UNIVERZITA PALACKÉHO V OLOMOUCI

Filozofická fakulta

Katedra anglistiky a amerikanistiky

Building and Exploring a Corpus of Academic Writing by Czech Students of English

Diplomová práce

Autor: Bc. Anna Boková, AF-AES

Vedoucí práce: Mgr. Michaela Martinková, PhD.

Olomouc 2015

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a uvedla úplný seznam citované a použité literatury.

V Olomouci dne 17. 8. 2015 ………………………

Poděkování

Ráda bych poděkovala paní Mgr. Michaele Martinkové, PhD. za její cenné rady, trpělivost a pomoc při vedení této práce.

Contents

1 Introduction

The corpus based research into academic writing (e.g. Biber et al. 1999, Cortes 2004, Hyland 2008) has shown that this register is characterized by the usage of recurrent conventionalized multi-word expressions ("lexical bundles", "clusters" or "n-grams") that signal "competent use of language in a particular context" (Dontcheva-Navratilova 2012, 38). Using these multi-word expressions correctly contributes to a more native-like writing and is therefore essential for learners of English. As Hyland argues, "if learning to use the more frequent fixed phrases of a discipline can contribute to gaining a communicative competence in a field of study, there may be advantages to identifying these clusters so as to help learners acquire the specific rhetorical practices of the texts they are asked to write" (2008, 42). Moreover, Hyland points out that clusters are not only central to the academic discourse, but they are a means of differentiating genres (2008, 41). In this thesis two corpora of English texts will be created and explored; the first will be compiled of texts produced by university students (L1 Czech), the second corpus will be compiled of texts produced by professional linguists. This thesis aims to investigate the following questions: firstly what structural types of lexical bundles are used by Czech learners of English, secondly, whether they use them more or less often than professionals and thirdly, whether they use the same lexical bundles as professional linguists or not. Building on Hyland's research, it is assumed that professional linguists may use different repertoire of lexical bundles than L2 students of English (L1 Czech), since there exist genre variation between students' writing and professional writing.

The chapters 2, 3 and 4 focus on the theory; namely learner corpora, lexical bundles and creation of the corpora in Sketch Engine are described.

In the chapter 2, learner corpora research is reviewed. The definition of a learner corpus is introduced as well as the criteria for the creation of such a corpus. Since there has been a growing interest in the learner corpora research in recent years, the major existing learner corpora are described as well. Two methodological approaches for the linguistic analysis of learner corpora are presented, i.e. Contrastive Interlanguage Analysis and Computer-aided error analysis.

Firstly, the chapter 3 introduces the characteristics of lexical bundles, i.e. recurrent multi-word expressions, and explains the term formulaic language. The main focus is on the lexical bundles found in academic prose and their structural and functional types. Finally, this chapter reviews the research on lexical bundles in academic writing. It should be noted that there exist two approaches to examining lexical bundles. First, a researcher defines target bundles which are supposed to be representative of a certain genre, e.g. professional writing (usually on the basis of the previous research) and then those target bundles are looked for in the corpus of students' writing. Second, lexical bundles are extracted from professional writing and students' writing and then the findings are compared. The second approach is taken in this thesis, since one of the aims is to find out what lexical bundles are used by Czech learners of English.

In the chapter 4, the software for the creation of the corpora is described, namely Sketch Engine.

The practical part (chapter 5) focuses on the compilation of the corpora, methods used for the extraction of data and on the data analysis. First, the creation of the two corpora (with the help of Sketch Engine) – **Research Articles Corpus** and **Students' Theses Corpus** – is described as well as the characteristics of each of the corpus. Second, the query for the extraction of the lexical bundles is presented as well as the methods for sorting of the data. Third, the structure of the lexical bundles extracted from both corpora is subjected to an analysis. The lexical bundles in both corpora are divided into corresponding tables according to their structure and followed by a discussion on findings.

2 Learner corpora

Generally, a corpus (for linguistic purposes) is defined as "… a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (Meyer 2002, xi). Learner corpora research has become popular among researchers during the last two decades. As a part of corpus linguistics, learner corpora research is a rather new field of study. The history of the term "learner corpora" dates back to 1980s when the learner corpus research "… has created an important link between the two previously disparate fields of corpus linguistics and foreign/second language research." (Granger 2002, 4). Meyer points out that in the beginning, learner corpora were in fact created in order to study second-language acquisition, i.e. how people learn English (or a different language) as a second or foreign language[1]. Learner corpora serve as a source of information for teachers who may develop a teaching strategy for a concrete group of students and this particular teaching strategy is based on the description of how the students actually use language. Corpora information can be used to detect common errors that learners of a foreign language make. Another contribution of learner corpora is a methodology called data-driven learning. Students are encouraged to use corpora of native speech and via its examples they may learn more about how the language works. In other words, they are encouraged to explore and investigate a given corpus on their own which may help them with the process of learning the language (2002, 26-27).

A learner corpus needs a more specific definition. One is given by Granger when she says that "Learner corpora (LC) are electronic collections of foreign or second language learner texts assembled according to explicit design criteria"

[1] Granger states that learner corpora belong to the non-native varieties of English which can be further divided into English as an Official language (EOL), English as a Second Language (ESL) and English as a Foreign Language (EFL). ESL is acquired in an English-speaking environment whereas EFL represents English acquired in a classroom environment. Both EFL and ESL are found in learner corpora (2002, 8).

(2009, 14). Such corpora are different from the common native corpora in several ways. Firstly, they require considerably extensive expertise on the researcher's side due to the nature of the learner input. When building a corpus, researchers have to bear in mind that many of the methods and tools usually used in corpus research are based on native corpora. The high rate of errors usually found in learner corpora is not taken into account. Secondly, hand in hand with the corpus linguistics expertise goes the knowledge of the linguistic theory that is necessary for a successful analysis of the collected data. Thirdly, it is vital to bear in mind that since the data are learner texts, the understanding of SLA is helpful, if not essential, and should improve the interpretation of the results. Lastly, Granger also mentions the applicability of the learner corpora research to different pedagogical purposes, thus connecting learner corpora with foreign language teaching. However, she stresses the fact that " it is particularly important to study the impact of contextual factors, as these will determine to whether and to what extent the results of learner corpus research can be integrated into teaching" (2009, 15-16).

Learner corpora research then can be regarded as an interdisciplinary field of linguistic research, implementing findings from general corpus research, linguistic theory, SLA and foreign language teaching.

2.1 Learner corpora typology

For the purpose of this thesis, Granger's learner corpora typology will be used. She proposes that a corpus is usually described in terms of dichotomies. Granger states that there are four dichotomies relevant to learner corpora:

- Monolingual vs. Bilingual
- General vs. Technical
- Synchronic vs. Diachronic
- Written vs. Spoken

Generally, learner corpora tend to be **monolingual** and they usually contain **non-specialist (i.e. general)** language. Learner corpora are usually **synchronic**, i.e. they describe learner language at a particular point of time. Longitudinal learner corpora are not very common, since the compilation of such corpora is time-consuming and learner population would have to be followed for a long time. As for the collection of the data, it is more difficult to gather oral data. This explains the fact that there exist more **written** than spoken learner corpora. It has to be said that there exist exceptions to these general assumptions and that there exists linguistic research in the less prominent features of the dichotomies (i.e. bilingual, technical, diachronic and spoken) (2002, 10-11).

2.2 Learner corpora around the world

The growing popularity of learner corpora research in the past two decades encouraged the development of several important learner corpora. Here, the established well-known corpora will be described.

One of the largest and most important corpora is the **International Corpus of Learner English** (ICLE) which was initiated in 1990 by Sylviane Granger at Louvain University, Belgium. The official websites[2] state that the corpus consists of argumentative essays written by higher intermediate to advanced learners of English. There are texts written by learners from different countries, e.g. Chinese, Czech, Dutch, Polish or Turkish. ICLE now contains approximately 3.7 million words that represent 16 mother tongue backgrounds. Each of the subcorpus has to consist of at least 200,000 words provided that each student can contribute only 1000 words at maximum and this condition has to be met. ICLE is the result of cooperation of several universities and it provides homogeneity of its data, since the participating partners have adopted the same criteria (Université Catholique de Louvain 2011). McEnery and Hardie state that

[2] www.uclouvain.be/en-cecl-icle.html

to secure the comparability of the corpora, the topic and the length of the essays are controlled. Background information – gender, proficiency level and L1 background are also documented (2012). Granger adds that learners have to fill a questionnaire with more than 20 task and learner variables. They are recorded in a database, thus enabling the researchers to compile subcorpora with different settings of the variables (2003, 539).

The systematic work on ICLE led to its growth and gradual expansion. More researchers became interested in learner corpora. For example, De Cock from Louvain initiated the compilation of LINDSEI corpus or the **Louvain International Database of Spoken English Interlanguage**. A control native-speaker corpus LOCNEC, the **Louvain Corpus of Native English Conversation** was also created (McEnery and Hardie 2012, 82). Simultaneously, the **Louvain Corpus of Native English Essays** (LOCNESS) was created in order to make a comparison of native language (L1) and learner language (L2) possible (Flowerdew 2014, 2). It is clear that Granger's work as well as work of other Louvain researchers made a great contribution to the promotion and popularization of learner corpora research and without doubt inspired the production of bigger, commercial learner corpora.

One of them is **the Longman Learner's Corpus**. The official websites[3] state that this corpus contains 12 million words and consists of students' essays and exam scripts. The typical learner mistakes are analysed and the data are used for creating learning materials (1998 – 2008). The corpus is composed of several subcorpora. Students have different L1 backgrounds and produce L2 English writing which serves as a source of learning materials that can be created specifically for the L1 background in question (McEnery and Hardie 2012, 83). The compilation of the corpus is a still in progress and the authors invite public to

[3] http://www.pearsonlongman.com/dictionaries/corpus/index.html

send them their students' material to support its growth. The corpus also serves as a source of data for learner dictionaries[4].

The Cambridge Learner Corpus published by Cambridge University Press is another large learner corpus – actually the world's largest learner corpus as its website states[5]. It comprises of exam scripts written by students from different countries who take the Cambridge English exams around the world, e.g. FCE, CAE, BEC etc. Those exams are based on the Common European Framework of Reference for Languages which is an international standard for describing learner's language skills and which has six levels – from A1 to C2. The Cambridge Learner Corpus is annotated for learner errors and those data are used for creating teaching materials (e.g. publications focused on exam preparation or common mistakes and sample tests) for learners that are appropriate for a certain level or exam and produces materials that are specifically targeted (e.g. specific L1 backgrounds – some errors can be the result of specific first language interference).

Not only English institutions are creating large learner corpora, a lot of research is currently in progress in Asia. For example, **the Chinese Learner English Corpus** (CLEC) has around million words, **the Chinese Academic Written Corpus** (CAWE) that is composed of dissertations written by Chinese undergraduates majoring in English linguistics or applied linguistics has approximately 400 000 words. **The Hong Kong University of Science & Technology** (HKUST) **Learner Corpus** has 25 million words. In Japan at Meikai University, there is **The Japanese English as a Foreign Language Learner** (JEFLL) **Corpus** which contains approximately 700 000 words. There are learner corpora composed of texts produced by Korean learners, Malaysian learners and

[4] http://www.pearsonlongman.com/dictionaries/corpus/learners.html

[5] http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-English-Corpus-Cambridge-Learner-Corpus/?site_locale=en_GB

other Asian countries. Corpora of learners’ texts whose L1 is an Arabic language are created as well.

It is obvious that learner corpora research is well-established in the field of corpus linguistics. The corpora may vary in their size or annotation, but it is without any doubt that their contribution to the other fields of linguistic enquiry is considerable and their importance is gradually growing. In the next subchapter, the basic characteristics of a learner corpus will be described

2.3 Characteristics of a learner corpus

There exist basic features that a learner corpus should have. According to Granger, those are explicit design criteria, authenticity, textual character of data and consistency in documentation. A learner corpus should be compiled according to previously set conditions (2002, 7).

The texts produced by learners of English are not randomly chosen, but they have to be compiled carefully with respect to previously defined characteristics. **Explicit design criteria** are the first distinctive feature of a learner corpus and the condition of defining them is in fact a very important one, since “the usefulness of a learner corpus is directly proportional to the care that has been exerted in controlling and encoding the variables” (Granger 2002, 9). Some of those variables can be the same for native corpora and learner corpora (e.g. gender, age), but some are specific only to learner corpora. The specific criteria can relate to the **learner** (e.g. what is his/her mother tongue, what is his/her level of proficiency, whether he/she has the knowledge of other languages) or to the **task** (the timing of the learner production - whether it was timed or not as well as whether the production was planned or not) (Granger 2009, 17; 2002, 9). The variables thus encoded and documented in a corpus can be further used in the following research, enabling the researchers to build specific subcorpora and to explore different phenomena based on the set criteria. For example, the language of female Czech learners of English with the advanced level of proficiency who were asked to write a 300-word essay about travelling could be examined.

