on, ran away, grow up.)* The mixture of styles reflects an inexperienced learner who is not very aware of such differences in register. This "random" selection of phrasal verbs takes place mainly in the upper-intermediate group. Another plausible explanation reflects the learners' effort to display their language skills, which are apparently on a higher level than that of pre-intermediate learners. As far as the use of some phrasal verbs is concerned, a few reflect the learner's sensitivity towards the topic. Several phrasal verbs are closely linked to the life of fictional characters (*grow up, bring up, the story goes on, the character wants to find out, ran away*); a few phrasal verbs (*point out, sum up)* reflect students' thoughts in relation to the text structuring (essay writing, book reviews).

As for the combination of lexical verbs with adverb particles, the results show that only three verbs are more productive than the rest but still combine only with two adverb particles: the verb *come* is followed by two adverb particles - *out, back; run* is followed by *away, down*; *get* is followed by *up, back*. The rest of the verbs follows only one adverb particle.

There are 21 lexical verbs which occur with adverb particles, and are represented mainly by relatively common lexical verbs. Seven of them (with the exception of 1) correspond with the most productive lexical verbs introduced by LGSWE (1999, 412).

The adverb particles which accompany the lexical verbs in the sample belong to the most frequent and productive adverb particles, according to LGSWE findings (1999, 412). Czech learners used 1. *up* (12); 2. *out* (11); 3. *back* (4); 4.-5. *down, on* (3); 6. *away* (2); 7. *in* (1).

5.7.1 Error and qualitative analysis of the Czech pre-intermediate, intermediate and upper-intermediate learner writing

As has been mentioned, 8 phrasal verbs *(pass out, wake up, close in, look up,* ***get sb up, set up****, cut out, find out,)* occur in the pre-intermediate level out of which two (in bold letters) are used in a non-standard way (see Table 20 above). The non-standard uses of the phrasal verbs are linked to the use of inappropriate particles and the use of a vague verb. The very low number of occurrences is explicable due to the relatively low learners' level.

Two phrasal verbs were not used appropriately (*set* ***up on*** *the journey;* ***get sb up***). Although their meaning can be easily deduced, they would not be produced by a native speaker. The first inappropriate use of phrasal verb ***set up*** *on the journey* is related to the inappropriate choice of the adverbial particle; *set off the journey or set out on the journey* would be more appropriate.

(1) *alf of his, in that time, an unknown father, they* ***set up on*** *the journey to kill the highest and worst man I*

In comparison with example (1), example (2) offers different explanations:

(2) *er. But he is looking for another woman who could* ***get him up****. A mistress, she will place in t*

The collocation *get him up* was neither attested in the BNC nor found in ODPV (2010). Therefore, the consultation with native speakers was necessary. The native speakers suggested that the collocation *get him up* could have a sexual connotation whereas the learner's intended meaning was *to help someone climb the career ladder.*

A look at the results in Table 18 obtained from the intermediate level learner group reveals that there are no dubious cases of phrasal verbs. There are 7 phrasal verb types, altogether 8 tokens (*get stg back, knock out, fall down, take one' s breath away, come out 2x, go back, come back).* Most of them belong to common phrasal verbs, with the exception of *take one' s breath away.*

The upper-intermediate learner's writing shows some differences compared to the pre-intermediate and intermediate groups. Even though the highest incidence of phrasal verbs was found in the upper intermediate students' writing (14 phrasal verb types: *slow down, end up, find out, ran away, wake up, grow up, bring up, turn out, point out, go on, set up, run sb down, sum up, come back), s*ome phrasal verbs occur more than once (20 tokens) and the total number is still not very high. Drawing on LGSWE (1999, 412) findings related to phrasal verbs, the phrasal verbs which occur in the upper-intermediate learner writing belong to those which occur quite frequently in the language. *Find out* as well as *grow up* occur three times, *go on* and *sum up* occur twice in the Czech sample. The rest of the phrasal verbs occur once only.

Some phrasal verbs in the Czech sample are apparently related to the semantic field of reading, in particular the life of the characters. Some phrasal verbs appear to be closely linked to the life of the characters (*grow up, bring up, the story goes on, find out, ran away*). Phrasal verbs such as *point out, sum up* are used by students because of the genre "review".

5.8 Phrasal verbs in the non-Czech learner sample

The data collected from the non-Czech learner sample are processed in the same way as in the Czech learner sample.

5.8.1 Quantitative analysis of phrasal verbs in the non-Czech learner sample

The group of 19 non-Czech learners produced 43 phrasal verb types, 57 tokens (see Table 22). Out of these 57 occurrences, 6 occurrences are regarded as dubious cases, they are marked in bold italics (***block out*** *a spell*; ***find out on*** *what will happen;* ***carry out*** *a definition;* ***carry out*** *a Gothic theme; hardships he* ***came through****;* ***turn up*** *the torture.*

ODPV (2010), the BNC database and native speakers were consulted to verify the accuracy of the learners ' use of the phrasal verbs. Some phrasal verbs do not occur in ODPV but are attested in the BNC (***rise up** the career ladder;* ***lure** Daisy **back**; (the boy)* ***sauntered up*** *(to the plate)).* The collocation ***rise up*** *the* (ADJ) *ladder* provides a very limited number of occurrences for this collocation attested in the BNC. These occurrences contain different adjectives and yet they belong to the semantic field of *jobs and career life (rise up the promotion/corporate ladder).* There are about 13 occurrences of the phrasal verb *lure back* in the BNC, prototypically found in the passive construction. The phrasal verb *saunter up* in the sense of *slow manner of walking* does occur in the BNC but there are only 6 occurrences.

The repertoire of phrasal verbs is more extensive with the non-Czech learners than with the Czech learners. The first positions in terms of the phrasal verb frequency are occupied by 1. *go on* (6 occurrences*)*; 2. *come back* (3); 3. *put down* (3); 4. *cut off*; *set up*; *get back* (2). The range of phrasal verbs reflects an informal style adopted by the learners.

Lexical verbs in the non-Czech sample (see Table 23) combine with the following adverb particles: 1. *come* is followed by 4 adverb particles (followed by *along, back, through, up)*; 2. *set* followed by 3 adverb particles (*out, back, up)* and *put* followed by *(down, in, forward)*; 3. *go (on, down)*.

The range of lexical verbs which form phrasal verbs is relatively broader in contrast with the Czech learners, but still not to a great extent. Non-Czech students produced altogether 32 lexical verb types complemented by adverb particles.

The most frequent adverb particles are 1. *up* (12); 2./3. *back , on* (8 *)*; 4. *out* (8) . Other adverbial particles, which come into a relation with the lexical verbs, occur but not so frequently (*down, off, in, trough, away, along, behind, forward*). No adverb particles such as *around, about etc.* occur.

Table 22 provides a list of all phrasal verbs detected in the non-Czech learner writing; they are sorted alphabetically. Inappropriately used phrasal verbs are marked in bold italics.

Table 22: Phrasal verbs in the non-Czech learner sample

No.	PV types	Meaning	Tokens per type
1	*beat off*	repulse	1
2	***block out (the spell)***	break a spell	1
3	*call in*	request, order	1
4	***carry out (a definition)*** ***carry out (Gothic theme)***	provide a definition; introduce a theme	2
5	*close up*	close temporarily	1
6	*come along*	arrive, turn up	1
7	*come back*	return	3
8	*come up*	rise	1
9	***come through (hardships)***	survive	1
10	*coop up*	confine	1
11	*cut off*	remove by cutting	2
12	*end up*	finish	1
13	*fight back*	return by struggling hard	1
14	*find out* ***find out on (what will happen)***	learn by study	2
15	*get back*	recover a possession	2
16	*give away*	reveal	1
17	*go on*	1. continue 2. happen, continue	6
18	*go down*	set, disappear below the horizon	1
19	*lure back*	attract/get back	1
20	*leak out*	become known	1
21	*leave behind*	leave as a sign	1
22	*pass on*	hand stg to another person	1
23	*pay off*	settle	1
24	*pick up*	hold, raise	1
25	*pull in*	attract	1
26	*pull through*	survive	1
27	*put down*	1. stop reading 2. suppress, silence sb	3
28	*put in*	install	1
29	*put forward*	advance, propose, suggest	1

30	*rise up*	climb the career ladder	1
31	*run away*	escape	1
32	*saunter up*	walk slowly	1
33	*set out*	begin to work’ with the intention	1
34	*set back*	place, situate	1
35	*set up*	1.establish 2. place in position	2
36	*speed up*	cause to go faster	1
37	*strip down*	remove all clothes	1
38	*take on*	undertake	1
39	*tear up*	destroy by pulling sharply	1
40	*turn in*	abandon, leave	1
41	***turn up (the torture)***	cause to face	1
42	*win back*	get back	1
43	*work oneself out*	be resolved, settled	1
Total			**57**

Table 23 shows the productivity of lexical verbs with different adverb particles, i.e. the combination of adverb particles with different verbs.

Table 23: Adverb particle productivity (combination with different verbs) in the non-Czech learner sample

Lexical verb	Adverb particle	No. of particles
come	*along, back,through, up*	4
set	*out, back, up*	3
put	*down, in, forward*	3
go	*on, down*	2

5.8.2 Error and qualitative analysis of phrasal verbs in the non-Czech learner sample

A few dubious cases were found. The errors are mainly due to collocational deviations, the choice of inappropriate lexical verbs or adverb particles and the use of rather a vague verb. From the total number of 43 phrasal verb types (57 tokens) found in the non-Czech learner writing, 6 occurrences require further elaboration.

The selection of the verb and adverb particle for *spell* in example (3) ***block out*** *a spell* is debatable. Prototypical nouns following the phrasal verb *block out* refer to *sun rays, sun light, noise. Remove/ break a spell* would be more appropriate:
(3) *d. In order to defeat him, Harry uses his mind to* **block out spells** *Voldemort casts on him, and since the type of wand*

Collocational deviations appear in examples (4) and (5): the learners used the phrasal verb *carry out* together with *a definition* and *Gothic theme w*hile the intended meaning in example (4) was *to perform, conduct.*

(4) *iterature. He used many themes and conventions to* **carry out** *the definition of Gothic writing. He deserves much*

Example (5) could be analyzed and categorized similarly - as the erroneous collocational range extension: there is no occurrence of such a collocation attested in the BNC.

(5) *rs. In his stories he uses a variety of themes to* **carry out** *the Gothic theme. In the story, "The Tell-Tale He*

Example (6) *come through* suggests the selection of the inappropriate lexical verb whereas the adverb particle is used correctly; *go through hardships* would be more appropriate.

(6) *ubt optimistic; having endured such hardships and* **came through** *it all as he did. The narrator was a very clever*

The preposition is superfluous in example (7) *find out on; find out* or just the simple verb *find* are the better alternatives.

(7) *ok was really addicting and I was always eager to* **find out on** *what will happen on the next page. Christopher Poa*

The last example (8) in the non-Czech learner writing *turn up the torture* was not found either in the BNC or ODPV. *Suffer* or *cope with hardships* would sound more authentic.

(8) *t also acted as an indication to his torturers to* ***turn up*** *the torture. His resourcefulness is the sole thing*

Overall, the analysis indicates that the errors are triggered by wrong collocations or the use of an inappropriate particle with the verb. The non-Czech learner sample contains several phrasal verbs, which correspond with the topic selection (*come back, fight back, go on, get back, find out)* or are closely associated with the life of characters, story-telling or some kind of reference to a book *(put the book down, pull the reader in).*

5.9 Phrasal verbs in the native speaker sample

Data collected from the native speaker sample follows the same procedures as both non-native samples.

5.9.1 Analysis of phrasal verbs in the native speaker sample

The native speaker material and data differ markedly as regards the frequency and variety of meanings of phrasal verbs, the scope of adverb particles. The analysis was carried out on 22 reviews written by native speakers, the sample totalling approximately 9 400 words. The analysis yielded 53 phrasal verb types and 64 tokens (see Table 24). Low repetition is obvious with the exception of the phrasal verb *find out* coming up 3 times and 8 phrasal verbs occurring twice (*come back, end up, get back, go down, grow up, pick up, throw away, wake up).* The rest of the phrasal verbs occur only once. From the LGSWE list of the most frequent phrasal verbs, only three phrasal verbs are present in the native sample: *find out, go on, come back*. The overall range of phrasal verbs suggests a somewhat informal style, with no traces of "academic" English phrasal verbs.

A number of less frequent phrasal verbs figure prominently. Most of these appear in ODPV (2010). Those not present in the dictionary were attested in the BNC. They are as follows: *brush away* (38); *lure back* (15); *push along* (14); *rush back* (98); *sweep out* (53); *talk off* (2); *travel around* (46); *walk away* (640)*, state back* (6); *shock out* was not attested in the BNC. The fact that these phrasal verbs are not listed among the ODPV entries has four possible explanations: some of the phrasal verbs are relatively new coinages, some of them are generally less frequent phrasal verbs and some of them are neologisms, a greater degree of creativity could be a possible factor due to the selected topic. Finally, the dictionary was compiled before the arrival of corpora. Some occurrences of less frequent phrasal verbs were found in the native writing: e.g. c*hurn out, piece together, brush away, sweep out.*

The variety of lexical verbs is also greater in the native sample than in the learner samples – it includes 43 different lexical verb types.

The combination of lexical verbs with different adverb particles provides the following results: *come* is the most productive - it is complemented by 4 adverb particles: *up, around, back, out*; *move* and *go* are followed by 3 adverb particles (*on, out, forward)*; *go (back, down, around*); *get* is followed by 2 adverb particles *(back, out),* (see Table 25).

Apart from the most common adverb particles (*out, up, back, down, away, off, on),* also phrasal verbs encompassing adverb particles such as *around (carry around, come around, go around, travel around)*; *forward (move forward)*; *together (piece together)*; *along (push along)* were found in the native speaker material. No occurrences with *through* and *in* or *about* were marked in the native data.

Table 24: Phrasal verbs in the native speaker sample

No.	PV types	Meaning	Tokens per type
1	*back up*	support	1
2	*break up*	disperse, go separate ways	1
3	*break down*	analyse in detail	1
4	*bring back*	remind one of stg	1
5	*brush away*	push aside	1
6	*calm down*	become calm	1
7	*carry around*	take from one place to a place	1
8	*carve out*	build (a career)	1
9	*clear away*	remove objects	1
10	*come around*	happen	1
11	*come back*	return	2
12	*come out*	happen	1
13	*come up*	arise	1
14	*check out*	go through	2
15	*churn out*	produce regularly in large amount	1
16	*cover up*	hide the real state of affairs	1
17	*end up*	finish	2

18	*find out*	discover	2
19	*get back*	return	2
20	*get out*	go away	1
21	*get caught up*	be involved in stg involuntarily	1
22	*go around*	go from one place to another	1
23	*go on*	continue	1
24	*go down*	1. reduce in force 2. come from a place	2
25	*grow up*	become adult	2
26	*head off*	start a journey	1
27	*lay out*	arrange	1
28	*lure back*	attract again	1
29	*make up (one's mind)*	decide	1
30	*move on*	progress	1
31	*move out*	leave the house you live in	1
32	*move forward*	progress to the front	1
33	*pick up*	take hold of, raise	2
34	*piece together*	assemble	1
35	*push along*	leave	1
36	*put down*	stop reading	1
37	*run out*	exhaust	1
38	*rush back*	return in a hurry	1
39	*set off*	start	1
40	*shock out*	surprise unpleasantly	1
41	*sit down*	be seated	1
42	*start off*	begin	1
43	*state back*	repeat what was said	1
44	*sweep out*	remove	1
45	*switch on*	connect an appliance	1
46	*switch off*	disconnect an appliance	1
47	*talk off*	divert the topic	1
48	*throw away*	get rid of	2
49	*travel around*	travel from 1 place to another place	1

50	*track down*	discover	1
51	*turn out*	appear, prove	1
52	*wake up*	become conscious	2
53	*walk away*	leave	1
Total			**64**

Table 25 presents the combination of lexical verbs with different adverb particles in the native speaker sample.

Table 25: Adverb particle productivity (combination with different verbs) in the native speaker sample

Lexical verb	Adverb particles	No. of particles
come	*up, around, back, out*	4
move	*on, out, forward,*	4
go	*back, down, around*	3
get	*back, out*	2

5.10 Comparison and summary of findings obtained from all three samples

The following aspects were investigated in all three samples: the range of phrasal verbs and their frequency, the variety of lexical verbs and the range of adverbial particles. Apart from that, the error analysis of non-standard occurrences in both non-native speaker samples was performed.

There are some differences between the two non-native groups: the range of phrasal verbs as well as the frequency in the Czech sample largely corresponds with the level of English the learners exhibit. Czech learners produced 25 phrasal verb types and 36 tokens (including 2 inappropriately used phrasal verbs). Such a low incidence of phrasal verbs is not surprising given that a prototypical pre-intermediate textbook presents approximately 30 phrasal verbs and intermediate language learners are supposed to be familiar with more than 60 phrasal verbs. In comparison with the Czech learner group, however, the non-Czech learner group produced generally twice as many phrasal verbs: 43 phrasal verb types, 57 tokens (including 6 inappropriately used phrasal verbs). Native speakers produced 53 phrasal verb types, 64 tokens (see Table 27). If the distribution of phrasal verbs in the native speaker sample is taken as the norm, then the Czech speakers' use of phrasal verb types is at 47.2 per cent, the use of phrasal verb tokens at 56.2 per cent and that of lexical verbs at 48.8 per cent of this norm. In other words, the distribution of phrasal verbs in the Czech sample is half that of the native speakers' in all respects. Further details are provided in Appendix 3a which shows all phrasal verbs found in all three samples together with their frequency.

The combination of lexical verbs with different adverb particles was also investigated. As regards the range of lexical verbs in the Czech, non-Czech learners and native speakers, a greater variety was obvious in the native sample. Czech learners produced 21 lexical verb types and non-Czech learners 32, native speaker sample comprises 43 different lexical verb types to form phrasal verbs.

Both learner samples contain phrasal verbs which are listed in LGSWE (1999) among the most frequent phrasal verbs in conversation and fiction. The Czech sample contains 6 such phrasal verbs: *find out, get back, set up, point out, turn out, go on;* the non-Czech sample 5 such phrasal verbs *come along, pick up, set up, find out, go on,* whereas only 3 phrasal verbs from LGSWE list were discovered in the native speaker sample (see Appendix 3b).

Despite the fact that non-Czech learners produced a high number of less common phrasal verbs, a native speaker's range of phrasal verbs is more diverse and idiomatic, non-native speakers produce phrasal verbs which tend to be more literal. The choice of phrasal verbs also reflects the style adopted. With very few exceptions related to academic prose - the phrasal verbs used mainly by the upper-intermediate Czech learners (i.e. *sum up, point out)*, the style which was adopted by both the native speakers and non native speakers is more colloquial than formal (see Table 26).

Table 26 shows the list of the most frequent phrasal verbs in all three samples. Some of the most frequent phrasal verbs in the samples belong to the commonest phrasal verbs in English : *find out, get back, set up, point out, turn out, go on, come along, pick up,*

Table 26: The most frequent phrasal verbs in the Czech, non-Czech and native speaker sample

Order	PV types CZL	Tokens per type CZL	PV types NCZL	Tokens per type NCZL	PV types NS	Tokens per type NS
1	*find out*	4	*go on*	6	*find out*	3
2	*fall in love*	4	*put down*	3	*wake up*	2
3	*grow up*	3	*come back*	3	*come back*	2
4	*go back*	2	*set up*	2	*check out*	2
5	*come out*	2	*carry out*	2	*end up*	2
6	*go on*	2	*cut off*	2	*get back*	2
7	*set up*	2	*find out*	2	*go down*	2
8	*sum up*	2	*get back*	2	*grow up*	2
9	*wake up*	2	-	-	*pick up*	2

Concerning adverb particles, Czech learners used quite a limited set of adverb particles (*up, out, back, down, on, away, in*). The particles are according to LGSWE (1999,412) the most common ones. The relatively high frequency of the particle *down* in the Czech sample might be explained due to the verbatim translation from Czech, rather than the learners' familiarity with the phrasal verb. Non-Czech learners produced phrasal verbs with even less common adverb particles. Apart from *up, back, on, out, down, off , in,* phrasal verbs with adverb particles such as *through, away, forward* and *behind* were detected and the variety of adverb particles in non-Czech learner writing even exceeded the native speaker repertoire by one particle. The range of native speaker adverb particles is also broad, though; it encompasses less common adverb particles: next to the commonly used adverb particles, such as *out, up, back, down, off,on* also *away, around, along, forward* and *together* occur (see Appendix 3b).

The combination of lexical verbs with different particles is in line with the initial assumptions. Czech learners produced only 4 lexical verb types which occur with two different adverb particles, i.e. *come, run, get, go.* Non-Czech learners as well as native speakers were more productive in this respect. Non-Czech learners used *come* with 4 adverb particles, *set/put* with 3*, go* and *turn* with 2 adverb particles. Native speakers used *come* with 4 adverb particles, *go* and *move* with 3, *break* and *get* with 2. Also the findings obtained from LGSWE gives evidence that *come* occupies the first place in terms of productivity in the language of conversation and fiction.

The differences have already been accounted for. A few similarities among the non-native samples can be found: the inappropriately used phrasal verbs in both non-native samples arise mainly due to the inappropriate extension of the collocational range, in particular, the use of rather a vague verb instead of a proper phrasal verb, and the choice of an inappropriate phrasal verb (either a wrong particle or a wrong lexical verb). On the other hand, the number of inappropriate uses of phrasal verbs is extremely low and thus sweeping generalizations should be avoided. Both non-native sample corpora share the following phrasal verbs: *grow up, bring up, come back, get back, go on, sum up, point out*. Several phrasal verbs are related to the topic, hence a certain degree of topic sensitivity can be seen in both learner samples.

