UNIVERZITA PALACKÉHO V OLOMOUCI

Filozofická fakulta

Katedra anglistiky a amerikanistiky

LEXICAL ISSUES IN L2 ENGLISH

Bakalářská práce

2019

Hana Mifková

Lexical Issues in L2 English

(Bakalářská práce)

Autor: **Hana Mifková**

Studijní obor: Anglická filologie

Vedoucí práce: **Mgr. Michaela Martinková, Ph.D.**

Počet stran (podle čísel): 45

Počet znaků: 80 763

Olomouc 2019

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

V Olomouci dne 29. 4. 2019 …………………….

Děkuji Mgr. Michaele Martinkové, Ph.D. za její trpělivost a cenné rady, které mi poskytla při psaní mé bakalářské práce.

Abstract

This bachelor thesis is concerned with the role of learner corpora in the second language acquisition research and their exploitation for lexicological analysis of university students' academic writing. The purpose of the thesis is to compile a learner corpus containing authentic language data gathered from bachelor theses written by students of English Philology at Palacký University in Olomouc and Masaryk University in Brno, and a control corpus consisting of research articles written by professional linguists. The second main objective of the thesis is to compare the lexical diversity of learner texts with the texts written by professionals. It is hypothesized that learner texts would have a lower rate of lexical diversity than professionally written texts, because English is not their native language and they are less experienced in academic writing.

Key words

Corpus linguistics, learner corpora, second language acquisition, SLA, English for academic purposes, academic writing, lexical diversity, TTR, STTR

Anotace

Tato bakalářská práce se zabývá problematikou studentských korpusů v kontextu výzkumu osvojování druhého jazyka a jejich využití pro lexikologickou analýzu akademického psaného projevu vysokoškolských studentů. Účelem této práce je vytvořit studentský korpus obsahující autentická jazyková data získaná z bakalářských prací studentů oboru anglické filologie z Univerzity Palackého v Olomouci a Masarykovy univerzity v Brně a kontrolní korpus složený z vědeckých článků napsaných profesionálními lingvisty. Druhým cílem je zjistit lexikální diverzitu studentských textů ve srovnání s texty psanými profesionály. Předpokládá se, že studentské texty budou mít nižší hodnotu lexikální diverzity než profesionálně psané texty, protože angličtina není jejich mateřský jazyk a mají méně zkušeností s akademickým psaním.

Klíčová slova

Korpusová lingvistika, studentské korpusy, osvojování druhého jazyka, angličtina pro akademické účely, akademická próza, lexikální diverzita, TTR, STTR

Contents

Introduction

During the last decades, second language acquisition (SLA) has become an extremely popular field of study and it is being investigated within all linguistic disciplines, including lexicology. In recent years, a new methodology has been trying to find its place in SLA research, namely computer learner corpora (CLC). Learner corpora contain language data produced by second or foreign language learners and as such they provide researchers with an authentic insight into learner writing or speech. However, most learner corpora are designed by corpus linguists who are not sufficiently informed by second language acquisition research and SLA researchers themselves are somewhat hesitant in using learner corpora and traditionally tend to use experimental data instead (Lozano and Mendikoetxea 2013, 65). Granger (2004, 134) states that "the contribution of CLC research to SLA so far has been much more substantial in description than interpretation of SLA data." Systematically and carefully designed learner corpora could potentially contribute to further SLA research, since they provide larger datasets with the possibility for generalizations. Corpus linguists can compile corpora focused on various kinds of learner language and focus their research on aspects which are particularly difficult for language learners, e.g. English for academic purposes (EAP). EAP poses many challenges for EFL learners and learner corpora present a valuable resource of learner EAP data. Most EAP studies based on learner corpora investigate various aspects of the EAP specific phraseology. They were concerned with the underuse, overuse or misuse of specific linguistic aspects. My thesis focuses on lexical diversity in L2 academic writing. Previous lexical diversity studies dealt with other genres or students of other mother tongue backgrounds. In my research, I focus on determining lexical diversity of academic texts written by Czech students of English in comparison with English-writing professionals. The analysis is based on corpora compiled from bachelor theses and research articles. The main aim of the practical part is to answer these research questions: What is the type token ratio of each corpus? Is there any difference between the lexical diversity of L2 students' and professionals' texts? What does the lexical diversity level suggest about the texts? It is hypothesized that the level of lexical diversity will be higher in research articles, since students' language proficiency and academic writing skills are lower.

The first section of the thesis provides a brief overview of the development of learner corpora, their typology, description of their characteristics and the specific criteria required for their compilation, and the methods used to analyze corpora. The main focus of the section is placed on the role of learner corpora in the context of second language acquisition research.

The second section of the thesis is concerned with the acquisition of L2 vocabulary and English for academic purposes. The main focus of the second section is placed on the concept of lexical diversity – how it is defined, what factors influence it and how it is measured.

In the methodology section, I describe the process of collecting the data for compilation of two corpora. There are many variables affecting the language data and it is important to select the textual data carefully to ensure validity of the analysis. The first corpus consists of linguistic bachelor theses written by students of English Philology at Philosophical Faculties at Palacký University in Olomouc and Masaryk University in Brno. The second corpus consists of research articles from linguistic journals written by professional linguists.

In the next part of the practical section, I measure the type token ratio (TTR) and the standardized type token ratio (STTR) of the two sets of texts and discuss their lexical diversity. Finally, the conclusions are drawn.

1 Learner corpora

This section introduces corpus linguistics in general and briefly describes the development of learner corpora. Thereafter, the definition of learner corpora is explained, followed by their typologies and the overview of some notable examples of learner corpora. Moreover, the prevailing methods for learner corpora analysis are presented. The final subsection focuses on the place of learner corpora in the context of the second language acquisition research. It is concerned with the variables affecting learner language output and the advantages and disadvantages of learner corpus data in comparison with experimental data, which are traditionally used in SLA research.

1.1 Characteristics of learner corpora

According to McEnery and Hardie (2012, 1), corpus linguistics differs from other linguistic fields in that it does not study any particular aspects of language, but rather "focuses upon a set of procedures, or methods, for studying a language". Thanks to these procedures various aspects of language can be investigated, which results in corpus linguistics having an enormous impact on the current approach to the study of language. Moreover, with the emergence of corpus linguistics new insight was brought not only into previous research, causing possible redefinition of language theories, but also entirely new research questions, previously difficult to explore, could be proposed (McEnery and Hardie 2012, 1).

For a long time, corpus-based research focused only on native speakers' language data, but in the late 1980s, linguists started collecting learners' language data as well. Granger (2002, 7) defines learner corpora as follows: "Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardized and homogeneous way and documented as to their origin and provenance". The fact, that the texts are collected electronically makes it possible to create corpora containing hundreds of thousands or millions of words. The authenticity of data, however, is often problematic in linguistic research. When people are aware of being observed, their behavior may change, which questions the authenticity. Moreover, according to Granger (2002, 8), learner language data do not even have the same degree of authenticity as native language data, since “foreign language teaching context usually involves some degree of ‘artificiality’ and (…) learner data is therefore rarely fully natural”. An essay, for example, is considered a result of an authentic classroom activity, and the corpus compiled of such texts can be regarded as authentic language data.

A necessary part of learner corpus compilation is meeting the strict design criteria in order to control the many variables, which influence learners’ language output. Another important feature of a well-designed corpus is annotation, which is a process of adding interpretative linguistic information to a corpus (Granger 2004, 128). The information facilitates and complements the conducted research and it can be added automatically, semi-automatically or manually. According to Granger (2002, 16–19), two profitable and most often used types of annotation are the part-of-speech (POS) tagging and error-tagging; the latter was developed specifically for the purposes of learner language research. Annotation should be following established convention in order to be comparable with other annotated learner and native corpora. If not meeting any of the conditions mentioned in Granger’s definition, the collected set of texts cannot be regarded as a corpus.

1.2 Typology of learner corpora

According to Granger (2002, 7) learner corpora can be categorized with respect to four dichotomies.

Monolingual	Bilingual
General	Technical
Synchronic	Diachronic
Written	Spoken

Monolingual corpora contain data of only one language, while bilingual corpora contain data from two languages. There are also several multilingual corpora, which contain either multiple L1s or L2s or both.[1] According to the genre of the collected texts, a corpus can be either general or technical (e.g. English for

[1] See https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

Academic Purposes). With respect to the span of time the language data come from, a corpus can be either synchronic (cross-sectional), i.e. consisting of data gathered at a single point in time from learners at different stages of language development, or diachronic (longitudinal). Longitudinal studies cover long periods of time during which it is possible to capture development stages of learners. This is a big advantage and disadvantage at the same time. Since the data come from the same group of learners they quite precisely map their development. However, the compilation of longitudinal corpora is demanding both because of time and money and therefore such corpora are still quite rare. Finally, regarding the mode of language, a corpus can be either written or spoken. The types of learner corpora listed in the left column, i.e. monolingual, general, synchronic and written corpora are far more common than their counter-types, mainly because they are easier to compile. However, it would be rewarding to face the difficulties that come with the compilation of the currently lacking types of corpora. Many linguists, e.g. Granger (2004, 138), Myles (2005, 388) or Lozano and Mendikoetxea (2013, 89), suggest that SLA and learner corpus researchers should focus on gathering more spoken or longitudinal learner data, since they could contribute to the investigation of learner speech and the development of interlanguage, i.e. language system learners develop during acquisition of an additional language, which bears features of both the native language and the target language (Gass and Selinker 2001, 12).

Granger (2004, 129) also distinguishes commercial and academic learner corpora. There are currently only two commercial corpora, *Cambridge Learner Corpus* and *Longman Learners' Corpus*, which were established by ELT publishers for the purposes of compiling dictionaries and other teaching aids designed to suit the specific learner needs (Pravec 2002, 88). Commercial corpora tend to be very large and multilingual, i.e. covering multiple L1s. The majority of academic learner corpora is monolingual and considerably smaller in size (although there are exceptions in both aspects, e.g. *International Corpus of Learner English* has 16 subcorpora of different L1s, *Hong Kong University of Science and Technology Learner Corpus* contains 25 million words[2]). According to Myles (2005, 375), the size of academic learner corpora is limited mainly because of the poor funding of the field — gathering such large collections of learner data is quite expensive,

[2] See http://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

therefore researchers often restrict to compile a corpus limited only to a specific research question. The main aim of academic corpora is to describe all aspects of learner language and to help better understand the factors that influence the development of interlanguage. Additionally, similarly to the commercial corpora, they contribute to the innovation of the design of teaching materials and methods to cater for the specific learners' needs.