The second fundamental characteristic of a learner corpus is its **authenticity** (Granger 2002, 8). Aston, Bernardini and Stewart say that "authenticity in this sense refers to a piece of text being "attested", having occurred as part of genuine communicative (spoken or written) interactions" (2004, 12). When this statement is applied to learner corpora, Granger argues that authenticity is problematic with respect to the genuineness of the learner language. It is due to the fact that learner interactions are almost always artificial to some degree and therefore not natural. Learner data are usually under some form of control – they are limited for example by the length, topic, timing or a certain task. Authenticity in learner corpora is therefore a kind of scale and can cover "genuine communications" as well as "authentic classroom activity" (2002, 8).

The third feature mentioned by Granger is the **textual character of a learner corpus**. She says that "to qualify as learner corpus data the language sample must consist of continuous stretches of discourse, not isolated sentences or words" (2002, 8). In other words, we may imagine for example essays and other pieces of written language or transcriptions of classroom communication that can qualify as learner corpus data, but not words or sentences in isolation which can be also found in the classroom interaction.

The last feature of a learner corpus that should be taken into consideration is what Granger calls **Standardization and documentation**. Granger stresses the fact that a researcher should use standardized annotation tools to make learner corpora comparable to native corpora. The design criteria set by a researcher should be documented in a consistent way, i.e. all the added information about learner and task variables should be accessible to other researchers (2002, 10).

2.4 Linguistic Analysis of Learner Corpora

The linguistic analysis of learner corpora is most often achieved with the use of one of the two methodological approaches. It is either Contrastive Interlanguage Analysis (CIA) or Computer-aided Error Analysis (CEA) (Granger 2002, 11-12).

2.4.1 Contrastive Interlanguage Analysis

As the title suggest, the Contrastive Interlanguage Analysis is a contrastive method of linguistic research. According to Granger, it carries out "quantitative and qualitative comparisons between native (NS) and non-native (NNS) data or between different varieties of non-native data" (2002, 12). In other words, CIA can either compare native language with learner language, or it compares data by learners with different L1s (Flowerdew 2014, 44), in Granger's words, it compares different varieties of learner interlanguage (L2 versus L2) (2009, 18). I will start with the former.

When comparing L1 and L2 varieties (NS VS NSS), the choice of control native corpus is crucial with respect to the variety of English (e.g. British English vs. American English) and proficiency level of native speakers (e.g. professional writers vs. students). Comparing L1 with L2 can be beneficial, since it can show the features of non-nativeness in learner written or spoken data. For example, it can show cases of under-use or over-use of words, phrases and structures (Granger, 2002, 13)

In opposition to the benefits of this approach, Granger states that the CIA approach has often been criticised for so called "comparative fallacy" (Bley-Vroman 1983, as cited in Granger 2009, 18). CIA in the critics' opinion fails to analyse learner interlanguage while comparing it to the native norm. In response to this criticism, it is argued that this kind of comparison is in fact beneficial in most cases, since it uncovers features of learner language that were not thought of before. Those features can be further analysed from the learner's perspective. Another argument in favour of CIA is the hypothesis that studies which compare learners with different proficiency levels are in fact based on L1 norm that is implicitly present. Granger calls this phenomenon as "comparative hypocrisy". Here, the importance of the quality and the comparability of the control (native) corpus is repeated once more, since it is essential for a successful analysis when the CIA methodology is applied. The approach that uses CIA in learner corpora research plays an important role in pedagogy as well. The L1 to L2 comparisons help to uncover differences between learner language and the L1 norm and this knowledge is consequently used for creating various pedagogical tools (2009, 18-19).

As to the comparisons of non-native learner corpora, these mainly contribute to the better understanding of the learner interlanguage. Specifically, the comparisons of the learner language with different mother tongue backgrounds can reveal features that are either developmental (i.e. shared by several learner populations) or possibly L1-dependent (specific for a certain learner population) (Granger 2002, 13).

2.4.2 Computer-aided Error Analysis

This methodology is based on former traditional Error Analysis that was popular in 1960s and 1970s. Without computerized data that are source for the CEA, Error Analysis lacked the advantages that are present now. The learner information was often insufficient as well as the information about the text type or the context of errors which was often lost. The collections of texts were usually used only to extract errors and were later destroyed, so the analysis could not be repeated. At present, the modern technology enables researchers to investigate various phenomena, such as overuse or underuse in learner language. The fact that the; data are computerized brings advantages, for example, certain searches are facilitated by the use of computer (Nesselhauf 2005, 40-41).

There are two methods of the CEA used most frequently. The first is based on selecting a problematic linguistic item and scanning the corpus to find all occurrences of misuse which are then the subject of the analysis. This method is fast, but dependent on the researcher's rather subjective choice of the particular item. The second method is based on the creation of a standardized error-tagging system that can find all the errors in a corpus, or at least errors from a specific category. The disadvantage of this method is its time-consuming nature (Granger 2002, 14), but as Granger states that "a thoroughly error-analyzed learner corpus is an invaluable resource which can inform most pedagogical tools" (2009, 24).

It is therefore clear that both methodologies used in learner corpora research, Contrastive Interlanguage Analysis as well as Computer Error-aided Analysis, provide researchers with sources of learner data that can be further analysed from various perspectives by using modern linguistic software.

3 Lexical bundles

This chapter will introduce the characteristics of lexical bundles that are mainly adopted from the *Longman Grammar of Spoken and Written English* (Biber et al. 1999). I will also introduce the term formulaic language and the previous research on lexical bundles produced by Czech learners of English will be summarized here as well.

3.1 Formulaic language

It is generally agreed and accepted that language is formulaic in nature (Wray 2002, 2012; Granger & Paquot 2012, Hyland 2008). However, as Qin points out, there is no strict definition of formulaic language is, and different terms are used to refer to it (2013, 220). Wray in her publication *Formulaic Language and the Lexicon* (2002) summarized what has been studied about formulaic language until then and proposed her own theoretical model. Although her findings are beyond the scope of this thesis, the basic facts about formulaicity of language are explained there. Wray states that words and phrases are called formulaic when we can process them without decomposing them to their lowest level. In other words, we can understand and use "prefabricated chunks" that are stored in our lexicon. Wray also proposes her own definition, but she does not use the term formulaic language. Instead, she uses the term **formulaic sequence** which is: "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray 2002, 3-9). Lexical bundles studied by Biber et al. (1999) fit this definition and for the purpose of this thesis, I will use the term lexical bundles which are, according to Qin "recurrent multi-word expressions identified by a computer program that occur at least 10 times per million words in a genre and found across at least five different texts in the genre" (2013, 221).

As for the terminology, there exist several terms (beside the lexical bundles) that refer to the notion of multi-word expressions which are extracted from corpora, frequency and distribution being the main criteria. Some of those terms are for example *clusters* or *n-grams*[6]. These terms "actually refer to continuous word sequences retrieved by taking corpus-driven approach" (Chen and Baker 2010, 30).

3.2 Characteristics of lexical bundles

Lexical bundles are characterized as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status. That is, lexical bundles are simply sequences of word forms that commonly go together in natural discourse" (Biber et al. 1999, 988). This kind of lexical patterning has to be differentiated from other kinds of multiword expressions, such as idioms and collocations. Biber et al. especially stress the difference between idioms and lexical bundles where the former are mostly invariable expressions whereas the latter can be regarded as **extended collocations.** In other words, there is a statistical tendency that some sequences of words will co-occur. These sequences must occur frequently to be considered a lexical bundle and for this purpose, Biber et. al introduced the operational definition of lexical bundles. According to them, for the sequences of four words (which will be analysed in this thesis as well) the minimal frequency is at least ten times per million of words plus the concrete lexical bundle has to appear in at least 5 different texts in a given register (1999, 988-990).

It is important to state that Biber et al. differentiate between lexical bundles in conversation and lexical bundles in academic prose. The focus of this

[6] This thesis will use the Sketch Engine software to extract lexical bundles. This software uses the term n-gram. Here, the definition of an n-gram is provided for the sake of clarity. The official website state that an n-gram is "a contiguous sequence of n items." It can be a sequence of morphemes, letters or words in a text or speech. A special type of n-gram is a unigram (a word), bigrams (sequences of two words) or trigrams (sequences of three words) can be created, too. In the word list of n-grams, the frequency of occurrence is what matters the most, so "there is not expected any relation among units in n-gram." With respect to this statement, it can be said that not all n-grams are collocations, but each collocation has to be an n-gram (2003-2013).

thesis will be on the latter. Here are some features typical for the lexical bundles in academic prose:

- Usually not a complete structural unit[7] => a lexical bundle in academic prose is often part of e.g. a noun phrase and the beginning of a prepositional phrase.
    - *the end of the*
    - *the presence of a*
- When it is a complete structural unit, it usually functions as a discourse signalling device.
    - *for the first time*
    - *in the same way*
- Composed mostly of nominal and prepositional elements. (Biber at al. 1999, 996-999)

3.2.1 Structural types of lexical bundles

Biber and Barbieri state that there exist three major aspects that differentiate lexical bundles from other formulaic expressions: they are extremely common, most of them "are not idiomatic in meaning and not perceptually salient"[8], and as was already mentioned above, they do not represent a complete structural unit (2007, 270). As for the structure of a lexical bundle, Biber et al. differentiate 12 types of lexical bundles (1999, 1014-1015). This classification is shown in Table 1 and will be used in this thesis.

[7] Lexical bundles in conversation are not complete structural units, too, but they are usually constructed from pronominal subject followed by a verb phrase plus complement clause (Biber et al. 1999, 991)

[8] In fact, Biber and Barbieri state that "most longer idioms are far too rare to be considered lexical bundles". Fiction is considered to be one of the few registers where idioms are used with moderately higher frequencies (2007, 269 -270).

Table 1 - Structural categories of lexical bundles

1	noun phrase with of phrase fragment
2	noun phrase with other post-modifier fragment
3	prepositional phrase with embedded of-phrase fragment
4	other prepositional phrase (fragment)
5	anticipatory it + verb phrase/adjective phrase
6	passive verb + prepositional phrase fragment
7	copula be + noun phrase/adjective phrase
8	(verb phrase +) that-clause fragment
9	(verb/adjective +) to-clause fragment
10	adverbial clause fragment
11	pronoun/noun phrase + be (+ …)
12	other expressions

3.2.2 Functional types of lexical bundles

Dontcheva-Navratilova states that "perhaps the most important condition for mastering lexical bundles is an understanding of their discourse functions …" (2012, 39). Previous research differentiates between three major discourse functions that lexical bundles can have. Cortes (2004), Biber and Barbieri (2007), Hyland (2008) point out that there exist three major discourse functions that lexical bundles can have:

(1) Stance expressions – attitudinal function
(2) Discourse organizers – discourse-organizing function
(3) Referential expressions – referential function

Attitudinal bundles (e.g. *may be due to, is likely to be, are more likely to)* express attitudes or frame some other proposition. Discourse organizers (e.g. *on the other hand, as a consequence of, on the basis of)* organize a text with respect to what has gone before and what is coming next – i.e. they establish logical relations in the text. Referential expressions (e.g. *at the beginning of, at the same time, as part of the)* refer to "physical or abstract entities, or to the textual context itself …" (Biber and Barbieri 2007, 270).

The functional typology of lexical bundles has no exact boundaries and the functions of a particular lexical bundle may sometimes overlap. This is supported by Dontcheva-Navratilova who states that there exist some discrepancies between the existing classifications. Firstly, she explains that it is due to their frequent structural incompleteness. In other words, it may happen that a certain lexical bundle can be included in more than one category and it is the researcher's choice to which class he/she includes it. Secondly, she argues that the sometimes ambiguous classification can be caused by the prevailing genre or discipline in a particular corpus, i.e. corpora compiled of hard-sciences texts are likely to have more referential expressions while humanities tend to have more discourse-organizing expressions (2012, 40).

Using lexical bundles correctly is important for learners of English, since "producing natural idiomatic English is not just a matter of constructing well-formed sentences, but of using well-tried lexical expressions in appropriate places" (Biber et. al 1999, 990). In other words, the correct use of lexical bundles contributes to a more native-like speech or writing. Hyland suggests that the use of certain lexical bundles can differentiate a particular register – i.e. the lexical bundle *in pursuance of* can signal a legal text. It then follows that mastering fixed phrases and multi-word expressions typical for a certain register means gaining a communicative competence in this register (2008, 42).

With respect to learner corpora, there has been growing interest in the use of formulaic language by L2 learners. As Granger and Paquot state: "...multiword units of all kinds (e.g., collocations, phrasal verbs, speech formulae) are notoriously difficult for learners, …" (2012, 130). This statement is supported by Qin's assertion that written production by L2 learners tends to be unnatural and non-native like. This is caused by wrong application or lack of formulaic language (2013, 221).