Only a small number of similarities appear in all three groups: a small number of phrasal verbs (5) which occur in all three samples: *come back, end up, find out, get stg back, go on.* Secondly, it is the adopted style in the selection of phrasal verbs suggests a relatively neutral tone in all three samples.

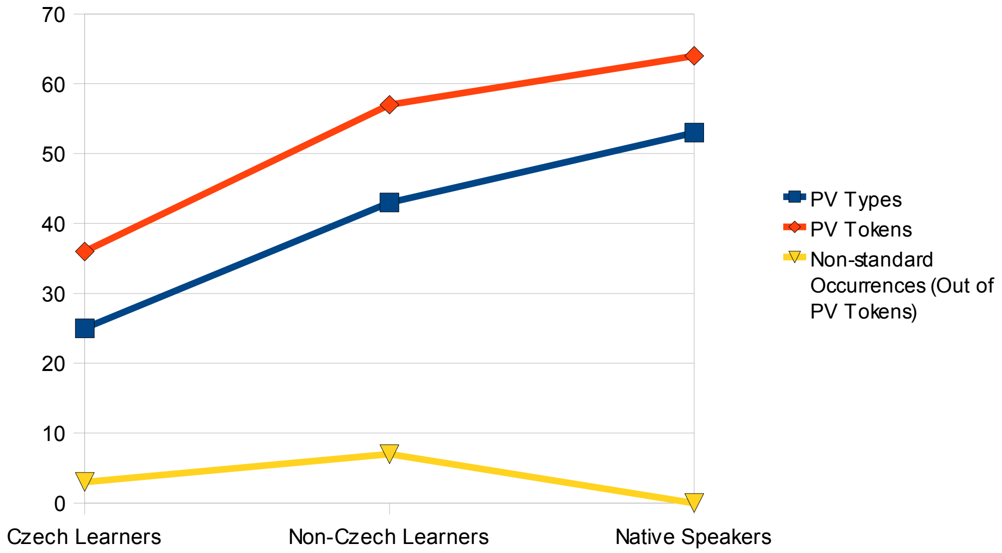
Table 27 shows the figures related to the phrasal verb types and tokens (column 2, 3 respectively) in all three sample corpora. The forth column presents the number of different lexical verbs, the last column contains dubious cases of phrasal verbs created by non-native speakers.

Table 27: Phrasal verb types, tokens and lexical verb types in all three samples. Dubious cases of phrasal verbs in the learner samples.

Sample	PV types	Tokens per type	Lexical verbs	Dubious cases
Czech learners	25	36	21	2
Non-Czech learners	43	57	32	6
Native speakers	53	64	43	-

Figure 2 reflects the frequency of phrasal verb types, tokens and non-standard occurrences in the Czech, non-Czech learner and native speaker sample: the number of phrasal verb types and tokens is the highest in the native speaker sample. Non-standard occurrences are represented by yellow colour, phrasal verb types are highlighted in blue colour, tokens in red colour. The number of types and tokens is the highest in the native speaker sample while the non-Czech learner sample contains most non-standard occurrences.

Figure 2: Phrasal verb types, tokens and non-standard occurrences in all three samples



5.11 Collocation, colligation, semantic and pragmatic associations of phrasal verbs

Using Sinclair's theory of the extended lexico-grammatical unit, the phrasal verbs *end up* and *churn out* will be discussed in this section to show the qualitative differences between the learner and the native speaker use of phrasal verbs. The phrasal verb *end up* occurs in all three samples, *churn out* does not occur in two of the samples. The reason why such a detailed analysis of these two phrasal verbs is carried out in this section is because, to quote Sinclair, (1991, 78) "Each sense of the phrase is co-ordinated with a pattern of choice that helps to distinguish it from other senses. Each is particular; it has its uses and its characteristic environment". This, of course, makes phrasal verbs very difficult for learners. Examples from ODPV (2010) will be presented and compared with the phrasal verb *end up* in both non-native and native samples. *Churn out* occurs only in the native sample and in this section it is used as another example of how very intricately language is organized.

End up is a relatively frequent phrasal verb; besides, it is one of the very few phrasal verbs which occurs in all three samples. The phrasal verb *churn out*, on the other hand, was used in the native speaker sample only, and so is a good example of a phrasal verb typically used by a native speaker but avoided by (or unfamiliar to) learners of English.

End up

The meaning of the phrasal verb *end up*, which occurs in all three samples, is paraphrased in ODPV (2010, 111*)* as 1. *finally be or do stg; finish as (dead, bankrupt, in jail, like everyone else;* e.g. *That' s how you will end up, my boy – black-hearted, evil-minded and vicious.; If we take her too seriously, we' ll end up in a mental home.; He ended up in prison.).*

The following occurrences comprising *end up* were retrieved from the individual samples – Czech, non-Czech and native speaker samples:

Example (9) was retrieved from the Czech sample:

(9) *lay was a traditional comedy, Wilder's characters* ***end up*** *tragically, yet the spirit of the book is optimist*

Example (10) was attested to the non-Czech learner sample:

(10) *goes on and the mysteries grow, but they finally* ***ended up*** *in a Swiss depository bank with a key from Sophie*

Examples (11) and (12) were obtained from the native speaker sample:

(11) *think is the point they're trying to make. You'll* ***end up*** *starting a lot of your sentences with*

(12) *aid, highly educated government official, but who* ***ended up*** *poor (this is his "poor dad"). His be*

Examples from the BNC were consulted for reference. Altogether, the search in the BNC yielded approximately 3166 occurrences of the phrasal verb *end up* reflecting the following prototypical characteristics of *end up:* it relates to possible future consequences or things which happened in the past. If the phrasal verb suggests a future event; it is usually preceded by a modal verb, modal idiom or semi-auxiliaries, suggesting that something negative or unpleasant might occur in the long run (*likely to, will, can, could*). However, a great number of occurrences comment on past events which in many cases were undesirable; very often participles or adjectives follow, as well as adjuncts of place (e.g. *she ended up crying, looking silly, committing suicide; in prison etc.).* According to Oxford Advanced Learner Dictionary (2005), this suggests the meaning *to find oneself in a place or situation* *that one did not expect to be in.* Pragmatically speaking, *end up* is very often used with a negative connotation. Despite this, some contexts where *end up* occurs do not necessarily have to be negative. The following examples listed in the OALD *after working her way around the world, she ended up teaching English as a foreign language* or *they are traveling across Europe by train and finally planning to end up in Moscow* are good examples of neutral context of this phrasal verb.

If the phrasal verb *end up* from the Czech sample *characters end up tragically* is compared with the dictionary definition and examples, its use corresponds with the definition *to find oneself in a place or situation that one did not expect to be in.* This is clearly demonstrated by the use of the adverb *tragically.* However, the phrasal verb is used neither in the past or future, and it is not preceded by a modal or semi-modal. In other words, its colligation is not very typical. *End up* in the native speaker sample fits the dictionary description: it is used in the past, it is followed by a negative adjective (*poor*). *End up* found in the non-Czech sample is used in a neutral sense *ended up in a Swiss depository bank.*

Churn out

The phrasal verb *churn out* was discovered only in the native sample. The BNC search yielded 98 occurrences.

The example retrieved from the native speaker sample:

(13) *s books, but I think this is the best one. He has* **churned out** *quite a few more books in the last fe*

The meaning of *churn out* given in ODPV (2010, 58) is *to produce something regularly in large amounts.* Simultaneously, the BNC concordance lines amount to 00 occurrences of this phrasal verb. The concordance lines reveal that the use of *churn out* does not change with time time. Although the highest number of occurrences is formed using the -*ing* form, either in the participle or gerund form, several examples are in the past or present perfect tense. Prototypical collocates which usually follow the phrasal verb refer mainly to books, records, new lines, letters etc. Semantically speaking, it conveys the meaning of s*omething being produced in large amounts and consistently* with the additional implicit meaning that *something is produced in large amounts but simultaneously something which is of low quality, not worth much money* (OALD 2005). All of these, as might be expected, are found in the occurrence of *churn our* in the native speaker sample.

5.12 Prepositional verbs and phrasal-prepositional verbs

The following sections present the data related to the prepositional and phrasal-prepositional verbs in all three samples.

5.12.1 Preliminary issues and procedures

Bearing in mind the different nature of phrasal and prepositional verbs, we shall focus on some aspects that will differ from those analysed in the phrasal verb section. In brief, one of the reasons why phrasal verbs present a major stumbling block for learners is that they usually have several meanings, the second problem is, at least in some cases, their opacity. It does not imply that learners do not consider prepositional verbs difficult but they do so for different reasons - it is the choice of the appropriate preposition which is not entirely trouble-free (Waibel 2007). Different traits of both types of prepositional and phrasal-prepositional verbs predetermine the focus of the analyses.

It is assumed that the number of prepositional verbs will be higher than phrasal verbs not only in the native but also in the non-native samples. Apart from the error analysis, which perhaps will be more elaborate and will present more errors than the phrasal verb section, a large number of prepositional verbs provide the opportunity to carry out a semantic taxonomy and allow to confirm or refute the assumption whether learners use the prepositional verbs which occur most commonly or whether the prepositional verbs they are familiar with belong to the less frequent in the language. LGSWE (1999) states that the most frequent semantic domains are activity prepositional verbs. The primary focus of this investigation is thus to find out whether learners produce a great number of activity prepositional verbs and whether they are used appropriately. Errors of different origin are likely to involve the following cases (Nesselhauf 2005):

4. prepositional verb is used for a different prepositional verb;
5. prepositional verb is used where simple verb would be more appropriate;
6. simple verb is used where prepositional verb would be better;
7. phrasal prepositional verb is used where prepositional verb would be more suitable;
8. prepositional verb is used instead of phrasal-prepositional verb.

The subsequent procedure of sorting the prepositional and phrasal-prepositional verbs manually was adopted in all three samples and all prepositional verbs and phrasal-prepositional verbs were checked with ODPV (1993). The initial investigation shows that there are almost twice as many prepositional verbs than phrasal verbs in all three samples. This result is in accordance with the findings presented in LGSWE. Namely, that prepositional verbs occur more frequently in the language than phrasal verbs. Additionally, as opposed to phrasal verbs, prepositional verbs occur frequently in all registers.

5.12.2 Prepositional verbs and phrasal-prepositional verbs in the Czech learner sample

The Czech sample (37 texts) contains 60 prepositional verb types, 111 prepositional verb tokens, out of which 6 occurrences are considered incorrect (see Appendix 3c). Some of the prepositional verbs cannot be found in ODPV (2010), they are attested in the BNC, though: *borrow from, get rid of.* Only one phrasal-propositional verb *run away from (1)* was found in the Czech sample. The almost total absence of phrasal-prepositional verbs in the sample confirms the claim (LGSWE 1999) that phrasal-prepositional verbs occur even more rarely than phrasal verbs.

As opposed to the phrasal verb analysis, prepositional verbs in the Czech sample are treated as a whole regardless of the level of the Czech speakers.

The majority of the prepositional verbs in the Czech sample occur once or twice. Some of them occur more than twice and relate to the language commonly used in reviews, some prepositional verbs relate to the characters: 1. *recommend to (9), live in (8), base on (7), write about (7), take place in (5), fall in love (4), go to (3), look for (3), talk about (3), listen to (3),* (see Table 23). According to the corpus findings (LGSWE 1999, 416-419), four prepositional verbs out of this list *base on, look for, talk about, listen to* belong to the most frequent verbs. The relatively high number of occurrences of the prepositional verbs *recommend to* could be explained by the choice of the topic since students abundantly use the phrase *I would like to recommend this book/film etc*. It is thus obvious that the phrase often represents “a filler” in the non-native essays while this phrase is almost missing in the native writing. The same is true of *write about, talk about.*

The primary concern of this analysis is therefore to establish whether the prepositional verbs pose the same kind of obstacle for learners as phrasal verbs do. In this respect, the number of prepositional verbs and the appropriate use of prepositions will be investigated. Additionally, drawing on the data, a semantic taxonomy of prepositional verbs will be constructed. Since LGSWE (1999) gives evidence that activity prepositional verbs are the most common verbs, this semantic analysis of prepositional verb is expected to show to what extent learners use the commonest prepositional verbs or whether they are familiar also with those occurring less frequently in the language. For example, the corpus findings presented in LGSWE (1999, 416-19) suggest that the prepositional verb *base on* is relatively frequent in newspaper English, it is also very frequent in academic English; *look for* occurs abundantly in fiction; *talk about* occurs very frequently in the language of conversation, fiction and quite frequently in newspaper English; *listen to* is especially frequent in the language of conversation, fiction, and it occurs relatively often in newspaper English. Apart from that, other 17 prepositional verb types found in the Czech sample (with lower frequency, though) occur in the language very frequently: *ask for, begin with, belong to, come from, deal with, fall into, get over, involved with, know about, look after, look at, put into, talk to, think about, think of, wait for, work on.* Overall, the selection of the prepositional verbs in the Czech writing indicates that the verbs occur both in the language of conversation as well as written language (see section 5.14).

Idiomatic multi-word verb combinations treated in LGSWE (1999) as a separate category scarcely occur in the Czech sample (*let the cat out of the bag, take place in).* However, since they contain a preposition and a verb, they were included in this investigation of prepositional verbs as well.

The following examples (14) – (19) are considered inappropriate and require further clarification:

(14) *e this very much. Although I don´t read so much I* ***dived into*** *it. It was very exciting.*

In example (14), the learner selected the prepositional verb inappropriately - a different prepositional verb would be required; one usually gets engrossed or immersed in a book rather than dives into a book.

In examples (15-16), the learner used a prepositional verb where a simple verb without the preposition *to* is necessary – the preposition is redundant; s*chools are attended* while *matters attended to*. In example (17), the preposition proves superfluous.

(15) *tory. Harry finds out that he is a wizard and is* ***attended to*** *the school of magic - The*

Hogwarts. In this boo

(16)'*s are simple and practical. Together they always* ***proceed with***. *Tom has imaginative plans. Toms Sawyer beli*

In example (17) the learner used a non-existing prepositional verb (*resist* is not followed by *from);* however, also the simple verb *resist* is considered inappropriate in this combination; people usually *can' t help laughing* rather than *resist laughing.*

(17) *to the world of imagination. I can't resist from laughing because it is very funny and also always actual. I*

The collocation in example (18) sounds odd:

(18) *of Huckleberry Finn" and Tom Sawyer* ***brings*** *their unique characteristics* ***into*** *this comical friend ship givi*

In example (19) the learner omitted the preposition *with:*
(19) *rry hated each other but than they* ***fell in love*** *each other. The rest of the film is abut their interesting re*

5.12.2.1 Semantic types of prepositional verbs in the Czech learner sample

In the following section, the semantic types of prepositional verbs found in the sample will be subject to investigation. Since LGSWE (1999) provides evidence that the most frequent type of multi-part verbs is prepositional verbs, this semantic analysis of prepositional verb used by Czech speakers is expected to show to what extent learners use the commonest prepositional verbs, whether they have a good command of these prepositional verbs or whether learners are familiar also with those less frequently ones occurring in the language.
Here is the outline of the semantic groups (according to LGSWE) of prepositional verbs found in the Czech learner sample:

1. Activity prepositional verbs (29 types, 37 tokens)
attend to (1), borrow from (2), bring into (1), come across (1), cut through (1), deal with (2),

dive into (1), escape from (1), get over (1), get rid of (1), get to (2), get out of (1), get control of (1), go to (3), introduce into (1), let the cat (1), look after (1), look at (2), look for (3), proceed with (1), release from (1), return to (1), search for (1), strap on (1), take out (1), trap in (1), wait for (1), work on (1), *put into*

2. Mental prepositional verbs (12 types, 20 tokens)
care about (1), fall in love (4), focus on (1), forget about (2), gather from (1), know about (2), listen to (3), resist (1), take into account (2), think about (1), think of (1), wish for (1)

3. Causative prepositional verbs (2 types, 3 tokens)
come from (2), make into (1)

4. Communicative prepositional verbs (7 types, 24 tokens)
ask for (1), introduce to (2), recommend to (9), talk about (3), talk to (1), tell about (1), write about (7)

5. Occurrence prepositional verbs (4 types, 15 tokens)
fall into (1), live in (8), take part in (1), take place in (5)

6. Existence/relationship prepositional verbs (5 types, 11 tokens)
base on (7), belong to (1), introduce into (1), involved with (1), share with (1)

7. Aspectual prepositional verbs (1 type, 1 token)
begin with (1)

LGSWE (1999, 419) provides statistics that the most frequent semantic group in the BNC is activity verbs which are evenly distributed across all registers; communication verbs and mental verbs are also used abundantly in the language of conversation, in fiction and journalistic English (with the exception of academic prose - the number of communicative verbs in this group is quite low). Causative and existence/relationship verbs occur mainly in academic prose. The findings obtained from this analysis confirm that activity verbs represent the most frequent category of prepositional verbs (approximately 33.1 per cent of all prepositional verb tokens in the sample are activity verbs). The second biggest semantic group in the Czech sample is the group of communicative verbs (21.6 per cent tokens), it is closely followed by the group of mental verbs (18.1 per cent of all tokens). On the other hand, the lowest incidence of prepositional verbs was found on the part of causative verbs (2.7 per cent of tokens), aspectual verbs even less (0.9 per cent). Some prepositional verbs can be placed into more than one semantic domain. For instance, even though *deal with* is listed in the category of activity verbs in LGSWE (1999), it could fall into the mental verbs group.

The survey and comparison of the distribution of semantic groups in the samples are given in Table 31 below.

5.12.3 Prepositional and phrasal-prepositional verbs in the non-Czech learner sample

The non-Czech sample consisting of 19 texts contains 90 prepositional verb types, 130 tokens (see Appendix 3d). The sample also contains 5 phrasal-prepositional verb types (*lead up to, make up for, look down on, come up with, look forward to)*, 5 phrasal-prepositional verb tokens (no inappropriate uses were found).

The most frequent prepositional verbs (see Appendix 3d) include the following prepositional verbs with the frequency of 2 and more: *go to (7), deal with (6), relate to (5), base on (4), discriminated against (4), look at (4)*. The high number of occurrences of *deal with* are used when the learners compare *the challenges characters need to face or a problem to be tackled to the contemporary issues we often have to deal with.*

According to LGSWE (1999), four of the previously mentioned frequent prepositional verbs in the non-Czech sample are very common in the BNC: *deal with* (especially in academic prose and fiction); *relate to* (commonly occurs in academic prose); *base on* (comparably frequent in newspaper, particularly frequent in academic prose); *look at* is the most frequent verb in all four registers. Apart from that, the following 25 prepositional verb types found in the non-Czech sample with a lower frequency belong to the most frequent verbs in the BNC according to LGSWE (1999): *agree with, begin with, believe in, come from, depend on, derive from, get into, go through, go on, happen to, include in, involve in, listen to, live with, look for, know about, occur to, result in, say about, speak of, speak to, suffer from, think about, wait for*, *compare with*.

There are 15 occurrences of incorrect phrasal verbs in the sample. Some of the inappropriate uses involve cases where the same phrasal verb is repeatedly used incorrectly. Dubious cases arise due to the use of a different prepositional verb than required by the context, the selection of a prepositional verb where a simple verb would be more appropriate, the phrasal-prepositional verb where the preposition is superfluous, or last of all, cases where a simple verb would fit in better than a prepositional verb.

There are 11 occurrences (examples 20-31) where an incorrect prepositional verb was selected instead of a more appropriate prepositional verb (the lexical element is used correctly in 6 occurrences while the preposition is incorrect; the last case involves also a wrongly selected verb).

The examples (20) and (21) are very similar, both were produced by the same learner: the replacement of a preposition is required in both cases, the lexical element is correct - *enter for tournaments* would be more appropriate.

(20) *Moody put Harry's name in the Goblet of Fire to* ***enter*** *him* ***in*** *the tournament, and when Harry's name was d*

(21) *izards that protested Harry's eligibility. He was* ***entered into*** *the tournament by "Mad-Eye" Moody, who was wo*

Set out on a journey or *go on a journey* would be more appropriate in example (22):

(22) *nged overnight and his life is shattered. He* ***goes onto*** *a journey with a storyteller, Brom, and his life o*

In example (23) *head for/towards* would be more appropriate:

(23) *rom the evil Balrog. They must go on without him,* ***heading*** *south,* ***into*** *Lorien, a forest of elves. The lady G*

In example (24) *people hold on to ideas* not *hold on ideas* :

(24) *of rejection of one's position and* ***holding on*** *the idea of a return illustrate the image of a man constant*

In example (25), the preposition *from* would suit the context better.

(25) *strong Dragon Rider to defeat King Galbatorix and* ***free*** *the empire* ***out of*** *his clutches. Theme: The theme*

Example (26) shows not only the preposition which is used inappropriately but also the verb;

the replacement of a completely different prepositional verb is required: *covered with blood* would be a good alternative:

(26) *which suggested the Red Death. "His vesture was* ***dabbled in*** *blood-and his broad brow, with all the features*

The following examples (27) – (30) refer to the same mistake and would be correct with the preposition *against*; all of them were produced by one learner:

(27) *r their gender. Today people of Arabic decent are* ***discriminated for*** *either their looks or their religion. This*

(28) *ed for their religion. Not so long ago women were* ***discriminated for*** *their gender. Today people of Arabic decent ar*

(29***)*** *discriminated for their skin color. Wiccans were* ***discriminated for*** *their religion. Not so long ago women were dis*

(30) *lates to today. Long ago African- Americans were* ***discriminated for*** *their skin color. Wiccans were*

The following examples (31- 34) show occurrences where a prepositional verb requires a change for a simple verb. A completely different lexical verb is required in some cases.