1.3 Overview of learner corpora

This subsection briefly describes the following learner corpora of English L2: *International Corpus of Learner English (ICLE)*, *Louvain International Database of Spoken English Interlanguage (LINDSEI)*, *Longman Learners' Corpus (LLC)*, *Cambridge Learner Corpus (CLC)*, and also mentions several learner corpora of languages other than English.

International Corpus of Learner English (ICLE) was launched in 1990 by Sylviane Granger at the University of Louvaine. It is an academic corpus compiled of (mostly) argumentative essays written by university students from non-English speaking countries. The first version of ICLE contained authentic written data produced by learners of 11 mother tongue backgrounds (Granger 2003, 540) and in the second version (ICLEv2, published in 2009) five L1s were added. To date, ICLE comprises of, namely, Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana and Turkish subcorpora. Currently, the ICLE team is working towards the third version of the corpus.[3] ICLE has two main research goals: first, "to collect dependable evidence on learners' errors and to compare them cross-linguistically in order to determine whether they are universal or language specific" and second, "to investigate aspects of 'foreign-soundingness' in non-native essays which are usually revealed by the overuse or underuse of words or structures with respect to the target language norm (Pravec 2002, 83)." Additionally, the founders of ICLE are "encouraging research into the potential applications of learner corpora to pedagogical materials and learning aids" (84).

At the same institution the first academic learner corpus of spoken language was established as well — the *Louvain International Database of Spoken English*

[3] See http://uclouvain.be/en/research-institutes/ilc/cecl/icle.html.

Interlanguage (LINDSEI). It contains oral data obtained by interviewing EFL students with eleven different mother tongue backgrounds.[4] Both LINDSEI and ICLE are accompanied by control native corpora for comparing learner language with native language — Louvain Corpus of Native English Essays (LOCNESS) for ICLE and Louvain Corpus of Native English Conversation (LOCNEC) for LINDSEI (McEnery and Hardie 2012, 82).

Longman Learners' Corpus (LLC) is a commercial learner corpus, which is part of the Longman Corpus Network (LCN). It contains 12 million words, but the whole LCN consists of 330 million words.[5] LLC is compiled of essays and exam scripts sent by students from all around[6] the world thus covering a wide range of L1s and proficiency levels. The results of the research conducted via LLC serve to provide better teaching and learning materials to teachers and students of ESL/EFL (Pravec 2002, 89). Specifically, LLC contributed to the design of several monolingual learner dictionaries (MLD). MLDs provide all information in the learners' target language, they offer more user-friendly definitions, focus on clear explanations of meanings and the syntactic, lexical and grammatical behavior of the words etc. (DeCock and Granger 2004, 72). Lexicographers explore the authentic learner data to determine the most frequent errors made by EFL students of various L1 backgrounds, and include them into the dictionary entry in the form of error notes, which provide learners with clear explanations of what causes the errors together with advice on how to avoid making them.

The largest learner corpus is the *Cambridge Learner Corpus* (CLC), which is part of the *Cambridge English Corpus* (CEC). CLC "currently contains over 50 million words taken from Cambridge exam scripts submitted by over 220,000 students from 173 countries."[7] A part of the corpus is error-annotated and the results of the learner error analyses serve to better design the English language teaching (ELT) tools, such as dictionaries and course books, which are specifically made to suit the needs of the selected group of students (Pravec 2002, 88). CLC is a commercial corpus available only for in-house use by authors working for

[4] See http://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html.

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

[6] In fact, there is a public call for teachers to send in their students' exam scripts and thus contribute to the constantly growing corpus. See http://www.pearsonlongman.com/dictionaries/corpus/learners.html.

[7] See https://www.englishprofile.org/cambridge-english-corpus.

Cambridge University Press, unlike LLC which is available for academic research as well.

The language data in the four above mentioned learner corpora were collected synchronically, i.e. at a single point in time. All of the described corpora are multilingual, i.e. containing multiple L1s and the target language (English) is the same for each of them, too. Initially, learner corpora contained data gathered mainly from EFL students, but in the recent years also other foreign languages are being investigated, including for example French (*French Interlanguage Database, French Learner Language Oral Corpora*), Spanish (*Corpus Escrito del Español L2*), Italian (*Lexicon of Spoken Italian by Foreigners*), German (*Fehlerannotiertes Lernerkorpus*), Croatian (*Croatian Learner Text Corpus*),[8] Czech (*Czech as a Second Language with Spelling, Grammar and Tags*)[9] and others, both written and spoken, with various L1 backgrounds.

1.4 Methods for analyzing learner corpora

There are two main methods for analyzing learner corpora: Computer-aided Error Analysis (CEA) and Contrastive Interlanguage Analysis (CIA).

CEA "focuses on errors in interlanguage and uses computer tools to tag, retrieve and analyze them" According to Granger (Granger 2002, 11), CEA is based on the traditional Error Analysis, popular in the 1970s, but it is distinct in one major aspect — the former EA examined decontextualized errors and did not consider the correct use of learner language, for which it is often criticized. In contrast, CEA focuses on the erroneous items within their context (both linguistic and situational), alongside the correct instances. Moreover, there are well-defined error categories, which are carefully documented. Granger (2002, 14) mentions two types of CEA analysis: the first method consists in selecting a linguistic feature, which is known to cause problems in learner use, and extracting all the misused data from the corpus: This method is quite fast but limited only to the issues already previously considered problematic. The second method requires creating a standardized system of error tags according to which all the errors in the corpus are tagged (or at least in a particular category, e.g. verb complementation etc.). The error-tagging

[8] See http://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

[9] http://ucnk.korpus.cz/czesl-sgt.php.

and the following analysis of the results are highly time-consuming, which is probably one of the reasons why there are less CEA based studies than studies using the CIA method (Granger 2004, 133). However, this method may uncover previously unknown learner errors and the results could be especially useful for designing teaching materials, such as textbooks or learner dictionaries.

CIA has its roots in the original Contrastive Analysis proposed by Lado. In his preface to *Linguistics across cultures* he claimed that "we can predict and describe the patterns which will cause difficulty in learning and those that will not cause difficulty" (as cited in Johansson 2003, 32). The similarities between languages account for positive transfer and the differences cause the negative transfer. The modern CIA is based on "quantitative and qualitative comparisons between native and non-native data or between varieties of non-native data" (Granger 2002, 12). The comparison of learner data with native data is especially useful for determining the non-native features of learner language, such as the underuse and overuse of particular patterns in learners' output. When comparing learner and native corpora, the control native corpus has to be chosen carefully — the researchers have to decide what should be the dialect, medium, level of formality, level of proficiency of the native speakers (students or professional writers) etc. (Granger 2002, 12). The comparison of multiple non-native corpora helps to better understand which interlanguage features are developmental, i.e. shared by learners from different mother tongue backgrounds, and which are probably the results of L1-transfer, i.e. those that are typical only for a group of learners having the same mother tongue background.

1.5 Learner corpora in SLA research

In the early period, learner corpora were used mainly for lexicographical purposes. So far, learner corpus research altogether was concerned predominantly with description of the data instead of their interpretation (Granger 2004, 134). However, recently, researchers have been investigating the possible ways in which learner corpora could contribute to the second language acquisition research. SLA research studies all aspects (linguistic but also psychological, sociological etc.) of the process of learning an additional language (L2). According to Lozano and Mendikoetxea (2013, 65), "the main aim of second language acquisition (SLA) research is to build models of the underlying representations of learners at a particular stage in the process of L2 acquisition and of the developmental constraints that limit L2 production." So far, SLA research has been rather hypothesis-building – researchers were conducting detailed studies involving only small numbers of respondents, but with the help of learner corpora they can test the original hypotheses on larger datasets and determine whether the findings can be generalized or not (Myles 2005, 376). The practice was similar in the first language acquisition research. Initially, first language researchers invented hypotheses which were then subjected to analyses using The Child Language Data Exchange System (CHILDES). Arguably, the emergence of learner corpora provided the missing link between SLA research and corpus linguistics (Granger 2002, 3). Nevertheless, as opposed to first language acquisition researchers who made huge progress in their domain thanks to corpus data, SLA researchers are usually reluctant to use learner corpora.

SLA research largely depends on quality learner language data, since the learners' output is the single most important evidence for the mental representations and developmental processes, which influence language production (Myles 2005, 374). Traditionally, SLA research draws on introspection (e.g. diary studies) or elicitation of experimental data rather than investigating natural language use data (Lozano and Mendikoetxea, 2013). Natural language data used to be disregarded because they do not reveal the whole scope of learner language. Learners prefer to use the kind of language they are confident in and tend to avoid the aspects, which cause them difficulties. When researchers want to study some particular linguistic aspects, they design experimental elicitations focusing on the particular areas. Moreover, they can elicit also infrequent linguistic aspects. Additionally, "presence of a particular structure/feature in learners' natural output does not necessarily indicate that the learners know the structure, and absence of a particular structure/feature in natural language use data does not necessarily indicate that learners do not know the structure" (Lozano and Mendikoetxea 2013, 67). However, as Granger (1998, 5) pointed out, "the artificiality of an experimental language situation may lead learners to produce language which differs widely from the type of language they would use naturally". Another problem with experimental elicitation is small dataset. It would be unbearably exhausting to study large number of respondents experimentally. The narrow elicited dataset provides valuable insights in the interlanguage development of individuals or small groups, especially if collected longitudinally, but it also prevents the generalizability of the results. In contrast, "[learner] corpora are usually quite large and therefore give researchers a much wider empirical basis than has ever been available before" (Gilquin et al. 2007, 322). The large dataset enables researchers to focus also on the infrequent linguistic aspects in the context of natural language use. Furthermore, "[learner corpus data] can be submitted to a wide range of automated methods and tools which make annotations (e.g., morpho-syntactic tagging, discourse tagging, and error tagging) and to manipulate them in various ways in order to uncover their distinctive lexico-grammatical and stylistic signatures" (322). In terms of representativeness and generalizability, natural language data stored in million-word learner corpora are definitely more suitable than the experimental data and the previously used natural language use data. However, controlling of the variables influencing learners in a non-experimental environment is difficult (Granger 2002, 6). Granger (2004, 125) warns that the "failure to control these factors greatly limits the reliability of findings in learner language research". Many learner corpus linguists strive to meet the strict design criteria to control the variables, yet it is not a matter of course for all of the existing public domain corpora and as Granger (126) admits, "there are so many variables that influence learner output that one cannot realistically expect ready-made learner corpora to contain all the variables" and the only way to ensure that the corpus contains all the relevant design parameters is preliminary theoretical analysis.