3.3 Lexical bundles in academic writing

The findings in *Longman Grammar of Spoken and Written English* (Biber et al. 1999) on lexical bundles in academic prose have inspired other linguists and since then, there has been an increasing amount of corpus-based research that focuses on academic writing. For example Hyland (2008) refers to lexical bundles as “clusters”. He compiled a corpus of research articles in four different fields (electrical engineering, applied linguistics, microbiology and business studies) and extracted 4-word clusters, his criteria being the frequency (at least 20 times per million words) and the breadth of use (a particular lexical bundle had to appear in at least 10 % of texts). Then he compared the data in his corpus with the data in a corpus of doctoral dissertations and master’s theses written in English by Chinese students, his main aim being to show that clusters can differentiate genres. His research showed that Master students had a wider range of clusters and used them with greater frequency than professional writers. One of the explanations of this phenomenon that Hyland suggests is that Master students rely more on formulaic expressions and are less confident or proficient in their writing. The other explanation he offers is that the wider use of clusters by students is due to the pedagogical genre where students are expected to show their research and disciplinary skills.

A very interesting study was published by Cortes who compares lexical bundles in two corpora; the first was compiled of articles from history and biology journals, the second consisted of student writing in the same fields. Cortes extracted 4-word bundles from the first corpus and used them as target bundles which she compared to the bundles found in the corpus of English-speaking students’ writing. She found out that students did not use the target bundles or their uses differentiated from the uses of bundles by professional authors (2004).

To sum up, there seem to be two approaches to examining lexical bundles. The first is to define target bundles that are found in texts written by professional writers in a certain field. Those target bundles are compared to bundles produced by L2 students at different levels, i.e. the specific target bundles are looked for in the corpora of students’ texts.

This approach was adopted e.g. by Dontcheva-Navratilova (2012) who created a corpus of diploma theses written by Czech students of English. There were 15 Master's theses in three different fields – linguistics, methodology and literature. The whole corpus had 254 000 words and was divided into three subcorpora, each representing one field. Dontcheva-Navratilova chose target bundles "on the basis of the results of previous research into lexical bundles in similar genres and disciplines …" (2012, 42-43). The target bundles were supposed to be representative of the particular genre or discipline. Beside the representativeness of the selected bundles, the choice of the target bundles was also based on the "native speakers' perception of appropriateness" (2012, 42). It should be noted that Dontcheva-Navratilova focused primarily on the functional analysis of the lexical bundles used by Master students of English (L1 Czech). Dontcheva-Navratilova's study found out that Czech Master students did not use a very wide repertoire of lexical bundles and that some of the defined target bundles were used only rarely. She concluded that it "seems to be due to an insufficient level of development of the rhetorical skills of the writers and to interference from L1 writing conventions" (2012, 56). However, it is necessary to stress the fact that Dontcheva-Navratilova did not investigate all the lexical bundles used by Czech students of English.

The second approach is to extract lexical bundles from the corpus of professional published writing and then extract lexical bundles from the corpus of students' texts. It is then possible to evaluate which lexical bundles are used by professionals and which are used by students. This is also the approach taken in this thesis: I will extract the lexical bundles Czech students of English actually use and compare those bundles with those used by professional writers. It is assumed (based on Hyland's research) that students – Czech L2 learners of English – actually use a certain repertoire of bundles which may differ from the repertoire of bundles used by professionals. The aim is to find out what bundles are there and possibly, what are the differences between them. The criteria used for the extraction of bundles will be partly based on Hyland's (2008) research; those are namely: a more conservative **frequency** (at least 20 times per million of words) with respect to the operational definition of lexical bundles and **breadth of use** (a lexical bundle has to appear at least in 5 different texts). The structural analysis of the lexical bundles used by Czech L2 learners will be the objective of this thesis, since Dontcheva-Navratilova concentrated on functional analysis in her research.

4 Creating corpora in Sketch Engine

The aim of this chapter is to describe the compilation of the learner corpus and the respective control corpus. First, the software tool used for the compilation of the corpora will be described, namely Sketch Engine. Then the source texts and the process of their collection and adjustment will be explained in detail. The characteristics of the learner corpus as well as the control corpus will be described, too.

Kilgarriff et al. state that the Sketch Engine "is a leading corpus tool, widely used in lexicography" (2014, 7). Sketch Engine is both the software and the web service; however, is necessary to differentiate between them. There are various functions that allow researchers to explore phrases, collocations, grammatical constructions and word concordances. Sketch Engine is used by lexicographers (creation of dictionaries with the use of very large corpora) university researchers (linguistic research, discourse analysis etc.), translators (identification of the terminology and phraseology), terminologists (e.g. large companies in a need of a consistent use of a particular term – Sketch Engine can be used for term-finding), language teachers and others.

Sketch Engine covers many languages and aims to grow. "The prerequisite for a basic resource for a language, is simply, a corpus". At this time, the corpora of languages such as Chinese, Arabic and other Asian languages are being developed (Kilgarriff et al. 2014, 18).

Firstly, I will describe the core functions of the software. Secondly, I will focus on the function that is crucial for this diploma thesis – the possibility of creating, uploading and managing a corpus.

4.1.1 Sketch Engine - Software

The functions covered in this section are the Word Sketch (this function has given the software its name), Concordance, Thesaurus and some of the new functionalities of the Sketch Engine, namely GDEX and bilingual sketches.

Word sketches are defined as an “automatic, corpus-derived summaries of a word’s grammatical and collocational behaviour” (Sketch Engine 2015). It is a one page summary of the information that is available about a word and it can actually serve as a draft dictionary entry. With the help of word sketches, a lexicographer’s work is facilitated (Kilgarriff et.al. 2014, 9-10)

Another basic function of the Sketch Engine is Concordance which enables the researcher to actually see what is in a corpus. It shows the raw data that can be accessed from a word sketch or from a simple search form. A researcher is able to search a specific query type, e.g. a lemma, a specific phrase or a word form. The Sketch Engine can even search for a character (in case of languages that do not put spaces between words – e.g. Japanese). Once the concordance is extracted, filtered and sorted, it can be further analysed (Kilgarriff et.al. 2014, 9-10).

The Sketch Engine is currently working on its own distributional thesaurus that is based on common collocations. If a pair of words shares many collocates, the two words in question will appear in each other’s thesaurus entry (Kilgarriff et al. 2014, 14).

There is a continuous work on improving the software as well as making it more user-friendly. Among the new functionalities that are being developed are for example GDEX, bilingual sketches and other. GDEX is an acronym for a Good Dictionary Examples function (added in 2008) that is applied in lexicography[9]. Bilingual sketches are mainly used by bilingual lexicographers (Kilgarriff et al. 2014, 28-30).

[9] The GDEX function uses an algorithm to show the best lines in the copus (examples) first. It is still questionable whether the examples found by the algorithm can appear in a dictionary entry without editing them first.

4.1.2 Sketch Engine – web service

As was mentioned before, The Sketch Engine is widely used in lexicography, translation or teaching. All of this is possible due to large corpora that are accessible through the Sketch Engine – this is the first possibility – to use the preloaded corpora that are managed by the Sketch Engine team. The second option is to create a corpus which is managed by the researcher/user (Kilgarriff et al. 2014, 23).

The official websites state that there are more than 200 different corpora in many languages that differ in size (i.e. the number of words found in a corpus). Through the Sketch Engine, researchers can access big corpora, such as BNC and their sub-corpora or for example TenTen Web Corpora that are available in different national languages. By big corpora at least 50 million-word corpora are meant. Corpora containing specific types of text are available, too. An example of such a corpus is the EUROPARL[10], a corpus extracted from the proceedings of the European Parliament which is used for studies in statistical machine translation and includes 21 European languages (Sketch Engine 2015). EUROPARL is a type of parallel corpus – its basic resources are texts and its translations. (Kilgarriff et al. 2014, 23). There are corpora focused on academic English – for example British Academic Written English (BAWE) and British Academic Spoken English (BASE) (The Sketch Engine 2003-2013). There are also learner corpora for Slovene, Czech and English. The Sketch Engine has historical corpora, too, such as LatinISE (Latin from B.C. period till 20th century) or corpora of children speech (e.g. CHILDES for 22 languages) Kilgarriff et al. (2014, 25).

It is clear that the possibilities of comparing different corpora are immense and the Sketch Engine is a tool that can be used in computational linguistics,

[10] http://www.statmt.org/europarl/

translational studies, learner corpora research and in other fields of linguistic research.

The Sketch Engine enables researchers to create their own corpora that can be further divided into sub-corpora. There are two possibilities of how to create a corpus. First it is possible to upload various texts in common formats (pdf, doc, txt, html etc.). The resulting corpus can be further managed by the users – more data can be added or deleted. Users can conduct a particular research in their corpus and are able to make their corpora available for other researchers. In this thesis, the corpora are created via uploading texts from the hard-drive. The second possibility is to use a tool called WebBootCaT, which creates a corpus from websites. Kilgarrifff et al. state that this tool is very efficient when a researcher aims to investigate terminology of a specific domain (2014, 26). I will now describe the compilation of the corpora that serve as a source of data in this study.

5 The practical part

In the practical part, first, the process of the compilation of the corpora will be described and a characteristic of each of the two corpora will be provided. Second, the methods used for the extraction and sorting of the data will be presented. Third, the extracted data will be analysed.

5.1 The Compilation of the corpora

I created two corpora of English texts. The first consisting of texts written by English L2 university students (L1 Czech) following English philology programmes at Palacký and Masaryk University. It is assumed that these students attended courses focused on academic writing during their studies, since these are usually compulsory. Moreover, when writing their Bachelor or Master’s theses students should attend diploma seminars focused on their field of enquiry, it is therefore expected that they have acquired some basic academic writing skills. The corpus of students’ texts was named **Students’ Theses Corpus**. The second corpus was compiled of research articles written by professional linguists and was named **Research Articles Corpus**. The characteristic of each one of them is needed in order to describe the different nature of the data.

5.1.1 Students’ Theses Corpus

The texts that were used for the compilation of the **Students’ Theses Corpus** were Bachelor and Master theses written in English by Czech university students studying at British and American departments. In this thesis, theses written by students of Masaryk and Palacký University were used. The texts were searched in the online databases of the universities. The majority of texts comes from the online database theses.cz[11], which serves as an online catalogue of Bachelor and Master theses. Several theses were written by my colleagues who either studied, or are currently studying English at Palacký University. The main criterion for the inclusion of a concrete text was its topic only. Theses focused on linguistics were selected (phrases such as *linguistics, English linguistics or English language* were in my search query). The choice of a concrete thesis was more or less random. I did not differentiate between the pdf format and word document (Sketch Engine can process both formats). However, some work had to be done before uploading the texts.

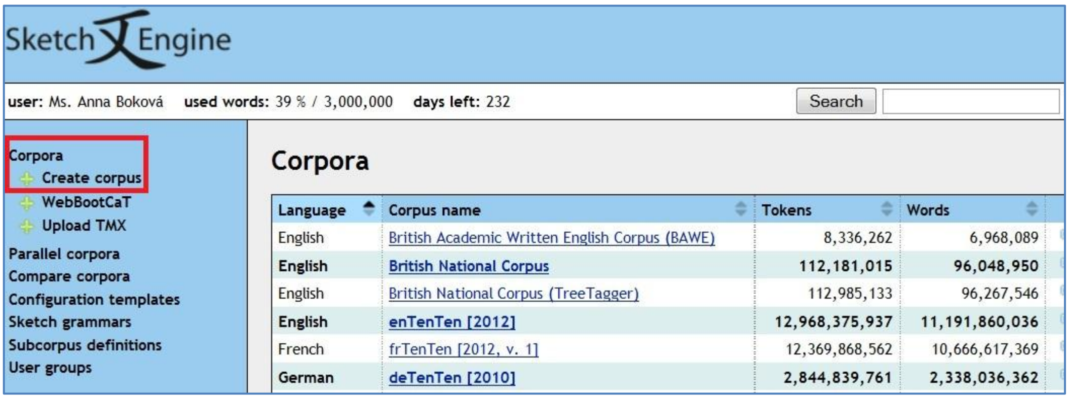
Firstly, it was necessary to adjust each individual file. Some parts were excluded from each file, such as front pages, acknowledgements, contents, Czech Resumés, annotations and appendices. It was assumed that lexical bundles possibly found in those parts could misrepresent the results. The adjustment of the doc format files was very simple, since it is possible to delete any part of the document. The pdf files were adjusted with more difficulties, since they normally cannot be modified. With the help of the Adobe Acrobat Professional programme it was possible to delete the unwanted pages. When all the files were prepared, the creation of the corpus could start.