Examples (31) and (32) require the replacement of a simple verb *avoid (being punished);* both examples were produced by one learner.
(31) *In the first chapter of the book, Tom tries to* ***keep*** *himself* ***from*** *being punished for eating the jam, by*

(32*) some kind of mischief, yet he somehow manages to* ***keep himself from*** *being punished, and rather seem like*
In example (33), the preposition is superfluous:

(33) *book. After finishing this, I am very anxious to* ***begin on*** *the fifth Harry Potter book: The Order of the P*

Not only the preposition is redundant in example (34); also a different simple verb would be subject to the alteration (*inflict pain*); *transfer onto* does not sound native like in the collocation with *pain.*

(34*) n some miraculous way,* ***transfer*** *their soul' s pain* ***onto*** *yours. These stories triggered Alex's mind to bel*

5.12.3.1 Semantic types of prepositional verbs in the non-Czech learner sample

Also the non-Czech learners' results confirm that activity verbs form the most abundant group in the language. Namely, 130 tokens include 58.4 per cent of activity verbs. The second largest group is represented by the existence/relation verbs (16 per cent of tokens), the third position is occupied by mental verbs (10.8 per cent of tokens).

The prepositional verbs in the non-Czech learner sample may be divided into the following semantic groups:

1. Activity verbs (49 types, 76 tokens)

derive from (1), go through (1), go to (7), go on (1), go onto (1), guide through (1), head into (1), head towards (1), hide behind (1), hide from (1), hold to (1), live with (2), look at (4), look for (2), keep from (2), keep out of (1), note for (2), prevent from (1), proceed on (1), protect from (1), resort to(1), retire from (1), saturate with (1), bring to life (2), cling to (1), come to the point (1), come through (1), dabble in (1), deal with (6), devote to (1), discriminate against (4), distract from (1), draft into (1), escape from (2), fall into (1), free out of (1), get into (1), get to (1), safe from (1), search for (2), stay with (2), strip of (1), stumble upon (1), transfer onto (1), wait for (2), look behind (1), pin on (1), rest upon (1), clear of (1).

2**.** Mental prepositional verbs (13 types, 14 tokens)

agree with (1), believe in (1), come to (1), dream of (1), focus on (1), grow in (1), listen to (1), know about (1), reflect on (1), suffer from (1), take into account (1), think about (2), thrust with (1)

3. Communicative prepositional verbs (5 types, 5 tokens)

confess to (1), describe as (1), say about (1), speak of (1), speak to (1)

4. Causative prepositional verbs (4 types, 4 tokens)

come from (1), contribute to (1), depend on (1), result in (1)

5. Occurrence prepositional verbs (4 types, 5 tokens)

embedded in (1), happen to (1), occur to (1), take place in (2)

6. Existence/relationship prepositional verbs (12 types, 21 tokens)

attribute to (2), base on (4), compare with (1), correspond with (1), develop into (1), distinguish between (1), include in (1), involve in (1), relate to (5), turn into (1), change into (1), share with (2)

7. Aspectual prepositional verbs (3 types, 5 tokens)

begin on (1), begin with (2), enter into (2)

For comparison of the distribution of semantic groups see Table 31 below.

5.12.4 Prepositional and phrasal-prepositional verbs in the native speaker sample

We found 101 prepositional verb types and 159 prepositional verb tokens in the native sample (22 texts). Further, 8 phrasal-prepositional verb types and 8 tokens (*look forward to, move away from, come up with, stand up for, pick up on, go on to, live up to, go down to)* were identified in the native speaker sample. Some idiomatic multi-word phrases (e.g. *come to a halt, come to an end, make sense of, put into action*) appeared in the sample as well and since they contain a preposition they are included in the prepositional verb analysis. For further details, see Appendix 3e, which provides the list of all prepositional verbs in the native sample.

The prepositional verbs with the frequency higher than two are more abundant than in the Czech and non-Czech samples. They include the 13 following prepositional verbs: *1. share with (9), 2. talk about (7), 3. deal with (6), invest in (6), 4. focus on (3), listen to (3), look at (3), look for (3), live in (3), happen to (3), put into action (3), think about (3), tell about (3)*. Out of these, *talk about, deal with, listen to, look at , look for, happen to, think* *about* belong to the most frequent prepositional verbs according to corpus findings (LGSWE 1999, 419). Apart from these, other 22 prepositional verbs found in the native sample occur very frequently in the language: *apply to, associate with, base on, begin with, come from, cope with, compose of, get into, go through, hear of, live with, provide for, relate to, result in, say about, send to, speak to, stand for, suffer from, talk to, think of, turn to.*

5.12.4.1 Semantic types of the prepositional verbs in the native speaker sample

The following results emerged from the semantic typology in the native speaker sample (see also table 31 below):

1. Activity prepositional verbs (49 types, 70 tokens*)*

act on (1), apply to (1), bump into (1), cling to (1), come across (2), come by (1), come out of (1), come to an end (1), come to a halt (1), conclude with (1), deal with (6), decide for (1), do with (1), endear to (1), engrave into (1), experiment with (1), fend for (2), get into (1), get on (1), go into(2), go through (1), go toward (1), invest in (6), lend to (1), live for (1), look at (3), look for (3), look to (1), move towards (1), play on (1), prepare for (2), protect from (1), provide for (1), put into action (3), react to (1), reach for (1), release from (1), return to (1), reveal to (1), root for (1), search for (2),, send to (1), sink into (1), struggle for (1), stumble across (1), take out (1), venture up (1), weave into (1), watch for (1).

2. Mental prepositional verbs (19 types, 29 tokens)

accustom to (1), cope with (1), dream about (1), focus on (3), fall in love (2), forget about (1), hear of (2), identify with (1), keep in mind (1), listen to (3), make sense of (1), muse on (1), reflect on (1), remind of (1), seek after (1), suffer from (1), think of (2), think about (3), , turn to (2)

3. Communicative prepositional verbs (10 types, 20 tokens)

convict of (1), say about (1), speak about (2), speak to (1), talk about (7),talk to (1), tell about (3), warn of (1), write about (2), introduce to (1)

4. Causative prepositional verbs (5 types, 6 tokens)

benefit from (1), come from (2), evolve from (1), result in (1), turn into (1)

5. Occurrence prepositional verbs (4 types, 8 tokens)
embedded in (1), flash across (1), happen to (3), live in (3)

6. Existence/relationship prepositional verbs (12 types, 24 tokens)
associate with (1) , base on (1), compare to (2), compose of (1), connect to(1), couple with (1), grow into (2), live with (2), relate to (2), stand for (1), thrust into (1), share with (9)

7. Aspectual prepositional verbs (2 types, 2 tokens)
begin with (1), embark on (1)

Not only both non-native samples, but also the native sample prove that activity verbs represent the largest semantic group of prepositional verbs. In the native sample, 44.1 per cent of tokens belong to this semantic group; the second position is occupied by mental verbs (18.3 per cent of tokens) followed by the existence/relationship group (15.9 per cent of tokens). The results show that not only native speakers but also both learner groups have a good command of prepositional verbs. The distribution of the prepositional verbs across different semantic domains used by the learners shows that prepositional verbs do not pose such an obstacle for them. Most importantly they show that learners are familiar with the commonest verbs, and use them more or less appropriately. They also produce some less frequent prepositional verbs from such semantic domains as existence/relationship, aspectual, causative and occurrence.

5.13 Comparison of prepositional verbs and phrasal-prepositional in all three samples

Both non-native samples and the native speaker sample comprise almost twice as many prepositional verbs in comparison with phrasal verbs. Such findings confirm the claim that, generally speaking, prepositional verbs occur more frequently in the language in comparison with phrasal verbs.

The Czech sample comprises 60 prepositional verb types, 111 tokens including only 6 occurrences of prepositional verbs used inappropriately. The non-Czech sample contains 90 prepositional verb types, 130 tokens including 15 inappropriately used prepositional verbs. There are 101 prepositional verbs types, 159 tokens in the native speaker sample (see Table 28, Figure 3).

Table 28: Prepositional verb types and tokens in all three samples

Sample	Prepositional verb types	Prepositional verb tokens
Czech learner	60	111
Non-Czech learner	90	130
Native speaker	101	159

All samples contain some prepositional verbs which, according to LGSWE (1999), occur very frequently in the language: the Czech sample contains 21 such prepositional verb types, the non-Czech sample 29, the native sample 29 (see Table 29). A closer look indicates that the distribution across genres is very similar to the findings presented in LGSWE: out of 21 very frequent prepositional verbs, the Czech sample offers 10 prepositional verbs which occur frequently in conversation, 11 in fiction, 13 in newspaper and 7 in academic prose. Out of 29 very frequent prepositional verbs in the language, the non-Czech learner sample includes 11 prepositional verb which occur very frequently in conversation, 16 in fiction, 15 in newspaper and 11 in academic prose; out of 29 very frequent prepositional verbs, the native sample comprises 12 prepositional verbs which are abundant in conversation, 15 in fiction, 17 in newspaper and 12 in academic prose (see Table 29). The adopted style of prepositional verbs thus cannot be specified. Moreover, LGSWE indicates only the most frequent prepositional verbs and the samples under scrutiny include prepositional verbs which are not listed in LGSWE, thus their register distribution remains questionable.

Table 29: Frequent prepositional verbs in the BNC also identified in all three samples

Sample	Frequent prepositional verbs in the BNC	CONV	FICT	NEWS	ACAD
Czech learner	21	10	12	13	7
Non-Czech learner	29	11	16	15	11
Native speaker	29	12	15	17	12

As regards the range of lexical verbs to form prepositional verbs, there are 45 lexical verb types in the Czech sample, 75 lexical verbs types in the non-Czech sample and 85 lexical verb types in the native sample. Again a greater diversity is obvious in the native sample.

The data concerning the productivity of the lexical verbs is not of much interest: the verbs *get, take* in the Czech sample combine with four different prepositions, *look* with three, *come, fall, introduce, talk, think* with two, the rest of the lexical verbs combines only with one particle. Similar results are obtained from the non-Czech sample: *go* combines with four different prepositions, *come, look* with three, *get, hide, keep, speak, take* with two, the rest only with one preposition. Perhaps surprisingly, the native sample produces similar results: *come* combines with 5 different prepositions, *live, look, go* with 3, *get, speak, talk, think, turn* with 2, the rest of the lexical verbs combine with only one preposition. The obtained results are in line with the data in LGSWE which says that almost no verbs are particularly productive to form prepositional verbs (LGSWE 1999, 421).

A few more differences deserve to be commented upon. In comparison with the non-native samples, the native sample contains a few idiomatic prepositional phrases such as *come to a halt, come to an end, make sense of, put into action* (the Czech and non-Czech samples contain only *take into account, take place in, let the cat out of bag).*

There are only few similarities that can be pointed out. Only a few prepositional verbs are shared by all three samples: *base on, begin with, come from, deal with, look at, look for, share with, search for, think about.* Some of them represent the most frequent prepositional verbs, e.g. *look at, think about* (LGSWE 1999, 416-19).

As far as the inappropriate use of prepositional verb, it has to be stated again that the number of inappropriately used prepositional verbs is not significant. Only very few prepositional verbs are considered incorrect in both non-native samples; the majority of such cases involve the use of an inappropriate preposition.

As regards the semantic types, a great degree of similarity can be observed in all three samples and the results are in accordance with LGSWE (1999). Table 31 below shows that activity verbs represent the largest group in all samples. Also mental verbs form quite a big group in all three samples. There is only a slight difference in view of communicative verbs, which occur relatively often in the Czech sample, in the native sample they form a medium-size group, however, they are almost missing in the non-Czech sample. The number of verbs in the remaining of the semantic groups is quite low. The outcome of the analysis can be supported by the data provided by LGSWE. It presents evidence that activity and mental verbs are abound in all registers, communicative verbs are also used very often in all registers except for academic prose. The same applies to the mental group, however, it occurs in academic prose relatively often. Regardless of the registers and the participants' nationality, the results correspond with the findings presented in LGSWE (1999, 419).

All in all, the results of comparing the semantic types found in the samples show both learner groups have a good command of prepositional verbs and are not inferior to the native speakers in this respect. The distribution of prepositional verbs across different semantic domains used by learners shows that prepositional verbs do not pose such an obstacle for learners. Even more importantly, they show that learners are familiar with the commonest verbs, and use them more or less appropriately. The learners also produce some less frequent prepositional verbs from semantic domains such as existence/relationship, aspectual, causative or occurrence.

Phrasal-prepositional verbs occur rather rarely in the three samples: only one occurrence was noted in the Czech sample, 5 occurrences in the non-Czech sample, the native sample contains 8 phrasal-prepositional verbs *look forward to, move away from, come up with, stand up for, pick up on, go on to, live up to, go down to*). Again, such a low number of occurrences confirm the claim made by LGSWE that phrasal-prepositional verbs are a marginal group. Table 30 gives the list of the most frequent prepositional verbs found in the samples (in the Czech sample, two idiomatic multi-word phrases *take place in* were included in the analysis). Both learner samples share the prepositional verb *base on,* Czech and native speakers have also the prepositional verb *look for* in common.

Table 30: Prepositional verbs with the frequency more than two occurrences in all three samples

Order	Prepositional verbs CZL	Tokens CZL	Prepositional verbs NCZL	Tokens NCZL	Prepositional verbs NS	Tokens NS
1	*recommend to*	9	*deal with*	6	*share with*	9
2	*live in*	8	*relate to*	5	*talk about*	7
3	*base on*	7	*base on*	4	*invest in*	6
4	*write about*	7	*discriminated against*	4	*deal with*	6
5	*take place in*	5			*focus on*	3
6	*fall in love*	4			*happen to*	3
7	*look for*	3				
8	*listen to*	3			*listen to*	3
9	*go to*	3			*live in*	3
10	*talk about*	3			*look at*	3
11					*look for*	3
12					*put into action*	3
13					*tell about*	3
14					*think about*	3

Activity verbs represent the most numerous group in all three samples. According to the findings presented in LGSWE, they are the most frequent in all registers (see Table 31 below). The semantic analysis gives evidence that both learner groups have a good command of the commonest prepositional verbs (see Table 29).

Table 31: Distribution of prepositional verb semantic categories in the Czech, non-Czech and native speaker sample

Semantic groups of prepositional verbs	CZL sample %	NCZL sample %	NS sample %
Activity	33.3	58.4	44.1
Mental	18.1	10.8	18.3
Communicative	21.6	3.9	12.6
Causative	2.7	3.1	3.8
Occurrence	13.5	3.9	5.0
Existence/ Relationship	9.9	16.0	15.9
Aspectual	0.9	3.9	0.3
Total %	100.0	100.00	100.00

5.14 Conclusions

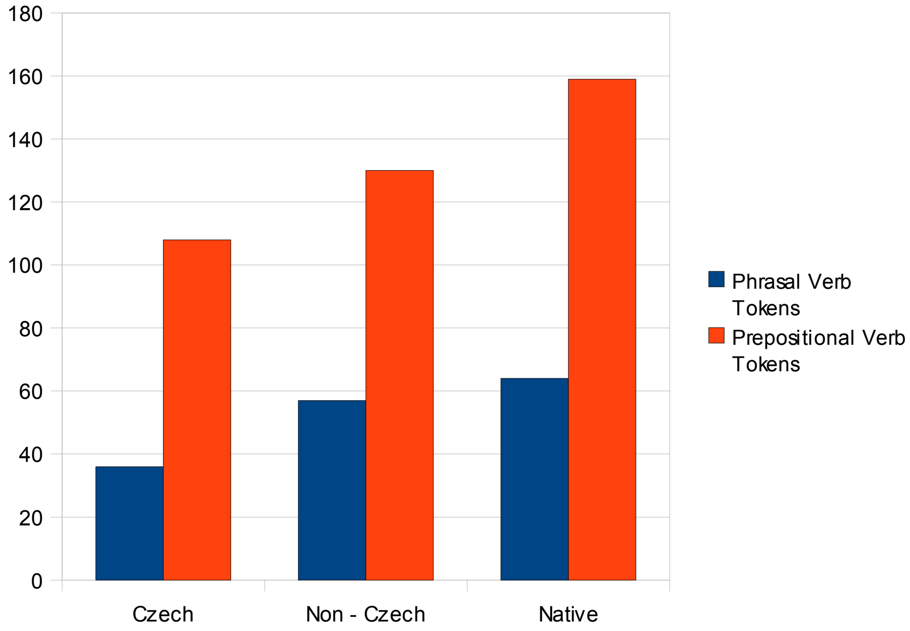
In summary, this chapter deals with the analysis of phrasal and prepositional verbs. The results confirm findings from previous research into multi-word verbs reported in LGSWE in several respects. In particular, the outcome of the analysis shows that phrasal verbs generally pose a major obstacle for learners whereas prepositional verbs appear less difficult for learners. The first part of the investigation confirms that learners produce fewer phrasal verbs than native speakers and that phrasal verbs used inappropriately by learners tend to occur even though to a very small extent. Further, it was expected that the range of phrasal verbs in the native sample would be wider compared with the non-native samples. The results confirm both assumptions: although the majority of the phrasal verbs found in the learner samples are used appropriately, the main stumbling block appears to be an extremely low incidence of phrasal verbs, especially in the Czech learner sample. Compared with the learner groups, native speakers produced twice as many phrasal verbs, also the range of phrasal verbs is wider in the native speaker sample.

The second concern of the chapter was the contrastive analysis of prepositional verbs. The investigation aimed to confirm that native speakers would produce the greatest number of prepositional verbs. Besides this, errors in terms of especially erroneously selected prepositions were expected in the non-native samples. The analysis confirms that prepositional verbs appear less problematic for learners. Data collected from the analysis shows that prepositional verbs are used more frequently in the learners' language than phrasal verbs. This result confirms the claim made by LGSWE. Indeed, the learners produce almost twice as many prepositional verbs than phrasal verbs and use them, in the majority of cases, appropriately. The distribution of prepositional verbs across semantic categories in both learner samples suggests that learners produce mainly the commonest phrasal verbs in the language and, with a few exceptions, have a good command of prepositional verbs. Despite the fact that learners produce a great number of prepositional verbs, the native sample contains almost twice as many prepositional verbs as there are in the Czech learner sample.

To conclude, the analysis of multi-word verbs in the samples highlights the differences between native speakers and learners in the use of this particular type of phraseological units. At the same time it shows that the difference depends on the specific subcategory of the phraseme.

Figure 3 shows the number of phrasal and prepositional verb tokens. The red colour concerns the number of prepositional verbs in the individual samples. It is apparent that native speakers used the greatest number of prepositional verbs of all. The blue colour highlights the number of phrasal verbs in all three samples. Again, it has to be said that the number of phrasal verbs is the lowest in the Czech learner sample. The figure clearly indicates that prepositional verbs are used more commonly than phrasal verbs, both by native speakers and non-native speakers.

Figure 3: Phrasal and prepositional verb tokens in all three samples



6. Collocation

Even though it is universally acknowledged that a learner's awareness and competence in actively using multi-word units improves with increasing proficiency, several studies and results from contrastive inter-language analysis (CIA) conclude that collocations appearing among different types of multi-word units are most problematic for learners. The following comparative study of non-native and native speakers' use of collocations first gives a theoretical outline of the term collocation: it focuses on two approaches towards collocations - phraseological and distributional. The theoretical part draws heavily on Granger and Paquet (2008). The research study on collocation investigates the level of collocational competence among learners of English in two ways. It seeks to find out to what extent learners are familiar with English collocations and to what degree their sense of collocation salience approaches native-speaker "collocational" sensitivity.

6.1 Theoretical background

The following sections provide a theoretical outline of the term collocation and discuss two approaches towards phraseology which influenced the concept of collocation. Additionally, related areas such as selectional preferences, selectional restrictions and lexical solidarities to collocation are briefly mentioned.

6.2 Two approaches to phraseology

Phraseology is now in its heyday, nonetheless, this state of affairs has been valid only for a relatively short period of time. The process of establishing phraseology as a field deserving an appropriate status has been impeded by two main factors - the wide-ranging and rather mixed-up terminology and the all-embracing scope of the field (Granger, Paquot 2008, 27). In general terms, phraseology primarily deals with different types of multi-word units which form a scale from "the least phraseological" to "the most phraseological" and where criteria must be formulated in order to set apart the different types of multi-word units from each other. During the evolutionary process phraseology was forced to undergo, two main streams emerged: the phraseological approach, established in the spirit of the Eastern European tradition, and the more recent corpus-based approach. Whereas the former predominantly concentrates on the comparably fixed multi-word units (e.g. idioms, proverbs, sayings, formulae), the latter subsumes all types of multi-word units regardless of their degree of opacity (Granger, Paquot 2008, 27). Just as phraseology has been developing, so has the notion of collocation. First, collocation was described in relatively vague and rudimentary terms which were later specified until two distinct concepts have been arrived at: collocation as a lexical syntagma, collocation as a statistical phenomenon.

6.3 The emergence of collocation

The number of definitions of collocation seems vast enough to frustrate any attempt to reach a consensus on what collocation actually is. Stubbs (2002, 57), for instance, argues that collocation "is the area where no generalizations are possible". In a similar vein, Čermák (2007) notes that the term collocation is not clearly defined and requires further specification.

In fact, Palmer (1925) was probably one of the first to draw attention to collocations preceding even Firth. In Palmer's *Second Interim Report on Collocation* (1933), collocation was defined as "a succession of two or more words that must be learned as an integral whole and not pieced together from its component parts". However, collocation followed a more elaborate and complicated path from the one introduced by Palmer (Cowie 1999, 54) and the concept of collocation later put forward by Firth (1957) paved the way for the two main approaches towards collocation. These stand at the forefront of current linguistics: the former is the phraseological approach, the latter the distributional approach (frequency-based).