The variability of learner data is caused by many linguistic, situational and psycholinguistic factors (Granger 2004, 125). Some of the variables are typical also for native corpora, but some are L2-specific. While compiling a corpus, all of the variables have to be included in order to produce a reliable source of learner data. Thanks to the elaborated system of variables, researches are then able to build a subcorpus according to chosen criteria and investigate various linguistic phenomena based on very concrete data selection. The basic variables include gender and age of the writers/speakers, and genre, length, medium, field and topic of the produced texts. These two groups of variables are labeled as learner variables and task variables, respectively. Learner corpora should contain information about learners' native language (alternatively also mother's and father's native language), geographical provenance, L2 proficiency (determined by self-rating, according to the Common European Framework of Reference for Languages or by filling a standardized placement test), type of school and field of study, L2 exposure (time spent in L2 speaking country, age of first immersion in the L2) or ability to speak other foreign languages. Regarding the task, the information about the type of exam, its timing and use of reference tools has to be included (Granger 2004, 126). While designing a corpus of L2 Spanish (CEDEL2), students were asked to provide information also about the place where the text was written (school, home or both) and whether they conducted prior research regarding the topic (via TV, internet etc.) (Lozano and Mendikoetxea 2013, 84). All of the mentioned factors are indispensable for correct interpretation of learner data. For example, it is presumable that a learner who has a lower level of proficiency and has never been to an L2-speaking country might be more prone to making mistakes than a learner who spent a few years in an environment where the foreign language was spoken daily. The knowledge of learners' mother tongue is necessary for investigating the difficulties which learners with the same mother tongue background face and determining which mistakes are the results of transfer from learners' mother tongue and which are shared with learners of different L1s. The task variables are very useful as they inform researchers about the amount of time spent on the task or whether learners had dictionaries (monolingual or bilingual) and other helping tools at their disposal, which too affect the production of the texts.

However, in addition to the already mentioned learner variables, traditional SLA research focuses also on other important factors influencing the second language acquisition. These include motivation, aptitude, learning strategies, personality factors or anxiety. All of these have major impact on learners and determine whether their language learning is successful or not. The strongest predictor of language-learning success is aptitude, followed by motivation (Gass and Selinker 2001, 349). Aptitude denotes students' potential to learn something new. Gass and Selinker refer to Gardner's work, who distinguishes two types of motivation – integrative and instrumental, which describe the need to learn a L2 to be able to communicate with the target language community and the need to study language as a means to achieve some other goal, respectively. Anxiety can make second language learning severely harder, as some learners suffer from social anxiety, which may cause their restraint and reluctance to improve themselves in speaking, others are anxious when taking tests, which then reflects in their results. The ability to learn a L2 with success can be predicted also by some personality factors, such as introversion and extroversion or willingness to take risks etc. It is very important to list as much information about learners as possible to ensure the interpretation of data is accurate and further applicable, however, documenting these variables within large learner corpora would probably be very problematic. During my research I did not come across any learner corpus which would take into account also these SLA variables. According to Tono (2003, 806), "learner corpus researchers should exchange ideas with SLA researchers in a more structured and systematic way. Many corpus-based researchers do not know enough about the theoretical background of SLA research to communicate with them effectively, while SLA researchers typically know little about what corpora can do for them." It is important that learner corpora were designed according to the second language acquisition theories and enable SLA researchers to test their hypotheses.

2 Lexical issues in learner English

The following section discusses acquiring L2 vocabulary and then focuses on English for Academic Purposes. Subsequently, it introduces the concept of lexical diversity, its definition and tools for measuring it, especially the type token ratio.

2.1 Acquiring vocabulary

Vocabulary is an important element of second language knowledge and its acquisition is actually never finished (in both first and second language). The lexical system is quite unstable as it undergoes constant changes with new words emerging, other words becoming obsolete or their senses narrowing, widening or transforming – even native speakers' mental lexicon undergoes many changes during their lives (Pietilä et al. 2015, 1). It is quite challenging to acquire a native-like knowledge of second language vocabulary – not only because learners have to learn what the words mean and possibly adapt to the gradual changing of these meanings, but also they have to learn how to use the words, i.e. they have to acquire the collocations, colligations, written vs. spoken forms, synonyms, antonyms etc. According to

Pietilä et al. (2015, 2), "the vocabulary of a language is sensitive to a wide range of co-textual and contextual considerations" – it has to meet various requirements regarding grammar, register, style, mode etc. For instance, features typical for spoken language are greater repetition, redundancy and inefficient vocabulary use, whereas written language is characterized by more diverse vocabulary and greater syntactic complexity. Not meeting the requirements inevitably results in unnatural output, such as using contractions and informal language or other notions typical for speech in academic writing. Sadeghi and Dilmaghani (2013, 328) refer to several second language studies, which "have indicated lack of vocabulary is what makes writing in a foreign language difficult, and that vocabulary proficiency is probably the best indicator of the overall text quality". Doró (2015, 57) makes the same point: "effective vocabulary use is an important indicator of quality writing and also makes a strong impression on the reader." For each discourse different linguistic features are typical, which present new challenges for learners. Academic written discourse is especially challenging.

2.2 English for academic purposes

English for academic purposes (EAP) is necessary not only for all researchers of any discipline who publish their papers but for many students as well. All students of higher education in English speaking countries and students in non-English speaking countries whose courses are taught in English are required to write essays and theses in academic English. However, English for academic purposes is highly conventionalized and thus it causes problems even for novice native writers and for second language learners it constitutes an especially great challenge (Gilquin et al. 2007, 321).

Chafe and Danielewicz (1987, 23-24) summarize the characteristics of academic writing as follows:

> Academic writers represent for us the extremes of what writing permits. Their vocabulary is maximally varied, and they avoid both hedges and inexplicit references. Their writing is maximally literary, with almost no colloquial items or contractions. Their intonation units are maximally long (…). Their sentences (…) are maximally coherent. They show little involvement with themselves, or with concrete reality (…). This kind of

> language represents a maximum adaptation to the deliberateness and detachment of the writing environment.

The previous corpus-based studies of EAP provided researchers with detailed descriptions of specific EAP phraseology, which is characteristic for semantically and syntactically compositional word combinations, e.g. the aim of this study, the extent to which, it has been suggested etc., and other specific aspects, such as frequent use of nouns and linking adverbials (Gilquin et al. 2007, 321). EAP pedagogy (when based on corpora at all) has mostly drawn on native corpora rather than learner corpora. However, learner corpora should prove particularly useful for designing EAP teaching materials. According to Mazgutova and Kormos (2015, 3), investigating L2 learners' academic writing skills has mostly focused on cohesion, coherence and organization, only recently have researchers started analyzing also the linguistic features of students' writing, their improvement and development of proficiency. Mazgutova and Kormos's study is concerned with the lexical and syntactic development in L2 academic writing. My research focuses on the lexical issues in learner academic writing, particularly on the lexical diversity of Czech students' EAP texts in comparison with academic papers written by professional linguists.

2.3 Lexical diversity

According to Crossley et al. (2011, 182), "lexical proficiency comprises breadth of knowledge features (i.e., how many words a learner knows), depth of knowledge features (i.e., how well a learner knows a word), and access to core lexical items (i.e., how quickly words can be retrieved or processed)." In their study of the relationship between human judgment of lexical proficiency and lexical indices, "lexical diversity was the most predictive index" (190).

There are different approaches to definition of lexical diversity and its relationship with the concept of lexical richness. Some researchers consider lexical diversity and lexical richness to be synonymous, while other researchers treat it as two separate concepts (Wang 2004, 66). For the purposes of this thesis I embraced the latter approach, since it is the more recent one. According to Jarvis (2013, 15), originally, lexical richness denoted the wealth of words in one's mental lexicon, but recently it became an umbrella term for all lexical measures thus including not only breadth but also the depth of vocabulary knowledge – "the current meaning of lexical richness thus applies broadly to everything from lexical diversity through lexical sophistication (…) to lexical density (…), and beyond". Laufer and Nation (1995) mention also lexical originality, and explain the calculation of these measures of lexical richness. Lexical originality is calculated by multiplying the number of tokens unique to one writer by hundred and dividing the result by the total number of tokens. It measures the learner's performance in relation to the group in which the text was written; the group factor, however, decreases its reliability. Lexical density is calculated by multiplying the number of lexical tokens by hundred and dividing the result by the total number of tokens. It is the percentage of lexical words in a given text – the more lexical words the text has, the denser it is considered to be. Lexical sophistication is calculated by multiplying the number of advanced tokens by hundred and dividing the result by the total number of lexical tokens. It focuses on the advanced words in the text. The definition of advanced words is influenced by the level of language proficiency, the educational system and the researchers themselves, so in consequence, the definition may differ significantly, which in result causes the instability of this measure. Finally, there is lexical variation, or lexical diversity.

Lexical diversity is sometimes used interchangeably with the terms lexical variation (Laufer and Nation, 1995), lexical variability (Mazgutova and Kormos, 2015) or vocabulary richness (Kubát and Milička, 2013) etc. Lexical diversity includes the breadth of vocabulary knowledge, i.e. the range of words used in a text. "[It] describes the quality of vocabulary content of the learner's output. (…) Higher lexical diversity is generally considered to indicate more advanced proficiency than lower lexical diversity" (Sadeghi and Dilmaghani 2013, 328).

