Univerzita Karlova

Filozofická fakulta

Ústav anglického jazyka a didaktiky

L2 Academic Writing Before and After the Advent of Assistive Writing Tools
Akademické psaní nerodilých mluvčích před nástupem asistenčních nástrojů a po něm

Bakalářská práce

Anna Melicharová

Praha 2023

Vedoucí práce: Mgr. Kateřina Vašků, Ph.D.

Prohlášení:

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, dne 2. ledna 2024

Anna Melicharová

Key Words

Lexical bundles, formulaic language, comparative study, corpus-based study, medical research articles, non-native speakers, assistive writing tools

Klíčová slova

Lexikální svazky, formulaický jazyk, komparativní studie, korpusová studie, odborné lékařské články, nerodilí mluvčí, asistenční nástroje

Abstract

The study focuses on formulaic language, specifically on the use of four-word lexical bundles in medical research articles written by non-native English speakers which were selected from three university journals written by non-native speakers English: *Acta Medica* (AM) by the Faculty of Medicine in Hradec Králové of Charles University, *Biomedical Papers* (BMP) by the Faculty of Medicine in Olomouc of Palacký University Olomouc, and *Prague Medical Report* (PMR) by the First Faculty of Medicine of Charles University. A large number of studies have been caried out in the past focusing on the different use of lexical bundles of native and non-native speakers, therefore such study would not provide any new information. In recent years, authors have significantly expanded their options to use various assistive writing tools for academic writing. These tools use predictive models to offer users suitable formulations or even compose the texts themselves. It can be expected that these tools will influence the use of four-word lexical bundles in medical research articles written by non-native English speakers and for this reason this topic has been chosen as the subject of research. The aim of this thesis is to determine whether the phraseology of the texts before the advent of assistive writing tools differs from that of the latest texts. The theoretical part will describe the formulaic language with focus on phraseology in medical research articles, the structural classification as well as the functional classification of four-word lexical bundles, the general issue of lexical bundles in the language of both native and non-native English speakers, learner corpora, and additionally, the most common assistive writing tools available at the time of the examined texts will be introduced. The analytical part is a comparative corpus-based study and aims to identify four-word lexical bundles in medical research articles written by non-native English speakers within two corpora that are divided into two periods – 1998-2010 and 2011-2022. The parting year has been chosen with respect to the rising presence of assistive writing tools which are nowadays a common part of every Internet user. Based on corpora created for the purpose of this thesis, the identified four-word lexical bundles will be described in terms of their frequency, structural classification and functional classification. The corpora will be mutually compared to find out whether there are any differences occurring in these periods with respect to the rising use of assistive writing tools which are expected to influence the frequency and the functional use of four-word lexical bundles. Although the potential results cannot be directly linked to the presence or absence of assistive writing tools, due to their inevitable occurrence, it can be expected that they play a part in the possible differences. Nevertheless, after analyzing the four-word lexical bundles,

no significant differences were discovered, as both the functional and structural classification remained unchanged within the two corpora. Regarding the frequency, it was slightly higher in the 1998-2010 corpus, however, these results do not show any significant differences as the number of lexical bundles differs in the corpora.

Abstrakt

Studie se zaměřuje na formulaický jazyk, konkrétně užití čtyřslovných lexikálních svazků (lexical bundles) v odborných lékařských článcích nerodilých mluvčích, které byly získány ze tří univerzitních časopisů, které jsou psány nerodilými mluvčími: *Acta Medica* (AM) lékařské fakulty Univerzity Karlovy v Hradci Králové, *Biomedical Papers* (BMP) lékařské fakulty Univerzity Palackého Olomouc v Olomouci a *Prague Medical Report* (PMR) 1.lékařské fakulty Univerzity Karlovy. V minulosti se velké množství studií zaměřovalo na srovnání rozdílného užité lexikálních svazků mezi rodilými a nerodilými mluvčími a další taková studie by nepřinesla žádné nové informace V posledních letech se výrazně rozšířily možnosti autorů využít pro psaní akademických textů různé asistenční nástroje, které využívají predikční modely k tomu, aby nabídly uživateli vhodné formulace, případně přímo texty samy komponují. Lze očekávat, že užití čtyřslovných lexikálních svazků v odborných lékařských článcích nerodilých mluvčích tyto nástroje ovlivní a z toho důvodu bylo toto téma zvoleno jako předmět výzkumu. Cílem práce je zjistit, jestli se frazeologie textů před nástupem asistenčních nástrojů liší od frazeologie textů nejnovějších. V teoretické části bude popsán formulaický jazyk zaměřený na frazeologii v odborných lékařských článcích, strukturální a funkční klasifikace lexikálních svazků, obecná problematika lexikálních svazků v jazyce rodilých i nerodilých mluvčích, žákovské korpusy a dále budou představeny nejčastější asistenční nástroje dostupné v době vydání zkoumaných textů. Analytická část je komparativní korpusová studie, která si klade za cíl identifikovat čtyřslovné lexikální svazky v odborných lékařských článcích nerodilých mluvčích angličtiny ve dvou korpusech, které jsou rozděleny na dvě časová období – od roku 1998 až do roku 2010 a od roku 2011 do roku 2022. Přelomový rok byl zvolen s ohledem na stoupající přítomnost asistenčních nástrojů, které jsou v dnešní době běžnou součástí každého uživatele internetu. Na základě korpusů odborných lékařských článků psaných nerodilými mluvčími angličtiny, vytvořených pro účely této práce, budou popsány užívané čtyřslovné lexikální svazky s ohledem na frekvenci, strukturální klasifikaci a funkční klasifikaci. Korpusy budou vzájemně porovnány za účelem zjištění, zdali se v těchto dvou obdobích objevují rozdíly, které by mohly být ovlivněny

stoupající přítomností asistenčních nástrojů. Lze předpokládat, že tyto nástroje ovlivní frekvenci a funkční užití čtyřslovných lexikálních svazků. Přestože nelze případné rozdíly přiřknout přítomnosti nebo absenci asistenčních nástrojů, díky jejich nevyhnutelné přítomnosti lze předpokládat, že budou hrát určitou roli v případných rozdílech. Nicméně, po následné analýze čtyřslovných lexikálních svazků nebyly nalezeny žádné signifikantní rozdíly. Funkční i strukturální klasifikace zůstala nezměněna v obou korpusech. Co se týče frekvence, v korpusu 1998-2010 byla o něco vyšší. Kvůli rozdílným počtům lexikálních svazků neprokazují tyto výsledky signifikantní rozdíly.

Table of Contents

LIST OF ABBREVIATIONS **9**

LIST OF TABLES AND FIGURES **10**

1 INTRODUCTION **11**

2 THEORETICAL BACKGROUND **12**

2.1 FORMULAIC LANGUAGE **12**

2.1.1 Phraseology in Medical Research Articles *13*

2.2 LEXICAL BUNDLES **14**

2.2.1 Structural Classification *14*

2.2.2 Functional Classification *15*

2.3 LEXICAL BUNDLES OF L2 LEARNERS **18**

2.4 LEARNER CORPORA **19**

2.5 MEDICAL WRITING **21**

2.5.1 Lexical Bundles in Medical Research Articles *21*

2.6 ASSISTIVE WRITING TOOLS **22**

3 MATERIAL AND METHOD **24**

3.1 SOURCES OF MATERIAL **24**

3.1.1 Criteria for Article Selection *25*

3.1.2 Creation of the Corpora *27*

4 ANALYTICAL PART **29**

4.1 RESULTS **29**

4.1.1 Comparison of the Frequency *34*

4.1.2 Structural Classification of Retrieved LBs *37*

4.1.3 Functional Classification of Retrieved LBs *41*

4.1.4 Structural Classification within the Functional Classification *46*

5 CONCLUSION **49**

REFERENCES **52**

RESUMÉ **55**

APPENDICES **58**

List of Abbreviations

AM – Acta Medica

BMP – Biomedical Papers

ELF – English as Lingua Franca

L1 – First language

L2 – Second language

LB – Lexical bundle

PMR – Prague Medical Report

List of Tables and Figures

Figure 1: File Organization

Figure 2: Example of the Corpus Frequency Comparison

Figure 3: Example of the Corpus Frequency Test Result

Table 1: Structural Classification According to Biber, Conrad, and Cortes (2004)

Table 2: Functional Classification of Lexical Bundles According to Biber, Conrad, and Cortes (2004)

Table 3: Functional Classification of Lexical Bundles According to Hyland (2008)

Table 4: Functional Classification of Lexical Bundles According to Dontcheva-Navratilova (2012)

Table 5: Number of Collected Articles per Year and the Total Number and Running Words from 1998 to 2010

Table 6: Number of Collected Articles per Year and the Total Number and Running Words from 2011 to 2022

Table 7: Comparison of Retrieved LBs from the 1998-2010 and the 2011-2022 Corpora

Table 8: Comparison of the Frequency of Matching Lexical Bundles

Table 9: Structural Classification of Retrieved Lexical Bundles

Table 10: Structural Classification of Retrieved Lexical Bundles and the Percentage Representation

Table 11: Functional Classification of Retrieved Lexical Bundles

Table 12: The Total Number of LBs Based on Functional Classification and the Percentage Representation in the 1998-2010 Period

Table 13: The Total Number of LBs Based on Functional Classification and the Percentage Representation in the 2011-2022 Period

1 Introduction

Language production, whether written or spoken, is omnipresent and it is no wonder that it has been a subject of research for a long time. Formulaic language, nowadays known to form up to "70 percent in native speaker´s daily communication" (Altenberg 1998), has been the centre of attention of linguists in the last couple of decades (Arani et al. 2015, 52). Multi-word units forming the formulaic language help "encoding work for the speaker and decoding work for the addressee, thus allowing for the construction of fluent spoken discourse" (Ibid, 52). Such combinations of words have been found under various names e.g. *clusters* (Scott 2004, 225), *n-grams* (Stubbs 2007a), or *lexical bundles* (Biber et al. 1999, 990) that will be used in this thesis as well. The frequent use of formulaic language of native speakers raises the question of its importance in L2 acquisition and it was considered that lexical bundles (LBs) are not required to be taught as they will be naturally picked up by learners (Biber & Barbieri 2007, 284). There have been arguments supporting the importance of frequency and salience (Ellis 2002, 178) but also salience over frequency (Gass & Mackey 2002). However, saliency and frequency do not always come together, so it is crucial to focus on LBs based on their discourse functions (Biber & Barbieri 2007, 284).

Apart from non-native speakers, the use of LBs of native speakers has been examined in different disciplines showing that each has its preferred LBs that help in the structural organization of the texts (Hyland 2008). The analytical part of this thesis will focus on medical research articles and the differences between two periods. There has not been much research regarding medical research articles in general (Arani et al. 2015, 61), that is one of the reasons this genre has been chosen. On the contrary, the comparison of native and non-native speakers has been covered in various studies; this thesis focuses on the difference between non-native speakers with regards to the rising presence of assistive writing tools. The aim is to identify LBs and their frequency, function, and structure in two corpora that are divided based on the rising presence of assistive writing tools. Such study cannot ensure results directly linked to the usage of such programs, as the quality of education, motivation or even the years spent in English-speaking countries may influence the writer. However, the presence of assistive tools is inevitable and more likely used in recent years and the results may show a possible change in the use of LBs, whether it´s their frequency or function.