The corpus creation itself starts with the function that enables a researcher to build his/her own corpus by clicking on the Create corpus button (see Figure 1). The choice of the language and the name of the corpus have to be set afterwards. This primary setting is followed by the uploading of the previously assembled documents (see Figure 2) which were saved on my hard drive. Each document had to be uploaded manually. When added successfully, the uploading of the file had to be finished. The name of each file includes information about the type of thesis (Bc. or Mgr), the year it was written in and the name of the thesis (see Figure 3).

[11] available from https://theses.cz/

When a new file was added, the corpus had to be re-compiled (the Sketch Engine offers this option automatically).

Figure 1. First step in the corpus creation

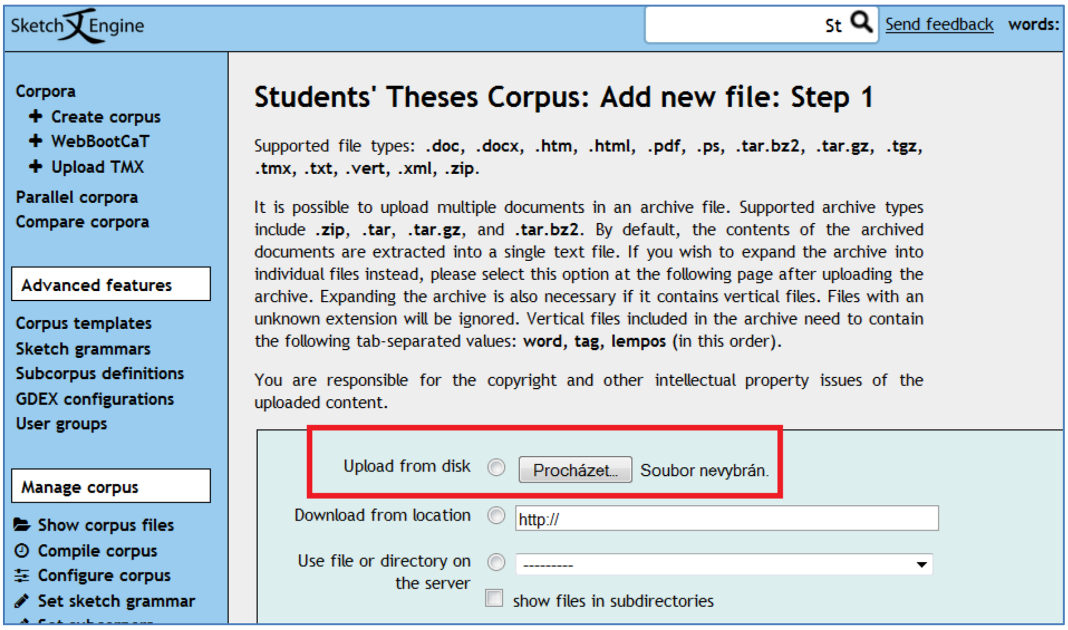


user: Ms. Anna Boková used words: 39 % / 3,000,000 days left: 232

Corpora

Language	Corpus name	Tokens	Words
English	British Academic Written English Corpus (BAWE)	8,336,262	6,968,089
English	British National Corpus	112,181,015	96,048,950
English	British National Corpus (TreeTagger)	112,985,133	96,267,546
English	enTenTen [2012]	12,968,375,937	11,191,860,036
French	frTenTen [2012, v. 1]	12,369,868,562	10,666,617,369
German	deTenTen [2010]	2,844,839,761	2,338,036,362

Figure 2. Uploading from disk



Students' Theses Corpus: Add new file: Step 1

Supported file types: **.doc**, **.docx**, **.htm**, **.html**, **.pdf**, **.ps**, **.tar.bz2**, **.tar.gz**, **.tgz**, **.tmx**, **.txt**, **.vert**, **.xml**, **.zip**.

It is possible to upload multiple documents in an archive file. Supported archive types include **.zip**, **.tar**, **.tar.gz**, and **.tar.bz2**. By default, the contents of the archived documents are extracted into a single text file. If you wish to expand the archive into individual files instead, please select this option at the following page after uploading the archive. Expanding the archive is also necessary if it contains vertical files. Files with an unknown extension will be ignored. Vertical files included in the archive need to contain the following tab-separated values: **word**, **tag**, **lempos** (in this order).

You are responsible for the copyright and other intellectual property issues of the uploaded content.

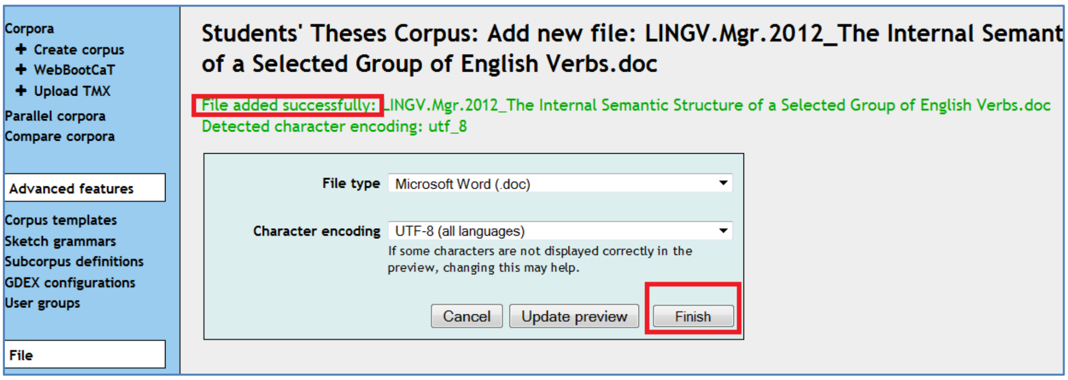
Upload from disk ○ Procházet... Soubor nevybrán.

Download from location ○ http://

Use file or directory on the server ○ ---------

□ show files in subdirectories

Figure 3. The file name



Students' Theses Corpus: Add new file: LINGV.Mgr.2012_The Internal Semant of a Selected Group of English Verbs.doc

File added successfully: LINGV.Mgr.2012_The Internal Semantic Structure of a Selected Group of English Verbs.doc
Detected character encoding: utf_8

File type: Microsoft Word (.doc)

Character encoding: UTF-8 (all languages)
If some characters are not displayed correctly in the preview, changing this may help.

The size of a particular thesis ranges from 9,270 tokens to 44,847 tokens; it assumed that Master theses tend to be longer as compared to Bachelor theses. The total amount of tokens in the **Students' Theses corpus** is 711,222 and the total number of words is 553,005[12]. The number of texts is 31 (15 Bachelor theses and 16 Master theses).

5.1.2 Research Articles Corpus

The second corpus – **Research Articles Corpus** – consists of texts written by scholars, i.e. professionals in the field of linguistics studies. All texts are written in English. Research articles were downloaded from the online database Cambridge Journals[13] that are accessible through Palacký University e-resources[14]. I selected three representative journals that focus on linguistics – *English Language and Linguistics*, *Journal of Linguistics and Language Teaching*. It is assumed that the articles are written by professional linguists who share a similar level of proficiency in English and language knowledge and are therefore comparable. It was not possible to find out whether the professional

[12] The difference between tokens and words is in punctuation. Every word and punctuation in a corpus is referred to as a token.
(https://www.sketchengine.co.uk/documentation/wiki/SkE/Help/JargonBuster)

[13]http://journals.cambridge.org/action/login;jsessionid=3E2CBB3A9C1D13D879B523F9A55FA0F4.journals

[14] http://ezdroje.upol.cz/

linguists are native speakers or not, but it was assumed that the editors were native speakers of English. The **Research Articles Corpus** will serve as a control corpus, since it is assumed that professionals are more experienced and skilled in academic writing; the **Students' Theses Corpus** is considered as a reference corpus.

The process of compilation of the **Research Articles Corpus** was identical to the compilation of the **Students' Theses Corpus** described above. To achieve the intended size of the corpus (around 700 000 tokens) 50 texts were used: 17 articles were taken from *English Language and Linguistics* journal, 14 articles from *Journal of linguistics* and 14 articles from *Language Teaching.* Five articles were chosen from other linguistic journals. All of the articles were downloaded in pdf format. It was necessary to upload more articles than during the compilation of the **Students' Theses Corpus**, since their length ranges from 3,594 to 26,240 tokens, i.e. research articles seem to be generally shorter than some of the Master theses found in the **Students' Theses Corpus**. The total number of tokens in the **Research Articles Corpus** is 679,263 and the total number of words is 534,155.

Table 2 shows the main properties of the corpora created by the use of the Sketch Engine.

Table 2: Summary of the corpora

	Number of Tokens	Number of Words
Students' Theses Corpus	711,222	553,005
Research Articles Corpus	679,263	534,155

5.2 Methods

In this chapter, I will present the methods used for the extraction of the data from the corpora I created, i.e. the **Students' Theses Corpus** and the **Research Articles Corpus**. I will introduce my query and I will describe how the data were sorted.

5.2.1 The query

The query was created on the basis of the previously defined criteria for extracting lexical bundles, i.e. the normalized frequency (at least 20 times per million of words) and the breadth of use (a lexical bundle has to appear at least in 5 different texts in a corpus). This investigation focuses on 4-word lexical bundles, since this approach has been taken in several previous studies focusing on lexical bundles (Cortes 2004; Hyland 2008; Dontcheva-Navratilova 2012) and therefore it is possible to compare the present study with the previous research.

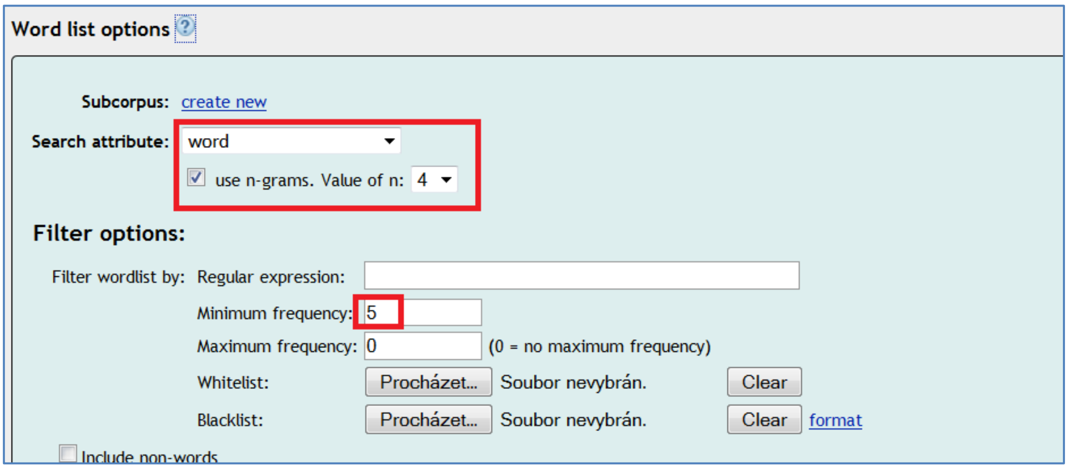
The corpora were not only created by the Sketch Engine, but they were also explored via its tools. The Sketch Engine's function *Word list* is able to create word lists based on different criteria – e.g. it can filter the most frequent words in a corpus or the most salient collocates for a chosen verb[15]. This study's first objective was to create a word list of 4-grams[16] (which are in fact 4-word lexical bundles) found in both corpora (the **Students' Theses Corpus** and the **Research Articles Corpus**).

In this study, the sequences of 4 words will be examined at first. For each of the corpora in question (the **Students' Theses Corpus** and the **Research Articles Corpus**), a separate query (see Figure 4) was created in order to obtain word lists of 4-grams. **Word** was chosen as a search attribute and the use of n-grams was allowed with the sequence of 4. Automatically, the highest frequency at which a 4-gram occurs was 5, i.e. the Sketch Engine did not generate 4-grams which occurred less than 5 times in a given corpus. Then the word list was made and consequently downloaded via the saving option (see Figure 5).

[15] http://www.sketchengine.co.uk/documentation/wiki/Website/Features#Wordlists

[16] For the sake of clarity, when a 4-gram is mentioned in this chapter it is considered a lexical bundle. 4-gram terminology is used here with respect to the software tool that is provided by Sketch Engine.