There are many factors affecting lexical diversity of a text. One of the major differences in lexical diversity levels is caused by the mode of language, i.e. whether the text is spoken or written. Generally, written texts have greater lexical diversity, because writers have more time to think about their lexical choices or change them afterwards, in speaking there is no such possibility, the time to think is limited and the speaker has to use the first word that comes to mind. Lexical diversity in writing can be further influenced by "familiarity with the topic, skill in writing and communicative purpose" Laufer and Nation (1995, 308). When people have to write about a topic which they are unfamiliar with they are likely to use a limited range of words and vice versa. Similarly, lexical diversity would be lower for people with poor writing skills and higher for experienced writers. Therefore, it can be expected that research articles analyzed in this thesis will be more lexically diverse than bachelor theses, since professional linguists write academic papers regularly, as it is part of their job. In contrast, undergraduate students do submit a few essays over the course of their studies but bachelor thesis is their first academic work of such an extent. Jarvis (2002, 75) adds that, lexical diversity of learners' texts can be affected by their mother tongue, L1–L2 proximity, L2 proficiency or age. The use of dictionaries or thesaurus can, too, highly affect the level of lexical diversity in learner texts. Doró (2015) mentions two studies, which compared essays written at school, i.e. controlled environment, and at home, i.e. with the access to teaching aids. "[Muncie] rightly pointed out that a lexically richer final essay may not reflect students' lexical development, but rather their access to various types of aids" (Doró 2015, 61). Additionally, also genre can affect the range of words in a text. Chafe and Danielewicz carried out a comprehensive study of four different genres, two spoken (conversations and lectures) and two written (letters and academic papers), in American academics. They studied various aspects of these genres, including also lexical diversity, which was found highest in academic writing (1987, 5). Their findings could imply that high lexical diversity levels can be expected also in the analyzed texts in the present thesis. However, they compare academic papers only with letters, admitting that they are a problematic and least homogeneous type of texts observed in their research. There are other studies, which focused on the comparison of lexical diversity in multiple genres. Sadeghi and Dilmaghani (2013) performed a research on the relationship of lexical diversity and genre in Iranian EFL students' texts and measured lexical diversity in texts written in argumentative, comparative and narrative genres. Johansson's (2008) conducted a developmental research on lexical diversity and lexical density of expository and narrative genres, both written and spoken, produced by four age groups (ten-, thirteen- and seventeen-year-olds and adults) of Swedish L1 speakers. Both studies found highest lexical diversity in written narratives. However, Kubát and Milička's paper on vocabulary richness in Čapek's writing did not determine significant differences between any of the genres with the exception of children's literature, differing from four of six other genres, which is understandable considering texts for children need to be written more simply for easier readability. The other genre, which scored TTR rate more similar to children's literature, was scientific texts. This would suggest, in contrast to Chafe and Danielewicz, that academic writing is not very lexically diverse. Nevertheless, there might be also differences between individual languages.

The most common way to measure lexical diversity is the type token ratio. The type token ratio is calculated as the number of types divided by the number of tokens (alternatively multiplied by a hundred for a result in percentage). "A token is any instance of a particular wordform in a text" and a "type is a particular, unique wordform" (McEnery and Hardie 2012, 50). In other words, the number of the different words is divided by the total number of the words in a text. Further explanation of the distinction between type and token is in Jarvis (2013, 15).

$$type\ token\ ratio = \frac{number\ of\ types}{number\ of\ tokens} \text{ or } \frac{number\ of\ types \times 100}{number\ of\ tokens}$$

The closer the ratio is to 1 (or 100 per cent when multiplied), the more varied the vocabulary is (McEnery and Hardie 2012, 50). Nevertheless, the degree of lexical diversity can never be exactly 1 or too close to 1, since there exist many linguistic items which naturally occur repeatedly in every text, for instance articles, auxiliaries, prepositions etc.

The type token ratio is a useful and popular measure, but there are some constraints placed on its reliability. "[It] has been shown to be unstable for short texts and [it] can be affected by differences in text length" (Laufer and Nation 1995, 310). The reason for the latter is that longer texts tend to have a lower value of type token ratio, because the longer a text is, the more repeating words occur. Therefore, comparing texts of different lengths produces unreliable results. This sensitivity to variance in text length is a persistent obstacle, to which various linguists have proposed several solutions. The first offered alternative was the mean segmental type token ratio, or MSTTR, "which involves splitting a text into several equally-sized segments, and using the mean TTR across all segments as the text's overall index of lexical variability" (Jarvis 2013, 16). Some linguists, for example Březina (2018), use the term STTR or standardized type token ratio for this measure. Březina mentions also the moving average type token ratio, or MATTR, which is similar to STTR (MSTTR) in that it divides the texts into equally-sized segments, “however, instead of dividing the text into successive non-overlapping segments, MATTR uses an overlapping window smoothly moving through the text” (Březina 2018, 58). Researchers have developed also other more sophisticated software tools, such as D or MTLD, whose use is affected by text length only to a “small or negligible degree” (Koizumi and In’nami 2012, 556).

3 Methodology

The central aim of the practical part is to determine and compare the lexical diversity of L2 student and professional academic writing. The research is based on language data contained in a L2 student corpus and a L1 corpus. First, I discuss the corpora compiled by Anna Boková, a former student of English Philology at Palacký University, and evaluate their possible utilization for my own research. Second, I describe the process of gathering the data and building of a new learner corpus compiled of Bachelor theses written by Czech students of English Philology, and a control corpus compiled of research articles written by professional linguists. Next, I describe the precautions, which had to be made in order to obtain a valid measurement of the type token ratio and the possible methodological problems.

3.1 Design criteria

In 2015, Anna Boková created two corpora for her master thesis entitled *Building and Exploring a Corpus of Academic Writing by Czech Students of English*. The *Research Articles Corpus* consists of research articles written by professional linguists. These texts were downloaded from three linguistic journals. The *Students’ Theses Corpus* is compiled of students’ bachelor and master theses written by Czech students of English Philology at Palacký University in Olomouc and Masaryk University in Brno. The fact that there are texts by students of both bachelor and master studies could pose a methodological problem in my lexical diversity research, since their language proficiency may differ.

Another problem lies in the comparability of text length. Figure 1 shows the total number of words and texts in both corpora. It is evident that the comparable size, i.e. total number of words, of the two corpora was achieved through different number of texts in each corpus, which suggests that the lengths of the texts in the two corpora are not of similar lengths. Students' texts are significantly longer than journal articles, which is why Boková had to download more texts into the second corpus. Moreover, figure 2 and 3 show that there are differences between the lengths of individual texts within each corpus, which are further illustrated by the graph in figure 4. The graph was created via the graph tool at Lancaster Stats Tools online (Březina, 2018). [10]

	texts	**tokens**
Students' Theses Corpus	31	553,005
Research Articles Corpus	50	534,155

Figure 1: Summary of Boková's corpora.

text ID	**tokens**	**text ID**	**tokens**
text 1	9,776	**text 16**	15,104
text 2	12,614	**text 17**	28,780
text 3	10,583	**text 18**	35,878
text 4	14,135	**text 19**	18,868
text 5	10,528	**text 20**	25,770
text 6	14,781	**text 21**	22,121
text 7	19,152	**text 22**	25,078
text 8	10,024	**text 23**	21,684
text 9	9,330	**text 24**	27,052
text 10	14,745	**text 25**	29,140
text 11	7,416	**text 26**	16,631
text 12	9,940	**text 27**	26,534
text 13	11,271	**text 28**	23,749
text 14	10,947	**text 29**	35,043
text 15	20,197	**text 30**	20,809
		text 31	11,296

Figure 2: Lengths of texts in Boková's Students' Theses Corpus.

text ID	**tokens**	**text ID**	**tokens**
text 1	9,314	**text 26**	13,187
text 2	9,248	**text 27**	16,066
text 3	10,156	**text 28**	13,906
text 4	7,905	**text 29**	11,438
text 5	11,464	**text 30**	20,992
text 6	6,108	**text 31**	2,876
text 7	10,435	**text 32**	10,126
text 8	9,124	**text 33**	17,032
text 9	7,677	**text 34**	4,672
text 10	17,618	**text 35**	13,236
text 11	8,732	**text 36**	7,876
text 12	9,692	**text 37**	6,423
text 13	6,074	**text 38**	7,004
text 14	9,540	**text 39**	15,984
text 15	9,372	**text 40**	14,137
text 16	12,823	**text 41**	8,556
text 17	9,381	**text 42**	8,579
text 18	15,596	**text 43**	7,190
text 19	17,492	**text 44**	15,049
text 20	8,480	**text 45**	5,900
text 21	16,106	**text 46**	9,368
text 22	13,202	**text 47**	8,194
text 23	17,595	**text 48**	7,648
text 24	16,138	**text 49**	9,465
text 25	12,623	**text 50**	6,596

Figure 3: Lengths of texts in Boková's Research Articles Corpus.

[10] http://corpora.lancs.ac.uk/stats/toolbox.php

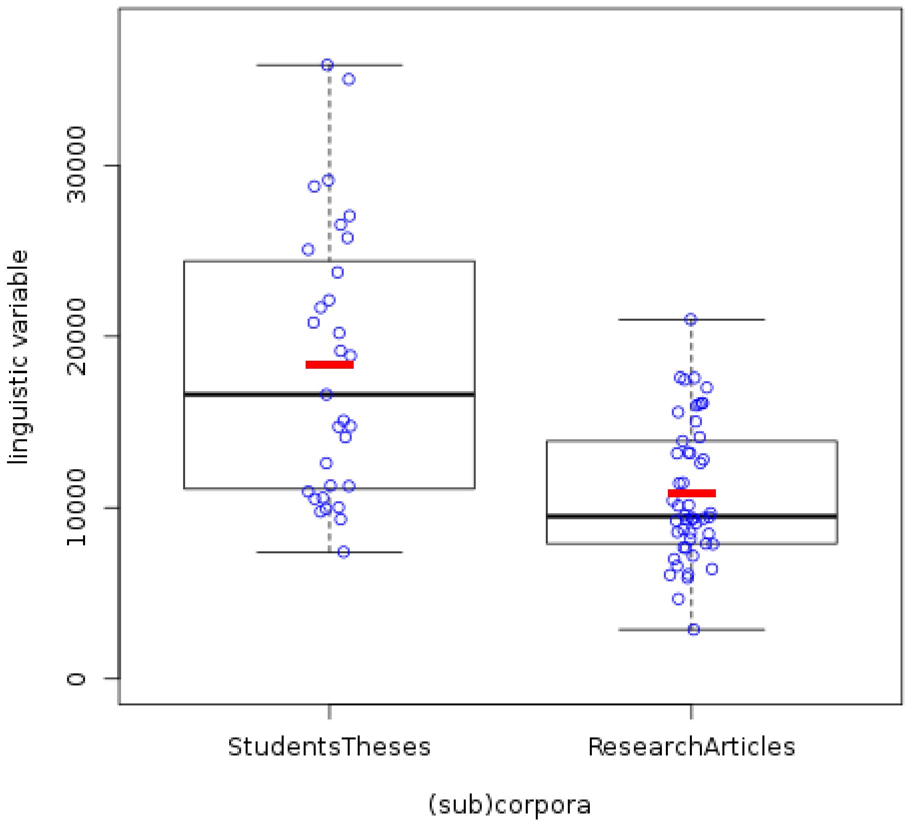


Figure 4: The comparability of text sizes of Boková's corpora.