2 Theoretical Background

The theoretical part of this thesis will outline the problem of formulaic language with focus on LBs, their classification based on their function and structure while providing the existing research in this field. The summary of the current findings and knowledge will provide a baseline for the understanding of LBs and their importance with emphasis on the connection between formulaic language and medical research articles as they provide the data for the analytical part. Since the resources on the influence of assistive writing tools on either non-native or native speakers are limited, it is not possible to provide data in this area, however, the available programs and their functions will be listed together with their impact within recent years.

2.1 Formulaic Language

Language is a man´s tool to express thoughts and attitudes and is undeniably present in everyday life. Whether it concerns a conversation between co-workers, business meeting or simply an advertisement on the subway, language is there, and it is therefore important to understand how it works and what elements make it so predictable and easy to consume. According to Kjellmer (1991, 112) language is highly dependent on "combinations of words that customarily co-occur". The term formulaic language is defined by Wood (2015, 693) as "multi-word expressions that have a single meaning or function, and that are prefabricated or stored and retrieved mentally as if a single word". It is important to understand how these formulaic expressions differ and what is their function.

Multi-word units forming the formulaic language are e.g. LBs, collocations, idioms, or multi-word verbs. There are differences between these units regarding their idiomatic nature – *in a nutshell* is idiomatic and needs to be learned as a single word as the literal meaning is not the same (Ellis et al. 2008, 377). Collocations are not idiomatic and based on a phraseological approach, their frequency is not as important as the "native-like co-selection of node and collocate(s); thus, *heavy rain* is a collocation even if it occurs only once, while *strong rain* is not" (Ebeling & Hasselgård 2015, 211). Other approaches focus on frequency; therefore, it is necessary to distinguish collocations from LBs. Collocations have a strictly defined node that is accompanied by collocates based either on their semantics or frequency but most importantly, individual parts of collocations do not have to be neighboring while LBs require such closeness (Ibid 2015, 211). The usage of such constructions serves to not only

distinguish native speakers from learners, but also indicates the advanced level in the native community (Ibid 2015, 216).

2.1.1 Phraseology in Medical Research Articles

Medical research articles are the target of this thesis, and as disciplines differ in specific terminology and phraseology, it is necessary to understand what defines this discipline. Medicine is a complex field and requires thorough knowledge of terms and the ability to connect pieces of information in larger scale. However, difficulties may occur even during communication with the patient. It is crucial to be able to express oneself properly when delivering bad news or talking about a sensitive topic. Similarly, authors of medical reports or research articles must be aware of the words they use and their proper meaning.

When it comes to formulaic language, it is possible to use rich words to express an opinion in fiction, but in medical writing it can lead to serious problems (Croft 2002). Even the opposite case might not be the best choice; using words like *never, always, impossible* or *certain* needs to be proceeded with caution as they carry a strong degree of certainty and if not used properly, the author´s credibility might be affected (Ibid 2002). These examples illustrate the necessity of carefulness; nevertheless, they do not rule out the presence of phraseology in medical research articles.

Jargon, or "special words and phrases that are used by particular groups of people, especially in their work" as defined by the online Cambridge Dictionary (n.d.), is present in the medical field as well. It can have different impact when comparing conversations with patients and writing research articles. A study conducted by Méndéz-Cendón and López-Arroyo (2003) compared the occurrence of phraseological units in medical research articles and abstracts. They focused on sub-technical terms, "words that have taken on a more restricted meaning and syntax in certain scientific and technical fields" (Ibid 2003, 251) and their distribution within the paper. The results showed e.g. different premodification of or by the word *study*: the first occurrence in the "Materials and methods" part was premodified by adjectives to describe the method, while the second occurrence premodified a noun group to illustrate a procedure; in the "Results" part, *study* was premodified by a number to display the results (Ibid 2003, 263-4). According to a contrastive analysis of phraseological devices in medical abstracts, the attitude of the writers differed – Spanish authors considered their non-academic audience and modified their writing accordingly, but English authors used complex language and specific terms (Méndéz-Cendón & López-Arroyo 2007, 514).

2.2 Lexical Bundles

According to Bieber et al. (2007, 264) LBs can be defined as "multiword sequences that occur most commonly in a given register", though the definition can differ from one linguist to another, e.g. Hyland (2008, 5) describes LBs as "words which follow each other more frequently than expected by chance". These expressions are crucial not only for discourse functions but also for fluent speech and language acquisition. The first instance of the term lexical bundles is present in the Longman Grammar of Spoken and Written English from 1999, that compared both spoken and written discourse (Biber & Barbieri 2007, 264) and since then multiple studies have been carried out, focusing particularly on these formulaic expressions and their function. Based on their function, LBs serve as an indicator of a genre and make the text more predictable and easier to read (Hyland 2008, 5). It is no surprise linguists perceive LBs "as important building blocks of coherent discourse and characteristic features of language use in particular settings" (Ibid 2008, 8). Although LBs turned out to be "the most frequent recurring sequences of words in any collection of texts" (Hyland 2012, 150), the issue of proper linguistic knowledge influences both non-native and native speakers as they perceive their native language differently.

The frequency of LBs is an important factor when doing a research study and its scope differs according to various linguists from 10 to 40 times per million words (Bieber et al. 2006, Bieber et al. 2004). In most studies, four-word bundles are chosen for analysis as they often include three-word bundles and are more frequent than five-or-more-word bundles (Hyland 2008, 8). There is however a common issue when analyzing LBs that concerns the overlapping of individual bundles. In this case, "two four-word bundles are actually a part of five-word string" (Hyland 2012, 151) and it is necessary to eliminate such occurrences. The following section focuses on the structural classification of LBs.

2.2.1 Structural Classification

Studies on LBs focus mostly on the functional classification, however Biber et al. (2004) carried out a study concerned with their structure. As shown in the Table 1, three types were identified – LBs with verb phrase fragments, dependent clause fragments and phrasal LBs that can be realized either by a noun phrase or a prepositional phrase (Ibid 2004, 380). In their study, they focused on identifying structural types of LBs across various registers and concluded that LBs in academic prose are mostly phrasal and almost 70 % included a noun phrase or "a sequence that bridges two prepositional phrases". Another discovery was made in

classroom teaching that used "twice as many different LBs as conversation, and about four times as many as textbooks" that is caused by the reliance on both spoken and written register, while textbooks did not tend to use LBs as much; non-formulaic expressions were preferred. (Ibid 2004, 382).

Lexical bundles that incorporate *verb phrase* fragments	a) 1st/2nd person pronoun + VP fragment b) 3rdperson pronoun + VP fragment c) Discourse marker + VP fragment d) VP (with non-passive verb) e) VP with passive verb f) *Yes-no* question fragments g) *Wh-* question fragments	a) *you don´t have to* b) *it´s going to be* c) *I mean you know, you know it was* d) *is going to be* e) *is based on the* f) *are you going to* g) *what do you think*
Lexical bundles that incorporate *dependent clause* fragments	a) 1st/2nd person pronoun + dependent clause fragment b) *Wh-* clause fragments c) *If-* clause fragments d) *To-* clause fragments e) *That-* clause fragments	a) *I want you to* b) *what I want* c) *if you want to* d) *to be able to* e) *that there is a*
Lexical bundles that incorporate *noun phrase* and *prepositional phrase* fragments	a) NP with *of*-phrase fragment b) NP with other post-modifier fragment c) Other NP expressions d) Prepositional phrase expressions e) Comparative expressions	a) *one of the things* b) *a little bit about* c) *a little bit more* d) *of the things that* e) *as far as the*

Table 1 Structural Classification According to Biber et al. (2004)

2.2.2 Functional Classification

The functional classification of LBs is not as straightforward due to their inability to form complete structural units; but it enables them to be multifunctional and fit into more categories (Dontcheva-Navratilova 2012, 40). LBs can convey more functions within one instance, e.g. *take a look at* can be both a directive and a topic introducer (Biber et al. 2004,

383). Hyland (2008, 5) also points out that although it is well known that LBs differ according to genres, their occurrence is dependent on disciplines as well and he analysed this relation in his paper. As the function of LBs is not strict, multiple examples of their classification are shown.

Stance expressions	a) Epistemic stance	a) *I think it was*
	b) Attitudinal/modality stance	
	b1) Desire	b1) *if you want to*
	b2) Obligation/directive	b2) *I want you to*
	b3) Intention/prediction	b3) *are we going to*
	b4) Ability	b4) *to be able to*
Discourse organizers	a) Topic introduction/focus	a) *what do you think*
	b) Topic elaboration/clarification	b) *has to do with*
Referential expressions	a) Identification/focus	a) *that´s one of the*
	b) Imprecision	b) *and stuff like that*
	c) Specification of attributes	
	c1) Quantity specification	c1) *there´s a lot of*
	c2) Tangible framing attributes	c2) *the size of the*
	c3) Intangible framing attitudes	c3) *the nature of*
	d) Time/place/text reference	
	d1) Place reference	d1) *in the United States*
	d2) Time reference	d2) *at the same time*
	d3) Text deixis	d3) *shown in figure N*
	d4) Multi-functional reference	d4) *the end of the*

Table 2 Functional Classification of Lexical Bundles According to Biber et al. (2004)

Research-oriented	a) Location	a) *at the same time*
	b) Procedure	b) *the use of the*
	c) Quantification	c) *a wide range of*
	d) Description	d) *the structure of the*
	e) Topic	e) *in the Hong Kong*
Text-oriented	a) Transition signals	a) *on the other hand*
	b) Resultative signals	b) *as a result of*
	c) Structuring signals	c) *in the present study*
	d) Framing signal	d) *in the case of*

Participant-oriented	a) Stance features b) Engagement features	a) *are likely to be* b) *it should be noted that*

Table 3 Functional Classification of Lexical Bundles According to Hyland (2008)

Referential bundles	a) Time/place/text-deixis bundles b) Attribute bundles c) Topic-specific bundles	a) *at the end of the* b) *a little bit of* c) *in the curricula of*
Discourse organizers	a) Logical relations bundles b) Intratextual reference bundles c) Framing bundles	a) *on the other hand* b) *in the present study* c) *in the case of*
Attitudinal bundles	a) Stance bundles b) Interactional bundles	a) *the fact that the* b) *as can be seen*

Table 4 Functional Classification of Lexical Bundles According to Dontcheva-Navratilova (2012)

Although there are some differences, three functional categories of LBs can be identified. Dontcheva-Navratilova´s (2012) Referential bundles cover referential LBs as described by Biber and Barbieri (2007) and also the Research-oriented by Hyland (2008); the Discourse organizing LBs refer to the Text-oriented from Hyland (2008) and its title is borrowed from Biber and Barbieri´s (2007) classification as well as the term Attitudinal bundles that corresponds with Hyland´s (2008) Participant-oriented and Biber and Barbieri´s (2007) Stance expressions (Dontcheva-Navratilova 2012, 40-41). Stance (Biber et al. 2004), Participant-oriented (Hyland 2008), and Attitudinal LBs (Dontcheva-Navratilova 2012) express the same function; they stress the attitude of the speaker towards other participants regarding the up-coming proposition (Biber et al. 2004, 389) and they provide a source of an interaction between the writer and the reader (Hyland 2008, 18). The most thorough description is by Biber et al. (2004) as they divided these bundles into Epistemic and Attitudinal and those into four subcategories. They concern Desire, Obligation/Directive, Intention/Prediction and Ability and enable more detailed classification.

The other corresponding categories are Discourse organizers (Ibid 2004), Text-oriented (Hyland 2008) and Discourse organizers (Dontcheva-Navratilova 2012). These bundles make the text more organized and available to the reader (Hyland 2008, 17) and indicate topic

change or its elaboration (Biber et al. 2004, 391). According to Hyland´s findings (2008, 16), Research-oriented LBs occur mostly in research articles, covering nearly two-thirds of present bundles. They are usually formed by the preposition and *of* structure and bring readers´ attention to a particular instance in the text. Although LBs can be distinguished by their function, they can overlap based on the current context.