6.4 Collocation and the phraseological approach

The Russian scholars Vinogradov (1947) and Amosova (1963) are considered to be the founders of the phraseological approach. The phraseological approach defines collocation as a word-combination which "is fixed but only to a certain extent" (Nesselhauf 2005, 11). There are numerous other definitions on collocation which highlight the aspect of arbitrariness: "Collocations are arbitrary word-combinations that are bound by mutual expectancy and predictability" (Crystal 1995). The phraseological approach sets apart collocations from other types of multi-word units and postulates the existence of a phraseological continuum that comprises multi-word units of different types with a varying degree of fixedness and opacity (Nesselhauf 2005). These multi-word units run the gamut of the most transparent, governed only by syntactic and semantic co-occurrence restrictions (free combinations), to the most opaque (pure idioms) and where rules which distinguish phraseological units from the non-phraseological ones must be formulated (Granger, Paquot 2008, 28). Cowie's (1981) typology is a good case in point. It consists of two main types of word-combinations: composites (restricted collocations, figurative and pure idioms) and formulae (having a pragmatic function). Two important criteria ascribed to composites are transparency (literal or non-literal meaning of the string) and ability to be substituted (the question of whether the combination can be substituted and to what extent such a replacement is restricted).

Accordingly, Cowie forms a phraseological continuum comprising categories which are not entirely clear-cut but creating a cline of:

1. Free combinations (e.g. *drink coffee*) in which substitution of the elements can be specified semantically and all elements in the string carry the literal sense, the combination is fully transparent. Free combinations go in line with the rules of grammar and its constituents can be freely substituted. It is the least cohesive type of the word-combinations.
2. Restricted collocations (e.g. *deliver a baby; overcome problems*) with a possible partial replacement of elements, arbitrary restrictions on the substitutability must be taken into account. Restricted collocations contain one element in both a literal and non-literal sense, but still guarantees the transparency of the combination.
3. Figurative idioms (e.g. *do a U-turn*) which seldom allow for the replacement of the elements, the literal sense is retained, though.
4. Pure idioms (e.g. *spill the beans* in the sense of *revealing one's secret*) have a completely opaque meaning, the elements are used in non-literal sense and cannot be substituted (Nesselhauf 2005, 14-15).

Occasionally, it is possible to come across two types of collocation distinguished by some authors and partly overlapping with Cowie's typology: the term "open collocation" refers to free combinations, while "restricted collocation" requires that one element must be used in non-literal sense. As indicated above, the meta-language in phraseology might cause confusion since a great variety of terms exist even though they represent the same concept. As a result, a certain degree of variation among the representatives of the phraseological approach can be seen, with some using the term collocation even when referring to free combinations (Lyons 1977). However, the term restricted collocation generally prevails.

Nonetheless, further issues need to be considered. Hausman (in Nesselhauf 2005) points out the arbitrarily restricted compatibility as the crucial factor, and this helps to differentiate collocations from free combinations. Some emphasize the transparency aspect, which allows for the distinction between collocations and idioms (Nesselhauf 2005, 16).

Others focus on the syntactic relation of the elements in collocations.

In view of this, the terms "lexical" and "grammatical collocations" are often brought up (Benson et al. 1986, Bahns 1993). Lexical collocation is made up of two lexical words arranged into the following structural types: noun + verb (*bees buzz)*, adjective + noun *(a compulsive liar*), verb + noun (*run business*), noun + of + noun (*a herd of elephants)*, adverb + adjective (*bitterly cold*), verb + adverb (*fork out handsomely)*. Grammatical collocation, on the other hand, comprises a lexical word together with a grammatical element - usually a preposition (e.g. *dream of*) or a structure (e.g. *mind* + *ing, manage* + *to*) and corresponds to Sinclair's colligation.

The usual number of elements in a combination is generally two or more and this view largely prevails. Some linguists consider the relationship between the constituents of collocation and argue that the elements of the collocation differ in nature (Nesselhauf 2005, 17). Melčuk (1995) for instance, holds that in *crack a joke,* the meaning of *joke* can be derived from general lexicon, while the meaning *crack* depends on the particular collocation.

6.5 Collocation and frequency-based/distributional approach

The beginnings of the frequency-based (distributional) approach date back to Firth, who was one of the first to touch upon the term collocation. In his paper *The Modes of Meanin*g (1951)*,* he made the claim "*you shall know the word by the company it keeps*" and attempted to deal with collocations using examples which illustrate that words habitually occur with a specific set of collocates. His concept of collocation was revolutionary in that he emphasized the importance of syntagmatic rather than paradigmatic aspect in lexical relations: "Meaning is an abstraction at the syntagmatic level" which he illustrated with the example where "one of the meanings of *night* is its collocability with *dark* and of *dark*, the collocation with *night* (Firth 1957b, 196). A decade later, Halliday (1961) pointed out the significance of the statistical aspect, which was gradually coming to the fore: "The syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at *n* removes (a distance of *n* lexical items) from an item *x*, the items *a,b,c*". Halliday aimed to provide a list of collocates which co-occur with the node within a short space; this co-occurrence will be statistically significant, the probability that the node will co-occur with its collocates will be higher than random co-occurrence, happening by chance. A more recent trend set by Sinclair, and the establishment of the distributional approach, has made people think about the way language works in a completely different way. First of all, Sinclair emphasized the importance of distinguishing word-forms (e.g. *sit)* from lemma (as set of word-forms e.g. *sit, sits, sitting, sat)* and then went on to clarify how, in fact, language is organized. In the light of Sinlair's (1991) idiom principle, collocation is crucial and stands for "the occurrence of two or more words within a short space of each other in a text but has a different value in the description of the two words" (Sinclair 1991). "Significant collocation" refers to such co-occurrences which "occur more often than their respective frequencies and the length of text in which they appear would predict" (Sinclair 1991). In order words, in examples such as *the girl, the* and *girl* cannot be classified as significant collocations since the definite article is so frequent that it could easily occur with a great number of other words in a text; by contrast, the words such as *shrug* and *shoulders* represent a significant collocation since *shrug* is likely to occur in a context where *shoulders* occur. In a similar fashion, the distinction between "upward" and "downward collocations" is also pointed out by Sinclair. The term upward collocation is the type of collocation in which words will habitually collocate with other words that are more frequently used than they are themselves in English. Downward collocation refers to the collocation whereby words will habitually co-occur with words that occur less frequently than they do. Downward collocation is important in view of the semantic analysis of the word. Sinclair's example (1991, 115) best illustrates the point: "When *a* is node and *b* is collocate, this is called downward collocation, i.e. is the collocation of *a* with a less frequent word. When *b* is node and *a* is collocate, this is the case of upward collocation".

Apart from that, Sinclair developed the terminology related to collocation. He introduces the terms: "the node", "the collocate" and "the span". The node represents the word under investigation; the collocate enters into collocation with the node; the span concerns the distance between words.

However, Sinclair and his co-researchers use the term collocation in several different ways. For instance, some of them regard all co-occurrences of all frequencies to be collocations whereas Stubbs (1995) insists that the term collocation refers to frequent occurrences only (Nesselhauf 2005,12). Opinions regarding how many words a collocation comprises and whether they must be adjacent vary among linguists. Generally, two words (occasionally three words) form a collocation, although Firth regarded the whole string as a collocation (in Nesselhauf 2005, 13). Very much unlike the phraseological approach, Sinclair and his colleagues pay little notice to a careful classification, and concentrate on the importance of co-occurrence, regardless of their type (Granger, Paquot 2008, 29). Virtually all types of multi-word units are treated within the distributional approach. Semantic criteria are not used in order to designate a particular type of a multi-word unit within the distributional approach, the emphasis is placed on a different view of meaning. The meaning is best accounted for using Sinclair's (1991) "model of the extended lexico-grammatical units" or Hoey's (2005) "theory of lexical priming": the distribution of meaning is scattered over all the elements of a multi-word unit rather than confined to a single word. A lexical item and its meaning participate in the interplay of lexical, grammatical, semantic as well as pragmatic layers which help to establish the multi-word unit with its prototypical collocations, grammatical structure, semantics and the pragmatic aspect. Hoey (2005) explains the concept of collocation in terms of lexical priming: speakers prime words with other words due to the previous encounters with the word. Hoey's claim is significant in that he regards collocation as the starting point which allows for grammar to emerge: "Grammar is created in the way we collect and associate collocational primings" (Hoey 2005).

6.6 A possible convergence between the traditional and the distributional approach

In as much as each of the two approaches has something to offer, Granger and Paquot (2008, 41) believe that their reconciliation would be the best solution. The traditional approach has developed a sophisticated classification of phraseological units, the corpus-driven approach has access to enormous amounts of objective linguistic data thanks to methods of automatic extraction and corpora.

6.7 Related concepts – selectional preferences, selectional restrictions and lexical solidarities

The terms "selectional restrictions" and "selectional preferences" are often brought up in connection with collocation. The fact that attention should be drawn to the link between collocation and grammar was first pointed out by Chomsky in his *Aspects of the Theory of Syntax* (1965, 114) and later developed by Katz and Fodor (1963) within the framework of decompositional generative semantics. The term selectional restrictions refers to "the conditions for the compatibility of elements which are a consequence of the meaning of a word and expressed by means of semantic features" (Nesselhauf 2005, 19). In the examples *(*e.g. *the idea cut the tree; I drank the bread)* selectional restrictions block the existence of such formations; *cut* requires a "concrete" subject and *drink* a "liquid" object (Palmer 1976, 100). The term selectional preferences relates to a word's tendency to co-occur with words that belong to certain lexical sets. For example, the adjective *impeccable* prefers to modify nouns that denote *manners* or *behaviour,* the verb *knit* requires subjects and objects. The subjects denote human beings whereas objects are inanimate. Palmer (1976, 97-8), who prefers the term "collocational restrictions", distinguishes three kinds of them: first, restrictions due to the meaning of the items (it is highly improbable to come across a collocation *sweet salt)*; second, every word has a set of items it co-occurs with, a so-called collocational range; even though a word's collocational range may often be extended, a communicatively competent language speaker is somehow well-aware to what degree the range allows for extension and still sounds natural for a native-speaker; third, restrictions due to specific lexical reasons - even almost absolute synonyms cannot combine freely with the same set of nouns (e.g. *rancid butter/bacon; addled eggs/brains*). The term "lexical solidarities" coined by Coseriu (1967) again describes the compatibility of lexemes in terms of their semantic features. However, whereas selectional restrictions carry a negative implication in order to prevent certain combinations occurring, lexical solidarities have positive implications and account for the compatibility of certain elements (Lipka 1990), and in this respect correspond to selectional preferences.

6.8 Non-native speakers and their phraseological performance in the area of collocations

Investigation of collocations involves either an elicitation test or is based on the learners' production (Nesselhauf 2005). As Nesselhauf (2005) points out elicitation tests (Bahns, Eldaw 1993, Herbst 1996, Shei 1999) include both cloze tests and translation tasks - a possible drawback is the small amount of data they provide. As Nesselhauf (2005, 8) observes, however, results obtained from a great number of studies focusing on collocations cannot confirm that the use of collocations is in all cases influenced by learners' proficiency. Similarly debatable is the aspect of mother tongue interference. Whereas several studies attempt to prove that mother tongue interference appears crucial, others do not seem to go along with this statement. Many researches, however, seem to agree on how non-native speakers and native speakers perceive collocations. Howarth (1998) observes that while native speakers regard collocations as ready-made sequences which go together and are not to be separated, a learner's idea of collocation often seems to be that of "separated items which have become paired" (in Nesselhauf 2005).

The following studies deal exclusively with collocations:

Bahns (1993) conducted a contrastive analysis of German learners and English speakers with a special focus on *noun* + *verb* and *verb* + *noun* combinations. Her main findings provide evidence that for a great number of lexical collocations, a direct translational equivalent in English or in the learners' mother tongue exists; there is no need to teach such collocations in the majority of cases. She concludes that only combinations which do not a have direct translational equivalent should be taught.

An especially relevant study for the present section on collocation is that of Sylviane Granger (in Cowie 2005), in which the collocational preferences of a group of French learners and that of a group of native-speakers are compared. In particular, Granger attempts to explore two types of multi-word combinations – collocations of *adverbial amplifier* + *adjective* type (e.g. *bitterly disappointed, blissfully unaware, totally amazed)* and *formulae* (e.g. *we can/could/should/might notice that...*; *I think that....).* Granger's study reveals that French learners produce a significantly lower number of amplifiers than native speakers. Further observations show that some of the collocations used by the French learners suggest a direct link with French, which points to mother tongue transfer (e.g. *highly developed/civilized/specialized* have direct French equivalents). Granger's investigation proves her initial hypotheses that French learners have a tendency to use amplifiers as building blocks rather than as parts of ready-made units. With sentence frames, she highlights two important findings: first, French learners produce fewer prefabs than their native counterparts. Second, an excessive use of sentence builders mostly entailing *think and say* (*I think, I would say that, I think that)* was noticed. Granger ascribes this to the intimate familiarity with these expressions, which following Dechert (1984), she calls "islands of reliability" (Cowie 2005, 155).

Similarly, Howarth (1996) carried out a contrastive study of collocations in non-native and native academic writing. Howarth's research involves very advanced foreign users, namely teachers of English from various linguistic backgrounds. The analysis confirms the native speakers' awareness of the need to adhere to the established academic norms whereas non-native writers apparently lack such knowledge. A greater incidence of non-standard formations was found in the non-native writing. Careful observation suggests that with increasing proficiency learners are aware of the distinctions between free combinations and phraseological combinations, which they seem to memorize and use in a satisfactory way but still, they show considerable limitations and inadequacy as far as the central point of the phraseolocical spectrum is concerned: the area of collocations. The conclusion is that the most problematic area is that of restricted collocations, where the highest incidence of errors was observed.

Sadeghi (2009) conducted a detailed investigation of Farsi (Persian) and English collocation in which 76 Farsi learners of English were asked to undergo an English collocations test. The results reveal that learners most commonly encounter difficulties in areas where a difference between mother tongue and target language word patterns can be observed.

The question arises why collocations pose such a difficulty for learners. It was briefly mentioned at the outset that the core problem lies in the way native and non-native speakers use collocations. Whereas collocations are used by native speakers as a single entity – as pairs rather than two items which normally occur separately, non-native speakers are prone to regarding the items in a particular collocation as two separated items rather than as a unity (Wray 2005, 211): "For native speakers, collocations are pairs which can become separated under certain circumstances while adult learners' collocations are to be seen as separate items which have become paired". Equally importantly, native speakers regard collocation as a formulaic combination, non-native speakers prefer the non-formulaic approach, even though this is a subconscious process.

6.9 Sample analysis – research into collocations

The investigation of the collocational competence of two groups of learners and native speakers is divided into two parts described in 6.9.1 and 6.14 respectively, each involving different procedures.

6.9.1 Preliminaries of the collocational analysis of all three samples

As mentioned above, the investigation of the collocational competence of two groups of learners and one group of native speakers is divided into two parts. The first part deals with the collocational behaviour of selected nodes in the three sample corpora: Czech learners (37 essays), non-Czech learners (19 essays) and native speakers (20 reviews). All three samples contain approximately 9300 - 9400 words. Czech learners are students from a grammar school in Prague; they are sixteen and seventeen year old students with pre-intermediate,

intermediate and upper-intermediate to FCE level. The other group of non-native speakers are students from various linguistic backgrounds; their essays were downloaded from the website http://bookrags.com/. The total of 22 book reviews written by native speakers, mainly professional review writers, were downloaded from the website available at http://happypublishing.com/ (for a detailed description see Chapter 3, Section 3.3). The collocational analysis attempts to find out to what extent learners produce frequently occurring collocations in the BNC (5 and more occurrences) and to what extent they produce collocations which fall into either the peripheral zone, occurring in the BNC to a little degree (1-5 occurrences) or which are not attested in the BNC at all. The structural types of *noun + adjective, noun + verb* come under scrutiny.

The preliminary investigation of word-classes included several obstacles. After the list of all nouns, adjectives, verbs and adverbs was retrieved from the individual sample corpora, the findings indicated that the analysis would have to be limited to the collocational behaviour of nouns only. The reasons are mainly the low frequencies of the individual adjectives, verbs, adverbs and/or the fact that they are not shared by all three sample corpora.

In particular, only very few adjectives exceeded the established minimum of 10 occurrences and mostly, they were represented by relatively general adjectives (e.g. *main, best, favourite*, *good, old, new, next, popular, famous).* This fact presupposed an upward collocation and thus little collocational richness. More "interesting" adjectives were found in the native sample, however, the frequency was still very low. The second obstacle was their complete absence in the non-native samples. The range of "collocationally interesting" verbs was also somewhat limited. The majority of the verbs with a higher frequency are very general verbs such as *make, go, have, set, call, get, want.* More interesting verbs occurred especially in the native sample where for example *attain, assort, ban* etc. might produce interesting results, however, as indicated, their presence was limited only to the native sample. Adverbs in the samples occurred only seldom and for this reason they were excluded from the analysis.

Accordingly, the list of nouns was reduced to those belonging to the semantic field of *reading and literature,* a few others with more than 10 concordance lines were added with the final list of nodes comprising the lemma of *book, novel, story, author, writer, time, people, world,* with their singular and plural forms treated separately (see Appendix 4a, b, c). The analysis is limited to the collocational types of *adjective + noun, noun + verb* and is carried out distributionally.

The subsequent procedure will involve three stages: retrieval of concordance lines containing the nodes from all three sample corpora, division of collocations into structural types and a check with the BNC. An arbitrary boundary of 5 and more occurrences in the BNC was decided on combinations considered frequent collocations, and so relevant, from those which fall into the "peripheral" zone (1- 5 occurrences) or do not occur at all. The term collocation will be ascribed only to such combinations, which have the frequency of occurrence in the BNC of 5 and more. The collocates of the nodes are, as far as the structural type *adjective* + *noun* concerned, in the majority of cases, adjacent in this part of the investigation and only the adjectives on the left side of the node will be analysed. As far as the structural type *noun* + *verb* is concerned, the collocates need not be necessarily adjacent and will occur in the horizon of 4 word on the right side. The concordance lines containing grammatical collocations or sequences such as *this girl, my book, the book is about a girl...* will not be subject to analysis.

The preliminary findings give rise to the following assumptions:

1. non-native speakers will use very common collocates more often than native speakers;

2. non-native speakers will be less familiar with the norms of co-occurrence and especially the structural type *noun* + *verb* might show signs of non-standard use of collocations;

3. native speakers' collocations will reflect the familiarity with standard usage, the norms of co-selection.

6.10 Czech learners' use of collocations

The following sections 6.10.1 – 6.10.2 deal with Czech learners'collocational competence. Two structural types of collocation, namely *adjective* + *noun* and *noun* + *verb,* are analysed in the Czech learner sample.

6.10.1 Structural type *adjective* + *noun* in the Czech learner sample

First, the structural type *adj +noun* came under scrutiny and the lemmas *novel, story, book, author, writer, people, world, time* were subject to thorough investigation. The analysis of the collocational behaviour of selected nouns yielded the following results: 71 types and 90 tokens were obtained from the analysis out of which only 8 occurrences (9 per cent of tokens) have no matches in the BNC, 19 occurrences (21 per cent of tokens) are attested in the BNC within the span of 1 to 5 occurrences, the rest of the 63 occurrences (70 per cent of tokens) occur 5 times and more, with some of them exceeding thousands in the BNC. Thus 70 per cent of tokens of the structural type *adjective* + *noun* collocation in the Czech learner sample are placed in the category "5 and more" (see Tables 32, 33).

The list of collocates is split into the three corresponding groups:

1. Combinations (8 types, 8 tokens) not attested in the BNC

horror-fiction novels, background stories, friend book, underworld people, non-existing world, today's world, Czech authors

2. Peripheral collocations (18 tokens, 19 tokens) having the frequency in the BNC of 1-5 occurrences

only novel, full-length novel, short novel, historic novel, fantasy story, little stories, breathtaking stories, seventh book, banned book, earliest books, magic world (2), unrealistic world, helpless people, pious people, interesting writer, unknown writer, classic authors, English authors

The last group encompasses collocations with the frequency of 5 and more occurrences in the BNC. The first number in brackets indicates the number of tokens in the Czech sample followed by the frequency in the BNC. They are arranged in descending order.

3. Very frequent collocations (45 types, 63 tokens) in the BNC

first time (1-8324), some time (3-4467), young people (1-3615), last time (1-2797), short time (1-1007), outside world (1-624), present time (1-416), new book (310-1); first book (4-242), short story (2-205), short stories (1-168), true story (1-153), free time (1-153), first novel (1-143), love story (1-94), ideal world (1-93), good book (2-92), different world (1-91), second book (2-76), whole book (1-61), fantasy world (2-41), detective stories (2-39), real story (3-38), horror story (1-38), American writer (1-36), wonderful world (1-35), last book (1-35), only book (1-21), popular book (1-19), English writer (1-19), fourth book (1-16), love stories (3-16), short-story writer (1 -16), American author (2-14), favourite book (3-13), favourite books (1-12), historical novels (1-12); interesting book (1-12), British writer (1-11), British author (2-9), favourite writer (1-8), favourite author (1-8), favourite authors (1-8), recent time (1-7), beautiful books (1-6).

A brief look at the collocations which occur in the BNC very frequently deserves a few comments. The range of adjectives is not especially “interesting” - the learners use relatively very common adjectives (*new, good, favourite, popular, beautiful),* several of these refer to the authors’ or writers’ nationality (*British author, British writer).* Learners seem to have followed a safe and secure strategy - hardly any collocate out of these could be labelled as atypical (for more details, see section 6.13).

Table 32 presents the summarized data on the structural type *adjective* + *noun* collocation in the Czech learner sample. The first column shows the presence/absence of the given collocations in the BNC, the second column contains the number of types, the third column gives the number of tokens.

Table 32: Types and tokens of *adj* + *noun* collocations in the Czech learner sample

Number of the same collocations in the BNC	Types	Tokens	
			%
zero	8	8	9.0
up to 5	18	19	21.0
5 and more	45	63	70.0
Total	**71**	**90**	**100.0**

Table 33 presents the list of the adjectives of the node found in the Czech sample. The second column provides the list of very frequent adjectival collocates, the third column shows the collocates not attested in the BNC, the last column lists peripheral collocates in the BNC.