These circumstances did not interfere with Boková's research on lexical bundles, but it prevents me from using her corpora for the lexical diversity research. As was already mentioned, the type token ratio is sensitive to variance in text length. According to Doró (2015, 61), "if texts are standardized in length or are very close in their number of tokens, the TTR is (…) still a useful and quick alternative." Laufer and Nation (1995, 310) also suggest to overcome this problem with only including texts of equal lengths. Therefore, I decided to build a smaller corpus of equally sized texts and aimed on including only 20—25 texts (as opposed to the 50 texts with more than half a million words in Boková's corpora). This precaution allowed me to meet the strict design criteria regarding the text length and thus guarantee more counterbalanced sets of texts. I downloaded more than a hundred of texts and then I selected those that were distinctly shorter or longer than the average and deleted them. I was looking for texts consisting of eight or nine thousand words and I tried to keep the difference between the shortest and longest texts around 2,000 words. The resulting TTR of an eight- or nine-thousand-word text can be low, but that does not have to mean that the text is of low quality. Additionally, I decided to support the analysis by also calculating the standardized type token ratio, which divides the analyzed text into equally long parts and measures the mean TTR of all the segments, thus making the resulting ratio more representative of authors' vocabulary range.

To make the comparison of the individual texts possible, it is necessary to control other variables too. Therefore, all of the students' textual data come from bachelor theses only. Master theses are usually much longer than bachelor theses and students' proficiency is higher too. Bachelor students share several features, e.g. the amount of time of L2 instruction (at least seven years, i.e. four years at high school and three years at university, but mostly more, since majority of students start learning English at elementary schools), they are at least 21 years old and their language proficiency should be C1. The majority of the students are Czechs, a few of them are Slovaks. However, no questionnaire was distributed to gather the information – this is just an assumption based on usual circumstances regarding undergraduate studies and there could be exceptions. Genre and topic are the same for all the texts. I decided to include only linguistic theses covering topics from phonology, morphology, stylistics, syntax, lexicology, corpus linguistics etc. All authors study English Philology at the Faculty of Arts at either Palacký University in Olomouc (17 theses) or Masaryk University in Brno (6 theses). The texts in the second corpus all come from linguistic journals and cover topics from linguistics too. The authors work at different universities all around the world. It may be assumed that the authors have native-like level of English language proficiency, since they are professionals of English linguistics and regularly publish in English. To avoid a misinterpretation of the data I used the Ethnea software[11], which can predict one's ethnicity. I typed in the names of every author and generated a prediction table, based on which I could determine whether a certain text was written by L1 or L2 professional.

[11] Available at http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py.

3.2 Compilation of the corpora

The corpora were created via Sketch Engine. Sketch Engine was established in 2004 by Adam Kilgarriff (Kilgarriff et al., 2004). It is used mainly by lexicographers working for dictionary publishing houses and by those involved in linguistic research, language teaching and translation. Sketch Engine contains 500 corpora in more than 90 languages.[12] Its users are allowed to perform linguistic research in the corpora and, which is most important for my thesis, to compile one's own corpora as well.

The texts for the Bachelor Theses corpus were downloaded from the Palacký University portal and the registry of theses at www.theses.cz. There are many formal requirements for Bachelor theses, including the Czech summary and annotation, acknowledgements, contents, front page, references and alternatively appendices or abbreviation overviews. All of these might later affect the results of research conducted in a corpus compiled from such texts, therefore they had to be deleted. Some of the theses were in doc/docx format, which can be easily edited, but majority of the texts were in pdf format, so I downloaded a free trial version of Adobe Acrobat Professional, which allows also editing of pdf files, and I removed all of the mentioned parts of texts in every thesis.

The texts for the Research Articles corpus were downloaded from the Cambridge Core website[13], which provides full texts of articles from journals on various subjects and to which students of Palacký University have free access through ezdroje.upol.cz. I selected the option "Browse by subject" and chose "Language and Linguistics". I used a few of the articles included in Boková's corpus which met the text length requirements, and then I downloaded some additional articles. Finally, I selected thirteen articles from the *English Language and Linguistics*, three from the *Journal of Linguistics* and seven from the *Studies in Second Language Acquisition*, which best met the length requirements. Like in the Bachelor Theses corpus, every document in Research Articles corpus had to be edited before uploading, however, the number of pages that had to be deleted was

[12] See https://www.sketchengine.eu/#blue.
[13] https://www.cambridge.org/core.

lower than in the previous set of texts – mostly it was only the bibliographies and sometimes appendices.

Figure 5 summarizes information about the two newly created corpora. It shows the number of texts, number of tokens and number of types in each corpus. Both the Bachelor Theses corpus and the Research Articles corpus are comprised of the same amount of texts, i.e. 23, and their total word counts differ only by 4,182 words. Similarly, the lengths of individual texts are comparable: the differences between the shortest and longest texts are 1,642 in Research Articles and 2,028 in Bachelor Theses. Figure 6 shows the numbers of words (tokens) in each text.

	Texts	**Tokens**	**Types**
Bachelor Theses	23	204,371	43,357
Research Articles	23	200,182	46,568

Figure 5: Summary of the corpora.

text_ID	**BT**	**RA**
text1	9,035	9,257
text2	8,501	9,459
text3	8,300	8,301
text4	7,972	9,121
text5	8,083	9,610
text6	8,820	8,562
text7	9,383	9,157
text8	9,507	8,312
text9	8,573	8,409
text10	8,045	8,885
text11	9,777	7,972
text12	9,486	8,774
text13	9,770	8,132
text14	9,715	8,723
text15	8,809	8,220
text16	8,732	8,228
text17	9,062	8,644
text18	9,511	9,098
text19	7,749	8,429
text20	9,457	7,968
text21	9,369	8,831
text22	7,936	9,292
text23	8,779	8,798

Figure 6: The lengths of texts in Bachelor Theses corpus and Research Articles corpus.

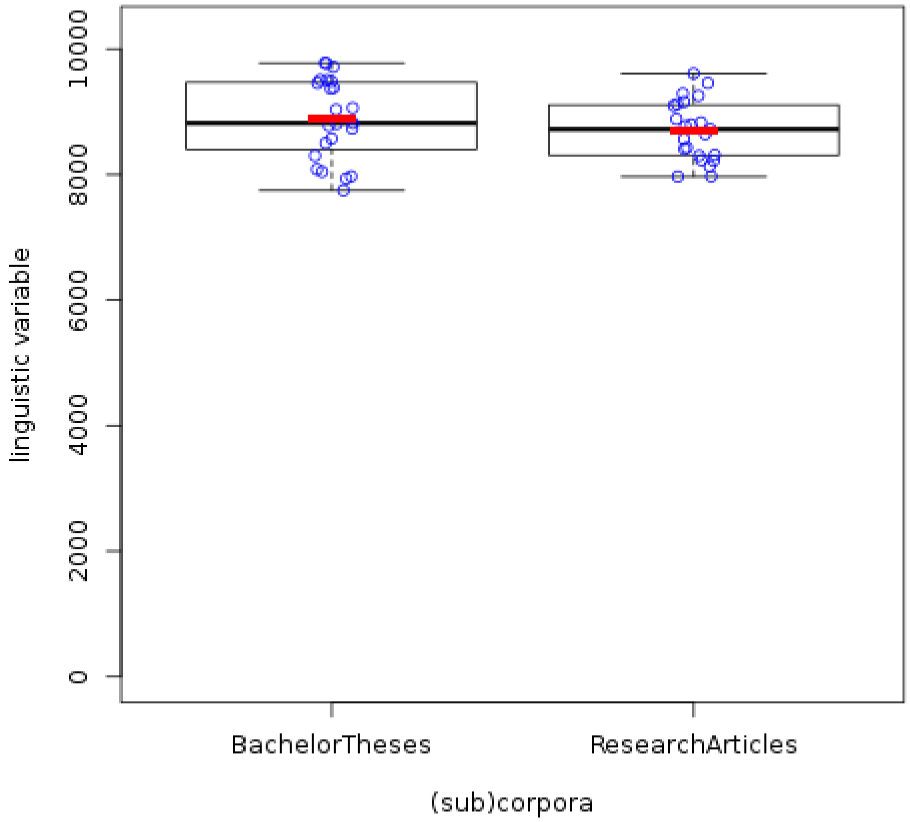


Figure 7: The comparability of the sizes of Bachelor Theses and Research Articles.

There are still differences between the individual texts, but to collect 46 samples with exactly the same number of words would not be possible. Figure 7 illustrates the comparability of the sizes of the two corpora. According to the graph, the differences between individual texts should not restrict the TTR measurement.

3.3 Measuring lexical diversity

First, lexical diversity of the two corpora will be measured by the type token ratio, i.e. by dividing the number of types by the number of tokens.

The number of types in Bachelor Theses corpus is 204,371 and the number of tokens is 43,357. The resulting mean type token ratio of the corpus is 0.21.

$$\frac{204371}{43357} = 0.21$$

The number of types in Research Articles corpus is 200,182 and the number of tokens is 46,568. The resulting mean type token ratio of the corpus is 0.23.

$$\frac{200182}{46568} = 0.23$$

The results show a higher level of lexical diversity in professional writing than in L2 student academic writing, as was hypothesized. However, the ratios are both lower than what could be expected in academic writing. This is caused by the length of the texts, which is around eight or nine thousand words on average. Naturally, the longer the text is the more repeating words occur. Therefore, I decided to find out the STTR values as well, and to do this for each text separately. Note again that TTR measures the ratio for the text as a whole, whereas STTR tool divides the texts into several sections of equal sizes and thus it avoids the instabilities caused by too long texts and length variance. For determining the STTR values I used the Lancaster Stats Tools online (Březina, 2018).[14] The TTR normalization basis was set on 1000. I uploaded and measured each text individually and included the result in the table below (figure 8).

text_ID	BT	RA
text1	0.38	0.37
text2	0.42	0.39
text3	0.36	0.44
text4	0.38	0.4
text5	0.32	0.41
text6	0.33	0.4
text7	0.31	0.4
text8	0.39	0.38
text9	0.38	0.43
text10	0.39	0.41
text11	0.36	0.4
text12	0.34	0.41
text13	0.4	0.41
text14	0.41	0.42
text15	0.31	0.38
text16	0.41	0.43
text17	0.34	0.36
text18	0.35	0.37
text19	0.36	0.41
text20	0.33	0.4
text21	0.37	0.39
text22	0.39	0.36
text23	0.45	0.4

Figure 8: STTR values of individual texts in Bachelor Theses corpus and Research articles corpus.

[14] http://corpora.lancs.ac.uk/stats/toolbox.php

The resulting mean STTR for the Bachelor Theses corpus is 0.36 and for the Research Articles corpus it is 0.39. The STTR values are higher than TTR values, but they are in accordance with the previous results, i.e. they show higher lexical diversity in journal articles.