The last triad consists of Referential (Biber et al. 2004), Research-oriented (Hyland 2008), and Referential LBs (Dontcheva-Navratilova 2012). In his research, Hyland discovered that these LBs formed a majority in Science and Engineering texts and consist of a noun phrase and *of* structure (Hyland 2008, 14). This type of LBs serves to describe real-world instances and research related objects or materials (Ibid 2008, 13-14). Biber et al. (2004) provided more detailed division including seven subcategories. Specification of attributes has three subcategories: Quantity specific describe a certain amount, Tangible framing attributes focus on the size and structure of the head noun, while Intangible framing attributes "identify abstract characteristics" (Ibid 2004, 395). Time/place/text reference can be divided into four remaining subcategories: the first three refer respectively to those included in the title of this category and the fourth one, Multi-functional reference, can cover more than one reference at once (Ibid 2004, 396).

2.3 Lexical Bundles of L2 Learners

One of the basic definitions of LBs is their common occurrence but they are not idiomatic and do not form complete structural units unlike idiomatic expressions – they are not as frequent in either spoken or written registers and they occur mostly in fiction (Biber et al. 2004, 377). LBs function as a link between these units (Biber & Barbieri, 2007). LBs such as *on the other hand* or *the fact that the* cannot form a clause or a phrase and this distinction from the other formulaic devices causes their lack in foreign language teaching (Dontcheva-Navratilova 2012, 39). Another reason LBs are not receiving enough attention in classrooms might be their salience; so, despite their commonness, they often remain unnoticed even by university students (Shin & Kim 2017, 81). Therefore, the acquisition of LBs cannot be reached by unconscious exposure through e.g., academic writing – the frequency was low, and the function associated with certain genre was often misused (Cortes, 2004, 417-420).

When studying the use of phraseological units in L2 learners, both quantitative and qualitative approach should be used as it is necessary to determine not only the frequency but also the grammatical correctness based on the context (Ebeling & Hasselgård 2015, 217). It is quite

common that learners take instances from their native language and try to make them work in L2, such as *considered as* that comes from French *considéré comme,* and these errors usually go unnoticed as phraseological (Ibid 2015, 217-218). The following examples by Ellis et al. (2008, 377) stress the importance of learning words not as a single unit but rather by its company: *describe about problem, get advantage of* or *did the mistake.* These examples show incorrect usage of common phrases that should be *describe the problem, take advantage of* and *made the mistake* and though fixed expressions are taught in schools, native-like collocation remains a problem. They have come up with Academic Formulas List which they hope to be used in teaching curriculum and that provides "formulaic sequences, […] that are significantly more common in academic discourse" (Ibid 2008, 378, 392).

A study carried out by Shin and Kim (2017) focused on teaching L2 learners article usage with the help of LBs. They contain a lot of articles as they serve as a connection between structural units (Biber & Barbieri 2007) and therefore their knowledge was assumed to improve the correct usage of English articles. According to previous studies, learners whose L1 system did not have articles tend to omit them in English and on the contrary, direct teaching of articles led to their excessive usage (Shin & Kim 2017, 80). In this three-week study, students were divided in groups based on their proficiency levels and further into a control group with no instructor and a treatment group that was provided with an instructor explaining "the use of articles in LBs as well as their functions as wholes in context" (Ibid 2017, 84). At the end of the study, this group of students was compared with those who did not undergo this learning course and their pre- and post-experiment tests were analyzed and as the results showed, the treatment groups turned out to improve their skills and overall score better than the control group (Ibid 2017, 84-85). Based on this study, it is apparent that teaching LBs can have direct positive impact on other parts of language learning as well. Dontcheva-Navratilova (2012, 55-56) suggests a few examples that could be used in teaching classes such as "recognition of LBs in academic discourse base on frequency of occurrence in texts; pattern practice to develop confidence […] [or] creative use in written performance".

2.4 Learner Corpora

Learner corpora can be defined as "electronic collections of texts produced by foreign or second language learners" (Paquot & Granger 2012, 3) and while learner corpora research as a discipline is relatively young, beginning in the late 80s and early 90s of the 20th century, it is useful for L2 acquisition and may contribute methods to the learning process (Granger 2008,

1). Language learners are "understood as foreign language learners, i.e. speakers who learn a language which is neither their first language nor institutionalized additional language in the country where they live" (Ibid 2008, 1). However, as the English language is considered Lingua Franca, i.e. a language used by advanced non-native speakers as a means of communication, it complicates the viewpoint of learner corpora (Ibid 2008, 1). Still, he (2008) believes "that learner corpus approach and ELF approach […] should rather be regarded as two sides of the same coin".

L2 differs based on "age, gender, mother tongue background, […] time spent in a country where the foreign language is spoken" (Paquot & Granger 2012, 3) and all these criteria need to be considered when gathering data. Learner corpora have electronic base, providing wide range of available data that form a continuous text rather than phrases taken out of context and therefore better reflection of the actual usage (Granger 2008, 2). He (2008) typologically divides learner corpora to six categories: commercial vs. academic, big vs. small, English vs. non-English, writing vs. speech, longitudinal vs. cross-sectional, and immediate vs. delayed pedagogical use. Some of these categories are self-explanatory but the rest will be provided with brief description.

Commercial corpora, though not as prominent, includes data with multiple L1 background and the main corpora are the *Longman Learner´s Corpus* and the *Cambridge Learner Corpus* (Ibid 2008). Academic corpora, on the other hand, are based in educational settings and usually provide one L1 background; the *International Corpus of Learner English* is an exception. Longitudinal corpora gather data from the same learners over a longer period and due to their long duration, so-called quasi-longitudinal corpora are used more frequently as they gather data from learners with various levels of knowledge within a single time. Cross-sectional corpora gather data in a single period but from learners coming from different areas. Immediate pedagogical use collects data from teachers during lessons and the subsequential results within the same group of students. Data collected for the delayed pedagogical use are used for different students with the same attributes and level of proficiency as the studied group.

Since the learner corpora is a written database, various programs can be used to identify the frequency and create a comparison. One of the widely used programs is *AntConc* that will be used in the analytical part as well. One of the studies concerning LBs showed, that the frequency of four-word bundles was higher in Chinese learners rather than in native speakers,

nevertheless this result does not automatically mean advanced level of English, quite the contrary – "less proficient learners seem to be more reliant on lexical bundles" (Ibid 2012, 16). Further analysis showed that the longer the learners spend in English-speaking country, the more similar is their frequency of two-word LBs to the native-level (Ibid 2012, 18).

2.5 Medical Writing

Some aspects of medical writing have been foreshadowed in the 2.1.1. section such as the need to express oneself accurately and with a precise knowledge of the semantics of words or the necessity control the amount of medical jargon that is used – this applies mostly on the spoken aspect. Written medical texts mostly include "case reports, research papers, abstracts, editorial, letter to the editor, prescriptions, experimental reports" (Méndéz-Cendón & López-Arroyo 2003, 248). Naturally, every written publication should have a certain structure and research papers are no exception as they should follow the so-called IMRAD structure, that contains at least these four sections: introduction, methods, results, and discussion (Ibid 2003, 249). When it comes to genre studies, the main target has usually been the rhetorical structure (Méndéz-Cendón & López-Arroyo 2007, 503) but the phraseology is equally important as it is one of the main aspects that help creating a coherent and structured text (Méndéz-Cendón & López-Arroyo 2003, 266). Although their study focused on the comparison between the micro and macro structure, i.e. phraseology and rhetoric, this thesis will focus solely on the phraseological aspect of medical research articles.

2.5.1 Lexical Bundles in Medical Research Articles

Formulaic language differs in various disciplines, however there is a limited number of studies on LBs in medical research articles (Arani et al. 2015, 61). Akbulut (2020), Dontcheva-Navratilova (2012) or Estaji and Monrazeri (2022), to name a few, focused on a comparison between native and non-native speakers and their usage of LBs in their studies. Arani et al. (2015, 64) examined their frequency, structure and function in medical research articles from the Science Direct Online and these features were identified and compared with the classification proposed by Hyland (2008). The most used LBs regarding the structural approach were those with prepositional phrases and regarding the functional approach Research-oriented bundles were the most frequent; nevertheless, Text-oriented bundles were not that different unlike Participant-oriented bundles that were very low, indicating the ignorance of their function (Arani et al. 2015, 64-5).

Considering the small number of studies carried out on medical research articles and the contrasting amount of those focusing on native and non-native speakers, this thesis aims to provide research on the differences between the frequency, function, and structure of LBs within two time periods of non-native speakers. Although native and non-native speakers are commonly compared, in recent years with the rise of technology and availability of multiple programs providing an instant correction or suggesting a better phrasing it gets more difficult to recognize the real authorship. Even though the usage of such programs cannot be proved with 100 % accuracy, its presence is undeniable and therefore it can be assumed that research articles from recent years and their phraseology might be affected by them.

2.6 Assistive Writing Tools

In 2023, the presence of technology is widespread and influences every aspect of our everyday lives. Education is no exception whether it concerns replacement of blackboards for interactive ones, the everyday use of laptops at universities or the wide range of possibilities when doing research. Although living in digital age is the only experience the younger generation has, it has not always been this way. The older generation might have even problems using modern cell phones, while their grand grandchildren cannot imagine their lives without them. Such a dependence on technology naturally rises a question of its impact on basic life activities, but for the purpose of this thesis, on LBs.

Well-known programs such as Microsoft Word or PowerPoint, commonly used by students, provide spell-checking functions that eliminate any unrequired errors. This can be useful as it corrects common typos such as *hte* to *the,* but it may cause more harm than good (Ismael et al. 2022, 233). The auto-correct spelling can lead to lowered attention towards spelling as it provides a security that the word will either be fixed according to the software or underlined to show a typo. This sort of reliance on external sources leads to worsened real-life ability of the writer that can cause problems not only during school years but also in the follow-up job applications (Ibid 2022, 234).

One of the individual programs is *Grammarly*, that has been used especially in the recent years, although it entered the digital world in 2009. It provides its users with “real-time writing suggestions” (Grammarly, n.d.), spelling and punctuation corrector and in the premium version there is also citation formatting and detection of plagiarism (Ibid, n.d.). Apart from software programs, there is also an online thesaurus that “provides users with over 550,000 synonyms and a suite of tools that simplify the writing process” (Dictionary.com,

n.d.), enabling users to avoid repetition in their papers. However, this source is not context dependent and for that reason, users cannot choose random synonyms. Considering this fact, technology alone presumably cannot improve students´ skills (MacArthur 2000, 86).

Although Internet and its associated benefits have been present for over two decades, research on assistive writing technology is limited due to the extremely fast improvement and growth of technology in general (Ibid 2000, 86). No such research has been made regarding its influence on LBs and it is not possible to assign the upcoming results directly to the influence of assistive writing tools. However, with the rising occurrence of technologies and digital tools providing simplifications, such as the auto correct function, it is likely that people would use such options to reach more native-like writing performance. The possible differences might be influenced precisely by the rise of assistive writing tools.

3 Material and Method

This part of the thesis will present the hypothesis that was set before starting the analysis itself, the material that was chosen for further examination and the collected data for the research, the methodology and criteria used for data selection as well as the detailed description of corpus creation.