Table 33: Adjectival collocates (types) of the nodes in the Czech learner sample

Node	Collocates with 5 and more occurrences	Collocates with zero occurrence	Collocates with up to 5 occurrences
novel	*first*	-	*only, full-length, short, historic*
novels	*historical*	*horror-fiction*	-
story	*real, true, short, horror, love*	-	*fantasy*
stories	*detective, short, love*	*background*	*little, breathtaking*
book	*good, first, whole, fourth, interesting, last, favourite, second, only, popular, new*	*friend*	*seventh, banned*
books	*favourite, beautiful*	-	*earliest*
author	*British, American, favourite*	-	-
authors	*favourite*	*Czech*	*classic, English*
writer	*short-story, favourite, British, English, American*	-	*interesting, unknown*
writers	-	*important*	-
people	*young,*	*underworld*	*helpless, pious*
time	*last, some, present, short, free, recent, first*	-	-
world	*different, fantasy, ideal, outside, wonderful*	*non-existing, today's*	*magic, unrealistic*

6.10.2 Structural type *noun* + *verb* in the Czech learner sample

The structural type *noun* + *verb* collocation is relevant for the analysis since the collocates are most mutually selective of all the structural collocational types. The analysis provided the following results: 59 types, 74 tokens were obtained out of which 17 tokens (23 per cent) have no matches in the BNC, 17 tokens (23 per cent) are the peripheral collocates having the frequency in the BNC of 1-5 occurrences, 40 tokens (54 per cent) have the frequency of occurrence in the BNC of 5 and more (see Tables 34, 35).

1. Combinations (16 types, 17 tokens) not attested in the BNC

book brings up, book keeps you in suspension, book looked nice, book made a good impression on me (2), book pictures, book exceeded public expectations, books chronicles, book leaves you thinking, novel released, story has a surprising end, story gets more serious, stories aided by, author criticizes, author took with light humour, authors include, writer lived

2. Peripheral collocations (15 types, 17 tokens) having the frequency in the BNC of 1-5 occurrences
book has a lots to offer, book talks, books help, books deal with, books made into films, novel reflects, story describes, story deals with, story gets complicated, story goes on, story takes place (3), stories show, author names, author introduced, writer wrote

The last group includes collocations which occur 5 times and more in the BNC. The first number in brackets indicates the number of tokens in the Czech sample, it is followed by the frequency attested in the BNC. The collocations are arranged in descending order. Again, learners aimed to be on the safe side and used, in the majority of cases, very general verbs.

3. Very frequent collocations (28 types, 40 tokens) in the BNC
people think (2-794), people live (1- 414), people know (1-380), people see (-, 186), book is called (2-153), people leave (1-125), story is going (1-93), book is written (5-89), people understand (1-73), world called (1-66), people read (3-52), book shows (1-38), book is based on (1-25), story started (1-22), book is divided (1-19), books provide (1-19) people discover (1-20), book ends (2-18), novel called (1-16), novel set in (1-16), story is set in (1-14), novel written (3-13), story based on (2-10), people survive (1-8), book develops (1-7), book focuses (1-7), author tried, book comprises (1-6)

Table 34 presents the findings obtained from the BNC. The first column shows the extent to which the collocation is attested in the BNC. The second and third column present number of types and tokens in the Czech sample respectively.

Table 34: Types and tokens of *noun* + *verb* collocations in the Czech learner sample

Number of the same collocations in the BNC	Types	Tokens	%
zero	16	17	23
up to 5	15	17	23
5 and more	28	40	54
Total	**59**	**74**	**100**

Table 35 lists the verbal collocates of the respective nodes. The collocates are divided into three groups according to their frequency in the BNC.

Table 35: Verbal collocates (types) of the nodes in the Czech learner sample

Node	Collocates with 5 and more occurrences	Collocates with zero occurrence	Collocates with up to 5 occurrences
novel	*be called, set in,* *be written*	*release*	*reflect*
novels	-	-	-
story	*be based on, go, set in, start*	*get more serious, has a surprising end*	*describe, deal with, get complicated, go on, take place*
stories	-	*aided by*	*show*
book	*comprise, end, have (lots to offer),* *be called, be written,* *be based on, be divided,* *develop, focus, shows, talks,*	*bring up, keeps you in suspension, look nice, make a good impression, picture, exceed public expectations, leaves you thinking*	*talk, have (lots to offer)*
books	*provide*	*chronicle*	*help, deal with, made into*
author	*try*	*criticizes, take with light humour*	*name, introduce*
authors		*include*	-
writer		*live*	*write*
writers	-	-	-
people	*understand, survive,* *see, discover,* *know, leave stg, live,think, read*	-	-
time	-	-	-
world	*be called*	-	-

Even though several examples are not attested in the BNC, they are possible combinations. Some of the combinations not attested in the BNC are atypical collocations. In the majority of cases either the verb is used incorrectly or the collocational range is extended inappropriately.

In example (35) the learner extended the collocational range inappropriately – *books are published* while *CDs are released*.

(35) *Walk" etc. "Carrie" is his first* ***novel*** *and it was* ***released*** *in 1974. It is the most popular book of his produc*

Also the example (36) *the book pictures* sounds a bit unusual – *the book describes/depicts* is a better alternative.

(36) *e. I am going to let the cat out of the bag. This* ***book pictures*** *the ideal world. Can*

children still see such as

6.11 Non-Czech learners' use of collocations

The following sections 6.11.1 – 6.11.2 deal with non-Czech learners' collocational competence. Again, two structural types of collocation, namely *adjective* + *noun* and *noun* + *verb,* are analysed in the non-Czech learner sample.

6.11.1 Structural type *adjective* + *noun* in the non-Czech learner sample

The same nodes (*book, books, novel, novels, story, stories, author, authors, writer, writers, people, world, time*) were selected for this analysis, the overall number of concordance lines was almost the same as the concordance lines in the Czech learner sample and the native speaker sample (see Appendix 4b). However, the number of concordance lines with the structural type *adjective* + *noun* was dramatically lower in comparison with the group of Czech learners. The search yielded a mere 27 types, 28 tokens (see Table 36), which suggests that Czech learners produced twice as many types and three times more tokens. Out of the 28 tokens in the non-Czech learner sample, 7 tokens (25 per cent) were not attested in the BNC, 5 tokens (18 per cent were retrieved from the BNC within the span of 1-5 occurrences), 16 tokens were attested in the BNC 5 times and more (57 per cent); see Table 37. For further details see Section 6.13.

1. Combinations (7 types, 7 tokens) not attested in the BNC
magnificent novel, universal novel, seamless story, fictional story, forth time, genius writer, amazing writer

2. Peripheral collocations (5 types, 5 tokens) having the frequency in the BNC of 1-5 occurrences
controversial novels, Gothic stories, thrilling book, good author, Gothic writers

3. Very frequent collocations (15 types, 16 tokens) in the BNC
same time (2-7 640), some time (1-4 467), good time (1-880), new world (1-631), whole story (1-226), short story (1-205), good book (1-92), second book (1-76), whole book (1- 61), dead people (1-43), talented people (1-30), third book (1-23), entire book (1-15), American author

(1-14), today people

Table 36 presents the data on the type and tokens of the collocational structures *adjective* + *noun* found in the non-Czech sample.

Table 36: Types and tokens of *adj* + *noun* collocations in the non-Czech learner sample

Number of the same collocations in the BNC	**Types**	**Tokens**	**%**
zero	7	7	25.0
up to 5	5	5	18.0
5 and more	15	16	57.0
Total	**27**	**28**	**100.0**

Table 37 presents the data on the structural type *adjective* + *noun* in the non-Czech sample. The second column contains adjectival collocates which occur in the BNC with the frequency of occurrence 5 and more, the third column shows collocates with zero occurrence in the BNC, the last column presents minor occurrences (1-5) in the BNC.

Table 37: Adjectival collocates (types) of the nodes in the non-Czech learner sample

Node	**Collocates with 5 and more occurrences**	**Collocates with zero occurrence**	**Collocates with up to 5 occurrences**
novel	-	*magnificent* *universal*	-
novels	-	-	*controversial*
story	*short, whole*	*seamless* *fictional*	-
stories	-	-	*Gothic*
book	*whole, entire, third, second, good*	-	*thrilling*
books	-	-	-
author	*American*	-	*good*
authors	-	-	-
writer	-	*genius, amazing*	-
writers	-	-	*Gothic*
people	*dead, talented, today*	-	-
time	*good, some, same*	*forth*	-
world	*new*	-	-

6.11.2 Structural type *noun* + *verb* in the non-Czech learner sample

The investigation into the structural type of *noun + verb* with the nodes *novel, story, book, time, world, people, writer* and *author* in the non-Czech sample yielded the following results: the sample contains only 29 types and 32 tokens of the collocations. There are 10 tokens (31 per cent) with no matches in the BNC, 15 tokens (47 per cent) which occur within the span of 1–5 occurrences in the BNC, 7 tokens (22 per cent) placed in the category "5 and more" occurrences in the BNC.

1. Combinations (10 types, 10 tokens) not attested in the BNC

novel set back, novel befitting, novel has issues, stories trigger, book guarantees, book entertain, writer created, people have schools, people were tortured, people were tried and tortured

Even though some combinations were not attested in the BNC, they make sense, other require a little clarification.

In example (37) the learner perhaps mixed up *set in* with *set back:*

(37) *a timeless and universal* ***novel****, even though it's* ***set back*** *during the antebellum era. I agree with this becau*

Example (38) is very unusual and the intended meaning is not entirely clear:

(38) *nitely an epic adventure and a magnificent* ***novel****,* ***befitting*** *of its status as a global phenomenon. I am positiv*

The verb *deals with or presents* would be a more appropriate collocate for *issues*. The collocation *the novel has many issues* does not make too much sense.

(39) *We fight it. Huck Finn is a* ***novel*** *that* ***has*** *many* ***issues*** *we deal with today in it. Fancy that. An old class*

2. Peripheral collocations (12 types, 15 tokens) with 1-5 occurrences in the BNC

novel progresses, story goes on, story describes, story takes place (3), book portrays, book attracts, book identifies, book is recommended, book is set in, people surround, people look down on, stories inspire

The following group refers to the collocations attested in the BNC 5 times and more. The first number in brackets refers to the number of tokens and it is followed by the BNC frequency. The collocations are arranged in descending order.

3. Very frequent collocations in the BNC (7 types, 7 tokens)
people try (1-149), book entitled (1-68), story begins (1-48), book shows (1-32), story is written (1-16), story is published (1-12) , book focuses (1-8)

Table 38 presents the number of types and tokens; further, it indicates whether they are or are not attested in the BNC.

Table 38: Types and tokens of *noun* + *verb* collocations in the non-Czech learner sample

Number of the same collocations in the BNC	**Types**	**Tokens**	**%**
zero	10	10	31.0
up to 5	12	15	47.0
5 and more	7	7	22.0
Total	**29**	**32**	**100.0**

6.12 Native speakers' use of collocations

The following sections 6.12.1 – 6.12.2 deal with native speakers' collocational competence. Once again, two structural types of collocation, *adjective* + *noun* and *noun* + *verb,* are analysed in the native speaker sample.

6.12.1 Structural type *adjective* + *noun* in the native speaker sample

The search in the native speaker sample produced 52 types and 58 tokens of the collocational structural type *adjective* + *noun* (in few cases the adjective is modified by an adverb). Namely, the native speaker sample contains 18 tokens (31 per cent) which are not attested in the BNC. Some of the combinations are typical of American English, which explains their complete absence in the BNC (*a darn good book, a soft-cover book, a high-priced book);* 8 tokens (14 per cent) were present in the BNC but only scarcely, 32 tokens (55 per cent) are placed in the category "5 and more". At first sight, the range of collocates of the nouns is wider compared with both learner samples (see Tables 39, 40) . Section 6.13 offers a detailed comparison of the collocational analysis in all three sample corpora.

1. Combinations (16 types, 18 tokens) not attested in the BNC

beautifully-written novel, highly-acclaimed novel, fictional story, frequently-published author, poetry writer, self-proclaimed writer, Brazilian writer, beautifully-crafted book, darn good book (3), soft-cover book, high-priced book, parenting book, overpriced book, actual book, grown-up world, conflicting worlds

2. Peripheral collocations (7 types, 8 tokens) having the frequency in the BNC of 1-5 occurrences

beautiful story, next story, charming book, funny book, helpful book, non-fiction book (2), slender book

3. Very frequent collocations (29 types, 32 tokens) in the BNC

same time (1-7640), short time (1-1007), real people (1-679), new people, (1-631), whole world (1-426), right time (1-412), new book (1-310), first book (1-242), short story (1- 205), short stories (1-168), first novel (1-143), difficult time (1-139), business world (1- 109), good book (2-92), life story (2-88), new novel (1-72), success stories (1-71), little book (1-88), valuable time (1-69), changing world (1-62), precious time (1-50), second novel (1-39), simple story (1-17), ordinary people (1-16), favourite book (1-13), different book (1-12), personal stories (2- 8), delightful book (1-8), narrow world (1-7),

Table 39 presents the data obtained from the native speaker sample. The first column refers to the presence/absence of the collocations in the BNC, the second and the third columns list the number of types and tokens.

Table 39: Types and tokens of *adjective* + *noun* collocations in the native speaker sample

Number of the same collocations in the BNC	**Types**	**Tokens**	**%**
zero	16	18	31.0
up to 5	7	8	14.0
5 and more	29	32	55.0
Total	**52**	**58**	**100.0**

Table 40 shows the distribution of the collocates found in the native speaker sample in the BNC. The second column contains very frequent collocates in the BNC (5 and more occurrences), the third column lists the collocates with no BNC frequencies, the forth column shows collocates which are in the peripheral zone of only 1-5 occurrences.

Table 40: Adjectival collocates (types) of the nodes in the native speaker sample

Node	Collocates with 5 and more 5 occurrences	Collocates with zero occurrence	Collocates with up to 5 occurrences
novel	*first, second, new*	*beautifully-written*	-
novels	*acclaimed*	-	-
story	*life, short, simple*	*fictional*	*beautiful, next*
stories	*personal, success, short*	-	-
book	*good, different, first, delightful, favourite, new, little*	*darn good, soft-cover, beautifully crafted,high-priced, parenting*	*charming, funny, helpful, non-fiction, slender*
books	-	*overpriced, actual*	-
author	-	-	*frequently-published*
authors	-	-	-
writer	-	-	*poetry, self-proclaimed, Brazilian*
writers	-	-	-
people	*real, ordinary, new*	-	-
time	*same, precious, right, valuable, difficult, short*	-	-
world	*business, whole, narrow, changing*	-	-
worlds		*conflicting*	

6.12.2 Structural type *noun* + *verb* in the native speaker sample

The analysis into the structural type of collocation *noun + verb* in the native sample produced the following results: 43 types, 46 tokens were obtained out of which 10 tokens (22 per cent) have no matches in the BNC, 11 tokens (24 per cent) occur in the BNC with the frequency of 1-5, 25 tokens (54 per cent) have the frequency of 5 and more occurrences (see Tables 41, 42) .

1. Combinations (10 types, 10 tokens) not attested in the BNC

book teaches, novel comes out, novel comes to an end, story has impacted, story serves, stories mark, stories unfurl, books make you feel, author muses, world nourished

2. Peripheral collocations (11 types, 11 tokens) having the frequency of 1-5 occurrences in the BNC

book feels personal, book makes it clear, story flows, story focuses, story changed, story shares, stories share, author covered (points), author presents, world treat, worlds collide

The following group includes the collocations with 5 and more occurrences in the BNC. The first number in the bracket is linked to the number of tokens, it is followed by the BNC frequency. The collocations are sorted in descending order.

3. Very frequent collocations (22 types, 25 tokens) in the BNC

people want (1-789), people talk (1-192), time passes (1-175), story is told (1-04), world seems (1-77), book contains (2-75), people read (1-62), people consider (1-53), story begins (2-49), book provided (1-40), people complain (1-38), book says (1-34), author wrote (1-17), people are placed (1-14), story unfolds (1-12), book claimed (1-9), book opens (1-17), book is filled with principles (2-7), story follows (1-7), people comment (1-8), people lead lives (1-7), author uses (1-7),

Table 41 presents the number of types and tokens of the collocations in the native speaker sample and their obtained frequencies in the BNC.

Table 41: Types and tokens of *noun* + *verb* collocations in the native speaker sample

Number of the same collocations in the BNC	Types	Tokens	%
zero	10	10	22.0
up to 5	11	11	24.0
5 and more	22	25	54.0
Total	**43**	**46**	**100.0**

Table 42 lists the verbal collocates of the nodes found in the native speaker sample. The first column gives the nodes, second column gives a list of collocates which occur in the BNC 5 and more times, the third column contains collocates which are not attested in the BNC, the fourth column show collocates which rarely occur in the BNC.

Table 42: Verbal collocates (types) of the nodes in the native speaker sample

Node	Collocates with 5 and more occurrences	Collocates with zero occurrence	Collocates with up to 5 occurrences
novel	-	*come out, come to an end*	-
novels	-	-	-
story	*begin, be told, unfold, follow*	*impact, serve*	*flow, focus, change, share*
stories		*mark, unfurl*	*share*
book	*contain, be filled with, open, provide, say, claim*	*teach*	*feel personal, make it clear,*
books	-	*make you feel*	-
author	*write, use*	*muse*	*cover points, present*
authors	-	-	-
writer	-	-	-
writers	-	-	-
people	*consider, complain, talk, read, lead lives, be placed, comment, want*	-	-
time	*pass*	-	-
world	*seem*	*nourish, treat*	-
worlds	-	-	*collide*

6.13 Comparison and summary of findings obtained from all three samples

The previous sections have dealt with the collocational behaviour of the selected nodes – lemmas of *book, story, novel, author, writer, time, world, people*. The minimum of concordance lines for the individual nodes was set at 10. All the nodes fulfilled this condition with the exception of the nodes *author* and *writer* in the native speaker and the non-Czech learner sample. Concordance lines containing the type *adjective + noun* and *verb + noun collocation* were investigated. The results in this part of analysis are worth discussing.

Concerning the structural type *adjective + noun,* the results show that Czech learner sample contains most types and tokens (71 types, 90 tokens). As regards the non-Czech learners, they produced 27 types, 28 tokens; the native speaker sample provides 52 types, 58 tokens. The overall counts as far as the distribution of collocations in the BNC is concerned suggests that Czech learners produced 70 per cent of tokens placed in the category “5 and more occurrences” in the BNC. The non-Czech learner sample provides 57 per cent of such collocations, the native speaker sample contains 55 per cent. Perhaps contrary to expectations, the total of 31 per cent of collocates found in the native speaker sample have no matches in the BNC whereas it is only 9 per cent in the Czech learner sample and 25 per cent in the non-Czech learner sample. There are several possible explanations, though: a few collocations in the native speaker sample, as already noted, are used exclusively in American English and thus can hardly be expected in the BNC (e.g. *a soft-cover book, a high-priced book).* Furthermore, since the language of reviews falls into the domain of written English genre where critics desire to draw readers' attention, they use collocations which sound perfectly natural to a native speaker (e.g. *overpriced, parenting, conflicting book*) and still, they are not used commonly in everyday life situations or colloquial language. Therefore, it is not surprising that the native speaker sample contains collocations of this kind. By contrast, both non-native speaker samples suggest that learners prefer "safe bets" and avoid creativity at the expense of collocational richness. The fact that a great number of very frequent collocations in the BNC are found especially in the Czech learner sample is explicable on the grounds that most of the adjectival collocates are very general adjectives (e.g *good, wonderful, short, beautiful, popular, different, ideal, whole etc.)* and thus can hardly be dismissed as atypical or impossible collocates. Similarly, even though the non-Czech learner sample provides only a small number of collocations, several of the adjectival collocates are again very general adjectives such as *good, same, whole, third, short*. The native speakers' repertoire of adjectival collocates occurring in the BNC resembles, to a certain extent, that of both non-native speakers'**.** However, collocations in the native speaker sample appear to be more varied (e.g. *delightful, acclaimed, precious*) in comparison with both non-native samples.

As for the structural type *noun + verb,* the results are as follows: despite the fact that Czech learners produced again most types and tokens (59 types, 74 tokens) and the number of the BNC high-frequency collocations found in the Czech learner sample is 54 per cent of tokens, the same is true of the native speaker sample wherein 54 per cent of tokens have their matches in the BNC. By contrast, non-Czech learners produced only 22 per cent of very frequent verbal collocates. Again greater collocational richness is observable in the native speaker sample. The collocations such as *stories unfold, the book is filled with, the book claims, the author covers points* etc. clearly illustrate the point (see Tables 35, 42).

All in all, a great number of the BNC high-frequency collocations found especially in the Czech learner sample suggest that Czech learners prefer using combinations they are familiar with and choose very general collocates at the expense of collocational richness. In other words, most of such combinations would not be treated as collocations within the phraseological approach. Appendix 4a, b, c gives the list of all collocations found in all three sample corpora.

6.14 Salient collocations

Cowie (2005, 13) explains the term "sense of salience" as "a sense of what constitutes a conventional ready-made collocation in English". In order to find out the learners' sense of salience, a total of 21 participants (15 secondary school students at intermediate level and 6 adult learners exhibiting FCE to CAE level) were asked to take the test of salient collocations created by Sylviane Granger. The term collocation is used by Granger to refer to "the linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its synonyms because of constraints which are not on the level of syntax or conceptual meaning but usage" (in Cowie 2005, 146).