4 Data visualisation

Using a graph tool at Lancaster Stats Tools online I created two graphs, which show the individual values of both TTR (figure 9) and STTR (figure 10) for the two sets of texts. Both graphs reflect the mean values, which unanimously demonstrate a greater lexical diversity for research articles. However, it also shows major differences between individual students. TTR of learner texts range from 0.15 to 0.28 and the blue dots are unevenly spread out across the graph. The linguists' TTRs are laid out more evenly, with the exception of two texts, which are distinctly higher. In the STTR graph, the differences between individual students are still visible, but they are not as extensive. Also, the two previously distinctive L1 texts are depicted lower. This is probably caused by the standardization of the texts, which takes place when using the STTR tool. The standardization also reflects in the higher STTR values when compared to TTR values. TTR rates are lower, because they consider the total number of words in the whole text. Regarding the exact degree of lexical diversity of a text, STTR is a more suitable tool than TTR, since thanks to the segmental measurement it is more representative of writer's breadth of vocabulary knowledge. TTR proved to be less reliable, since repetition of words in long texts is inevitable and lowers the resulting ratio of different words. However, when considering the question of which group of texts shows greater lexical diversity, either type of measuring is possible and reliable. Both TTR and STTR demonstrated higher lexical diversity in journal articles than in bachelor theses.

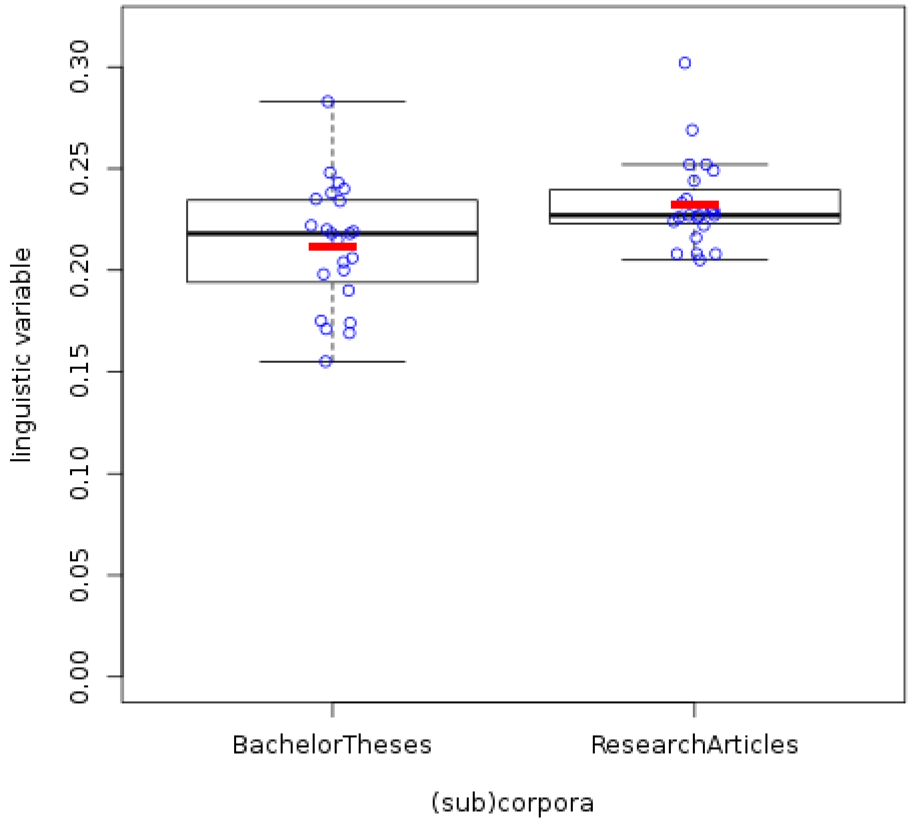


Figure 9: TTR of Bachelor Theses corpus and Research Articles corpus.

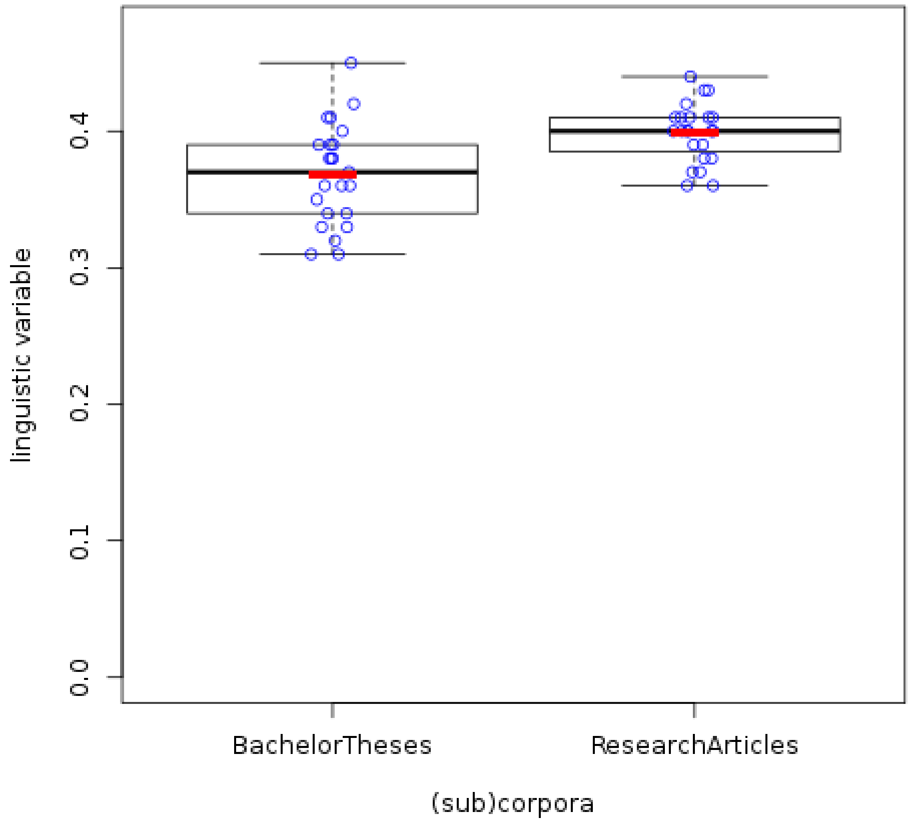


Figure 10: STTR of Bachelor Theses corpus and Research Articles corpus.

Figures 11 and 12 show the range of measured TTR and STTR values for each corpus and more clearly demonstrate the differences between lexical diversity levels of the two corpora. When the two vertical lines do not overlap, it means that the difference is significant. Figure 11 shows that the differences between TTR values of the two corpora are not significant, since the lines do overlap. However, figure 12 shows significant differences between lexical diversity of the two sets of texts.

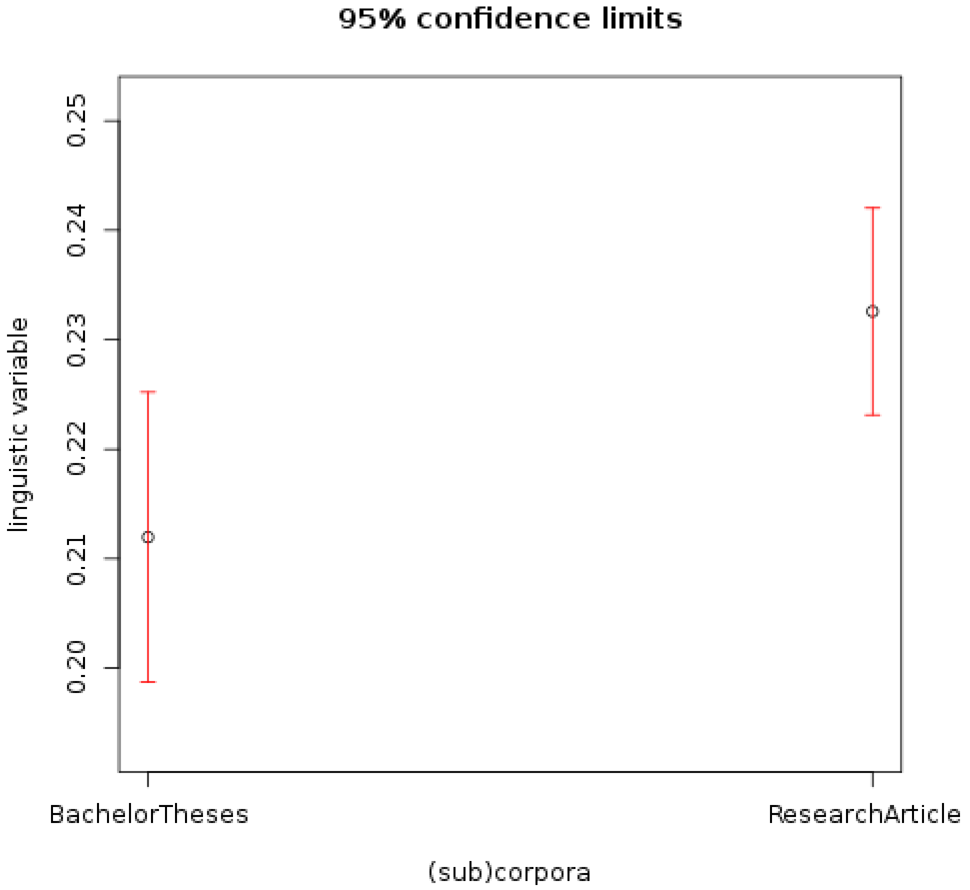


Figure 11: Inference graph of TTR values of Bachelor Theses corpus and Research Articles corpus.