Figure 4. The Query



Word list options

Subcorpus: create new

Search attribute: word

☑ use n-grams. Value of n: 4

Filter options:

Filter wordlist by: Regular expression:

Minimum frequency: 5

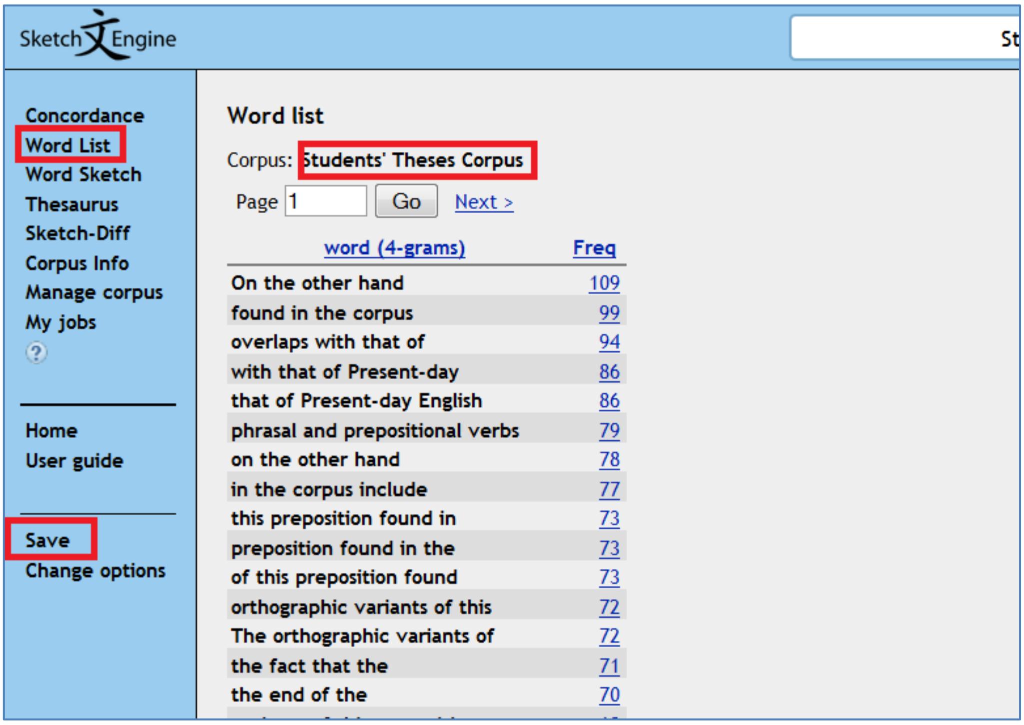
Maximum frequency: 0 (0 = no maximum frequency)

Whitelist: Procházet... Soubor nevybrán. Clear

Blacklist: Procházet... Soubor nevybrán. Clear format

□ Include non-words

Figure 5. The word list of 4-grams found in the Students‘ Theses Corpus



Word list

Corpus: Students' Theses Corpus

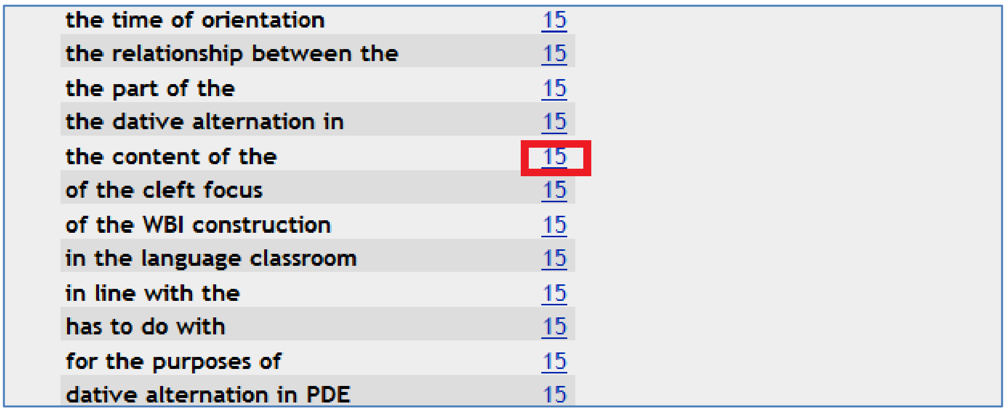
Page 1 Go Next >

word (4-grams)	Freq
On the other hand	109
found in the corpus	99
overlaps with that of	94
with that of Present-day	86
that of Present-day English	86
phrasal and prepositional verbs	79
on the other hand	78
in the corpus include	77
this preposition found in	73
preposition found in the	73
of this preposition found	73
orthographic variants of this	72
The orthographic variants of	72
the fact that the	71
the end of the	70

5.2.2 Sorting the data

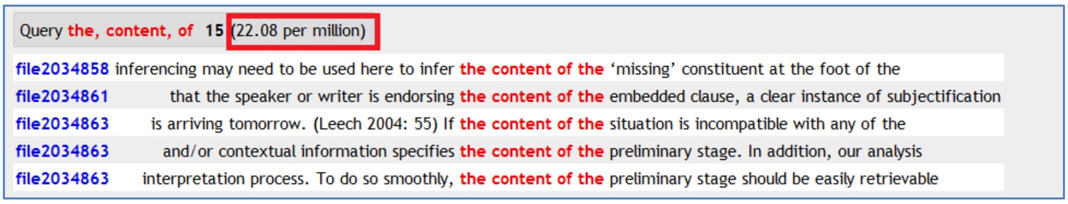
With respect to the criteria that were set for the lexical bundles in this research, the data had to be sorted. First, the criterion of normalized frequency per million of words had to be satisfied, i.e. it was necessary to check the normalized frequency via the hyperlink (see Figure 6) that is available for each one of the n-grams. After clicking on it, the immediate context of an 4-gram is shown as well as the normalized frequency (see Figure 7) per million of words. Only 4-grams with the normalized frequency over 20 per million were included. Each wordlist was saved in a table, so the data could be further sorted.

Figure 6: The hyperlink



the time of orientation	15
the relationship between the	15
the part of the	15
the dative alternation in	15
the content of the	15
of the cleft focus	15
of the WBI construction	15
in the language classroom	15
in line with the	15
has to do with	15
for the purposes of	15
dative alternation in PDE	15

Figure 7: Normalized frequency



Query the, content, of 15 (22.08 per million)

file2034858	inferencing may need to be used here to infer the content of the ‘missing’ constituent at the foot of the
file2034861	that the speaker or writer is endorsing the content of the embedded clause, a clear instance of subjectification
file2034863	is arriving tomorrow. (Leech 2004: 55) If the content of the situation is incompatible with any of the
file2034863	and/or contextual information specifies the content of the preliminary stage. In addition, our analysis
file2034863	interpretation process. To do so smoothly, the content of the preliminary stage should be easily retrievable

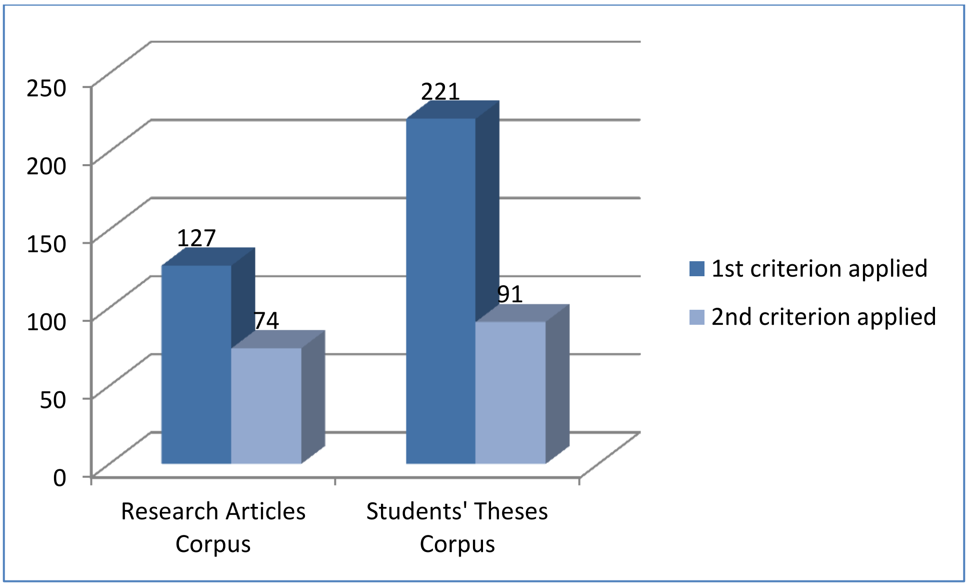
The total number of 4-grams was the subject of interest. In the **Research Articles Corpus**, 406 4-grams were found. However, when the data were observed more closely, some discrepancies appeared. Among sequences of 4-grams emerged sequences of letters such as *T I O N, M E N T, G L I S, O L O G*. The context of these sequences was checked again with the use of the hyperlink. It was found out that the sequences of letters appeared in the source texts, but they were interpreted by the Sketch Engine as words because the letters were separated by individual spaces. Since this thesis is interested only in the sequences of words, the sequences of letters were manually excluded from the list of 4-grams. 127 4-grams remained after this exclusion.

To satisfy the second criterion for identifying a lexical bundle, namely the breadth of use, the context of each of the 127 4-grams had to be explored through the hyperlink. By clicking on it, it is possible to see not only the immediate context, but also information about the original file. As stated already, if there were at least 5 different files in which a particular 4-gram occurred, this 4-gram was identified as a lexical bundle. After meeting both set criteria, **74** lexical bundles remained (53 bundles were excluded) that were extracted from the **Research Articles Corpus** and that will be subject to an analysis.

Data in the **Students' Theses Corpus** were sorted in the same way. There were 221 4-grams which satisfied the first criterion (normalized frequency per million of words). When the second criterion (the breadth of use) was applied, **91** lexical bundles remained (130 bundles were excluded).

It should be noted that in the **Research Articles Corpus,** 41 % of bundles (53 bundles) were excluded when the second criterion (breadth of use) was applied; in the **Students' Theses Corpus,** it was 59 % of bundles (130 bundles). Figure 8 represents the decrease in the number of lexical bundles in both corpora.

Figure 8: Comparing the corpora



It is obvious that the decrease in the number of 4-grams is much more prominent in the **Students' Theses Corpus** than in the **Research Articles**

Corpus. One of the plausible explanations would be simply the fact that students repeat themselves more while writing about a certain topic. 4-grams such as *variants of this preposition* (repeated 68 times in a single text) or *prototypical meaning overlaps with* (61 times in a single text) are examples that were excluded because they occurred in less than 5 texts. This phenomenon is supported by Hyland who suggests that students seem to rely more on prefabricated lexical bundles when they propose their arguments. Moreover, *"repetition of strings has been recognised as a problematic feature of academic texts by second language writers"* (Hyland 2008, 50).

To sum up, there were 74 lexical bundles extracted from the **Research Articles Corpus** and 91 lexical bundles extracted from the **Students' Theses Corpus** that will be subjected to the analysis.

5.3 Data analysis

The corpora subjected to analysis – **Students' Theses Corpus** and **Research Articles Corpus** – represent two types of academic written English, namely the writing produced by L2 students and the writing produced by professionals. The writing of students and professionals differ with respect to their readers and generally, their purpose. Theses produced by students represent one of the conditions of gaining a university degree and students are expected to demonstrate a certain level of writing skills. Professional linguists, on the other hand, write their papers for the community of professionals in the same academic field. It is assumed that the latter are more skilled and the **Research Articles Corpus** therefore serves as a control corpus.

For the sake of clarity, the Table 3 is included, which shows 50 most frequent lexical bundles in the **Research Articles Corpus** and **Students' Theses Corpus.** The blue-shaded cells represent the lexical bundles that were found in both corpora; only 21 bundles out of the first 50 were used both by Czech students and professionals.

Table 3: The 50 most frequent lexical bundles in both corpora

Rank	Research Articles Corpus	Number of occurrences	Students' Theses Corpus	Number of occurrences
1	on the other hand	98	On the other hand	109
2	on the basis of	82	on the other hand	78
3	in terms of the	70	the fact that the	71
4	the use of the	64	the end of the	70
5	in the case of	63	on the basis of	57
6	the extent to which	55	at the end of	57
7	at the same time	54	it is necessary to	52
8	with respect to the	48	part of the thesis	43
9	at the end of	44	it is possible to	41
10	as well as the	40	is one of the	41
11	the fact that the	39	to the fact that	39
12	as a result of	39	one of the most	37
13	in the context of	34	in the form of	37
14	in the use of	33	as well as the	37
15	On the other hand	33	the total number of	36
16	the nature of the	32	can be found in	36
17	the end of the	29	As far as the	36
18	in the sense of	28	As can be seen	36
19	a wide range of	27	at the same time	34
20	the basis of the	26	can be seen in	32
21	that there is a	25	I would like to	32
22	at the time of	25	the beginning of the	31
23	should be noted that	24	of the fact that	31
24	to the fact that	22	is based on the	31
25	to refer to the	22	by the fact that	31
26	per cent of the	22	in the context of	30
27	on the one hand	22	at the beginning of	30
28	the semantics of the	21	when it comes to	28
29	is one of the	21	for the purposes of	28
30	in the form of	21	can be seen from	28
31	in terms of a	21	it is important to	27
32	at the beginning of	21	can not be used	27
33	to do with the	20	to be able to	25
34	on the part of	20	the usage of the	25
35	does not seem to	20	a wide range of	25
36	can be found in	20	the nature of the	24
37	presence or absence of	19	in the field of	24
38	At the same time	19	be seen in Figure	24
39	used to refer to	18	the results of the	23
40	the case of the	18	the analysis of the	22
41	is likely to be	18	that there is a	22
42	in the present study	18	for the purpose of	22
43	in relation to the	18	the practical part of	21
44	as part of the	18	part of this thesis	21
45	In the case of	18	in the sense of	21

46	I would like to	18	in the process of	21
47	to be able to	17	as a result of	21
48	that the use of	17	the first part of	20
49	it should be noted	17	practical part of this	20
50	in the absence of	17	of the thesis is	20

Here, it is clearly visible that many lexical bundles used by professional linguists do not appear at all in the students' repertoire or appear far less frequently, and vice versa, some of the lexical bundles used by students were never, or only rarely, found in the corpus of professional writing.