The test includes the combinations of the *adverb* (in the function of amplifier) + *adjective* type in which the learners are asked to select, from a list of 15 adjective in each case, the acceptable collocates of 10 adverbs, by underlining all the adjectives which they think co-occur with the amplifier. Secondly, if they felt that one adjective co-occurs with the amplifiers more than the rest of the adjectives, they were asked to circle it. The learner data is contrasted with the collocational frequencies retrieved from the BNC. The arbitrary limit of at least 3 occurrences in the BNC is established and the combinations reaching a lower frequency are dismissed as atypical collocations. Greater tolerance as far as the minimum arbitrary limit is concerned was opted for since the collocations under scrutiny represent so-called restricted collocations – these are far less frequent in the language. The term "incorrect collocate", which may be used throughout the following sections, refers to all the adjectives "incorrect" in the sense that they do not occur in the BNC with the amplifiers listed in Granger' test.

6.14.1 The collocation salience test: the most salient collocations (circled)

Since the first part of the analysis focuses on learners' collocation salience, in this part, learners are asked to circle the collocation which, in their opinion, is the most typical of the given node (see Appendix 4d).

Table 43 gives an overview of the data obtained from the students. The first column gives the list of amplifiers, the second column refers to the most salient collocate for the amplifiers (listed in the Granger**'s** collocation salience test), the third column shows the number of correct responses - learners who selected the most salient collocation correctly, the fourth column contains the learners' responses with a collocate which can co-occur with a given amplifier but is not the most salient one. The last column shows learners' use of collocations which do not occur in the BNC at all (incorrect responses).

Table 43: The collocation salience test and learner responses

Amplifier	**BNC frequency of the most salient collocation**	**Responses with the most salient collocation %**		**Responses with a different correct adj for the amplifiers**	**Incorrect adjectives for the amplifiers**
1. highly	*highly significant* (156)	3	14.0	*highly important* (7) *highly reliable* (2) *highly aware* (1)	*highly impossible* (2) *highly available* (2) *highly happy* (1)
2. seriously	*seriously ill* (227)	14	66.0	-	
3. readily	*readily available* (426)	2	9.5	*readily aware* (1)	*readily cold* (1) *readily different* (3) *readily difficult* (2) *readily significant* (2) *readily happy* (1) *readily reliable* (1)
4. blissfully	*blissfully happy* (11)	7	33.0	*blissfully ignorant* (3) *blissfully clear* (1)	*blissfully different* (2) *blissfully essential* (1) *blissfully miserable* (1)
5. vitally	*vitally important* (191)	8	39.0	*vitally significant* (1) *vitally essential* (2)	*vitally happy* (5)
6. fully	*fully aware* (239)	3	14.0	*fully clear* (1) *fully available* (6) *fully reliable* (2)	*fully different* (2) *fully essential* (1) *fully ignorant* (1) *fully impossible* (1)
7. perfectly	*perfectly clear* (117)	11	52.0	*perfectly reliable* (1) *perfectly happy* (2) *perfectly aware* (3)	-
8. bitterly	*bitterly cold* (102)	6	29.0	*bitterly aware* (1)	*bitterly miserable* (2) *bitterly essential* (2) *bitterly ignorant* (1) *bitterly difficult* (1) *bitterly clear* (1)
9. absolutely	*absolutely clear* (149)	3	14.5	*absolutely impossible* (5) *absolutely different* (4) *absolutely reliable* (4) *absolutely happy* (2) *absolutely ignorant* (1)	
10. utterly	*utterly different* (29)	0	0.0	*utterly ignorant* (3) *utterly miserable* (6) *utterly impossible* (1) *utterly clear* (2)	*utterly available* (2) *utterly aware* (1)

Table 43 shows that out of the possible 210 correct answers, the analysis yielded only 162 collocations that were marked by the learners as salient, out of which 57 responses (27 per cent) are correct. In other words, if the overall number of possible correct answers (210) is taken as the native speakers' norm, then the Czech learners' use of the most salient collocations is at 27 per cent. The other 54 responses (25.7 per cent) are not the most salient collocations, though, the search in the BNC confirmed that they do occur at least three times. However, there are few other combinations selected by students (8 responses) which contain collocates occurring in the BNC very scarcely (one or two occurrences).These include *highly aware* (1)*, blissfully clear* (1)*, readily aware* (1)*, vitally essential* (2)*, fully reliable* (2) and *absolutely ignorant* (1). Additionally, some of the collocations which should have been circled were underlined instead but they are not the most salient ones (see the Section 5.12). For more detailed information on the salient collocation frequencies, see Appendix 4e.

To comment on some collocations individually, *highly significant* (156 occurrences in the BNC) was marked as the most salient collocation only by 3 learners (14 per cent), 7 learners circled *important* as the most salient collocate for *highly,* 2 learners selected *reliable* as the most salient collocate for *highly.* Apart from *highly significant* (156 occurrences) which is the most salient*, highly important* (38) and *highly reliable* (8) belong to salient collocations also. The rest of the collocates circled by learners for *highly* (*highly impossible, highly available, highly happy)* are clearly atypical combinations not occurring in the BNC. For more details, see Appendix 4f which provides the list of all adjectival collocates of *highly.*

Similarly, the collocation *readily available* with 426 occurrences attested in the BNC, was circled only by 2 learners (9 per cent) as the most salient one, the rest of the circled collocations with the node *readily* are obviously atypical combinations not attested in the BNC, for instance *readily happy* (1)*, readily reliable* (2)*, readily different* (3). The search in the BNC indicated that no other adjectival collocates from the list collocate with the node *readily* (with the exception of *readily aware,* which occurs only once). One collocation worth mentioning is *seriously ill,* which was correctly marked by 14 learners (66 per cent). A plausible explanation for such a great number of correct responses might be that it translates very nicely into Czech (*být vážně nemocen*). Also *perfectly clear* was marked by half of the learners - 11 (52 per cent) even though we cannot claim that it has a direct equivalent in Czech. No correct response was acquired for the most salient collocate of *utterly* where the most salient collocation is *utterly different,* including *utterly impossible* (11 occurrences in the

BNC) and *utterly miserable* (10) as other typical collocations. Even though 6 learners circled *utterly miserable* as the most salient collocation and 1 learner *utterly impossible* as the most salient collocation, *utterly different* was not circled even once. The overall results confirm the learners' weak sense of salience. Whether learners' native language plays a considerable role is not so much obvious and even though some combinations point to a certain degree of a mother tongue interference (e.g. *fully ignorant, fully impossible*) as well as transfer (e.g. *absolutely happy, absolutely impossible*) and students have a tendency to translate word for word especially in the cases of words they are familiar with, this aspect cannot be considered the main stumbling block in this investigation. The main impediment seems the insufficient knowledge of some amplifiers and their collocates respectively which results in atypical collocations and confirm the learners' weak sense of salience.

6.14.2 The collocation salience test: underlined collocations

In the second part of the test on salient collocation, the learners were required to underline all possible adjectives which collocate with the amplifiers (nodes). As in the previous section, the cases with fewer than 3 occurrences in the BNC were dismissed, the cut-off limit (strictly arbitrary) of at least 3 occurrences was set before a combination could be called a collocation. At this stage, the analysis was divided into 3 parts.

The first part (see section 6.14.2.1) seeks to find out how many students out of 21 underlined the typical collocates for the amplifiers in question. However, 100 per cent accuracy is not the aim of this investigation since it would literally approach zero. In other words, to analyse how many students underlined correctly all possible adjectives and the corresponding amplifier would be pointless - there would be no such students.

The second investigation of underlined collocates (see Section 6.14.2.2) is confined to the nodes *readily* and *seriously.* These two amplifiers are specific in that according to the BNC, they do not co-occur with other collocates than the most salient ones - *seriously* ***ill*****,** *readily* ***available*** (with the exception of *readily aware* which occurs in the BNC only once and thus is dismissed in this analysis as atypical). What becomes the primary concern of this investigation is the number of students who did not underline any adjective from the set of available adjectives. In this case, 100 per cent accuracy is focused on.

The third part of the analysis (see Section 6.14.2.3) focuses on the number of students, who underlined the collocates which were supposed to be circled and thus ascribed them less importance.

The erroneously underlined collocates are plentiful. They will be commented upon only very briefly since this would go beyond the scope of the analysis.

6.14.2.1 Correct responses with underlined collocates

Table 44 presents the number and percentage of students who underlined the possible collocates of the nodes. The second column contains all possible collocates (not the most salient ones) with the corresponding BNC frequencies, the third column shows how many students out of 21 underlined the right collocates.

Table 44: Learner responses - underlined collocations

Amplifier	Possible collocates (BNC frequency)	Correct responses (out of 21)	%
1. highly	*important* (38) *reliable* (8)	*important* (9) *reliable* (11)	43.0 52.0
2. seriously	-	-	
3. readily	-	-	
4. blissfully	*ignorant* (6)	*ignorant* (5)	24.0
5. vitally	*aware* (5) *significant* (3)	*significant* (5)	24.0
6. fully	*clear* (12) *available* (6)	*available* (4) *clear* (2)	19.0 9.5
7. perfectly	*happy* (96) *aware* (17) *reliable* (6)	*reliable* (4) *happy* (3) *aware* (3)	19.0 14.0 14.0
8. bitterly	*aware* (6)	*aware* (2)	9.5
9. absolutely	*essential* (122) *impossible* (119) *reliable, different* (5) *happy* (3)	*different* (12) *reliable* (10) *happy* (10) *impossible* (9) *essential* (5)	57.0 48.0 48.0 43.0 24.0
10. utterly	*miserable* (10) *impossible* (11) *clear, happy,* *ignorant* (3)	*ignorant* (3) *clear* (1) *miserable* (1) *impossible* (1)	14.0 4.7 4.7 4.7

To comment briefly on some examples, a great number of the correct responses concern the node *highly* in which case 9 out of 21 learners (43 per cent) opted for *highly important* and 11 learners (53 per cent) for *highly reliable* as typical collocations. Small wonder Czech learners underlined these collocations since both translate very nicely into Czech. With the collocates of *blissfully,* for instance*,* the learners were clearly confused by or perhaps unfamiliar with the meaning of the amplifier since they selected combinations which are evidently contradictory e.g. *blissfully miserable, cold, important* (1). Only 5 learners (24 per cent) opted correctly for *blissfully ignorant.* With *vitally aware,* there was not even 1 correct underlined collocation. The highest number of correct responses was in the case of *absolutely,* in which 12 learners (57 per cent) opted for *absolutely different,* 10 learners (48 per cent) decided to underline *absolutely happy* and *absolutely reliable;* 9 learners (43 per cent) *absolutely impossible;* 5 learners (24 per cent) selected *absolutely essential.* Again most of them have direct Czech equivalents. On the other hand, the amplifier *utterly* proved very problematic. In the majority of cases in terms of *utterly*, only one student opted for the right alternative. The students may have been either unfamiliar with the amplifier or at least they did not know that this amplifier tends to convey meaning with negative connotations.

6.14.2.2 Underlined collocates of *readily, seriously*

The amplifiers *readily* and *seriously* will be commented upon separately. In this case, students with correct responses are those who underlined no collocate from the list of adjectives since no other collocations than *readily available (426)* or *seriously ill (227)* occur in the BNC. The only exception is the "collocation" *readily aware* which does occur in the BNC but only once. The results obtained from the students show that slightly more than one third of the students were correct in their responses - they did not underline any collocate. In other words, the total of 8 students (38 per cent) did not underline any other adjective for both *readily* and *seriously.*

6.14.2.3 Underlined collocations that should have been circled

Table 45 presents the results of the most salient collocates. Instead of being circled, they were underlined by learners as one of those possible collocates. In other words, the students marked the collocate as less typical than it actually is and gave the collocate less significance. The first column provides a list of the amplifiers, the second column lists the underlined collocates which were supposed to be circled, the third column gives the number of learner responses.

Table 45: A survey of the most salient collocations underlined instead

Amplifier	Underlined collocates that were to be circled	Learner responses	%
1. highly	*significant*	5	24.0
2. seriously	*ill*	3	14.5
3. readily	*available*	2	9.5
4. blissfully	*happy*	3	14.5
5. vitally	*important*	3	14.5
6. fully	*aware*	2	9.5
7. perfectly	*clear*	5	24.0
8. bitterly	*cold*	1	5.0
9. absolutely	*clear*	10	48.0
10. utterly	*different*	2	9.5

6.15 Summary of the collocation salience test findings

To summarize our findings, the results from all parts of the collocation salience analysis confirm that the learners' sense of collocation salience is weak. Namely, in the first part of this investigation focusing on the most salient collocations, the analysis yielded only 27 per cent of correct responses. In other words, if the number 210 (the possible correct answers) is taken as the native speakers' norm, then the Czech learners' use of the most salient collocations is at 27 per cent of this norm (approximately one third that of native speakers'). For more details see Table 43.

As regards the amplifiers *readily* and *seriously,* a similar result was obtained – approximately one third of the students (38 per cent) were correct in their responses. In this particular case, students with correct responses are those who did not choose, from the Grangers' list of adjectives, any other collocate for the amplifier *readily* than *available* and for the amplifier *seriously* the collocate *ill* (no other collocations than *readily available* or *seriously ill* from the collocation salience test occur in the BNC). The only exception is *readily aware* with one occurrence in the BNC.

The part of the collocation salience investigation where learners were asked to underline other possible collocates for the listed amplifiers shows that correct responses are not plentiful. The exceptions are amplifiers *highly* and *absolutely* (see Table 44). Given that the number of 21 correct responses for one collocation is taken as the native speakers' norm, then Czech learners achieve 57 per cent of correct responses in terms of the collocation *absolutely different,* 52 per cent of correct responses with *highly reliable* and 48 per cent of correct responses with *absolutely reliable* and *absolutely happy.* However, in the majority cases, it is only around 20 percent (*vitally significant, blissfully ignorant, fully available).* The amplifier *utterly* is the most problematic for learners– only one learner selected *utterly impossible, utterly miserable* or *utterly clear* as salient collocations.

All in all, a*bsolutely* and *highly* are two amplifiers where learners had more courage to select more collocates, presumably due to the fact that both amplifiers are familiar for learners (transfer from Czech) as opposed to, for instance, *utterly.* A great number of clearly atypical collocations, inadmissible for native speakers, were marked by the learners, however, these are not cases of mother tongue interference - they have no counterparts in Czech.

To answer the question whether salient collocations pose a problem for learners is obvious. The analyses show that even several advanced students who participated in this project find such collocations very challenging. The immediate implication of the replication test is that Czech learners are not able, in the majority of cases, to opt for “the correct” collocation and their sense of collocation salience is weak.

7. Conclusion

The research reported in the thesis explores the degree of authenticity of the formulaic language used by NNSs and the extent to which a learner's L1 interferes in the production of multi-word units. Drawing on Granger's Contrastive Interlanguage Analysis (CIA 1996), which compares not only NSs and NNS, but also learners with different language backgrounds and focuses on features both common and unique to these learners, the investigation was conducted on two different learner sample corpora and subsequently contrasted with a native sample corpus. Different types of evidence (based on the BNC, the PIE, existing dictionaries and native speakers' introspection) were used in the evaluation of the findings.

The aim of the study was to confirm the hypothesis that multi-word units present a challenge for non-native speakers for several reasons. In general terms, it was assumed that the learners would be more inclined to the application of what Sinclair calls the open-choice principle - their language production would largely proceed on a "slot-and-filler" basis and be less idiomatic than that of the native speakers. This assumption was independently tested on three types of phraseological combinations, lexical bundles, multi-word verbs and collocations. In the chapter on lexical bundles or non-idiomatic recurrent word-combinations, contrary to the assumption, learners were expected to produce more types and tokens of these non-idiomatic sequences and adopt a more repetitive pattern of expression than native speakers. Although lexical bundles represent single choices (and therefore come under the heading of Sinclair's idiom principle), in this particular case learners were assumed to follow a safe and secure strategy. The native speakers, on the other hand, were expected to demonstrate more creativity in their reviews and so use fewer recurrent sequences. Regarding the phrasal verbs, the non-native speakers were thought to produce a smaller number of phrasal verbs than native speakers. In the chapter focusing on collocations, a weak sense of collocation salience was expected in the non-native speakers. Even though the results obtained in each chapter generally tend to support the initial hypothesis, they also indicate that it is unwise to draw premature conclusions about a non-native speaker's use of multi-word sequences. Indeed, the previous studies focusing on all kinds of multi-word units have come up with many conclusions which vary to some degree.

Unlike most other studies, though, this pilot probe, by examining several kinds of multi-word units at once, serves a different purpose: it attempts to develop a composite methodology that will show the degree of idiomaticity used by (Czech) learners in their English language production. Due to investigating as many as three types of multi-word units in four ways the samples had to be restricted and the results are to be taken as tentative. As the methodology appears to have proved feasible it opens the way for studies using larger samples with more ambitious and specific goals.

After reviewing the major findings of the study in the first part of the conclusions it is perhaps fitting to consider some perspectives for future research and ELT learning in connection with phraseological language in the light of our findings about multi-word combinations.

7.1 Review of major findings

The body of the thesis consists of three main chapters analysing (a) non-idiomatic recurrent word combinations or lexical bundles (Chapter 4), (b) multi-word verbs with a special focus on phrasal and prepositional verbs (Chapter 5), and (c) collocations (Chapter 6) whose findings will be reviewed in this order.

Starting with Chapter 4 devoted to contrastive analysis of three- and four-word combinations, we had two objectives. First, to seek confirmation that both learner groups will be more repetitive in the number of word-combination types and tokens whereas native speakers will be more creative in the use of word-combinations. Second, it was expected that learners would produce fewer distinct lexical bundles in the strict sense than native speakers and their word-combinations would mainly include sequences created on an ad hoc basis. These were expected to have some matches in the PIE, but not enough to qualify as lexical bundles in the strict sense. Two terms were adopted for the three- and four-word sequences in this chapter: “word-combinations” and “lexical bundles”. The frequency threshold was set at least at ten occurrences per million words, a criterion adopted by Biber et. al. (1999). The term “word-combination” was used for any three- or four-word sequence regardless of its frequency in the PIE whereas the term “lexical bundle” was used only for such word-combinations which occur more than ten times per million words.

After the three- and four-word sequences were identified using the software program Collocate, their frequencies were checked against the PIE. The findings obtained confirm that the non-native and native speaker word-combinations show considerable differences and only some similarities. Both learner groups produced almost twice as many word-combinations than the native speaker group. Namely, Czech learners produced 127 four-word combinations types. Similarly, non-Czech sample provides 119 four-word combination types whereas native speaker sample only 54 four-word combination types. As regards the three-word combination types, 370 three-word combination types were identified in the Czech sample, 320 three-word combination types in the non-Czech sample. However, the search yielded only 220 three-word combination types in the native sample. Such results support the initial assumptions about greater repetitiveness in the learner samples and the aspect of creativity in the native sample. It is also worth mentioning that the frequency of types was relatively stable in the native sample. The majority of word-combinations occur twice. The Czech sample shows the opposite, though. The range of the frequency of types is from twelve to two and the uneven distribution is observable especially as regards three-word combinations. Subsequent analyses revealed interesting findings about lexical bundles in the strict sense and the word-combinations with either low frequencies in the PIE or the word-combinations with zero occurrence in the PIE. The investigation in the PIE yielded more three-word true lexical bundles than four-word true lexical bundles in all three samples. In fact, four-word bundles in the strict sense were almost missing in all samples. Even though Czech learners produced the highest amount of true lexical bundles from all three groups, the number was only slightly higher (7.1 per cent of four-word bundles; 22. 7 per cent of three-word bundles) than in the native sample (5.6 per cent of four-word bundles; 19. 1 per cent of three-word bundles). Therefore, it would not be reasonable to claim that Czech speakers produced most lexical bundles of all. The validity of such findings could be either confirmed or refuted only by using a larger sample. However, if we assume for the moment that a larger sample would yield a similar result, then a possible explanation for a greater use of lexical bundles by Czech learners could be that learners are more inclined to use sequences they are familiar with rather a than more creative approach. Apart from the group of true lexical bundles, the examination also revealed that a large number of word-combinations in all three samples had no matches in the PIE. These sequences include mainly word-combinations consisting of book titles, names of authors, films (e.g. *book Harry Potter and, The Adventures of Huckleberry*). Also, many of the word-combinations, though, attested in the PIE were not frequent enough to qualify as lexical bundles (e.g. *the book I read, recommend a book which*). Most of these are topic-bound, mainly related to the semantic field of reading and thus it is not surprising that they have become prominent in the sample.

Chapter 5, dealing with phrasal-verb and prepositional-verb use, focuses on the range and the frequency of phrasal and prepositional verbs. Since the difficulties learners usually encounter with phrasal verbs differ from those they have with prepositional verbs the two classes of verbs were analysed separately. Numerous studies have proved that learners often struggle with phrasal verbs for several reasons: a great number of phrasal verbs often carry several meanings, out of which some can be completely opaque, some learners perceive particular phrasal verbs as problematic for their complete absence in their L1. Further, the specific context in which these verbs must be used is also not entirely easy for learners to master. As Sinclair (1991) observes, each phrasal verb carries its own lexical, semantic, syntactic as well as pragmatic implications. By contrast, prepositional verbs pose a challenge for learners primarily from the point of view of choosing the appropriate preposition. The choice of the preposition which is a matter of learning “by heart ” seems, in the majority of cases, to be the main impediment for learners. The findings obtained in this investigation are in keeping with all the observations made in the literature: there is only a small incidence of phrasal verbs in the written language, they are characteristic of the spoken language.

In particular, 25 types and 36 tokens were identified in the Czech learner sample; 43 types and 57 tokens were found in the non-Czech learner sample; 53 types, 64 tokens were produced by native speakers. If the distribution of phrasal verbs in the native speaker sample is taken as the norm, then the Czech speakers’ use of phrasal verb types is at 47.2 per cent and the use of phrasal verb tokens at 56.2 per cent. In other words, the distribution of phrasal verbs in the Czech sample is half that of the native speakers’ in both respects.