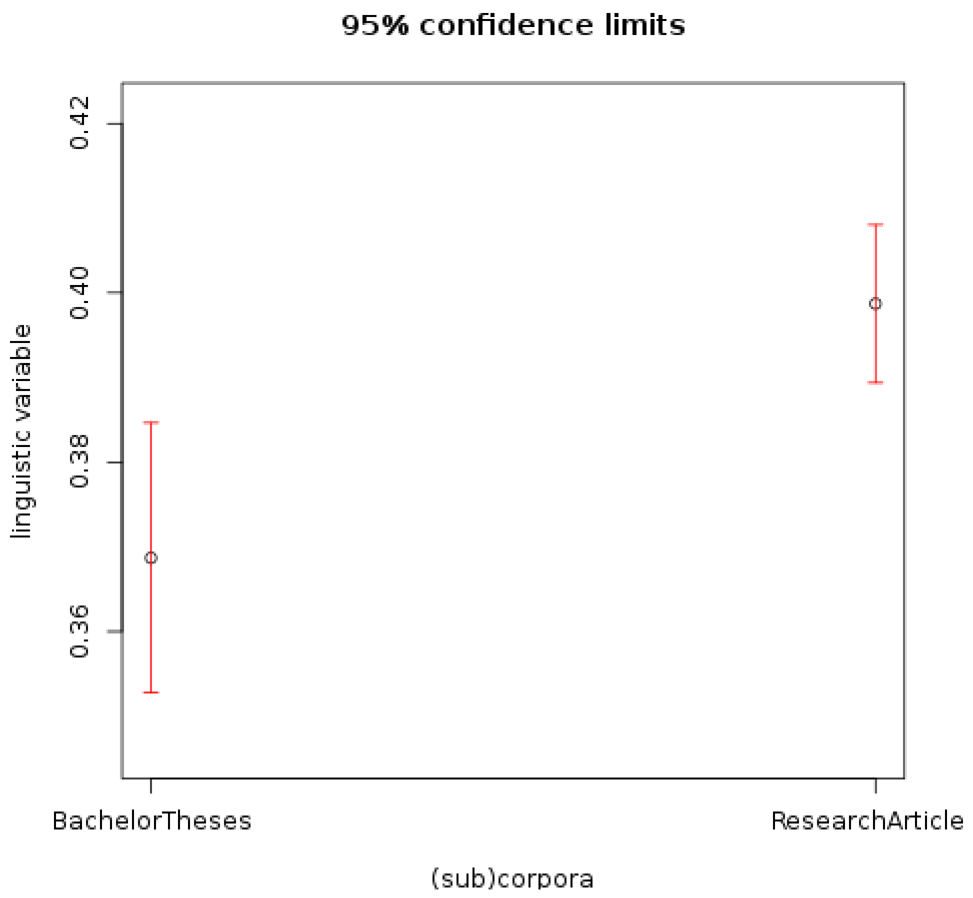


Figure 12: Inference graph of STTR values of Bachelor Theses corpus and Research Articles corpus.

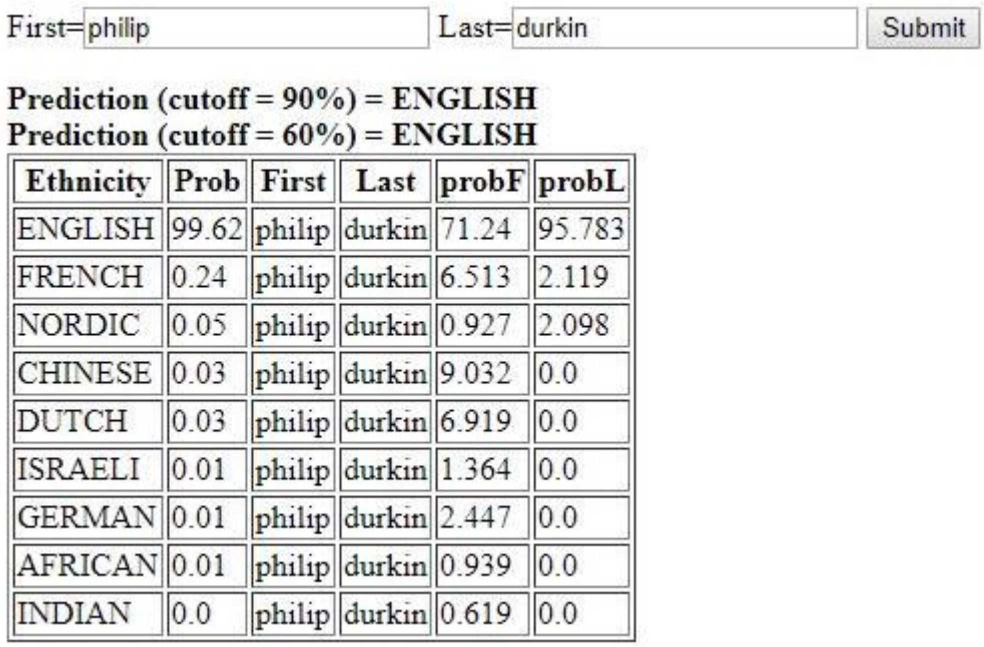
5 Methodological problems

As stated above, there are many factors, which can influence lexical diversity of a text. The texts included in the two corpora are comparable in many aspects, which narrows down the number of possible influences and accounts for better evaluation of the results. L2 learners often score lower levels of lexical diversity in their texts when compared to L1 writers. However, it is not certain that all of the analyzed journal articles were written by native speakers of English. I used Ethnea to determine whether English is authors' L1 or L2. It was found, that out of the 23 texts in Research Articles corpus there are only 9 articles written by English native speakers, 4 of which were co-authored by linguists of other L1s. The remaining 14 texts were written by linguists of German, Nordic, Korean, Japanese, Arab or French origin (see figure 13). Figure 14 shows an example of ethnicity prediction table for the name Philip Durkin.

	name1	ethnicity	name2	ethnicity	name3	ethnicity	name4	ethnicity
text1	Islam Youssef	Arab						
text2	Minna Nevala	Nordic (Finnish)						
text3	Carol Percy	English						
text4	Claudia Felser	German						
text5	Warren Maguire	English						
text6	Yasuaki Ishizaki	Japanese						
text7	Turo Vartiainen	Nordic (Finnish)						
text8	Sigrid Beck	German	Remus Gergel	Slavic/Romanian				
text9	Britta Mondorf	German	Ulrike Schneider	German				
text10	David Lorenz	English/German[15]						
text11	Günter Rohdenburg	Dutch/German						
text12	Philip Durkin	English						
text13	Claire Childs	English	Christopher Harvey	English	Karen Corrigan	English	Sali Tagliamonte	Italian/Arab
text14	Lieselotte Anderwald	German/Nordic						
text15	Saleh Batais	French/Arab	Caroline Wiltshire	English				
text16	Larry Hyman	English						
text17	Kristen Kennedy Terry	English						
text18	Jeong-eun Kim	Korean	Hosung Nam	Korean				
text19	Jens Schmidtke	German						
text20	Gholam Hassan Khajavy	Arab	Peter McIntyre	English	Elyas Barabadi	Arab		
text21	Laurent Dekydtspotter	French	Hyun-Kyoung Seo	Korean				
text22	Holger Hopp	German						
text23	Brent Walter	English	Junko Namashita	Japanese				

Figure 13: Ethnicities of research articles authors.

[15] Ethnea predicted ethnicity of David Lorenz as English or German, with higher probability of him being English. However, Google search revealed he works at university in Germany so it is assumed he is German and I excluded him from the L1 professionals group.



First=philip Last=durkin

Prediction (cutoff = 90%) = ENGLISH
Prediction (cutoff = 60%) = ENGLISH

Ethnicity	Prob	First	Last	probF	probL
ENGLISH	99.62	philip	durkin	71.24	95.783
FRENCH	0.24	philip	durkin	6.513	2.119
NORDIC	0.05	philip	durkin	0.927	2.098
CHINESE	0.03	philip	durkin	9.032	0.0
DUTCH	0.03	philip	durkin	6.919	0.0
ISRAELI	0.01	philip	durkin	1.364	0.0
GERMAN	0.01	philip	durkin	2.447	0.0
AFRICAN	0.01	philip	durkin	0.939	0.0
INDIAN	0.0	philip	durkin	0.619	0.0

Figure 14: Ethnea prediction table.

text_ID	ResearchArticles
text1	0.37
text2	0.39
text3	0.44
text4	0.4
text5	0.41
text6	0.4
text7	0.4
text8	0.38
text9	0.43
text10	0.41
text11	0.4
text12	0.41
text13	0.41
text14	0.42
text15	0.38
text16	0.43
text17	0.36
text18	0.37
text19	0.41
text20	0.4
text21	0.39
text22	0.36
text23	0.4

Figure 15: STTR rates for L1 and L2 professionals.

Figure 15 again shows STTR rates of research articles. Texts written by English L1 professionals are marked by blue, texts written by both L1 and L2 professionals are marked by green and texts written by L2 professionals are white. 7 of the 9 texts written by L1 professionals scored 0.4 or higher, whereas only 8 of the 14 texts written by L2 professionals scored 0.4 or higher. The results show greater lexical diversity in L1 professionals’ texts than in L2 professionals’ texts. It also implies that the difference between STTR rates of research articles and bachelor theses would probably be more extensive if all the research articles were written by native speakers.

The higher level of lexical diversity in Research Articles corpus is caused mainly by the fact that some of the authors are English L1 and thanks to their greater experience and writing skills. Professional linguists publish academic papers regularly as opposed to L2 undergraduate students whose first proper academic work is the bachelor thesis. It would be desirable to include into the comparison also academic texts written by L1 undergraduate students, who have the same level of academic writing skills as Czech undergraduate students. However, they possess the advantage of English being their mother tongue, which would probably reflect in higher lexical diversity when compared to L2 students of English. Regarding the academic genre, the experience of both L1 and L2 students is lower than professional linguists’ experience. Nonetheless, the literature review in this thesis does not say for certain what to expect regarding lexical diversity in relation to genre.

Conclusion

Vocabulary acquisition is an essential part of acquiring a second language and rich vocabulary influences the quality of language output. The range of lexical knowledge, i.e. lexical diversity, can be measured via the type token ratio (TTR), i.e. dividing the number of different words in a text by the total number of words. However, TTR is often unreliable. The main issue with this measure is its sensitivity to text length, which causes inaccuracies when comparing texts with different number of tokens. This obstacle can be overcome by selecting only texts of equal lengths or by using some of the more sophisticated tools, such as the standardized type token ratio. STTR tool divides the analyzed text into equally sized segments, calculates their TTR and then determines the mean value.

The central aim of the thesis was to answer these research questions: What is the type token ratio of each corpus? Is there any difference between the lexical diversity of L2 students' and professionals' texts? What does the lexical diversity level suggest about the texts? Two corpora were created in Sketch Engine as the basis for the lexical diversity research. The first corpus, named Bachelor Theses, was compiled from linguistic bachelor theses written by students of English Philology from the department of English and American studies at Palacký University in Olomouc and Masaryk University in Brno. The second corpus, named Research Articles, consists of research articles written by professional linguists for three journals, namely *English Language and Linguistics*, *Journal of Linguistics* and *Studies in Second Language Acquisition*. There are 23 texts in each corpus, all of which have around eight or nine thousand words. The similar sizes of the texts should account for the validity of the type token ratio measurement. The ratio was determined by dividing the total number of types by the total number of tokens. For a more accurate result, the standardized type token ratio was measured too. Both values were higher for research articles. It was hypothesized that texts written by professional linguists would have higher lexical diversity, which was confirmed. It was found that nine of the twenty-three research articles were written (or co-written) by L1 professionals and the rest by L2 professionals. The higher level of lexical diversity in research articles is credited to their greater experience in academic writing and to the fact that some of the authors are native speakers of English.