The aim of this thesis was to examine LBs in medical research articles in two time periods with respect to the advent of assistive writing tools which may have had an impact on the frequency and functional and structural classification of LBs. Within the scope of this analysis, it is not possible to assign the potential difference directly to the assistive writing tools as the authors have not been questioned; therefore, the only material used is their written work. However, with the rising influence of the Internet and its available assistive tools, the probability of their usage increases over the time. Moreover, the accessibility of such tools and programs is nowadays a common thing for every user and for this reason, it can be expected that one of the reasons writing of non-native speakers changes, and with it the formulaic language as well, is precisely that.

This thesis aims to answer the following questions:

1. What are the most frequent four-word LBs in the 1998-2010 corpus and in the 2011-2022 corpus?
2. What is the structure of four-word LBs in the 1998-2010 corpus and in the 2011-2022 corpus?
3. What functions these four-word LBs have in the 1998-2010 corpus and in the 2011-2022 corpus?
4. Are there any significant differences in frequency, function and structure within the two corpora?

The analytical part is a comparative study so therefore, the findings in each corpus will be compared to reveal possible differences. The hypothesis for this research is that in the 2011-2022 period, the four-word LBs will be used more frequently than in the 1998-2010 period due to the autocompletion and word-suggesting programs while at the same time not serving the same function since said programs do not take into account the context in which LBs occur.

3.1 Sources of Material

The focus of the research of this thesis were lexical bundles in medical research articles written by non-native English speakers. The reasoning behind this choice is that the

comparative research on native and non-native speakers have been done thoroughly in the past and therefore this study would not result in any new information regarding this topic. Thus, it was necessary to find a source of materials which would provide articles with authors of said criterion. For this reason, three journals from different Czech universities were chosen as the source. They are *Acta Medica* (AM) by the Faculty of Medicine in Hradec Králové of Charles University, *Biomedical Papers* (BMP) by the Faculty of Medicine in Olomouc of Palacký University Olomouc, and *Prague Medical Report* (PMR) by the First Faculty of Medicine of Charles University. All these three journals have available online archives which enabled access to the oldest published articles and since they are available to the public, the data collection lied in downloading said articles from the database.

3.1.1 Criteria for Article Selection

Simply downloading every single article was nevertheless not possible as the criteria for the research did not allow native English speakers. Naturally, each article provided the names of the authors, which were usually groups of three or more students, as well as their universities, and such information was useful for identifying the language background of the authors and therefore enabling easy selection of those significant for the research and eliminating those which were not. The majority of the students were either Czech or Slovak, but there were also plenty of other nationalities, probably due to the exchange student programs such as Erasmus etc., and it was therefore necessary to eliminate those whose native language was English or those authors from countries where the English language could occur alongside the official language of the country. All the journals have the articles divided into reviews, original articles and case reports and in order to maintain a consistent approach, only the original articles were used as the subject of further analysis. Despite this selection, there was still a sufficient amount of material that could be used to create two corpora with appropriate number of running words.

The websites providing online archives also state that each article undergoes an evaluation by the editorial board. However, after further examination, the corrections made do not include any major changes within the article nor do not influence the author´s original thoughts or writing ability, and therefore, it should not have any significant impact on the text originally written.

In order to create two corpora for comparison, it was necessary to choose an appropriate year to divide the available material into two periods. The oldest volume from AM was from 1998,

BMP´s oldest volume was from 1998 while PMR´s oldest volume was from 2004. Despite this gap, the oldest volumes from all journals were used. Although all the journals continue publishing new material to the present day, the year 2022 was chosen as the last one to be included in the research data, as the current year 2023 would not be able to provide a full database. The year 2010 was chosen to be the last in the old period as it would create more or less similar groups regarding material but also because the Internet and its tools were already available to the majority of people for some time and therefore a regular part of people´s lives.

After all these criteria have been settled, all the suitable articles were downloaded in their original pdf file. As this kind of file is not compatible with the chosen program called *AntConc,* it was necessary to convert each article into proper one, that being the txt file. This process also enabled further changes made within the articles which were necessary to create a text suitable for further analysis. Each article was modified so it did not include unnecessary information such as "Acknowledgements" and "Resources". The reason for this omission was the repetitive writing pattern within these sections throughout the whole journal as well as the fact that the information included did not reflect the author´s ability to express oneself in written English.

For example, the "Acknowledgement" from PMR from 2022, volume 123, no. 1 states that "*the authors thank to Marie Fayadová (Institute of Geochemistry, Mineralogy and Mineral Resources, Faculty of Science, Charles University, Prague) for her help with mineralization of samples".* In the same journal from 2012, volume 113, no. 2 there was not a section titled "Acknowledgements"; instead there was the following information between "Abstract" and "Introduction": *"this study was supported by grants: NS/9831-4 of the Internal Grant Agency of the Ministry of Health of the Czech Republic, CZ0123 from Norway through the Norwegian Financial Mechanisms, and NT 12342-5/2011 of the Ministry of Education, Youth and Sports of the Czech Republic"* However, considering the subject matter of given information, it was not essential for the purpose of this study.

The articles also included the date when the text was received and accepted and a particular department of each author – once again these inputs are not carrying required information. Every page of an article included the name of the journal together with volume number details etc. which were also eliminated. Therefore, the last part of each text that was kept was titled either "Conclusion" or "Discussion" depending on the article´s structure.

3.1.2 Creation of the Corpora

The downloaded articles were consequently sorted out into individual files according to the journal and then, depending on the year of publishing, into two separate periods. Therefore, each journal had two files, one containing data for the old period and the other one for the new period. Next, there was a division based on the year and each article was named in the same manner to keep the files organized. First, there is an abbreviation of the name of the journal, then the year of publication, volume number, issue number, and finally the starting page number. The example in the Figure 1 below shows the "new" file of PMR.

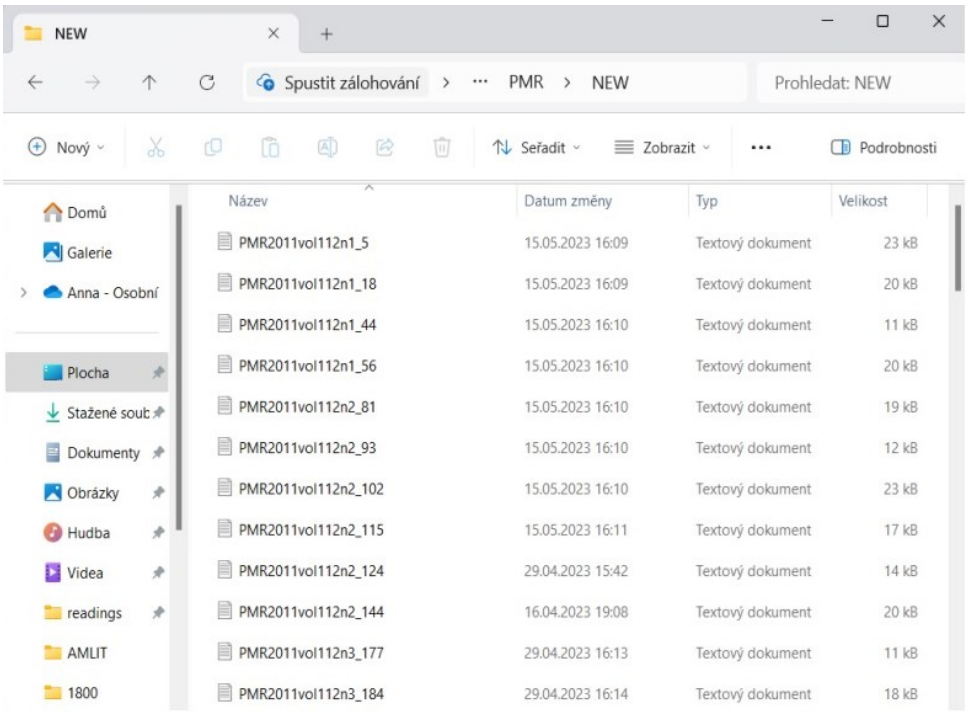

Figure 1 File Organization

In the Tables 6 and 7 below is an overview of the number of articles per each year and the total number of articles per year from all three journals as well as the number of tokens i.e. running words, to provide a better picture of what was worked with. Despite the inequality of available records from PMR within the old period, the final number of articles was not that distinct. The 2011-2022 period contains less articles due to the increased number of foreign authors, which could not be used as they did not fit with the criteria, and some numbers of articles may also differ because the journals had less volumes within particular years. However, the final results do not cause any disproportion. As the Table 5 shows, the period from 1998 to 2010 ended up having the total number of 533 articles with 1,187,667 running words and the period from 2011 to 2022 had 412 articles with 1,138,870 running words, ensuring there was enough material to continue with the analysis. Once this procedure was

over, it was possible to create a corpus. All of the articles were opened in *AntConc,* providing the information on running words and it was possible to begin with retrieving the LBs. The results along with their proper analysis will be discussed in the following analytical part of the thesis.

Journal	1998	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	N. of Articles	Tokens
AM	16	16	11	12	11	16	15	18	7	18	18	16	7	6	187	433304
BMP	0	14	19	16	6	18	19	17	17	21	5	4	22	18	196	384525
PMR	0	0	0	0	0	0	0	17	21	31	18	12	27	24	150	369838
Total	**16**	**30**	**30**	**28**	**17**	**34**	**34**	**52**	**45**	**70**	**41**	**32**	**56**	**48**	**533**	**1,187,667**

Table 5 Number of Collected Articles per Year and the Total Number and Running Words from 1998 to 2010

Journal	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	N. of Articles	Tokens
AM	11	13	11	8	13	10	8	7	6	6	8	5	106	281798
BMP	21	16	16	27	20	16	14	15	14	13	13	13	198	600777
PMR	19	14	10	5	15	13	6	4	7	4	6	5	108	256295
Total	**51**	**43**	**37**	**50**	**48**	**39**	**28**	**26**	**27**	**23**	**27**	**23**	**412**	**1,138,870**

Table 6 Number of Collected Articles per Year and the Total Number of Running Words from 2011 to 2022

4 Analytical Part

The analytical part of this bachelor thesis will provide the retrieved LBs from both corpora and their further analysis based on their structural and functional classification. The two corpora will be compared, and the aim is to point out any differences that may occur. Furthermore, the final results of the research, which focused on the difference between LBs from medical research articles from 1998-2010 and 2011-2022 and the possible influence of assistive writing tools, will be presented and discussed.

4.1 Results

The *AntConc* program, which was used as a tool for extracting LBs, was created by Laurence Anthony, professor at Waseda University in Japan (Laurence Anthony, n.d.) and is widely used for linguistic corpus analysis as it provides not only LBs lists but also other language patterns such as lemmas, clusters or collocates. The aim of this study was to examine four-word LBs, therefore in the N-Gram section, the N-Gram size was always four. However, several other criteria, such as frequency and range within the articles, had to be made to get the list of LBs. The frequency in each period differed based on the number of tokens. In the period from 1998 to 2010, the minimal frequency was twenty and the minimal range was ten.

This resulted in 311 LBs which is a fairly large number and therefore only the first one hundred LBs were used in order to maintain a reasonable number for analysis. Still, the one hundred bundles had to be further reduced due to the fact that a lot of the bundles were overlapping and therefore creating five-word or even six-word bundles which were not suitable for the analysis. Thanks to the KWIC (key word in context) option, it was possible to see the individual bundles within the sentences and hence identify the actual scope of the bundle. For example, *AntConc* identified LBs *the aim of our* and *of our study was* which turned out to be two much longer LBs – *the aim of our study was to* with frequency of 30 and *aim of our study was to* with the frequency of 31, which were not sufficient enough to be used, let alone were not four-word bundles. Such process eliminated nineteen bundles, which led to seventy-five remaining LBs to be analysed and compared to the new period.