Further findings worth mentioning support the hypothesis that phrasal verbs occur with half the frequency of prepositional verbs in the non-native samples and a comparably low incidence of phrasal verbs was marked in the native sample too. Still, the native sample contains twice as many phrasal verbs as well as prepositional verbs than the Czech sample and the range of phrasal verbs found in the native sample is much wider in comparison to the non-native samples. That phrasal verbs pose a potential pitfall for language learners, something that has been argued in numerous studies before, is patently obvious in both learner samples. Although most phrasal verbs produced by the learners in the present study are used appropriately, pre-intermediate as well as intermediate learners seek alternative ways of expressing the meaning and avoid phrasal verbs altogether. The dubious cases involve mainly the inappropriate extension of collocational range, the use of a simple lexical verb instead of a phrasal one or the use of an inappropriate phrasal verb where a different phrasal verb is necessary.

The second area of interest investigated in the chapter on multi-word verbs was the use of prepositional verbs. As was mentioned above, prepositional verbs are relatively frequent in English and occur equally importantly in all registers in comparison with phrasal verbs. The sample analysis confirms this: all three samples, Czech, non-Czech and native, contain almost twice as many prepositional verbs than phrasal verbs, with the native speakers producing the most prepositional verbs of all three samples. Namely, Czech learners produced 60 types, 111 prepositional verb tokens; 90 types, 130 tokens were found in the non-Czech learner sample, 101 types, 159 tokens were identified in the native speaker sample. Another concern of this chapter was to find out whether the semantic groups of prepositional verbs used by learners and native speakers differ in distribution. The survey of semantic types proves that in all three samples the largest semantic group of verbs is the group of activity verbs, the group which is reported to be the most frequent in the language (LGSWE 1999). This semantic analysis shows that despite a lower number of prepositional verbs in the learner samples, the learners use and are familiar with a great number of prepositional verbs which belong to the most frequent ones in English and that the style they adopt is more or less neutral.

Chapter 6 describing the investigation concerning collocations subsumes two types of analyses. The first one examines the collocational behaviour of selected words functioning as nodes (in the node-collocate pair). The objectives of this investigation were to find out to what extent learners produce collocations which occur frequently in the BNC; it also focused on the range of collocates used by the learners and the native speakers. Two obvious conclusions emerge from the investigation: learners produce such collocations easily enough unless specifically asked to match the node with the possible collocates. Both learner groups produced a great number of collocations that occur with a high frequency in the BNC and in the Czech learner sample it was even 70 per cent of tokens in terms of the *adjective + noun* collocation. Second, most of the collocates in the non-native samples are very general and thus have only little information value (e.g. *first, same, good, favourite*). As Klégr points out (2005, 91) there are appropriate methods to obtain statistically significant collocates, however, “focusing on statistically significant collocates will not provide a comprehensive enough picture of the node’s combinability”. The non-native samples provide evidence that in the majority of cases collocates are represented by very common adjectives (in the type *adjective + noun)* and very common verbs (in the type *noun + verb)* and practically none of these collocates could be dismissed as atypical. Nevertheless, such collocations are so general that they can hardly be assigned the status of phraseological expressions. The fact remains that the native speakers' repertoire of collocations is wider. Apart from this, the native sample contains unusual collocations with no occurrences in the BNC (e.g. *self-proclaimed writer, highly-acclaimed novel, story unfurl*) and still, native speakers find them perfectly natural, although for instance "self-proclaimed writer" has quite a number of hits on the web. This is not true of some of the collocations in the non-native samples having zero occurrence in the BNC and being rejected by native speakers as "unnatural" (e.g. *novel released, stories aided by*). Especially the type of collocation *noun + verb* is more mutually selective and the analysis confirms that both groups of non-native speakers produced a number of collocates of this type which are not found in the BNC, with several of them sounding distinctly odd to native speakers.

The second type of collocational analysis described in Chapter 6 is a replication of Sylviane Granger's (2005) collocation salience test of restricted collocations of *adverb + adjective* type. The analysis confirms that the learners' sense of salience is weak, i.e. they have difficulties assigning typical adverbs to the adjectival nodes. It is worth noting that only 27 per cent of learner responses were correct in the test assessing the most salient collocation. Furthermore, it emerges that learners find not only the most salient but other salient collocations extremely challenging. It is true that 57 per cent of Czech learners correctly opted for the collocation *absolutely different* or that there are 52 per cent of correct responses in terms of *highly reliable.* Nevertheless, these correct responses are minor exceptions since it is only around 20 percent in the majority of cases (e.g. *vitally significant, blissfully ignorant, fully available* etc.). As regards the amplifier *utterly,* for instance, it is only 4.7 per cent of correct responses.

The *adverb + adjective* type of collocation is usually encountered by learners who have reached the advanced or proficiency level. Even at this stage, such collocations often present a formidable challenge for learners (Granger 2005). The immediate implications of the replication test is that even Czech learners are often not able to distinguish between the "good" and "bad" collocations, which points to their unfamiliarity with and poor knowledge of such restricted collocational pairs and consequently, their weak sense of collocation salience.

7.2 Prospects for future research

While it is hoped that the present study has at least partly clarified and outlined the divergences in the non-native and native production of multi-word units and sufficiently confirmed the initial hypotheses, the limitations following from such small-scale research, based on relatively small sample corpora less than 30 000 words altogether are obvious. As has been mentioned above, the results are viewed as preliminary and the main thrust of this contrastive study of the use of multi-word units in non-native and native speakers was to develop and test a methodology that will assess the degree of idiomaticity in learners' language production. Having established that such assessment is possible, a much more detailed account that would provide evidence on the differences between non-native production and native production of multi-word units is called for, drawing on larger samples. Since only two groups of non-native speakers participated in the investigation, further research could include learners from different linguistic backgrounds and language families in the investigation. It is assumed that each group of language learners would have, in the words of Pawley and Syder (1983), a specific foreign flavour. Further, different levels of language learners could participate in the investigation. The most typical pitfalls that learners face could be specified as well as the type of multi-word units that proves the least or the most problematic. The learners' (appropriate) use of diverse multi-word units will undoubtedly depend on the register under scrutiny, and the results obtained from such analyses will be influenced accordingly. Given that essay writing represents quite a specific text type and that some multi-word units focused upon in the present study will have specific distribution, further analyses in different registers and text types can be expected to yield different statistical counts. Nevertheless, the difference between non-native and native production will certainly be in evidence and the results could show a more obvious gap between non-native and native multi-word unit production than revealed by this study.

Further research into multi-word units could focus especially on the following:

1. recurrent non-idiomatic word combinations (lexical bundles) produced by learners in different registers and at different learner levels (classroom language);
2. an in-depth analysis of phrasal verbs seeking confirmation that with increasing proficiency learners tend to use more phrasal verbs and use them effectively in the appropriate contexts;
3. since the majority of the participants in this research included learners with lower

levels of English proficiency, it would be useful to focus on phrasal verb use by learners exhibiting advanced to high proficiency levels;

4. factors which influence the increase of use of phrasal verbs (exposure to the language – direct or indirect contact with native speakers, study stays etc.);
5. the learners' production of phrasal verbs in different registers of spoken language since written language, let alone essay writing, does not presuppose much use of phrasal verbs in general;
6. investigation of phrasal verbs in terms of the extended lexico-grammatical unit framework; the resultant phrasal-verb "profiles" – collocational, colligational, semantic and pragmatic – in learners and native speakers could be then compared; it will be interesting to see whether advanced learners are able to follow (subconsciously) the grammar, and semantic and pragmatic patterns specific to phrasal verbs;
7. comparison of phrasal-verb use between students exhibiting the same level of English but coming from different language environments;
8. several studies claim that learners' production of phrasal verbs largely depends on the mother tongue language family, hence a study should be made whether language proficiency influences the salient use of phrasal verbs even with learners whose mother tongue comprises phrasal verbs;
9. contrastive investigation of collocations involving a wide range of nodes of specified word-class status with a sufficient number of concordance lines that would allow comparison of their collocates and all structural types of collocations as used by learners with different language backgrounds;
10. the collocation salience test that would involve other structural types of restricted collocations than just *adverb* + *adjective*;
11. detailed analysis of lexical priming that would explore how much learners are primed for selected words in comparison with native speakers.

7.3 Prospects for ELT learning

Apart from the suggestions for further research into multi-word units listed above, the question arises what the implications of the results such as produced by our study are for ELT learning. While the present study confirms that multi-word units occur in non-native writing even at lower levels of language proficiency, considerable differences exist between non-native and native speakers' language production. One of the immediate implications is that there is an urgent need to raise learners' awareness of ready-made sequences (especially low-proficiency learners) and the importance of these sequences in language production, to find a way of teaching them to learners effectively and encourage the appropriate use of such ready-made sequences.

From my own teaching experience, most learners (especially those at initial levels of English) still seem to be somewhat doubtful about the significance of chunk-based language. They hardly realize that ready-made sequences are the synonym for native-like fluency. Leafing through a typical English textbook, it is possible to observe that phrasal verbs, prepositional verbs, collocations or even idioms receive some treatment. However, in the majority of cases, the amount of attention given to formulaic language especially in the textbooks aimed at learners with low levels of proficiency is far from sufficient. Even though it is possible to come across several phrasal verbs, collocations or idioms in these textbooks, they are usually integrated into the texts without being focused upon separately in extended sections devoted to idiomatic language. EFL teachers would be well-advised to take great care to persuade learners about the significance of chunk-based language and do their utmost to provide learners with as much idiomatic language as they are likely to encounter in everyday life. Apart from spending time on textbook activities featuring an adequate amount of idiomatic language, teachers would do well to increase their students' motivation so that the students themselves make use of a wide range of opportunities offering authentic English language material brimming with idiomatic language. Of course, there are no specific guidelines as to what multi-word sequences take priority over others, which ready-made sequences should particularly be incorporated into language learning and which not. At a guess, a good strategy might be to include such sequences which proliferate in the language and which learners are likely to come across in everyday life situations in the English speaking environment. Regardless of the answer to the question of whether or not it is plausible to achieve native-like fluency, the foremost experts in linguistics rarely find fault with Hoey's claim (2005): "A key factor in naturalness is collocation. Naturalness comes when there is a regular exposure to authentic material".

References:

ALTENBERG, B. (1998). “On phraseology of spoken English: The evidence of recurrent word combinations”. In A. P. Cowie (ed.) *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press, 101-122.

BAHNS, J. (1993). “Lexical collocations: A contrastive view”. *ELT Journal,* 47, 56-63.

BAKER, M. (1992). *In other words: A coursebook on translation*. London: Routledge.

BARLOW, M. (2000). “Parallel texts in language teaching“. In S. P. Botley, T. McEnery, A. Wilson (eds.) *Multilingual corpora in teaching and research.* Amsterdam: Rodopi, 106 -115.

BENSON, M., BENSON, E., ILSON, R. (1986). *The lexicographic description of English*. Amsterdam: John Benjamins.

BIBER, D., CONRAD, S., CORTES, V. (2003). “Lexical bundles in speech and writing: An initial taxonomy”. In A.Wilson, P. Rayson, T. McEnery (eds.) *Corpus linguistics by the Lune: A festschrift for Geoffrey Leech*. Frankfurt am Main: Peter Lang, 71-92.

BIBER, D., CONRAD, S., CORTES, V. (2004). “If you look at....: Lexical bundles in university teaching and textbooks”. *Applied Linguistics,* 25, 371-405.

BIBER, D., BARBIERI, F. (2007). “Lexical bundles in university spoken and written registers”. *English for specific purposes*, 26, 263-286.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S., FINGEGAN, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

BLOOMFIELD, L. (1933). *Language and linguistics*. New York: Henry Holt and Co.

BOWKER, L. (2001). ”Towards a methodology for a corpus-based approach to translation evaluation”. *Meta: Translator's journal,* 46-2, 345-364.

BOLINGER, D. (1971). *The phrasal verb in English*. Cambridge: Harvard University Press.

BOLINGER, D. (1976). “Meaning and memory”. *Forum Linguisticum,* 1, 1-14.

BUTLER, C.S. (1997). “Repeated word combinations in spoken and written texts: some implications for functional grammar”. In C. S. Butler, J. H. Connolly, R. A. Gatward, R. M.Vismans (eds.) *A fund of ideas: Recent developments in functional grammar.* Amsterdam: University of Amsterdam, 60-77.

CLARIDGE, C. (2002). “Multi-word verbs in early modern English. A Corpus-based study”. Amsterdam-Atlanta, GA: Rodopi, B.V, 27-51.

CORNELL, A. (1985). “Realistic goals in teaching and learning phrasal verbs”. *IRAL*, XXIII, (4), 269-280.

CARTER, R. (1998). *Vocabulary: Applied linguistics perspectives.* London: Routledge.

CHEN, YU-HUA, BAKER, P. (2010). “Lexical bundles in L1 and L2 academic writing”. *Language Learning and Technology,* 14-2, 30-49.

CORTES, V. (2002). “Lexical bundles in academic writing in history and biology”. Doctoral Dissertation. Northern Arizona University.

CORTES, V. (2004). “Lexical bundles in published and student disciplinary writing: Examples from history and biology”. *English for Specific Purposes,* 23, 397-423.

COSERIU, E. (1967). “Lexikalische Solidaritäten”. *Poetica,* 1, 293-303.

COWIE, A. P. (1999). *English dictionaries for foreign learners: A History.* Amsterdam and Philadelphia: John Benjamins.

COWIE, A.P. (ed.), (2005). *Phraseology: Theory, analysis, and applications*. Oxford: Oxford University Press.

CRYSTAL, D. (1995). *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press.

CRUSE, D. A. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

CRUSE, D. A. (2000). *Meaning in language*. Oxford: Oxford University Press.

ČERMÁK, F., BLATNÁ, R. (eds.), (2006). *Korpusová lingvistika: Stav a modelové přístupy.* Praha: Lidové noviny.

ČERMÁK, F. (2007). *Frazeologie a idiomatika česká a obecná.* Praha: Karolinum.

DARWIN, M., GRAY, S. (1999). “Going after the phrasal verbs: An alternative approach to classification”. *TESOL Quarterly,* 33-1, 65-83.

DECHERT, H. (1984). “Second language production: Six Hypotheses”. In H. Dechert, D. Mohle, M. Raupach (eds.) *Second language productions*. Tübingen: Gunter Narr Verlag, 211-230.

DUŠKOVÁ, L. (1994). *Mluvnice současné angličtiny na pozadí češtiny.* Praha: Academia.

FIRTH, J. R. (1957). “Papers in linguistics, 1934-51”. Oxford: Oxford University Press.

FRANCIS, W. N. (1958). *The structure of American English.* New York: Ronald Press.

FRAZER, B. (1976). *The verb-particle combination in English*. New York : Academic Press.

GRANGER, S. (2005). “Prefabricated patterns in EFL Writing: Collocations and Formulae”. In A. P. Cowie (ed.) *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press, 145-160.

GRANGER, S., PAQUOT M. (2008). “Disentangling the phraseological web”. In S. Granger, F. Meunier (eds.) *An interdisciplinary perspective.* Amsterdam/Philadelphia: John Benjamins, 27-49.

GRANGER, S., DAGNEAUX E., MEUNIER, F., PAQUOT, M. (2009). *International Corpus of Leaner English V2 (Handbook + CD-ROM).* Louvain-la-Neuve:Presses universitaires de Louvain.

HALLIDAY, M. A. K. (1961). “Categories on the theory of grammar”. *Word,* 17, 241-92.

HALLIDAY, M. A. K. (1966). “Lexis as a linguistic level”. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, R. H. Robins (eds.) *In memory of J.R. Firth* (1957*)*. London: Longman, 148-162.

HALLIDAY, M. A. K. (1978). *Language as a social semiotics*. London: Edward Arnold.

HOEY, M. (2005). *Lexical priming*. London: Routledge.

HOWARTH, P. (1996). “The phraseology of learners’ academic writing”. In A. P. Cowie. (ed.), (2005). *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press, 161-186.

HUDDLESON, R., PULLUM, G. (2002). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

HUNSTON, S. , FRANCIS, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English.* Amsterdam and Philadelphia: John Benjamins.

HYLAND, K. (2008). “As can be seen: Lexical bundles and disciplinary variation”. *English for Specific Purposes,* 27, 4-21.

JESPERSEN, O. (1924). *The philosophy of grammar.* London: Allen & Ulwin.

KATZ, J., FODOR, J. A. (1963). “The structure of a semantic theory”. *Language*, 39-2, 170-210.

KLÉGR, A. (2005). “Sadness/Smutek: a comparison of the verbal collocates”. In J. Čermák et al. (eds.) *Patterns: A festschrift for Libuše Dušková.* Modern Language Association (KMF), 91-105.

KRASHEN, S. (1981). *Second language acquisition and second language learning*. Oxford: Pergamon Press.

LEWIS, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove: LTP.

LIPKA, L. (1972). *Semantic structure and word-formation. Verb-particle constructions in contemporary English*. Munchen: Wilhelm Fink.

LIPKA, L. (1990). *An outline of English lexicology.* Tübingen: Gunter Narr.

LIVE, A. H. (1965). "The discontinuous verb in English". *WORD,* 21, 428-451.

LYONS, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

MASON, O. (2007). "Multi-word units as a model of grammar". *Proceedings of CL 2007.* Birmingham.

Mc ENERY, T., WILSON, E. (2007). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

MELČUK, I. (1998). "Collocations and lexical functions". In A. P. Cowie (ed.) *Phraseology: Theory, analysis, and applications.* Oxford: Oxford University Press, 23-53.

MELČUK, I. (1988). "Semantic description of lexical units in an Explanatory Combinatorial Dictionary: basic principles and heuristic criteria". *International journal of Lexicography,* 1-3*,* 165-188.

MELČUK, I. (1995). "Phrasemes in language and phraseology in linguistics". In M. Everaert, et al. (eds.) *Idioms: Structural and psychological perspectives.* Hillsdale, New Jersey: Lawrence Erlebaum Associates, 167-231.

MOON, R. (1998). *Fixed expressions and idioms in English*. Oxford: Oxford University Press.

NATTINGER, J., De CARRICO, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

NESSELHAUF, N. (2005). *Collocations in a learner corpus*. Amsterdam and Philadephia: John Benjamins.

NESSELHAUF, N. (2004). "Learner corpora and their potential for language teaching". In J. Sinclair (ed.) *How to use corpora in language teaching*. Amsterdam-Philadelphia: John Benjamins.

NESI, H., BASTURKMEN, H. (2006). "Lexical bundles and discourse signaling in academic lectures". *International Journal of Corpus Linguistics,* 11-3, 283-304.

PALMER, H. E. (1933). *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.

PALMER, F. R. (1976). *Semantics*. Cambridge: Cambridge University Press.

PAWLEY, A., SYDER, F. H. (1983). "Two puzzles of linguistic theory: native-like fluency and native-like selection". In J. Richard, R. Schmidt (eds.) *Language and Communication.* London: Longman, 191-227.

QUIRK, R. et al. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

ROBINS, R. H. (1964). *General linguistics: An introductory survey.* London – New York.

SADEGHI, K. (2009). "Collocational differences between L1 and L2: Implications for EFL Learners and Teachers". *TESL, Canada Journal*, 26-2, 100-124.

SALEM, A. (1987). *Pratique des Segments Répétés.* Paris: Institut National de la Langue Francaise.

SCHMITT, N. (2000). *Vocabulary in language teaching.* Cambridge: Cambridge University Press.

SCHMITT, N. (ed.), (2004). *Formulaic sequences: Acquisition, Processing and Use*. Amsterdam/ Philadephia: John Benjamins.

SCHMITT, N., GRANDAGE, S., ADOLPHS, S. (2004). "Are corpus-derived recurrent clusters psycholinguistically valid?". In N. Schmitt (ed.) *Formulaic sequences: acquisition, processing and use.* Philadelphia: John Benjamins, 173-189.

SCOTT, M., TRIBBLE, CH. (2006). "Textual patterns: Key words and corpus analysis in language education". Amsterdam/Philadelphia: John Benjamins.

SINCLAIR, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

SINCLAIR, J. (2004). *Trust the text.* Oxford: Oxford University Press.

STUBBS, M. (2002). *Words and phrases.* Blackwells Publishing Ltd.

STUBBS, M. (2004). "On very frequent phrases in English: Distributions, functions and structures". *A plenary lecture given at ICAME 25, the 25th anniversary meeting of the International computer archive for Modern and Medieval English.* Italy: Verona.

STUBBS, M., BARTH I. (2003). "Using recurrent phrases as text type". *Functions of Language,* 10-1, 65-108.

TAHA, A. (1960). "The structure of two-word verbs in English". *Language Learning,* 10, 115-122.

WAIBEL, B. (2007). Phrasal verbs in learner English: A corpus-based study of German and Italian students. Albert-Ludwigs-Universität: Freiburg.

WRAY, A. (2005). *Formulaic language and the Lexicon*. Cambridge: Cambridge University Press.

Internet and sources:

The British National Corpus (BNC) – accessed from http://ucnk.ff.cuni.cz/

The Phrases in English (PIE) – available at http://pie.usna.edu

Non-Czech learner essays – available at http://bookrags.com

Native speaker book reviews – available at http://happypublishing.com

Applications and concordancers:

Collocate 1,0 by Michael Barlow (2004). Athelstan.

Concgram 1, 0 by Chris Greaves (2009). John Benjamins Pub Co.

Dictionaries:

COWIE, A. P, MACKIN, R. (2010). *Oxford Dictionary of Phrasal Verbs: Oxford Dictionary of Current Idiomatic English, Volume 1.* Oxford: Oxford University Press.

Oxford Advanced Learner's Dictionary. (2005). Oxford: Oxford University Press.

Oxford Collocations Dictionary for learners of English. (2001). Oxford: Oxford University Press.

Cambridge Advanced Learner's Dictionary. (2008). Cambridge: Cambridge University Press.