An interesting explanation of this phenomenon is proposed by Hyland when he claims that these differences are caused by the genre variation. He stresses the fact that master's thesis is a pedagogic genre; therefore students have to demonstrate their research skills and familiarity with the discipline. Research articles, on the other hand, serve to the academics as a means of establishing reputation and exhibiting the novelty and relevance of their work (2008, 57). Hyland therefore suggests that students' use of lexical bundles does not necessarily have to be wrong, but genre specific (2008, 59).

It can be assumed that mistakes in the structure of lexical bundles may appear in the students' writing, since they are always influenced by their mother tongue. Dontcheva-Navratilova states that "there are some occasional grammatical errors in the use of lexical bundles" by students. She gives an example of omitting the definite article in the phrases *in the case of*, *on the one hand* or inserting the definite article unnecessarily, i.e. in ~~the~~ *terms of* (2012, 46). It was only the phrase *in case of the* that was found in the **Students' Theses Corpus**, but with lower normalized frequency (14 per million of words). The present analysis has shown that the structural inaccuracy is not such a common phenomenon and it seems that students do not make these errors in the most frequent lexical bundles (no errors of this type were noticed in the list of bundles subjected to the present analysis) and it can be assumed that students are aware of their fixed structure.

The structural distribution is very similar in both corpora and agrees with the distribution typical for academic prose, however, the individual lexical bundles used by students differ from the lexical bundles found in the **Research Articles Corpus** that serves as a control corpus in this thesis. In the following chapter, the structural distribution in the **Research Articles Corpus** and **Students' Theses Corpus** will be examined more closely.

5.3.1 Structural distribution of the lexical bundles in the Research Articles Corpus

For the sake of clarity, Table 4 presents again the different structural categories of lexical bundles (Biber et al. 1999):

Table 4: Structural categories of lexical bundles

1	noun phrase with of phrase fragment
2	noun phrase with other post-modifier fragment
3	prepositional phrase with embedded of-phrase fragment
4	other prepositional phrase (fragment)
5	anticipatory it + verb phrase/adjective phrase
6	passive verb + prepositional phrase fragment
7	copula be + noun phrase/adjective phrase
8	(verb phrase +) that-clause fragment
9	(verb/adjective +) to-clause fragment
10	adverbial clause fragment
11	pronoun/noun phrase + be (+ …)
12	other expressions

Almost all the structural categories were present in the **Research Articles Corpus** with two exceptions: lexical bundles with the structure *adverbial clause fragment* and *pronoun/noun phrase + be + (+ ...)* were not found in the list of bundles that were subjected to the analysis. These categories are therefore left out in the Table 4. *Longman Grammar of Spoken and Written English* (Biber et al. 1999) was used where all the structural categories are listed with examples of lexical bundles in academic prose. Each one of the lexical bundles from the **Research Articles Corpus** was compared with the examples from *Longman Grammar of Spoken and Written English.* The majority of the typical examples found in the **Research Articles Corpus** were present in this publication as well, so the sorting itself was easier, because it was clear to which category a certain lexical bundle should be included. Table 5 shows the structural distribution of lexical bundles in the **Research Articles Corpus**.

Table 5 - Structural distribution in the Research Articles Corpus

	Structural category followed by the bundles found in the Research Articles Corpus	**number of cases**	**% of lexical bundles**
1	noun phrase with of phrase fragment	13	18
	the use of the, the nature of the, the end of the, a wide range of, the basis of the, per cent of the, the semantics of the, presence or absence of, the case of the, the structure of the, one of the most, the part of the, the content of the,		
2	noun phrase with other post-modifier fragment	3	4
	the extent to which, the fact that the, the relationship between the		
3	prepositional phrase with embedded of-phrase fragment	20	27
	on the basis of, in terms of the, in the case of, at the end of, as a result of, in the context of, in the use of, in the sense of, at the time of, in the form of, in terms of a, at the beginning of, on the part of, In the case of, in the absence of, in the presence of, in the development of, On the basis of, for the purposes of, in terms of their		
4	other prepositional phrase (fragment)	14	19
	on the other hand, at the same time, with respect to the, On the other hand, to the fact that, on the one hand, At the same time, in the present study, in relation to the, as part of the, by the fact that, in line with the, in the sense that, in the next section,		
5	anticipatory it + verb phrase/adjective phrase	3	4
	it is important to, it is clear that, it should be noted,		
6	passive verb + prepositional phrase fragment	1	1
	can be found in		
7	copula be + noun phrase/adjective phrase	3	4
	is one of the, may be due to, is part of the		
8	(verb phrase +) that-clause fragment	3	4
	that there is a, that there is no, should be noted that		
9	(verb/adjective +) to-clause fragment	9	12

	to refer to the, to do with the, is likely to be, to be able to, can be used to, are more likely to, has to do with, does not seem to, used to refer to		
10	other expressions	5	7
	as well as the, may or may not, I would like to, would like to thank, that the use of		
Total		**74**	**100**

The data showed that the lexical bundles in the **Research Articles Corpus** consisted mostly of noun phrase components and prepositional phrase components (50 cases), which represent 67 % of the lexical bundles in question. Those structures generally dominate in academic prose; Biber et al. states that 60 % of lexical bundles found in academic prose are parts of noun phrases or prepositional phrases (1999, 995). The top 10 lexical bundles in the list are prepositional phrase constructions or noun phrase constructions with one exception – the lexical bundle *as well as the* which as Biber et al. state does "not fit neatly into any of the other categories" (1999, 1024). The most frequent lexical bundle – *on the other hand*[17] (in the middle of a sentence) – occurs more than 144 times per million of words. This lexical bundle is considered to be one of the most common 4-word lexical bundles in academic prose as well as the phrase *in the case of* which is 5th in the list (Biber et al. 1999, 994).

Lexical bundles *in the present study* and *in the next section* are used to refer to a particular discourse context (Biber et. al 1999, 1019).

The constructions with a verb component represent 27 % of the cases; they are generally used less frequently and are less repeated in the corpus when compared to lexical bundles with prepositional phrase component or noun phrase component. Three lexical bundles – *it is important to, it is clear that, it should be noted* represent the anticipatory it + verb/adjective pattern. Biber et al. state that

[17] The lexical bundle *On the other hand* (at the beginning of a sentence) occurs 33 times in the **Research Articles Corpus.**

these patterns often express writer’s stance –importance (1), (2) or necessity. They can also signal that something is taken for granted, in other words that the information in the *that* clause is a fact (3) (1999, 1020).

(1) Now, **it is important to** see whether the other major substrate language in the contact community for CSE –Malay – also exhibits person agreement in the form of the blocking effect. [JL.2014_11]

(2) Although in this review we focus on the home language and literacy practices of immigrant families, **it is important to** keep in mind that these practices exist … [LT.2014_8]

(3) However, **it should be noted** that even in languages of the latter type, clefts can be used as answers to wh-questions … [JL.2014_12]

Verb phrase components were mostly represented in the group with a to-clause fragment. There were 9 instances found in the corpus that matched the criteria. When an adjective is present in the predicate which is followed by a to-clause, it often expresses ability or possibility. The lexical bundle *is likely to be* seems to express probability (4).

(4) This in turn, **is likely to be** reflected in a decrease of overall sentence length. [ELL.2012_3]

Lexical bundles *I would like to* and *would like to thank* that were found in the **Research Articles Corpus** are worth mentioning as well. They were not found in the lists of lexical bundles in *Longman Grammar of Spoken and Written English* (Biber et al. 1999). This does not mean that those bundles are not used by academics, but they may be used less frequently. When the immediate context was checked, it was found out that these expressions are used in two ways.

First, those bundles serves as expressions of gratitude, e.g. to a fellow colleague or researcher (5) and (6). The lexical bundle *I would like to* is also followed by phrases such as *offer my sincere thanks*, *express my thanks* that express the same. However, it is necessary to state that those bundles expressing gratitude appeared in the footnotes of a particular research article.[18]

(5) Last but not the least, **I would like to** acknowledge the constructive and challenging criticism of two anonymous ELL referees and the excellent editorial support. [ELL.2014_15]

(6) **I would like to** offer my sincere thanks to Terttu Nevalainen and Matti Kilpiö for their comments on the earlier versions of this article [ELL.2013_8]

Second, *I would like to* can be followed by other verbs such as *acknowledge, convince, discuss* that can express researcher's goals (7) and (8).

(7) The explanation of ablaut that **I would like to** submit relies on lexical statistics and has a psycholinguistic slant. [JL.2014_14]

(8) The last area of SLA research that **I would like to** discuss is not easily characterized as a misapplication of SLA research but is probably best discussed in terms of its relevance (or dare I say irrelevance) to L2 pedagogy. [LT.2015_14]

The data in the **Research Articles Corpus** showed that the structural distribution of the lexical bundles matches the findings in the previous research. It can be said that this corpus of professional writing can be used as a control corpus to the **Students' Theses Corpus**. The examples introduced in this chapter (1), (2), (3), (4), (7) and (8) do not represent any unexpected findings, but serve as a confirmation that lexical bundles in the **Research Articles Corpus** are used in a

[18] The footnotes in the research articles used for the compilation of the **Research Articles Corpus** could not be excluded. The case of *I would like to* and *would like to thank* could be considered as misrepresentation of the data, but in my opinion, these examples show a phenomenon that is typical for academic prose, i.e. expressing gratitude for comments or suggestions to a particular study.

typical way. In the next chapter, the language of university students will be discussed in greater detail.

5.3.2 Structural distribution of the lexical bundles in the Students' Theses Corpus

In the **Students' Theses Corpus**, only one structural category was not present; namely the category with a *pronoun/noun phrase + be + (+ ...)*. Again, *Longman Grammar of Spoken and Written English* (Biber et al. 1999) was used for the checking of each particular lexical bundle. In 33 cases, lexical bundles produced by the Czech students were not found in this publication, for example the phrases such as *part of this thesis, As far as the, As can be seen, when it comes to, a wide range of, I would like to, from the point of, used in order to, to find out whether, in order to provide* etc. Some of them will be discussed in this chapter. This may suggest that students either use different lexical bundles than professional linguists or they may use them incorrectly. Table 6 shows the structural distribution in the **Students' Theses Corpus.**

Table 6 - Structural distribution in the Students' Theses Corpus

	Structural category followed by the bundles found in the Students' Theses Corpus	**number of cases**	**% of lexical bundles**
1	noun phrase with of phrase fragment	26	28
	the end of the, part of the thesis, one of the most, the total number of, the beginning of the, the usage of the, a wide range of, the nature of the, the results of the, the analysis of the, the practical part of, part of this thesis, the first part of, practical part of this, the use of the, the point of view, a result of the, The aim of this, the rest of the, point of view of, a part of the, the form of the, the basis of the, first part of the, this part of the, the case of the		
2	noun phrase with other post-modifier fragment	1	1
	the fact that the		
3	prepositional phrase with embedded of-phrase fragment	16	18

	on the basis of, at the end of, in the form of, in the context of, at the beginning of, for the purposes of, in the field of, for the purpose of, in the sense of, in the process of, as a result of, in the middle of, from the point of, of the number of, in the course of, In the case of		
4	other prepositional phrase (fragment)	13	14
	on the other hand, On the other hand, to the fact that, as far as the, at the same time, of the fact that, by the fact that, of the thesis is, in the practical part, in relation to the, as a basis for, in the same way, of this thesis is,		
5	anticipatory it + verb phrase/adjective phrase	6	7
	it is necessary to, it is possible to, it is important to, it is not possible, It is important to, is not possible to		
6	passive verb + prepositional phrase fragment	11	12
	can be found in, can be seen in, is based on the, can be seen from, be seen in Figure, can be used as, can be used to, can be used in, be used as a, be found in the, be seen from the,		
7	copula be + noun phrase/adjective phrase	2	2
	is one of the, is used in the,		
8	(verb phrase +) that-clause fragment	2	2
	that there is a, can be concluded that,		
9	(verb/adjective +) to-clause fragment	6	7
	to be able to, due to the fact, it comes to the, used in order to, can not be used, to find out whether,		
10	adverbial clause fragment	2	2
	As has already been, As can be seen		
11	other expressions	6	7
	as well as the, I would like to, when it comes to, I am going to, in order to provide, in order to be,		
Total		91	100

The data showed that the **Students' Theses Corpus** is dominated by the lexical bundles with a prepositional or noun phrase element which is similar to the findings in the control corpus. These patterns were detected in 56 cases (approximately 61 %) out of the total of 91 cases. The percentage is lower than in the **Research Articles Corpus**. Nonetheless, it still corresponds to the findings from the previous research – i.e. these structures normally represent over 60 % of cases in academic prose (Biber et al., 1999). On the other hand, the percentage of the cases with a verb element is higher than in the **Research Articles Corpus.**

Verb patterns represent 30 % of the cases. I will now describe some of the differences between the lexical bundles used by professionals and the lexical bundles used by Czech students that were noticed during the analysis of the data.