The frequency for the period from 2011 to 2022 was set to 25 and the range was kept at the same number. Once again, the total number of bundles was large, 238, so the starting point was settled at one hundred LBs. After doing the same procedure of eliminating LBs which were in reality part of bigger word cluster, the final number of LBs ended up being 61. The most frequent LB to be eliminated due to its real length was *the aim of this study was to*

which also appeared in the previous period. It appeared with the frequency of 111, but *the aim of this* with a different following word than *study* appeared thirty-five times and the bundle *of this study was to* thirty-five times. Therefore, the final frequency of the individual bundles was not adequate, and they had to be deleted. However, another issue occurred in this period, as the *AntConc* program identified *Fisher´s exact test* as a four-word bundle, mistaking the possessive *´s* for a separate word.

In this period, there was also an occurrence of bundles mentioning patients such as *a total of patients, the group of patients, group of patients with, of patients treated with, patients with and without, a group of patients,* and *the number of patients.* In the previous period, no such bundles were identified which presented an interesting change as both corpora were created from the same journals and yet the *AntConc* program selected these bundles in the 2011-2022 period only. However, the elimination process showed that some of these bundles were either part of a bigger LB or there has been included a number in the word sequence, e.g. *group of 100 patients,* and these bundles could not be used. Therefore, only *the group of patients, patients with and without,* and *the number of patients* were kept in the final number of LBs. No bundles including the word *patient* were selected in the 1998-2010 period despite the fact that the total number of bundles was higher.

In both of these columns, there are some five-word bundles as well. The reason for this choice was the frequency of the bundles. In the 1998-2010 period, there can be seen both *at the end of* and *at the end of + (the).* Unlike the previously mentioned bundles which had to be replaced from the list for their lack of frequency, these bundles still provide an interesting pattern despite being out of the set range of words. The same case can be seen in the 2011-2022 period even with the same bundles, which does not point out anything particular as it can be coincidence, yet it is worth commenting on. Another case, where it could be debatable whether these bundles are two different or just a various realization of the same bundle, are with the singular and plural verbs such as *was found to be* and *were found to be.* nevertheless, due to the inconsistent frequency, these bundles were kept as separate.

In the Table 7 below, there is a comparison of the retrieved bundles along with their frequency. The matching bundles are highlighted in yellow color for better orientation within the list. Those bundles which appear within the same line are also in bold. Although the functional classification of bundles will be discussed further in the thesis, context or topic dependent LBs are marked as they are not as defining as the rest of the bundles. The majority

of them describes the location, such as Czech Republic or the university hospital or the proper names of procedures made within the research. Topic dependent bundles which match within the two periods are highlighted in dark orange color, while those which appear only in one of the periods are highlighted in a lighter color.

1998-2010			2011-2022		
	LB	F		LB	F
1.	**On the other hand**	304	1.	**On the other hand**	225
2.	**In the case of**	270	2.	**In the case of**	200
3.	**At the time of**	119	3.	**At the time of**	194
4.	In the course of	107	4.	In the Czech Republic	175
5.	In the Czech Republic	105	5.	In accordance with the	86
6.	As well as in	100	6.	Are shown in table	84
7.	It is necessary to	100	7.	At the same time	80
8.	As well as the	99	8.	In our study we	79
9.	In the group of	98	9.	The results of the	79
10.	On the basis of	96	10.	In the control group	77
11.	In the presence of	93	11.	In the presence of	75
12.	In comparison with the	90	12.	In the treatment of	73
13.	At the age of	87	13.	As a result of	70
14.	At the same time	86	14.	As well as in	69
15.	In the control group	78	15.	As well as the	68
16.	At the end of + (the)	77	16.	In the present study	63
17.	It is possible to	72	17.	On the basis of	63
18.	As a result of	69	18.	Was approved by the	61
19.	Of the dentate gyrus	68	19.	At the department of	60
20.	In the present study	66	20.	For the treatment of	60
21.	The results of the	63	21.	Is one of the + (most)	57
22.	An important role in	62	22.	At the end of	56
23.	Of the left ventricle	62	23.	Are summarized in table	55
24.	In the development of	60	24.	Was used for the	55
25.	The central nervous system	60	25.	Mann Whitney U test	54
26.	One of the most	58	26.	In the pathogenesis of	52

27.	Is based on the	57	27.	Of the university hospital	52
28.	In the number of	56	28.	It is necessary to	51
29.	At the beginning of + (the)	56	29.	In the course of	48
30.	At the end of	56	30.	The quality of life	48
31.	At the level of	53	31.	The total number of	48
32.	In the form of	52	32.	In the development of	47
33.	Is one of the	52	33.	In the form of	45
34.	Are shown in table	50	34.	Is shown in table	45
35.	The course of the	50	35.	With the use of	44
36.	The aim of the	50	36.	The group of patients	43
37.	In combination with atropine	49	37.	Patients with and without	43
38.	To be the most	49	38.	Is one of the	42
39.	Were found in the	49	39.	Of the Czech Republic	42
40.	And the number of	47	40.	With the exception of	42
41.	In accordance with the	46	41.	One of the most	41
42.	In the treatment of	46	42.	The time of diagnosis	41
43.	In the area of	45	43.	There was a significant	41
44.	The influence of the	45	44.	No significant difference in	40
45.	At the department of	44	45.	The declaration of Helsinki	40
46.	Hradec Králové Czech Republic	44	46.	The number of patients	40
47.	Dose of mg kg	43	47.	Was not statistically significant	40
48.	In the absence of	43	48.	At the end of + (the)	39
49.	One way anova test	43	49.	BMI body mass index	39
50.	The same as in	42	50.	For the development of	39
51.	A significant decrease in	39	51.	In the number of	39
52.	In our study we	39	52.	The fact that the	39
53.	In the pathogenesis of	39	53.	There were no significant	39
54.	It is important to	39	54.	Was found in the	39
55.	Was found in the	39	55.	No statistically significant differences	38

56.	Is one of the + (most)	39	56.	Sensitivity and specificity of	38
57.	For the treatment of	38	57.	Was found to be	38
58.	Has been shown to	38	58.	At the age of	38
59.	With the use of	38	59.	And the presence of	37
60.	The size of the	37	60.	In comparison with the	37
61.	For the detection of	36	61.	In the diagnosis of	37
62.	The total number of	36			
63.	There were no significant	36			
64.	Was used for the	36			
65.	With respect to the	36			
66.	The fact that the	36			
67.	The beginning of the	36			
68.	In rats exposed to	35			
69.	Was found to be	35			
70.	Are summarized in table	34			
71.	As a marker of	34			
72.	In relation to the	34			
73.	As a consequence of	33			
74.	Is considered to be	33			
75.	The results of our	33			

Table 7 Comparison of Retrieved LBs from the 1998-2010 and the 2011-2022 Corpora

The total number of matching bundles was 39 and they were: *on the other hand, in the case of, at the time of, in the course of, in the Czech Republic, as well as in, it is necessary to, as well as the, on the basis of, in the presence of, in comparison with the, at the age of, at the same time, in the control group, at the end of + (the), as a result of, in the present study, the results of the, in the development of, one of the most, in the number of, at the end of, in the form of, is one of the, are shown in table, in accordance with the, in the treatment of, at the department of, in our study we, in the pathogenesis of, was found in the, is one of the + (most), for the treatment of, with the use of, the total number of, there were no significant, was used for the, the fact that the, was found to be,* and *are summarized in table.*

Since the number of LBs differed within these two periods, the percentage of matching bundles had to be calculated for each period respectively. The 1998-2010 period showed that 52 % of bundles matched with the 2011-2022 period, while the 2011-2022 period resulted in

64 % match of bundles. This provides an interesting realization – the new period contains less LBs which naturally leads to higher percentage within the matching bundles, while the old period is the other way around. Therefore, the overall result is very similar. Another intriguing fact is that the first three bundles not only do match, but they also appear in the same order which indicates that these bundles remained the most used in both corpora. Those bundles are *on the other hand, in the case of* and *at the time of.* In the 1998-2010 period, *on the other hand* had a frequency of 304, while in the 2011-2022 period it was only 225. The same pattern was found with *in the case of* where the higher frequency was in the old period. *At the time of* on the other hand, was more frequent on the new period. It is important to keep in mind that the total number of articles differs and therefore the frequency cannot be directly compared. Interestingly, the first two bundles fall into the category of Text-oriented bundles, though the subcategory differs, while the third one belongs to Research-oriented bundles, according to the Hyland (2008). More detailed division into both functional and structural classification will be provided later on. The following section will focus on the significance of frequency within the matching bundles.

4.1.1 Comparison of the Frequency

The following step after identifying matching LBs was to determine whether their frequency difference was statistically significant. In order to get the results, an online calculator called *Corpus Frequency Test Wizard* was used. This website enables to examine either one sample or a comparison between two samples which was the aim of this research. As can be seen in the Figure 2 below, both frequencies had to be put in as well as the sample size, that is the number of running words per each corpus. The Figures 1 and 2 show an examination of the bundle *on the other hand,* resulting in statistically significant sample at p<.01.

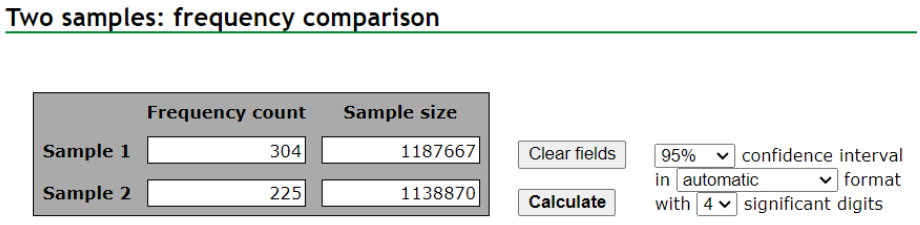

Figure 2 Example of the Corpus Frequency Comparison

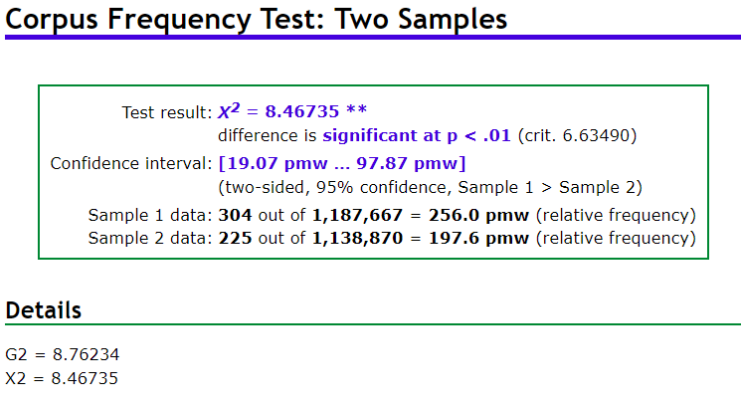

Figure 3 Example of the Corpus Frequency Test Result

The result of every comparison is shown in the following Table 8. Statistically significant difference is highlighted in yellow color and those bundles which turned out to be statistically significant have the higher frequency in bold. Although the first half indicates that the 1998-2010 period has more prominent frequency, the final results demonstrate ten dominant bundles in 1998-2010 and eight in 2011-2022. In general terms, the resulting numbers do not determine any major findings.