Additionally, lexical diversity of learner texts could have been positively influenced by the use of various reference tools.

Lexical diversity is not the only factor determining the quality of a text. Especially in English for academic purposes it is important to acquire correct phraseology and it has been observed in the past studies that L2 students tend to misuse, underuse or overuse some structures. It could be of interest to perform other analyses using the two corpora compiled for the purposes of this thesis and focus on various aspects of the EAP specific phraseology in the texts written by Czech students from our university.

Czech resumé

Korpusová lingvistika je velmi specifickým lingvistickým oborem, který nemá svůj vlastní výzkumný cíl, ale využívá korpusy jako nástroj k testování nejrůznějších lingvistických hypotéz. V osmdesátých letech minulého století se korpusy začaly tvořit nejen z textů rodilých mluvčích, ale také z textů studentů cizích jazyků. Vzhledem k tomu, že korpusy v dnešní době existují v elektronické podobě, mohou snadno obsahovat i miliony slov, což umožňuje generalizování výsledků. Právě zobecňování vždy představovalo problém v tradičním výzkumu osvojování druhého jazyka (SLA), neboť skupiny testovaných subjektů byly příliš malé. Sběr experimentálních dat je velmi časově náročný, proto se výzkumní pracovníci SLA zaměřovali spíše na jednotlivce či malé skupiny a popis vývoje jejich „interlanguage“, tedy jazykového systému, který si studenti vytvoří v průběhu osvojování dalšího jazyka (tento systém nese známky mateřského i cílového jazyka). Studentské korpusy sice nabízejí mnohonásobně větší vzorek dat, ale jejich kompilace je podmíněna přísnými pravidly. Psaný i mluvený jazykový výstup studentů totiž ovlivňuje obrovské množství proměnných, které je nutno do korpusů zahrnout a systematicky popsat. Mezi tyto proměnné patří například věk, mateřský jazyk, jazyková úroveň, schopnost mluvit dalšími cizími jazyky, místo pobytu a místo studia, typ školy, expozice cílovému jazyku, ale také téma, časový limit či překladové pomůcky. Kontrola všech těchto proměnných je nesmírně obtížná, ale její zvládnutí znamená cenný zdroj dat pro výzkum osvojování druhého jazyka.

Nedílnou součástí osvojování druhého jazyka je akvizice slovní zásoby. Bohatá slovní zásoba je významným faktorem ovlivňujícím kvalitu jazykového výstupu. Rozsah slovní zásoby, tedy lexikální diverzita, lze měřit podílem typů a tokenů v daném textu („TTR“), tj. počet různých slovních tvarů v textu se vydělí celkovým počtem slov. Tento způsob zjišťování lexikální diverzity je však v některých případech nepřesný. Hlavním problémem u TTR je citlivost na rozdílnost v délce textů, což v případě srovnávání více textů o různé délce vede k neplatným výsledkům. Tomuto problému lze předejít buď výběrem stejně dlouhých textů, nebo využitím některého ze sofistikovanějších nástrojů k určení lexikální diverzity, například STTR.

Dílčím cílem této práce bylo vytvořit studentský korpus z lingvistických bakalářských prací, jejichž autory jsou studenti anglické filologie z filozofických fakult Univerzity Palackého v Olomouci a Masarykovy Univerzity v Brně, a referenční korpus, který je vytvořen z odborných článků psaných pro lingvistické časopisy. Tyto korpusy mohou být dále k dispozici dalším studentům Univerzity Palackého pro jejich vlastní výzkumy v oblasti studentské akademické angličtiny. Angličtina pro akademické účely je specifická svou frazeologií. Studenti angličtiny jako druhého jazyka si musí osvojit různé typické kolokace, spojovací výrazy a další lexikální, sémantické i syntaktické aspekty akademické angličtiny, aby jejich texty splňovaly specifika tohoto diskurzu. Studentské korpusy mohou pomoci odhalit případy, které studentům dělají největší problémy a tyto poznatky mohou být využity k vytvoření vhodných učebních materiálů.

Hlavním cílem mé práce však bylo zjištění lexikální diverzity studentských akademických textů a její srovnání s akademickými texty psanými profesionálními lingvisty. Lexikální diverzita byla zjištěna výpočtem TTR a STTR, přičemž všechny průměrné hodnoty byly naměřeny vyšší pro odborné články. Hypotéza, která byla stanovena v úvodu, se tak potvrdila. Lze očekávat, že texty rodilých mluvčích budou lexikálně variabilnější než texty studentů. Nicméně je nutno podotknout, že není zcela jisté, že autoři zkoumaných vědeckých článků jsou rodilí mluvčí. Lze však předpokládat, že vzhledem k jejich působení v oboru jsou jejich jazykové schopnosti na úrovni rodilých mluvčích a zároveň se lze domnívat, že jejich články jsou rodilými mluvčími korigovány. S ohledem na tyto skutečnosti lze vyvozovat, že texty profesionálních lingvistů jsou lexikálně různorodější zejména díky jejich bohatším zkušenostem v akademickém psaní. Nicméně lze říci, že bakalářské práce jsou napsány velmi kvalitně, jelikož jejich průměrné hodnoty TTR i STTR byly pouze nepatrně nižší než u odborných článků. Vysoká hodnota lexikální diverzity u studentských textů však mohla být ovlivněna také neomezeným přístupem ke slovníkům či thesauru. Z přiložených grafů lze však vyčíst, že mezi naměřenými hodnotami lexikální diverzity jednotlivých studentských textů jsou znatelné rozdíly, které pravděpodobně svědčí o různých úrovních jejich jazykových schopností.

Works cited

Boková, Anna. 2015. “Building and Exploring a Corpus of Academic Writing by Czech Students of English.” Mgr. thesis, Palacký University in Olomouc.

Březina, Václav. 2018. *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge: Cambridge University Press.

Chafe, Wallace and Jane Danielewicz. 1987. Properties of spoken and written language. In *Comprehending oral and written language*, edited by Rosalind Horowitz and S. Jay Samuels. New York: Academic Press.

Crossley, Scott A., Salsbury, Tom, McNamara, Danielle S., and Scott Jarvis. 2011. “What is Lexical Proficiency? Some Answers from Computational Models of Speech Data.” *TESOL Quarterly* Vol. 45, No. 1, 182–193.

De Cock, Sylvie, and Sylviane Granger. 2004. “Computer Learner Corpora and Monolingual Learners’ Dictionaries: The Perfect Match.” *Lexicographica* 20, 72–86.

Durán, Pilar, Malvern, David, Richards, Brian and Ngoni Chipere. 2004. “Developmental Trends in Lexical Diversity.” *Applied Linguistics* 25/2, 220–242.

Gass, Susan M. and Larry Selinker. 2001. *Second Language Acquisition: An Introductory Course*. Mahwah; London: Lawrence Erlbaum Associates.

Gilquin, Gaëtanelle, Granger, Sylviane, and Magali Paquot. 2007. “Learner corpora: The Missing Link in EAP Pedagogy.” *Journal of English for Academic Purposes* 6, 319–335.

Granger, Sylviane. 1998. The computerized learner corpus: A versatile new source of data for SLA research. In *Learning English on Computer*, edited by Sylviane Granger, 3–14. London: Addison Wesley Longman.

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 & Philadelphia: 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. 2004. “Computer learner corpus research: Current status and future prospects.” In *Applied Corpus Linguistics: A Multidimensional Perspective*, edited by Ulla Connor and Thomas A. Upton, 123–145. Amsterdam: Rodopi.

Jarvis, Scott. 2002. “Short texts, best-fitting curves and new measures of lexical diversity.” *Language Testing* 19 (1), 57–84.

Jarvis, Scott. 2013. “Defining and measuring lexical diversity.” In: *Vocabulary Knowledge: Human ratings and automated measures*, edited by Scott Jarvis and Michael Daller, 13–44. Amsterdam/Philadelphia: John Benjamins.

Johansson, Stig. 2003. “Contrastive linguistics and corpora”. In: *Corpus-based Approaches to Contrastive Linguistics and Translation Studies,* edited by Sylviane Granger, Jacques Lerot, and Stephanie Petch-Tyson. Amsterdam/New York: Rodopi.

Johansson, Victoria. 2008. “Lexical diversity and lexical density in speech and writing: a developmental perspective.” *Working Papers* 53, 61–79.

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.

Koizumi, Rie and Yo In’nami. 2012. “Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens.” *System* 40, 554–564.

Kubát, Miroslav and Jiří Milička. 2013. “Vocabulary richness measure in genres.” *Journal of Quantitative Linguistics* 20 (4), 339–349.

Laufer, Batia and Paul Nation. 1995. “Vocabulary Size and Use: Lexical Richness in L2 Written Production.” *Applied Linguistics* Vol. 16, No. 3, 307–322.

Lozano, Cristóbal, and Amaya Mendikoetxea. 2013. “Learner corpora and Second Language Acquisition: The design and collection of CEDEL2.” In *Automatic Treatment and Analysis of Learner Corpus Data*, edited by Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, 65–100. Amsterdam: John Benjamins.

Mazgutova, Diana, and Judit Kormos. 2015. “Syntactic and lexical development in an intensive English for Academic Purposes programme.” *Journal of Second Language Writing* 29, 3–15.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice.* New York: Cambridge University Press.

Myles, Florence. 2005. “Interlanguage corpora and second language acquisition research.” *Second Language Research* 21 (4): 373–391.

Pietilä, Päivi, Doró, Katalin, and Renata Pípalová. 2015. *Lexical Issues in L2 Writing*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Pravec, Norma A. 2002. “Survey of learner corpora.” *ICAME Journal* 26. 81–114.

Sadeghi, Karim and Sholeh Karvani Dilmaghani. 2013. “The Relationship between Lexical Diversity in Genre in Iranian EFL Learners’ Writings.” *Journal of Language Teaching and Research*, Vol. 4, No. 2, 328–334.

Tono, Yukio. 2003. “Learner corpora: design, development and applications.” *Proceedings of the Corpus Linguistics 2003 conference*. 800–809.

Université catholique de Louvain. n.d. “ICLE.” Accessed October 14, 2017. https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

Université catholique de Louvain. n.d. “Learner corpora around the world.” Accessed March 4, 2018. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Université catholique de Louvain. n.d. “LINDSEI.” Accessed October 14, 2017. https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html

Wang, Xuan. 2014. “The Relationship between Lexical Diversity and EFL Writing Proficiency.” *University of Sydney Papers in TESOL* 9, 65–88.