The most frequent lexical bundle in the **Students' Theses Corpus** is the structure *On the other hand* (rank 1, found at the beginning of a sentence) and *on the other hand* (rank 2, in the middle of a sentence). The former has 109 instances (normalized frequency per million of words is 153) in the corpus, the latter has 78 instances (normalized frequency is 109). When the rank of the lexical bundle *On the other hand* was compared to the control corpus, it was found out that in the **Research Articles Corpus,** this lexical bundle has rank 15; it is therefore used less frequently by professional linguists. This may suggest that students' usage differs from professionals'. Tazegül states that "the use of connectives has always been a trouble spot for second or foreign language learners (SSL/FLL) of English" (2015, 118). In her study of the connective *on the other hand* in Turkish learner corpus, Tazegül confirmed that Turkish doctoral students tend to overuse this connective. She also found out that native speakers used *on the other hand* in company with *on the one hand* while non-native speakers did not (2015, 126). The data in the **Students' Theses Corpus** showed similar findings, since both *On the other hand* and *on the other hand* are the most frequent lexical bundles and the connective *on the one hand* does not appear in the list of lexical bundles extracted from the **Students' Theses Corpus** subjected to the analysis.

The lexical bundle *in the case of* (in the middle of a sentence) which has rank 5 in the **Research Articles Corpus** was not found in the **Students' Theses Corpus.** However, the same phrase used at the beginning of a sentence – *In the case of* – was found in the corpus, but there were only 15 instances (rank 90, normalized frequency being only 21 per million).

As was already mentioned in the beginning of this chapter, there were instances of lexical bundles that were not found in the control corpus. Lexical bundles that contain the phrase *in order to* may serve as an example. There are three different bundles: *used in order to, in order to provide and in order to be. In* *order to* is usually used to introduce adverbial clauses expressing purpose (Biber et al. 1999, 89). Since this phrase is often taught on the lower levels of education, it can be assumed that students use it because, again, they are familiar with it. Although the lexical bundles containing *in order to* have lower normalized frequencies (over 20) in the **Student' Theses Corpus**, they are not found in the **Research articles corpus** at all.

Another phrase that does not seem to be typical for academic discourse is the lexical bundle *when it comes to.* Students use it quite often – there are 28 instances found in the corpus (normalized frequency per million words is 39). This phrase was not found in the **Research Articles Corpus** at all, even when bundles with lower frequencies were checked.

It was observed that students seem to refer much more to their writing/thesis than professionals do in the research articles. The need for a kind of framing device can be considered as a plausible explanation. In other words, students need to express what will be discussed in their thesis, what will come first and what will come next. They use great variety of phrases – the lexical bundles often contain expressions such as *thesis*, *part*, *first part*, *practical part* etc. The 4-word lexical bundles are often part of 5 or 6-word bundles, e.g. *in the practical part of this thesis/my thesis/the thesis/this work* etc. When these findings were compared to the control corpus, it was found out that these bundles do not appear there. Only the bundles *on the part of* and *as part of th*e contained the word *part* in the control corpus, but those bundles were not used to refer to the particular study. There are 11 lexical bundles that are used as a framing device as opposed to only 2 such bundles in the **Research Articles Corpus** – *in the present study, in the next section.* As for the structure, the bundles used as a framing device in the **Students' Theses Corpus** are mostly noun phrases with an of phrase fragment or prepositional phrases. These structures are used heavily by students – the lexical bundle *part of the thesis* is in the top ten most frequent bundles. Interestingly, reference to e.g. *second part* or *theoretical part* can be found in the **Students' Theses Corpus,** but with much lower frequencies.

Another difference that was noticed in the **Students' Theses Corpus** is in the usage of the lexical bundle with the anticipatory *it*. Two of such phrases appeared in the top ten lexical bundles – *it is necessary to* (73 per million words) and *it is possible to* (57 per million words). At the same time, these bundles were not found in the **Research Articles Corpus** (there were three structures with the anticipatory *it* found, but with lower normalized frequencies – less than 25 per million of words). This finding may suggest that students tend to use these bundles more than professional linguists. The lexical bundle *it is necessary to* is often preceded by a discourse marker such as *therefore, moreover, first of all, however, To begin with, Nevertheless, In addition,* and followed by a verb. It seems that this phrase serves as a means of turning reader's attention towards the writer's statement. Students may also want to stress the information that follows their proposition. The examples in (9), (10) and (11) illustrate this phenomenon.

(9) First of all, though, **it is necessary to** go back to the roots.
[LINGV.Bc.2014_24]

(10) Therefore, **it is necessary to** explain why they were included in this category. [LINGV.Mgr.2013_2]

(11) In addition, **it is necessary to** state in advance what in particular such an assessment should monitor. [LINGV.Mgr.2014_17]

The data also showed that Czech students seem to use structures that contain passive verbs and prepositional phrase fragments. These structures are used in 12 % of all cases in the **Students' Theses Corpus** as opposed to only 1 % of all cases in the **Research Articles Corpus.** The lexical bundle *can be seen in* (normalized frequency per million of words is 45) is followed by the word *Figure, appendix* or *table.* This seems to be genre-specific, since students often use tables and figures in their theses and frequently refer to them. Since there was no lexical bundle which would contain the word *figure* found in the list of lexical bundles extracted from the **Research Articles Corpus**, the bundles with a lower normalized frequency were checked in this corpus. Only one lexical bundle was found – *as shown in figure* – there were only 5 instances in the whole corpus. However, this may suggest that researchers do not use the phrase *can be seen in* in the same way as students.

The lexical bundle *can be found in* (normalized frequency per million of words is 50) is often used by students in order to refer to grammar books (12), dictionaries (13) or to refer to appendices (14).

(12) More detailed distinction of phrasal verbs **can be found in** Comprehensive Grammar of English Language (2000) where we can found six types of multi-word verbs (see Table 4). [LINGV.Bc.2013_2]

(13) In this section I am going to focus on general information about the noun experience that **can be found in** dictionaries. [LINGV.Bc.2014_25]

(14) Example of the blank document **can be found in** the Appendix of this thesis (see Appendix 3); … [LINGV.Mgr.2014_19]

The fact that students tend to use verbal structures more than professionals may suggest that students are less experienced and are not aware of how to form their statements in a more academic way. They are therefore using structures they are more familiar with – passives, constructions with anticipatory *it* etc.

6 Conclusions

The first objective of this thesis was to create two comparable corpora of professional and students’ writing. Via the use of Sketch Engine, this goal was achieved and the **Research Articles Corpus** and **Students’ Theses Corpus** were created. Table 7, once again, summarizes the main properties of both corpora.

Table 7: Summary of the corpora

	Number of Tokens	Number of Words
Students’ Theses Corpus	711,222	553,005
Research Articles Corpus	679,263	534,155

When the corpora were compiled, the lexical bundles were extracted from both corpora according to previously set criteria, namely the **normalized frequency per million words** (over 20) and **breadth of use** (a particular lexical bundle have to occur in at least 5 different texts in a given corpus). There were 74 bundles in the **Research Articles Corpus** and 91 bundles in **Students’ Theses Corpus** that satisfied the above mentioned criteria and that were subjected to the structural analysis.

Table 8 shows the structural distribution of lexical bundles in both corpora.

Table 8: The structural distribution of lexical bundles in both corpora

		Research Articles Corpus		**Students‘ Theses Corpus**	
	Structural type of lexical bundles	**Number of cases**	**% of lexical**	**Number of cases**	**% of lexcical**
1	noun phrase with of phrase fragment	13	18	26	28
2	noun phrase with other post-modifier fragment	3	4	1	1
3	prepositional phrase with embedded of-phrase fragment	20	27	16	18
4	other prepositional phrase (fragment)	14	19	13	14
5	anticipatory it + verb phrase/adjective phrase	3	4	6	7

6	passive verb + prepositional phrase fragment	1	1	11	12
7	copula be + noun phrase/adjective phrase	3	4	2	2
8	(verb phrase +) that-clause fragment	3	4	2	2
9	(verb/adjective +) to-clause fragment	9	12	6	7
10	adverbial clause fragment	x	x	2	2
11	pronoun/noun phrase + be (+ …)	x	x	x	x
12	other expressions	5	7	6	7
Total		**74**	**100**	**91**	**100**

In both the **Research Articles Corpus** and **Students' Theses Corpus**, almost all structural types of lexical bundles were found. The structure consisting of a *pronoun/noun phrase + be (+ ...)* is not present in both corpora and the structure with an *adverbial clause fragment* is missing in the **Research Articles Corpus.**

The structural distribution of lexical bundles in both corpora corresponds to the structural distribution typical for academic prose, i.e. that over 60 % of lexical bundles that are found in academic prose consist of noun phrase and prepositional phrase components (Biber et al. 1999). In the **Research Articles Corpus**, 67 % of bundles have a noun phrase or a prepositional phrase element; in the **Students' Theses Corpus**, it is 61 %. The lexical bundles with a verbal component represent 27 % of cases found in the **Research Articles Corpus** and 30 % of cases found in the **Students' Theses Corpus.**

The most frequent lexical bundles in both corpora are the bundles *on the other hand* and *On the other hand.* The analysis showed that there is a difference in the use of *On the other hand* in the beginning of a sentence by professionals and by students. This lexical bundle has normalized frequency 153 per million of words in the **Students' Theses Corpus**, whereas in the **Research Articles Corpus,** the normalized frequency is 48 per million of words. This may suggest that students overuse this connective in the beginning of a sentence.

The analysis of the data in the **Students' Theses Corpus** showed that there exist differences between the language of Czech students of English and professional linguists. Although the structural distribution is very similar in both corpora and agrees with the distribution typical for academic prose, the individual lexical bundles used by students differ from the lexical bundles found in the **Research Articles Corpus** that is considered to be a control corpus in this thesis. In the top 50 lexical bundles, only 21 were used both by the students and professionals.

During the analysis, *Longman Grammar of Spoken and Written English* was used for the control of the structure of lexical bundles found in the corpora. The lexical bundles listed in this publication are the most frequent and the most typical for academic prose. Approximately one third (33 cases) of lexical bundles in the **Students' Theses Corpus** were not found in this publication and did not appear in the control corpus. Among these structures are for example lexical bundles containing the phrase *in order to* or words and phrases such as *thesis, first part* or *theoretical part.* Another such structure is the lexical bundle *when it comes to* that is also used only by students.

It was also observed that students use the structures with anticipatory *it* with higher normalized frequencies per million of words as opposed to the usage of these structures by professionals.

Other differences with respect to the control corpus were observed in the students' use of lexical bundles, e.g. Czech students use structures that contain passive verbs in 12 % of cases in the **Students' Theses Corpus** as opposed to only 1 % of cases in **Research Articles Corpus**. Students also seem to refer to their writing by using framing devices that contain expressions such as *thesis, part, first part, theoretical part* etc. and this structures do not appear in the control corpus.

The findings have shown that students do use lexical bundles, but they use structures not typical of academic discourse. This finding is in opposition to what Dontcheva-Navratilova states – that Czech students use a limited repertoire of lexical bundles. On the other hand, she also states that Czech students did not acquire the lexical bundles typical for academic discourse (2012, 55-56) which has proved to be true, since one third of lexical bundles were not found in *Longman Grammar of Spoken and Written English*.

However, the Hyland's proposition that lexical bundles can differentiate genres seems to be confirmed, since the findings in this thesis show variation in the structure of the lexical bundles across student and professional writing. It should be noted that the writing of a thesis has different goals, purposes and audience than the writing of a research article; therefore the lexical bundles in found different genres naturally differ.