No.	Matching LBs	Frequency 1998-2010	Frequency 2011-2022	Significance	p-value
1.	On the other hand	**304**	225	Yes	p<.01
2.	In the case of	**270**	200	Yes	p<.01
3.	At the time of	119	**194**	Yes	p<.001
4.	In the course of	**107**	48	Yes	p<.001
5.	In the Czech Republic	105	**175**	Yes	p<.001
6.	As well as in	**100**	69	Yes	p<.05
7.	It is necessary to	**100**	51	Yes	p<.001
8.	As well as the	**99**	68	Yes	p<.05
9.	On the basis of	**96**	63	Yes	p<.05
10.	In the presence of	93	75	No	
11.	In comparison with the	**90**	37	Yes	p<.001
12.	At the age of	**87**	38	Yes	p<.001
13.	At the same time	86	80	No	
14.	In the control group	78	77	No	
15.	At the end of + (the)	**77**	39	Yes	p<.01

16.	As a result of	69	70	No	
17.	In the present study	66	63	No	
18.	The results of the	63	79	No	
19.	In the development of	60	47	No	
20.	One of the most	58	41	No	
21.	In the number of	56	39	No	
22.	At the end of	*56*	*56*	No	
23.	In the form of	52	45	No	
24.	Is one of the	52	42	No	
25.	In accordance with the	46	**86**	Yes	p<.001
26.	In the treatment of	46	**73**	Yes	p<.01
27.	At the department of	44	60	No	
28.	In our study we	39	**79**	Yes	p<.001
29.	In the pathogenesis of	39	52	No	
30.	Was found in the	*39*	*39*	No	
31.	Is one of the + (most)	39	57	No	
32.	For the treatment of	38	**60**	Yes	p<.05
33.	With the use of	38	44	No	
34.	The total number of	36	48	No	
35.	There were no significant	36	39	No	
36.	Was used for the	36	**55**	Yes	p<.05
37.	The fact that the	36	39	No	
38.	Was found to be	35	38	No	
39.	Are summarized in table	34	**55**	Yes	p<.05

Table 8 Comparison of the Frequency of Matching Lexical Bundles

Of the 39 matching lexical bundles, the following 18 turned out to be statistically significant: *on the other hand, in the case of, at the time of, in the course of, in the Czech Republic, as well as in, it is necessary to, as well as the, on the basis of, in comparison with the, at the age of, at the end of + (the), in accordance with the, in the treatment of, in our study we, for the treatment of, was used for the,* and *are summarized in table.* In percentage terms, that accounts for 46 %. The 1998-2010 period has a slightly higher agency of ten following bundles: *on the other hand, in the case of, in the course of, as well as in, it is necessary to, as well as the, on the basis of, in comparison with the, at the age of,* and *at the end of + (the).*

The 2011-2022 period included the remaining eight bundles which does not cause any major disproportions*: at the time of, in the Czech Republic, in accordance with the, in the treatment of, in our study we, for the treatment of, was used for the,* and *are summarized in table*. The highest difference in frequency was naturally found in the first lexical bundle, where the difference was 79. Within the remaining bundles whose distinctness was not classified as significant were two bundles with the same frequency. *At the end of* with frequency 56 and *was found in the* with frequency 39; both these numbers are in italics. Interestingly, those bundles which were not found to be significant, there was not any supremacy in any period either. The 2011-2022 period had ten bundles with higher frequency, while the 1998-2010 period had only nine. Once again, no major difference. Overall, based on these results, there was no significant discovery within the difference as the representation of statistically significant LBs was very similar in both periods. These results lead to a conclusion that there are no notable differences that could be caused by using assistive technology.

4.1.2 Structural Classification of Retrieved LBs

Once the test showing whether the bundles were statistically significant or not was run, the original 75 and 63 LBs were divided into categories based on their structural classification. This division was based on the research made by Biber et al. (2004) and the results are shown in Table 9 below.

Structure	1998-2010	2011-2022
Noun phrase with *of*-phrase	the course of the, **on the basis of, the results of the**, the aim of the, the influence of the, the size of the, **the total number of**, the beginning of the, the results of our, one of the most	**on the basis of, the results of the, the total number of**, the quality of life, the time of diagnosis, the declaration of Helsinki, the number of patients, the group of patients, one of the most, sensitivity and specificity of
Noun phrase with other post-modifier fragment	an important role in, **the fact that the**, a significant decrease in	**the fact that the**, no significant difference in, no statistically significant differences, patients with and without
Prepositional phrase with *of*-phrase	**in the case of, at the time of, in the course of**, in the group of, **in the presence of, at the age of, as a result of, in the development of, in the number of, at the end of, in the form of,**	**in the case of, at the time of, in the course of, in the presence of, at the age of, as a result of, in the development of, in the number of, at the end of, in the form of, in the treatment of, at the department**

	in the treatment of, in the area of, **at the department of**, in the absence of, **in the pathogenesis of, for the treatment of, with the use of**, as a marker of, as a consequence of, at the level of, for the detection of, at the beginning of + (the), **at the end of + (the)**	**of, in the pathogenesis of, for the treatment of, with the use of**, with the exception of, for the development of, in the diagnosis of, **at the end of + (the)**
Other prepositional phrase expressions	**on the other hand, as well as in, in the Czech Republic, in comparison with the, in the present study, in the control group**, of the dentate gyrus, of the left ventricle, in combination with atropine, **in accordance with the, in our study we**, with respect to the, in rats exposed to, in relation to the, **at the same time**	**on the other hand, as well as in, in the Czech Republic, in comparison with the, in the present study, in the control group, in accordance with the, in our study we**, of the university hospital, of the Czech Republic, **at the same time**
Verb phrase with passive verb	is based on the, **are shown in table**, were found in the, **was found in the**, has been shown to, **was used for the, was found to be, are summarized in table**, is considered to be	**are shown in table, was found in the, was used for the, was found to be, are summarized in table**, was approved by the, is shown in table
Pronoun + *be*	**there were no significant**	**there were no significant**, there was a significant
***Be* + noun/adjective**	**is one of the, is one of the + (most),** to be the most	**is one of the, is one of the + (most)**, was not statistically significant
Anticipatory *it*	**it is necessary to**, it is possible to, it is important to	**it is necessary to**
Other expressions	**as well as the**, the central nervous system, and the number of, Hradec Králové Czech Republic, dose of mg kg, one way anova test, the same as in, and the presence of	**as well as the**, BMI body mass index, mann whitney u test

Table 9 Structural Classification of Retrieved Lexical Bundles

As the Table 9 shows, the total of nine categories were identified within the two periods. When comparing those categories with the original division by Biber et al. (2004), only the verb phrase, noun phrase, and prepositional phrase fragments were identified, while the

dependent clause fragments did not appear. Due to the occurrence of several bundles that did not fit any of the original categories, new division were created such as those with anticipatory *it* based on Hyland´s research (2008) or *be* + noun/adjective. The remaining bundles that could not be placed in any original category were put into a category called *other expressions*; otherwise, it would lead to an excessive number of categories with little bundles in them. The matching bundles which appeared in both periods are in bold.

In the following Table 10, the total number of bundles in each period and category is presented, while the number of matching LBs is in parentheses. The adjacent column shows the percentage these bundles make from the original list of retrieved bundles. As can be seen, the percentage results are very similar to each other, showing more or less similar representation of LBs in each category given the fact that the starting number of bundles is different for each period. The category with the highest representation of bundles ended up being prepositional phrase with *of*-phrase in both corpora, containing 32 % in 1998-2010 period and 31 % in 2011-2022 period. The results of Hyland´s study (2008) presented the noun phrase with *of*-phrase fragment as the most common LB in all considered disciplines while the prepositional phrase with of-phrase was identified as the third most common in biology.

The second group with the highest representation was other prepositional phrase expressions which accounted for 20 % in 1998-2010 period and 18 % in 2011-2022 period. On the other hand, the category with the lowest representation was pronoun + *be* in 1998-2010, resulting in only 1 %, and anticipatory *it* in 2011-2022 which accounted for 2 %. Both these categories contain only one bundle in the particular period.

The category which differed by the most bundles was *other expressions* which is significantly higher in the 1998-2010 period, nevertheless as this category contains different types of bundles, this result does not provide any considerable distinction. Verb phrase with passive verb resulted in the same percentage in both periods. Regarding the number of bundles, noun phrase with *of*-phrase and *be* + noun/adjective consisted of the same numbers of bundles, although the percentage turned out to be different which is naturally due to the different number in each list of bundles. Nevertheless, the noun phrase with *of*-phrase was significantly higher and was placed as the third most common category.

Structure	Number of bundles (matching bundles)		Percentage	
	1998-2010	2011-2022	1998-2010	2011-2022
Noun phrase with *of*-phrase	10 (3)	10 (3)	13 %	16 %
Noun phrase with other post-modifier fragment	3 (1)	4 (1)	4 %	7 %
Prepositional phrase with *of*-phrase	24 (16)	19 (16)	32 %	31 %
Other prepositional phrase expressions	15 (9)	11 (9)	20 %	18 %
Verb phrase with passive verb	9 (5)	7 (5)	12 %	12 %
Pronoun + *be*	1 (1)	2 (1)	1 %	3 %
Be + noun/adjective	3 (2)	3 (2)	4 %	5 %
Anticipatory *it*	3 (1)	1 (1)	4 %	2 %
Other expressions	8 (1)	3 (1)	11 %	5 %

Table 10 Structural Classification of Retrieved Lexical Bundles and the Percentage Representation

In conclusion, the most frequent lexical bundles were those formed by prepositional phrase with *of*-phrase and as they were the most frequent ones in both corpora, it indicates that those bundles are highly used in medical research articles. In general, the number of bundles as well as the resulting percentage do not demonstrate any major differences and it can be therefore assumed that even if there were some assistive writing tools used in writing these articles, it had not any significant impact on their structural classification. Moreover, two categories contained the exact number of bundles and when looking closer to the remaining eight

categories, in the 1998-2010 period seven of them included more bundles. The only category in 2011-2022 period which held more bundles was pronoun + *be* which not only consisted of merely two bundles, but it was also a category which was created solely for the purpose of this study. This classification clearly indicates that the bundles in 1998-2010 period were much more frequent.

4.1.3 Functional Classification of Retrieved LBs

This section focuses on the division of lexical bundles based on their functional classification. The Table 11 below shows the identified functional classification of bundles in both corpora using the categories as presented by Hyland (2008). Each column includes bundles from the said corpus.