RESUMÉ

Studie zkoumá míru autenticity jazykového projevu nerodilých mluvčích při vytváření různých typů víceslovních jednotek. Studie vychází z Kontrastivní mezijakové analýzy Sylviane Grangerové (1996), srovnává dva korpusy nerodilých mluvčích, tj. českých sedmnáctiletých studentů gymnázia a další skupiny nerodilých mluvčích, studentů s rozdílným původem, se vzorkem rodilých mluvčích, kteří jsou profesionálními autory recenzí. Všechny tři vzorky dosahují velikosti přibližně 9 400 slov. První vzorek tvoří 37 esejů studentů jednoho pražského gymnázia, druhý vzorek obsahuje 19 esejů skupinky nerodilých mluvčích různého jazykového původu; jejich eseje byly staženy z webových stránek http://www.bookrags.com//. Vzorek rodilých mluvčích tvoří 22 recenzí na knihy psanými profesionálními autory recenzí, dostupných na webových stránkách http://www.happypublishing.com//. Při zkoumání víceslovných jednotek bylo využíváno různých zdrojů: v prvé řady Britského národního korpusu (BNC), dále Frazeologické databáze (PIE), nejrůznějších slovníků, rovněž konzultací s rodilými mluvčími. Studie vychází z takzvaného korpusově-založeného přístupu („a corpus-based approach“). Důvodem užití různých typů zdrojů bylo stanovit maximální objektivitu výsledků. Tato studie si kladla za cíl potvrdit a ukázat, že tvorba různých druhů víceslovných jednotek bude pro nerodilé mluvčí obtížná z několika důvodů, přičemž stupeň obtížnosti bude souviset s typem víceslovné jednotky. Obecně se od počátku předpokládalo, že nerodilí mluvčí budou při tvorbě frazeologických jednotek využívat Sinclairova (1991) takzvaného „open-choice principu“, tedy budou mít tendenci vytvářet různá víceslovná spojení neidiomaticky, budou inklinovat k doslovnému přeložení, které může být gramaticky bezchybné, nicméně nese pro rodilého mluvčího často známku atypičnosti, nepřirozenosti. Na druhou stranu bylo možné očekávat, že produkce rodilých mluvčích bude v souladu s takzvaným, rovněž Sinclairem zavedeným termínem, „idiom-principem“, tj. spojení budou idiomatická, pro rodilé mluvčí přirozeně znějící. Výsledky studie víceméně tyto předpoklady potvrzují, stejně tak do značné míry potvrzují výsledky předchozích studií. I ty se však často liší. Je ovšem nutné upozornit na to, že tato pilotní studie si, na rozdíl od jiných studií, převážně kladla za cíl zmapovat produkci nerodilých mluvčích a vytvořit nosnou metodologii, která by objektivně zachycovala míru autenticity a idiomatičnosti v jazykové produkci nerodilých mluvčích. Vzhledem k tomu, že studie zkoumá tři různé typy víceslovných jednotek a to čtyřmi různými způsoby, bylo nutné pracovat na menším vzorku. Výsledky je tak nutno brát jako orientační. Jelikož

se způsob zvolené metodologie ukázal jako úspěšný, otvírá se tak cesta pro další studie, které budou užívat větších vzorků a jejichž cíle budou ještě více specifické. Jiným důvodem malé velikosti vzorku bylo i relativně krátké časové období, kdy bylo možné získat materiál od českých studentů a potažmo pak nový materiál od další skupinky nerodilých mluvčích a rodilých mluvčích na stejné téma. Původně se předpokládalo, že vzorky budou obsahovat eseje se třemi tématy, které se podařilo získat od českých mluvčích. Po důkladném pátrání po esejích se stejnými anebo velmi podobnými tématy u další skupinky nerodilých mluvčích a rodilých mluvčích se však ukázalo, že dostatečné množství esejů na stejné téma lze získat pouze u tématu recenze na oblíbenou knížku či film.

Práce zahrnuje hlavní předmět zkoumání ve třech kapitolách: kapitola čtvrtá se zaměřuje na častá neidiomatická troj- and čtyř-slovní spojení a zjišťuje, za prvé, do jaké míry se ta samá spojení budou v jednotlivých vzorcích opakovat. Předpokládalo se, že spojení vytvořená nerodilými mluvčími se budou opakovat častěji, zatímco spojení vytvořená rodilými mluvčímu budou kreativnější. Dalším předmětem zájmu bylo stanovit, do jaké míry mezi těmito víceslovnými neidiomatickými kombinacemi vyskytují Biberovy tzv. lexikální svazky, někdy nazývané shluky (lexical bundles), což jsou vysoce frekventovaná víceslovná spojení neidiomatického charakteru. I když neexistuje jednoznačný koncenzus, kolikrát se daný výraz musí vyskytovat v milionů slov, abychom takovéto spojení mohli označit za lexikální svazek a stanovení takovéto hranice je nutně arbitrární, využívá tato práce Biberova přístupu (1999). To znamená, aby slovní kombinace mohla být považována za skutečný lexikální svazek, musí se vyskytovat alespoň desetkrát v jednom milionu slov a v pěti různých textech. Práce rozlišuje mezi dvěma termíny, kterými jsou jednak „neidiomatická opakující se slovní spojení“ (non-idiomatic recurrent word-combinations) a „lexikální svazky“ neboli „shluky“ (lexical bundles). Lexikální shluky představují pouze taková víceslovná spojení, která splňují výše zmíněnou minimální hranici deseti výskytů. Ověření statutu lexikálního svazku v našich vzorcích umožnila Fletcherova Frazeologická databáze (Phrases in English). S ohledem na skutečné lexikální svazky se očekávalo, že takovýchto lexikálních svazků v pravém slova smyslu vytvoří více rodilí mluvčí. Na druhou stranu se dalo předpokládat, že spojení vytvořená nerodilými mluvčímu budou sice repetitivní, ne ovšem natolik, aby mohly být nazvány skutečnými lexikálními svazky. Poté, co aplikace Collocate našla troj- a čtyř-více slovná neidiomatická slovní spojení, výsledky jednoznačně prokázaly, že obě skupiny nerodilých mluvčích vytvořily téměř dvakrát tolik slovních kombinací než rodilí mluvčí a tudíž potvrdily daleko větší stupeň repetitivnosti. Na druhou stranu rodilí mluvčí prokázali, že jsou schopni být daleko kreativnější, jednotlivá spojení u nich nejsou tak častá. U čtyř-kombinací český vzorek skýtá 127 typů, další vzorek nerodilých mluvčích 119, přičemž vzorek rodilých mluvčích pouze 54 typů; u trojčlenných spojení jsou výsledky ještě průkaznější – čeští studenti vytvořili 370 typů, další skupinka nerodilých studentů 320 a skupina rodilých mluvčích pouze 220 typů. Rovněž je možné tvrdit, že tento aspekt repetitivnosti je patrný i u frekvence typů v českém vzorku, zejména u troj-slovních kombinací. Zatímco vzorek nativních mluvčích ukazuje, že frekvence typů je relativně stabilní, pouze několik spojení se vyskytuje 6, 5, 4, 3 (nejvíce dvakrát), oba vzorky nerodilých mluvčích, zejména však ten český, ukazuje nestabilní frekvenci typů, která se pohybuje v rozmezí od 12 – 2. Další fáze ukázala, že jednotlivé vzorky (všechny tři) obsahují velmi malé množství čtyř-slovních lexikální svazků v pravém slova smyslu, tj. slovních kombinací, které se vyskytují minimálně desetkrát v jednom milionu slov. U troj-kombinací byla situace poněkud jiná; ukázalo se, že čeští mluvčí vytvořili lexikálních svazků nejvíce – 22,7 procent, druhá skupinka nerodilých mluvčích 17,2 procent, přičemž rodilí mluvčí vytvořili 19,1 procent. I přes aspekt větší repetitivnosti se zpočátku předpokládalo, že rodilí mluvčí vytvoří lexikálních svazků v pravém slova smyslu více, zatímco čeští mluvčí a druhá skupina nerodilých studentů vytvoří spíše častěji repetitivní kombinace specifické pro kontext. Nicméně možná vysvětlení existují dvě: vzorek je velmi malý a v tomto ohledu ukazuje nejednozačné výsledky. Na druhou stranu by se výsledek dal interpretovat jako fakt, že čeští mluvčí jsou obeznámeni s takto četnými neidiomatickými spojeními a využívají je proto jako spojení „bezpečná", která se nebojí aplikovat. Analýza ovšem ukázala, že kromě lexikálních svazků jako takových existují ve vzorcích ještě další dvě skupiny spojení. První skupinu tvoří slovní kombinace, které se sice ve Fletcherově databáze Phrases in English (PIE) vyskytují, ovšem velmi sporadicky. Druhou skupinou jsou spojení, která se v PIE nevyskytují vůbec. První skupinka se vztahuje (ve všech) vzorcích k takovým slovním kombinacím, které jsou kontextově specifické, vztahují se k danému tématu, nicméně nejsou v jazyce natolik běžné, spíše se jedná o spojení vytvořená ad hoc a většina z nich spadá do sémantické oblasti čtení a knih (*I like reading).* Druhá skupina s nulovým výskytem v PIE je rovně významově specifická, na rozdíl od první však obsahuje spojení vztahující se ke konkrétním hrdinům, autorům, názvům knih a filmů (*The Da Vinci Code, The book Harry Potter*). Tato spojení proto nemohou být považována za běžnou součást jazykového rejstříku rodilého mluvčího.

Předmětem zkoumání této kapitoly byla i otázka strukturních typů. Strukturní taxonomie typů byla opět inspirována Biberovou klasifikací (1999). Podrobná syntaktická klasifikace do jednotlivých strukturních typů měla ukázat, zda rodilí mluvčí vytvoří více syntaktických typů těchto spojení. Výsledky v tomto ohledu se však nepotvrdily. U čtyř-kombinací rodilí mluvčí vytvořili méně strukturních typů než obě skupiny nerodilých mluvčí. I když by se hypoteticky dalo polemizovat, že při větším počtu slovních kombinací bylo možné předpokládat i narůstající počet strukturních typů, je tato domněnka čistě hypotetická a nelze na ni spoléhat. Opět je tedy nutno podotknout, že vzorek je příliš malý a neumožňuje vytvořit v tomto ohledu objektivní závěr.

Kapitola pátá zabývající se frázovými a předložkovými slovesy potvrzuje, že pro nerodilé mluvčí frázová slovesa představují značně obtížné jazykové jednotky. Důvody jsou různé. Jedním je například netransparetní povaha (některých) frázových sloves, jejich vícero významů s ohledem na kontext, dalším důvodem může být i absence frázových sloves v jazyce nerodilého mluvčího. Z tohoto důvodu pak nerodilý mluvčí hledá jinou možnost, jak význam vyjádřit. Používá například jednoslovný ekvilent daného frázového slovesa, i když kontext spíše preferuje užití frázového slovesa. Poté, co ze seznamu sloves poskytnutého programem Concgram byla ručně vytříděna slovesa, z nich následně ručně vytříděna všechna frázová a předložková slovesa, jejich status ověřen prostřednictvím Cowieho slovníku frázových sloves (ODPV 1993, 2010) a Britského národního korpusu (BNC), výsledky analýzy potvrzují, že vzorek českých mluvčích obsahuje pouze 25 typů a 36 tokenů frázových sloves. Situace se jeví poněkud lépe pro druhou skupinu nerodilých studentů, kteří vytvořili 43 typů a 57 tokenů frázových sloves. Vzorek rodilých mluvčích nabízí nejvíce frázových sloves, 53 typů a 64 tokenů. Je však nutné konstatovat, že obě skupiny nerodilých mluvčích používají frázová slovesa víceméně dobře až na malé množství výjimek. Pokud byla frázová slovesa užita nevhodně, jednalo se o nejčastěji buď o zvolení nevhodného kolokátu k danému frázovému slovesu, použití jednoslovného ekvivalentu místo frázového slovesa či užití nevhodné adverbiální částice. Velmi malé množství frázových sloves má několik příčin: prvním je bezesporu fakt, že vzorek je příliš malý. Na druhou stranu je nutné zvážit i rovinu stylistickou. Fakt, že frázová slovesa jsou typická pro mluvený jazyk a vzorky jsou tvořeny z esejů, vysvětluje skutečnost, proč ani vzorek rodilých mluvčích neposkytl dostatečné množství frázových sloves k analýze. I když na první pohled není mezi druhou skupinou nerodilých mluvčích a rodilými mluvčími tak markantní rozdíl, repertoár frázových sloves se přesto ve vzorku rodilých mluvčích ukazuje bohatší. Větší pestrost frázových sloves u vzorku rodilých mluvčích dokreslují i další skutečnosti, které byly předmětem zkoumání. Konkrétně se jedná o lexikální slovesa, která spolu s adverbiální ásticí tvoří slovesa frázová. Zatímco frázová slovesa jsou u vzorku českých studentů tvořena pouze 21 typy lexikálních sloves a u druhé skupinky nerodilých mluvčích je to 32 lexikálních sloves, rodilí mluvčí tvořili frázová slovesa z celkem 43 typů lexikálních sloves. Stejně tak rozsah adverbiálních částic se liší: čeští mluvčí užili limitovaný výběr částic (*up, out, back, down, on, away, in)* na rozdíl od druhé skupiny nerodilých mluvčích a rodilých mluvčích, kteří užili kromě výše uvedených i další adverbialní částice, jako např. *through, away, forward, behind* (nerodilý mluvčí) a *off, away, around, along, forward, together.* Všechny tři vzorky vykazují do malé míry i určité podobnosti, jakými je v prvé řadě neutrální styl a pět frázových sloves, které se vyskytují ve všech třech vzorcích: *come back, end up, find out, get something back, go on.*

Jelikož předložková slovesa působí nerodilému mluvčímu těžkosti zejména s ohledem na užití správné předložky, nikoli jejich užití vůbec, v kapitole věnované předložkovým slovesům byly zkoumány jiné aspekty. Výsledky analýz předkládají přesvědčivé výsledky s ohledem na počet předložkových sloves v jednotlivých vzorcích. Jasně se ukazuje, že předložková slovesa nerodilí studenti užívají ve dvakrát tak větší míře než frázová slovesa, stejně tak i rodilí mluvčí. Vzhledem k Biberově korpusovému svědectví (1999) se dal vyšší počet předložkových sloves očekávat – předložková slovesa jsou typická nejen pro konverzaci, ale vyskytují se hojně i v akademické próze, žurnalistice, fikci. Množství nesprávně užitých předložkových sloves je překvapivě malé v obou vzorcích nerodilých mluvčích: čeští mluvčí vytvořili 60 typů a 111 tokenů předložkových sloves, z nichž pouze 6 případů tvořila nesprávně použitá předložková slovesa; u druhého vzorků nerodilých mluvčích to bylo 11 případů (vzorek nabízel 90 typů a 130 tokenů). Vzorek rodilých mluvčích skýtá 101 typů a 159 tokenů. Důvodem pro malý počet nesprávně použitých předložkových sloves u nerodilých mluvčích může být i fakt, že značná část předložkových sloves užitá studenty jsou vysoce frekventovananá slovesa, se kterými se studenti setkávají a užívají je velmi často, mají je dostatečně zažitá. S tímto zdůvodněním souvisí i částečně další předmět zkoumání, jímž byla sématická klasifikace předložkových sloves. V analýzách byla pozornost věnována tomu, zda nerodilí mluvčí používají předložková slovesa, která patří k nejfrekventovanějším v jazyce, či k méně frekventovaným až okrajovým jevům. K zajištění objektivity byly využity Biberovy (1999) korpusové studie, které dokládají, že jak čeští studenti, tak i druhá skupina nerodilých studentů užívají nejčastěji nejvíce se vyskytující předložková slovesa v jazyce, tvořící skupinu tzv. dějových sloves (activity verbs).

Výzkum kolokací rozebírá kapitola šestá, v níž jsou kolokace předmětem dvou různých typů analýz. První analýza vychází z takzvaného korpusově-založeného přístupu (frequency- based approach). Cílem prvního zkoumání bylo stanovit, do jaké míry nerodilí a rodilí mluvčí vytvářejí kolokace, které se vyskytují v BNC hojně, konkrétně s četností pět a více. Pro analýzu byly vybrány dva strukturní typy kolokací, jednak adjektiva ve spojení se substantivy, dále pak substantiva ve spojení s verby. Podle výsledků lze konstatovat, že nerodilí mluvčí produkují hojně se vyskytující „kolokace" v BNC snadno. Velké procento takovýchto kolokací se vyskytuje v obou vzorcích nerodilých mluvčích: v kombinaci s adjektivem a substantivem vytvořili čeští mluvčí dokonce nejvíce vysoce frekventovaných kolokací v BNC (70 procent), další skupina nerodilých mluvčích vytvořila 57 procent takovýchto kolokací, rodilí mluvčí 55 procent. Situace je poněkud odlišná v kombinaci substantiva se slovesem, kde čeští mluvčí vytvořili 54 procent kolokací s četností vyšší než pět v BNC, další skupina nerodilých mluvčích 22 procent, rodilí mluvčí 54 procent. Hlubší analýza však potvrzuje, že zejména v kombinaci *adjektivum* + *substantivum* užívají nerodilí mluvčí velmi běžná adjektiva, jakými jsou například *good, favourite, popular* atd.. Mají nízkou vypovídací hodnotu a v tomto případě je proto nemožné nazvat převážnou většinu takovýchto adjektivních kolokátů jako nevhodné pro daný nod. Repetoár adjektiv rodilých mluvčích je naopak pestřejší, což lze pozorovat na adjektivech, jakými jsou například *delightful, acclaimed, precious, valuable, narrow* atd. Možné závěry, které lze z těchto výsledků vyvodit, jsou patrné: značná část „kolokací" vytvořených nerodilými (i rodilými) mluvčími nejsou restriktivními kolokacemi, jedná se naopak o vysoce frekventované „kolokace". Mnohé z nich by ovšem nebyly nazývány kolokacemi v tradičním slova smyslu, pokud by v rámci analýzy byl uplatňován frazeologický přístup ke kolokacím, který považuje za kolokaci jen taková spojení slov, která jsou vzájemně prediktabilní a očekávatelná.

V analýze zaměřující se na kombinaci podstatného jména spolu s verbem se ukázalo, že nerodilí mluvčí vytvořili více kolokací, které se v BNC nevyskytovaly (český vzorek 23 procent, druhý vzorek nerodilých mluvčích 31 procent).

Druhá část výzkumu věnovaná kolokacím vycházela ze studie provedené Sylviane Grangerové (2005) zahrnující test tzv. kolokační salience. Jinými slovy, kolokační salience poukazuje na schopnost jedince vybrat ze seznamu kolokátů ten nejvíce prototypický pro dané slovo, posléze i další možné kolokáty, které lze s daným výrazem spojit a stále budou nazývány pouvažovány za typické či přijatelné kolokáty daného slova. Test obsahuje strukturní typ kolokací adverbia a adjektiva. Test měl prokázat, do jaké míry jsou nerodilí studenti schopni rozpoznat z výběru adjektiv ta z nich, která jsou pro daný amplifikátor prototypická, a dále pak adjektiva, která lze také kombinovat s daným amplifikátorem. Této části výzkumu se zúčastnilo 15 studentů gymnázia a 6 dospělých jedinců, kteří navštěvují firemní kurzy angličtiny. BNC posloužilo jako kontrolní vzorek. Výzkum zahrnoval několik úrovní. V první fázi se stal předmětem zkoumání prototypický kolokát daného amplifikátoru, který měli studenti ze seznamu zakroužkovat. Z možných 210 odpovědí bylo získáno 162 odpovědí, z nichž pouze 27 procent tvořily správné odpovědi. Pokud bychom tedy 210 možných správných odpovědí brali jako normu rodilých mluvčích, pro české studenty by to znamenalo, že dosahují pouze 27 procent této normy. Další fáze si kladla za cíl zjistit, kolik studentů z počtu 21 podtrhne další možné kolokáty k danému adverbium. Stoprocentní úspěšnost, tzv. zjišťování, kolik studentů podtrhne všechny kolokáty správně, nebyla cílem, rovnala by se totiž nule. Výsledky stojící za zmínku jsou následující: relativně vysokou úspěšnost českých studentů lze zaznamenat u kolokací *highly important* (43 procent) a *highly significant* (53 pro cent), stejně tak jako kolokace *absolutely different* (57 procent)*, absolutely reliable, happy* (48 procent). Takovéto procento úspěšnosti lze však vysvětlit faktem, že se zmíněné kolokace se velmi dobře překládají do češtiny. Naopak ostatní amplifikátory *vitally, utterly, fully, bitterly, blisfully* nebo *perfectly* mají naopak procento úspěšnosti velmi nízké, pohybující se v rozmezí 4-24 procent. Čistě specifické jsou amplifikátory *seriously* a *readily,* které podle BNC nepřipouštěly z výběru jiné kolokáty než *seriously ill* a *readily available.* Předmětem zkoumání bylo tudíž zjistit, kolik studentů zvolí pouze tyto varianty, získaný výsledek opět potvrzuje nízký stupeň kolokační saliance (schopnosti poznat „správnou“ a „nesprávnou“ kolokaci) českých mluvčích – pouze 8 studentů z 21 potrhlo (zhruba jedna třetina) pouze tyto dva kolokáty, přičemž ostatní studenti podtrhávali daleko více adjektivních kolokátů.

Studie se pokusila zmapovat míru autenticity a idiomatičnosti v jazykové produkci dvou skupin nerodilých mluvčích. K zajištění komparitivních analýz byly oba korpusy nerodilých mluvčích porovnávány se vzorkem nativních mluvčích. Objektivita zjištění byla zaručena různými zdroji, Britským národním korpusem, Frazeologickou databází, slovníky a konzultacemi s rodilými mluvčími.