From a pedagogical perspective, it would be advisable to create learning materials for students based on the corpus findings. It would be helpful to summarize of which lexical bundles students should be aware of. From my experience as a student, it is expected of me to be familiar with the writing conventions in the academic discourse, but usually there are neither materials, nor courses that would make it possible. In this respect, I agree with Dontcheva-Navratilova when she states that exposure to lexical bundles through reading does not result in their acquisition (2012, 55).

The functional analysis of lexical bundles was not an objective in this thesis. Nevertheless, it may be one of the suggestions for further research. Dontcheva-Navratilova has already examined the functional distribution in a smaller corpus of Czech students' theses, but she used a different approach (see chapter 2.3). One of the suggestions for further research would be to examine the lexical bundles found in the **Students' Theses Corpus** from the functional perspective and compare the results to Dontcheva-Navratilova's findings.

7 Czech Summary

Tato diplomová práce se zabývá otázkou, zda používání tzv. lexikálních svazků (lexical bundles) dokáže rozlišit žánr v akademické próze. Lingvistická literatura uvádí, že správné používání lexikálních svazků u nerodilých mluvčí přispívá k plynulosti a přirozenosti vyjadřování v dané oblasti. Prvním cílem této diplomové práce je vytvořit dva korpusy anglických testů. První je sestaven z textů psaných v anglickém jazyce studenty českých vysokých škol, tedy nerodilými mluvčími, jejichž mateřským jazykem je čeština, druhý je vytvořen z textů výzkumných článků v oblasti lingvistiky, jejichž autory jsou buďto rodilí mluvčí, nebo se předpokládá, že jsou texty editovány rodilými mluvčími. Druhým cílem je analýza čtyřslovných lexikálních svazků nalezených v těchto korpusech, konkrétně jaké strukturální typy lexikálních svazků jsou používány studenty a jaké profesionály, zda tyto svazky používá více či méně daná skupina a zda obě skupiny používají stejné lexikální svazky. Na základě předchozích výzkumů se předpokládá, že repertoár lexikálních svazků se u obou skupin může lišit.

Teoretická část (kapitoly 2, 3 a 4) se zabývá studentskými korpusy (learner corpora), charakteristikou a klasifikací lexikálních svazků a v neposlední řadě také softwarem pro vytváření korpusu.

Studentské korpusy jsou definovány jako elektronické soubory textů, jejichž pisatelé užívají anglický jazyk nebo jiný jazyk jako druhý nebo cizí jazyk. Tyto korpusy se od korpusů rodilých mluvčích liší zejména ve vyšším počtu chyb, což je nutné brát v potaz při analýze dat. Studium studentských korpusů může mít dopad in na výuku daného jazyka, protože získané poznatky lze aplikovat v pedagogice. Kapitola 2 také pojednává o významných studentských korpusech, jejich typologii a pravidlech vytváření. Jsou představeny také dva hlavní metodologické přístupy pro lingvistickou analýzu studentských korpusů, jmenovitě *Contrastive Interlanguage Analysis* a *Computer-aided Error Analysis.*

V kapitole 3 je pojednáno o lexikálních svazcích, které jsou definovány jako opakující se výrazy, které se vyskytuj pohromadě v přirozeném diskurzu. Přitom se musí vyskytovat často, aby mohly být považovány za lexikální svazek (alespoň 10 výskytů na milion slov), a objevit se ve více než 5 různých textech ve zkoumaném korpusu (Biber et al. 1999). Tato diplomová práce se zaměřuje na lexikální svazky v akademické próze, pro niž jsou typické struktury tvořené podstatným jménem (noun phrases) nebo předložkovými vazbami (prepositional phrases).Tyto struktury převažují ve více než 60 % případů, v diplomové práci však bude celkově vyhledáváno 12 strukturálních typů lexikálních svazků. Lexikální svazky lze zkoumat také z hlediska jejich funkcí v diskurzu. V této kapitole je také shrnut dosavadní výzkum o lexikálních svazcích, pozornost je věnována zejména výzkumu, který zkoumá studentský korpus tvořený diplomovými pracemi českých studentů. Tato studie však používá jinou metodologii k získání a zkoumání lexikálních svazků, než která je použita v této diplomové práci.

Praktická část je zaměřena zejména na vytváření obou korpusů a na metody, které byly použity pro získání dat a jejich třídění. Značný prostor je poté věnován samotné analýze dat a jejich diskuzi.

S pomocí softwaru Sketch Engine byl vytvořen Korpus vědeckých článků (Research Articles Corpus) a Korpus studentských diplomových prací (Students‘ Theses Corpus). Na základě Hylandovy studie (2008) byla stanovena dvě kritéria pro získání potřebných dat. První kritérium normalizované frekvence stanovilo, že minimální frekvence výskytu daného lexikálního svazku v korpusu je 20 na milion slov. Druhé kritérium je stanoveno jako minimální šíře užití v daném korpusu, tzn., že daný lexikální svazek se v korpusu musí objevit alespoň v pěti různých textech. Na základě takto stanovených kritérií bylo získáno 74 lexikálních svazků z Korpusu vědeckých článků a 91 lexikálních svazků z Korpusu studentských diplomových prací, které byly dále podrobeny analýze.

Z celkové analýzy dat se ukázalo, že z hlediska struktury jak studenti, tak profesionálové používají lexikální svazky typické pro akademickou prózu. 67 % lexikálních svazků v Korpusu vědeckých článků a 61 % lexikálních svazků v Korpusu studentských diplomových prací bylo tvořeno strukturami s předložkou nebo podstatným jménem Lexikální svazky obsahující slovesný komponent představují 27 % v Korpusu vědeckých článků a 30 % v Korpusu studentských diplomových prací.

Nejčastějším lexikálním svazkem s nejvyššími normalizovanými frekvencemi je v obou korpusech konektor *on the ohter hand* a *On the other hand* (na začátku věty). Ukazuje se, že *On the other hand* studenti používají mnohem více než profesionálové.

Přestože jsou oba korpusy, co se týče strukturálního zastoupení, vyrovnané, ukázalo se, že konkrétní případy lexikálních svazků se v korpusech liší. Bylo zjištěno, že v prvních 50 lexikálních svazcích se v obou korpusech shoduje pouze 21 z nich. Studenti tedy ve svých diplomových pracích používají jiné lexikální svazky, které se v kontrolním korpusu (Korpus vědeckých článků) nenacházejí. Jako příklad jsou uvedeny lexikální svazky obsahující spojení *in order to,* dále pak *when it comes to, to find out whether* atd. Dále se ukazuje, že studenti velmi často odkazují na svou práci přímo v textu a vytvářejí tak logické vazby mezi tím, co už bylo řečeno a co bude následovat. Například: *part of the thesis, the practical part of, part of this thesis, the first part of.*

Analýza dat ukázala, že struktury lexikálních svazků a jejich rozložení v korpusu, které studenti používají, odpovídá strukturám, které se obvykle nacházejí v akademické próze. Jednotlivé případy lexikálních svazků nalezených v Korpusu studentských diplomových prací už ale neodpovídají těm v Korpusu vědeckých článků. Toto zjištění potvrzuje původní předpoklad, že se zde setkáváme s odlišnými žánry. Diplomové práce mají jiné zaměření, cíle i publikum než publikované vědecké články, repertoár lexikálních svazků se tedy liší.

Výsledky analýzy mohou mít přesah i do pedagogické oblasti, a to v podobě jasnějších pravidel pro používání lexikálních svazků v rámci univerzitních studií, tak aby jejich psaná produkce byla co nejvíce podobná konvencím akademické prózy. Co se týče dalšího výzkumu, nabízí se možnost prozkoumat vytvořené korpusy z hlediska funkčního užití lexikálních svazků a jejích analýzy.

8 Works cited

Meyer Charles F. 2002. English corpus linguistics: an introduction. Cambridge: Cambridge University Press.

Aston, Guy, Silvia Bernardini and Dominic Stewart. 2004.”Introduction: Ten years of TaLC.” In Corpora and Language Learners, edited by Guy Aston, Silvia Bernardini and Dominic Stewart. Amsterdam: John Benjamins.

Biber, Douglas and Federica Barbieri. 2007. “Lexical bundles in university spoken and written registers.” English for Specific Purposes 26, 263–286.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. Longman Grammar of Spoken and Written English. Harlow: Pearson Education Limited.

Bley-Vroman, R. 1983. “The comparative fallacy in interlanguage studies: The case of systematicity.” Language Learning 33, 1–17.

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

Dontcheva-Navratilova, Olga. 2012. “Lexical Bundles in Academic Texts by Non-native Speakers.” Brno Studies in English, 38(2), 37-58. doi: 10.5817/BSE2012-2-3

Flowerdew Lynne. 2014. “Learner corpus research in EAP: some core issues and future pathways.” English Language and Linguistics 20(2), 43-60.

Granger, Sylviane. 2002. “A bird’s-eye view of learner corpus research.” In Computer learner corpora, second language acquisition and foreign language teaching, edited by Sylviane Granger, Joseph Hung, and Stephanie Petch-Tyson, 3-33. Amsterdam: John Benjamins.

Granger, Sylviane. 2003. “The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research.” TESOL Quarterly 37(3), 538-546.

Granger, Sylviane. 2009. “The contribution of learner corpora to second language acquisition and foreign language teaching.” In Corpora and Language Teaching, edited by Karin Aijmer, 13-32. Amsterdam: John Benjamins.

Hyland, Ken. 2008. “Academic clusters: text patterning in published and postgraduate writing.” International Journal of Applied Linguistics 18(1), 41-62.

Chen, Yu-Hua and Paul Baker. 2010. “Lexical bundles in L1 and L2 academic writing.” Language Learning & Technology 14(2), 30-49.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. “The Sketch Engine: ten years on.” Lexicography 1, 7-36. doi: 10.1007/s40607-014-0009-9

McEnery, Tony and Andrew Hardie. 2012. Corpus Linguistics. Cambridge: Cambridge University Press.

Nesselhauf, Nadja. 2005. Collocations in a Learner Corpus. Amsterdam: John Benjamins.

Paquot, Magali and Sylviane Granger. 2012. “Formulaic Language in Learner Corpora.” Annual Review of Applied Linguistics. 32 130-149.

Pearson ELT. 1998-2008. “Longman Corpus Network”. Accessed August 16. http://www.pearsonlongman.com/dictionaries/corpus/index.html

Qin, Jingjing. 2014. “Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics.” System 42, 230-231.

Sketch Engine. 2015. Accessed August 16. https://www.sketchengine.co.uk/glossary-of-terms/

Tazegül, Assiye B. 2015. “Use, misuse and overuse of “on the other hand”: a corpus study comparing English of native speakers and learners.” International online Journal of Education and Teaching 2(2), 53-66. http://iojet.org/index.php/IOJET/article/view/70/109

Université catholique de Louvain. 2011. “Centre for English Corpus Linguistics – ICLE”. Accessed August 16. http://www.uclouvain.be/en-cecl-icle.html

Wray, Alison. 2002. Formulaic Language and the Lexicon. Cambridge: Cambridge University Press.

Wray,Alison. 2012. “What Do We (Think We) Know About Formulaic Language? An Evaluation of the Current State of Play.” Annual Review of Applied Linguistics 32, 231-254.

The examples used in this thesis come from these research articles and students' theses.

[JL.2014_11] – *Journal of Linguistics*, A Cognitive Grammar account of the semantics of the English Present Progressive.

[LT.2014_8] – *Language Teaching* – Home language and literacy practices among immigrant second-language learners.

[JL.2014_12] – *Journal of Linguistics*, Argument ellipsis in Colloquial Singapore English and the Anti-Agreement Hypothesis.

[ELL.2012_3] – *English Language and Linguistics*, Relative complexity in scientific discourse.

[ELL.2014_15] – *English Language and Linguistics,* The emergence of English refexive verbs - an analysis based on the Oxford English Dictionary.

[ELL.2013_8] – *English Language and Linguistics,* Subjectivity, indefiniteness and semantic change.

[JL.2014_14] – Journal of Linguistics, The regularity of the irregular verbs and nouns in English.

[LT.2015_14] – *Language Teaching* – SLA research and L2 pedagogy-Misapplications and questions of relevance.

[LINGV.Bc.2014_7] – British and American situational comedies.

[LINGV.Mgr.2013_2] – A Comparative Study of English, Czech and Slovak Weather Lore Sayings.

[LINGV.Mgr.2014_17] – Harry Potter teaches English – An English Summer Camp Design.

[LINGV.Bc.2013_2] – Analysis of Phrasal Verbs in pre-intermediate English Textbooks.

[LINGV.Bc.2014_25] – Countability of the Noun Experience.

[LINGV.Mgr.2014_19] – Language Advising of an English Teacher Helping Young Learners Develop their Learning Strategies and Facilitate the Process of Studying English Vocabulary.