Category	**Subcategory**	**1998-2010**	**2011-2022**
Research-oriented	Location	at the time of, in the course of, in the control group, at the end of + (the), at the beginning of + (the), at the end of, at the level of, the course of the, the beginning of the, at the same time, in the present study	at the time of, in the course of, in the control group, at the end of + (the), at the end of, at the same time, the time of diagnosis, in the present study
	Procedure	an important role in, in the development of, the aim of the, in accordance with the, in the treatment of, with the use of, was used for the, for the treatment of, the influence of the	in the development of, in accordance with the, in the treatment of, with the use of, was used for the, was approved by the, for the development of, for the treatment of, in the diagnosis of
	Quantification	one of the most, in the number of, is one of the, to be the most, and the number of, a significant decrease in, is one of the + (most), the total number of, there were no significant, the size of the	one of the most, in the number of, is one of the, is one of the + (most), the total number of, there were no significant, there was a significant, the number of patients, no statistically significant differences, was not statistically significant, no significant difference in
	Description	the fact that the, as a marker of, is considered to be, the same as in	the quality of life, sensitivity and specificity of
	Topic	of the dentate gyrus, of the left ventricle, the central	in the Czech Republic, at the department of, in the

		nervous system, in combination with atropine, Hradec Králové Czech Republic, dose of mg kg, one way anova test, in the pathogenesis of, in rats exposed to, at the department of	pathogenesis of, mann whitney u test, of the university hospital, of the Czech Republic, BMI body mass index
Text-oriented	Transition signals	on the other hand, as well as in, as well as the, in comparison with the, at the same time, in accordance with the	on the other hand, as well as in, as well as the, in comparison with the, at the same time, in accordance with the
	Resultative signals	as a result of, were found in the, was found in the, was found to be, as a consequence, the results of our, in accordance with the	as a result of, was found in the, was found to be, in accordance with the
	Structuring signals	in the present study, are shown in table, in our study we, has been shown to, are summarized in table	in the present study, are shown in table, in our study we, is shown in table, are summarized in table
	Framing signals	in the case of, in the group of, in the presence of, at the age of, is based on the, in the form of, in the area of, in the absence of, with respect to the, in relation to the, for the detection of	in the case of, in the presence of, at the age of, in the form of, with the exception of, and the presence of, patients with and without, the group of patients
Participant-oriented	Stance features	it is possible to, the fact that the	
	Engagement features	it is necessary to, it is important to	it is necessary to

Table 11 Functional Classification of Retrieved Lexical Bundles

Several of the bundles were detected as fulfilling the criteria for more than one category based on the context they appeared in. All of the bundles appeared in both corpora. Those bundles are provided below together with all the identified categories and examples which were acquired from the KWIC function in *AntConc.* Each example is followed by appropriate abbreviation of the article they were found in. There were four such bundles: *in accordance with the, the fact that the, at the same time,* and *in the present study.* With the exception of *in accordance with the,* which turned out to be part of three different functional categories, all of them were identified as falling into two categories. *The fact that the* was the only bundle which appeared in the Participant-oriented category as it was part of a sentence where the

personal pronoun *I* plays an important role because it indicates the author´s intervention in the writing.

In accordance with the occurred as:

1. Research-oriented in the subcategory procedure: *"Treatment of animals was **in accordance with the** Declaration of Helsinki Guiding Principles on Care and Use of Animals (DHEW Publication, NHI 80–23)."* (AM2003vol46n4_153)
2. Text-oriented in the subcategory resultative signals: *"Our results of serum IL-2 and s IL-2R levels are **in accordance with the** above cited findings, these parameters not being able either to discriminate between the different severity of asthmatic symptoms or to differentiate atopic and nonatopic asthmatics."* (AM 1998vol40n3_61)
3. Text-oriented in the subcategory transition signals: *"**In accordance with the** facts described above and with the results of our previous study (6) we have found significantly higher level of neopterine in exposed group of welders and grinders, too."* (AM2003vol46_31)

The fact that the occurred as:

1. Research-oriented in the subcategory description: *"This is also reflected by **the fact that the** in number of RB in patients with isolated haematuria at our site was followed by a decrease in number and percentage of IgA nephropathies and an increase in TBM."* (AM2009vol52n4_141)
2. Participant-oriented in the subcategory stance features: *"I am aware of **the fact that the** HIV test must be performer twice and am willing to undergo a repeated blood sampling."* (AM2000vol43n4_139)

At the same time occurred as:

1. Research-oriented in the subcategory location: *"Our study was unique in two ways – first, it is a wide range of surface markers combined with humoral factors measured **at the same time**."* (AM2007vol50n4_229)
2. Text-oriented in the subcategory transition signals: *"Its employment represents two advantages: the forearm subcutaneous venous system itself is not used for the anastomosis and, **at the same time**, the connection with the deep venous system is broken."* (BMP2004vol148_85)

In the present study occurred as:

1. Research-oriented in the subcategory location: *"The last category of displacement behavior could not be evaluated because it did not occur at low doses of MA used* ***in the present study****."* (PMR2012vol113n3_223)
2. Text-oriented in the subcategory structuring signals: *"Therefore,* ***in the present study****, we were unable to demonstrate autonomic responses specific for the Vojta Therapy."* (BMP2018vol162n3_206)

Once the functional classification was done, it was possible to compare the two periods percentagewise. As the functional classification includes only three main categories, the final percentage was derived from each main category and not the subcategories. Each Table shows the percentage results of each category respectively as well as the total number of bundles that were found in the particular subcategories. The total number of bundles marked with yellow colour is higher than the original number of retrieved LBs since some of them were placed in multiple categories as discussed previously.

1998-2010					
Research-oriented		Text-oriented		Participant-oriented	
Location	11	Transition signals	6	Stance features	2
Procedure	9	Resultative signals	7	Engagement features	2
Quantification	10	Structuring signals	5		
Description	4	Framing signals	11		
Topic	10				
Number of bundles	44		29		4
Percentage	58,66 %		38,66 %		5,33 %

Table 12 The Total Number of LBs Based on Functional Classification and the Percentage Representation in 1998-2010 Period

2011-2022					
Research-oriented		Text-oriented		Participant-oriented	
Location	8	Transition signals	6	Stance features	0
Procedure	9	Resultative signals	4	Engagement features	1
Quantification	11	Structuring signals	5		
Description	2	Framing signals	8		
Topic	7				
Number of bundles	37		23		1
Percentage	60,65 %		37,70 %		1,64 %

Table 13 The Total Number of LBs Based on Functional Classification and the Percentage Representation in 2011-2022 Period

The Table 12 shows that over half of the bundles fit into the category of Research-oriented LBs, accounting for 58,66 %. The subcategory with the highest agency was Location which describes not only place but also time (Ibid 2008, 13) and it included 11 bundles such as *at the time of* or *in the present study.* Quantification and Topic subcategories resulted in the same number of 10 bundles. The Topic subcategory includes a lot of medical terms e.g. *of the dentate gyrus, of the left ventricle, the central nervous system* or *in combination with atropine* which is expected in medical research articles as they are a specific filed of research. They were followed by Procedure subcategory with 9 bundles. Regarding the nature of examined articles, procedure is a necessary part of every research, especially when discussing topics that include experiments of various forms. The subcategory with least bundles was Description which may come off as surprising since medicine is rather descriptive discipline.

Text-oriented category includes 38,66 % of bundles, the Framing signals subcategory which "situate arguments by specifying limiting conditions" (Ibid 2008, 14) having the highest number of 11 bundles. *In the case of, in the group of,* or *in the presence of* are some of the examples. The second were Resultative signals with 7 bundles which corresponds with the essence of research. The Structuring signals contained the least bundles which may be surprising considering the fact that the articles chosen for this analysis included a large number of tables or graphs which could be referred to.

The Participant-oriented category contains only 5,33 % with the same number of bundles in both Stance and Engagement features. This category is not a prominent as it provides a certain perspective of the authors or refers to the readers. As it was mentioned before, medicine is a descriptive field and therefore does not leave much space for such interventions and it is necessary to provide facts rather than assumptions. Stance features include *it is possible to* and *the fact that the,* while Engagement features include *it is necessary to* and *it is important to.* As can be seen, only *it is possible to* suggests certain level of speculation.

In the Table 13 the results were similar; the Research-oriented bundles ended up in the first place accounting for 60,65 %. In this case, the Quantification subcategory contained the most bundles, and they were e.g. *the number of patients, no statistically significant differences*, or *one of the most*. The remaining subcategories such as Procedure, Location, and Topic did not differ much, but interestingly, Description included only two LBs: *the quality of life* and *sensitivity and specificity of.* Compared to the Table 12, it differs only by two bundles, but still, it is not a considerable number taking into account the character of medical articles.

The Text-oriented category was the second in order, just like in Table 12, resulting in 37,70 % of LBs. The Framing signals subcategory was the most frequent as well as in the previous Table 11, however, the subcategory with lowest occurrence of bundles were Resultative signals. The difference between the rest of the subcategories was not as significant, since they differed by four bundles, yet it provides rather unexpected results. Overall, the number of bundles in Text-oriented category was lower than in Table 12.

The last category was once again Participant-oriented LBs with only 1,64 %. It was already apparent from the Table 11 that Participant-oriented bundles would result with the lowest percentage as the 2011-2022 period contained merely one LB. This result supports the argument that it is not ordinary for medical research articles to focus on the reader or the author´s attitude. The only bundle that occurred was *it is necessary to* in the Engagement features subcategory. Comparing these results with Hyland´s findings (2008), the Participant-oriented category was the least frequent in all disciplines, including Biology which is the closet to medical research articles.

4.1.4 Structural Classification within the Functional Classification

The separate identification of functional and structural classification of LBs could be sufficient enough, nevertheless, further comparison was made between these two categories.

All three divisions of functional categories were additionally defined based on the structural classification to see if there was any consistent pattern, and these results were subsequently compared with Hyland´s findings (2008). Needless to say, his research (2008) focused on several different disciplines and therefore it cannot serve as a general criterion, but rather as a chosen subject for the comparison for this particular thesis. He found out that the majority of Research-oriented bundles were realized by noun phrase with *of* structure (Ibid 2008, 14), which did not apply for neither of the two corpora as both of them resulted in having the most bundles realized by prepositional phrase with *of* phrase. In the 1998-2010 period, there were 14 bundles while in the 2011-2022 period, there were 13 bundles. The second most frequent structure were other prepositional phrase expressions in both corpora.

According to Hyland (2008), the Text-oriented category was formed mostly by prepositional phrase with *of* bundles*,* which applied to a certain level for the 1998-2010 corpora as well as the prepositional phrase with *of* and other prepositional phrase expressions were the most common since both these categories included ten bundles. In the 2010-2020 corpora, other prepositional phrase expressions were the most frequently used, forming eight LBs. Although this result did not comply with the Hyland´s (2008), in general it was still a prepositional phrase structure, showing that this was the most frequent one.

The most frequent bundles in the Participant-oriented category had the same structural pattern in both corpora, corresponding with the previous results (Ibid 2008, 14), and it was anticipatory *it.* In the 1998-2010 period, three of four bundles were realized by this structure, and since the 2011-2022 period included only one, which also matched with the previous period, it was the only possible outcome.

The goal of these two sections was to identify and compare the functional classification of LBs in both corpora based on the template described in the theoretical part by Hyland (2008) and further determine their structure. As shown in Tables 12 and 13, this resulted in the Research-oriented category being the most frequent in both corpora and in both cases, it formed over the half of the bundles. In 1998-2010 corpus, it accounted for 58,66 % while in the 2011-2022 corpus it accounted for 60,65 %. Even if the number of bundles in each corpus was the same, it would not illustrate any major differences. The second most frequent was the Text-oriented category – in the 1998-2010 corpus it formed 38,66 % and in the 2011-2022 corpus 37,70 %. Once again, no significant differences were discovered, on the other hand, it is interesting that the categories were so similar within these two corpora. Even when

considering the fact that the subcategories differed slightly in representation, it did not cause any notable distinction. The least represented was the Participant-oriented category accounting for 5,33 % and 1,64 % in each corpus respectively which was not surprising, taking into account the aim of medical research articles and how they approach the topic and the readers´ audience. Interestingly, on both corpora, several bundles were identified as having more than one function.

Additionally, the LBs in functional categories were identified based on their structure. The Research-oriented category was mostly realized by prepositional phase with *of* structure in both corpora. In the 1998-2010 period, there were 14 bundles with this structure, accounting for 31,81 %, while in the 2011-2022 period there were 13 bundles, therefore forming 35,13 %. These findings were compared with those made by Hyland (2008) who discovered that the noun phrase with *of* was the most frequent. However, when looking into other studies cited in the theoretical part, such as the one carried out by Arani et al. (2015), prepositional phrases were the most frequently used, including both prepositional phrase with *of* and other prepositional phrase expressions. This study focused on medical research articles in various areas, providing the same focus as the one of this thesis, and therefore a better subject of comparison.

The most frequent structure in the Text-oriented category in the 1998-2010 corpus were prepositional phrase with *of* and other prepositional phrase expressions, while in the 2011-2022 corpus it was other prepositional phrase expressions. These results show that in general terms, the prepositional phrase was the most frequently used one in both Research and Text-oriented category and even in both corpora, supporting the results of Arani et al. (2015).

5 Conclusion

The aim of this thesis was to identify four-word lexical bundles in medical research articles written by non-native English speakers with respect to the rising usage of assistive writing tools. In order to make a comparison, two corpora covering different time periods were created. The first corpus started in 1998 and ended in 2010, which was chosen as the parting point. The second corpus continued in the following year and ending in 2022, as the current year 2023 did not have a sufficient number of articles. Furthermore, in each corpus, the frequency, function and structure of LBs were identified and compared. The hypothesis set at the beginning of the research was that in the later corpus, four-word LBs will occur with higher frequency, but different functional use since the assistive writing tools do not consider the context. The thesis aimed to answer the following questions:

1. What are the most frequent four-word LBs in the 1998-2010 corpus and in the 2011-2022 corpus?
2. What is the structure of four-word LBs in the 1998-2010 corpus and in the 2011-2022 corpus?
3. What functions these four-word LBs have in the 1998-2010 corpus and in the 2011-2022 corpus?
4. Are there any significant differences in frequency, function and structure within the two corpora?

The first step was to select those articles which were written by non-native speakers and create two corpora with sufficient number of running words. Each corpus ended up having over one million running words, which ensured that the results would be substantial. Next, LBs were identified with the help of the *AntConc* program, which resulted in 311 LBs in the 1998-2010 period and 238 LBs in the 2011-2022 period. As the total number of bundles was too high, only the first one hundred LBs were considered in this research. Subsequently, this number was lowered, as some of the bundles turned out to be either part of a bigger LB or contained a numeral. In the end, 75 and 61 LBs respectively were identified and further analysed.

The first question was immediately answered, as the most frequent LB in each corpus was *on the other hand* with frequency of 304 in the 1998-2010 corpus and 225 in the 2011-2022 corpus. Not only there was no difference between the most frequent bundle, but also the next two bundles matched as well. They were *in the case of* in the second place and *at the time of* in the third. This did not show any differences, quite the opposite as the first three bundles

were the same. Although the frequency differed in each corpus it is important to keep in mind that each corpus had different number of running words as well as the number of articles.

Once the bundles were identified, those which matched in both corpora were detected. The total number of matching bundles was 39, accounting for 52 % in the 1998-2010 corpus and 64 % in the 2011-2022 corpus. It is apparent that the later corpus resulted in higher percentage which is a result of different number of LBs. However, in both corpora the matching bundles formed over half of the total number. The next step was to determine whether the frequency of matching bundles showed any significant difference. An online *Corpus Frequency Test Wizard* was used which enabled the comparison between the two corpora. Of the 39 LBs, only 18 proved to be statistically significant, accounting for 46 %. Therefore, not even half of them showed considerable distinction in frequency. The first nine bundles were all statistically significant and interestingly, seven of them were more frequent in the 1998-2010 corpus, including *on the other hand* and *in the case of,* which were the two most frequent ones in both corpora. Overall, LBs with higher frequency were found in the 1998-2010 corpus, although the difference was not crucial. However, this finding disproves the hypothesis that the LBs in the newer corpus will occur with higher frequency. When looking at the total number of that were identified, the frequency was undoubtedly higher in the 1998-2010 corpus.

The next goal was to identify the structural classification of the 75 and 61 LBs based on the division made by Biber et al. (2004). Out of the three original categories, only two of them were realized in both corpora. Verb phrase fragments and noun phrase and prepositional phrase fragments occurred, while the dependent clause fragments were not identified. Due to the wide range of LBs, two categories were added, such as anticipatory *it* by Hyland (2008) or *be* + noun/adjective. As soon as the bundles were organized, it was possible to calculate their percentage representation. In the 1998-2010 corpus, 24 bundles were realized by the prepositional phrase with *of* phrase, accounting for 32 %. In the 2011-2022 corpus, the same structural classification was the most frequent, having 19 bundles and therefore accounting for 31 %. Some of the bundles identified as having this structure are *at the time of, in the course of* or *in the group of.* The second most frequent category in both corpora was other prepositional phrase expressions which included bundles like *on the other hand, in the present study* or *in the control group*. In the 1998-2010 corpora, 15 bundles were identified as such, resulting in 20 %, while the second corpus contained 11 bundles and therefore 18 %. The most contrasting categories were prepositional phrase with *of* phrase and other expressions;

both categories differed by 5 bundles. Despite the fact that the structural classification does not provide any major differences, the results are not any less significant. They present that despite the fact that the articles were written in different times, the prepositional phrase fragments remain the most frequently used in medical research articles.

The following question deals with the functional classification of LBs. The same categories that Hyland (2008) identified were used. Both corpora had over 50 % of Research-oriented bundles. To be more specific, the 1998-2010 corpus accounted for 58,66 % while the later for 60,65 %. Once again, these results do not display any major differences. Text-oriented category was the second most frequent and the percentage results were 38,66 % and 37,70 % respectively. The least frequent category was Participant-oriented one, which corresponds with the nature of medical research articles and the limited interaction between the writer and the reader. The hypothesis claimed that the 2011-2022 corpus would result in different functional use of LBs, as the assistive writing tools do not consider the context and therefore cannot estimate the correct usage of LBs. However, this was not proved as the function did not differ. Moreover, LBs in both corpora turned out to be suitable for more than one category. These bundles were *in accordance with the, the fact that the, at the same time,* and *in the present study.*

To answer the last question of the thesis and conclude the overall results of the research, despite some minor differences in percentage results, there were no significant differences considering functional or structural classification of LBs. When considering the frequency, the 1998-2010 corpus was more prevalent, but the comparison of frequency within the matching bundles did not show any significant differences. Therefore, no new discoveries were made. The comparative analysis showed, that although the LBs were different, over half of them were matching and both the structure and function correspond within the two corpora. Therefore, the rising presence of assistive writing tools did not project into the use of LBs. As it was mentioned, the potential changes could not be directly associated with the use of such tools, but as their presence is so common nowadays, their occurrence cannot be eliminated either. The results of this thesis show that the function and structure remain more or less the same over the years, while the frequency was found to be slightly lower in the present days.

References

Akbulut, F. D. (2020). A Bibliometric Analysis of Lexical Bundles Usage in Native and Non-native Academic. *Journal of Language and Lingusitic Studies, 16*(3). 1146-1166. https://doi.org/10.17263/jlls.803583

Altenberg, B. (1998). On the Phraseology of Spoken English: The Evidence of Recurrent Word-combinations. In A.P. Cowie (Ed.), *Phraseology. Theory, Analysis, and Applications* (pp. 101-122). Oxford University Press.

Anthony, L. (2023). AntConc (Version 4.2.0.0.) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Biber, D., & Barbieri, F. (2007). Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes, 26i*(3), 263-286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Cortes, V. (2004). "If You Look at … Lexical Bundles in University Teaching and Textbooks." *Applied Linguistics, 25*(3), 371-405. https://doi.org/10.1093/applin/25.3.371

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

Cambridge Dictionary. (n.d.) *Jargon.* *https://dictionary.cambridge.org/dictionary/english/jargon*

Cortes, V. (2004). Lexical bundles in Published and Student Disciplinary Writing: Examples from History and Biology. *English for Specific Purposes*. *23*(4). 397-423. https://doi.org/10.1016/j.esp.2003.12.001

Dictionary.com. (n.d.). *About.* https://www.dictionary.com/e/about/

Dontcheva-Navratilova, O. (2012). Lexical Bundles in Academic Texts by Non-native speakers. *Brno Studies in English, 38*(2), 37-58. https://digilib.phil.muni.cz/handle/11222.digilib/126942.

Ebeling, S., & Hasselgård, H. (2015). Learner corpora and phraseology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 207-230). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.010

Ellis, N. C. (2002). Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C., Simpson-Vlach, R. & Maynard, C. (2008). Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375-396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Estaji, M., & Montazeri, M. R. (2022). Native English and Non-native Authors' Utilisation of Lexical Bundles: A Corpus-Based Study of Scholarly Public Health

Papers. *Southern African Linguistics and Applied Language Studies,* 40(2), 177-199. https://doi.org/10.2989/16073614.2022.2043169

Gass, S. M., & Mackey, A. (2002). Frequency Effects and Second Language Acquisition: A Complex Picture?. *Studies in Second Language Acquisition*, *24*(2), 249–260. http://www.jstor.org/stable/44486616

Grammarly. (n.d.). *About Us.* https://www.grammarly.com/about

Granger, S. (2008). Learner Corpora. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook. Volume 1* (pp. 259-275). Walter de Gruyter.

Hyland, K. (2008a). As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes, 27*(1), 4-21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K. (2012). Bundles in Academic Discourse. *Annual Review of Applied Linguistics, 32*, 150-169. https://doi.org/10.1017/S0267190512000037

Ismael, K. O., Saeed, K., A., Ibrahim, A. S., & Fatah, D. S. Effects of Auto-Correction on Students' Writing Skill at Three Different Universities in Sulaimaneyah City. *Arab World English Journal (AWEJ) Special Issue on CALL, 8*, 231-245. https://dx.doi.org/10.24093/awej/call8.16

Jalali, Z. S., Moini, M. R., & Arani, M. A. (2015). Structural and Functional Analysis of Lexical Bundles in Medical Research Articles: A Corpus-Based Study. *International Journal of Information Science and Management*, *13*(1), 51-69.

Kjellmer, G. (1991). A Mint of Phrases. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 112-127). Longman.

Laurence Anthony (n.d.). *Contact.* https://www.laurenceanthony.net/contact.html

MacArthur, Ch. (2000). New Tools for Writing: Assistive Technology for Students with Writing Difficulties. *Topics in Language Disorders*, *20*(4), 85-100. https://doi.org/10.1097/00011363-200020040-00008

Méndez-Cendón, B. & López-Arroyo, B. (2003). Intralinguistic Analysis of Medical Research Papers and Abstracts: Rhetorical and Phraseological Devices in Scientific Information. *Terminology*, *9*(2), 247-268. https://doi.org/10.1075/term.9.2.06men

Méndez-Cendón, B. & López-Arroyo, B. (2007). Describing Phraseological Devices in Medical Abstracts: An English/Spanish Contrastive Analysis. *Meta*, *52*(3), 503-516. https://doi.org/10.7202/016735ar

Paquot, M., & Granger, S. (2012). Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics, 32,* 130-149. https://doi.org/10.1017/S0267190512000098

Pawley, A., & Syder, F. H. (1983). *Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency.* In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication* (pp. 191-225). Longman.

Scott, M. (2007). *WordSmith Tools version 5.0*. Lexical Analysis Software. https://is.muni.cz/el/1421/podzim2007/NJII_1369/WordSmith.pdf

Shin, Y. K., & Kim, Y. (2017). Using Lexical Bundles to Teach Articles to L2 English Learners of Different Proficiencies. *System, 69*, 79-91. https://doi.org/10.1016/j.system.2017.08.002

Stubbs, M. (2007a). An Example of Frequent English Phraseology: Distribution, Structures and Functions. In R. Facchinetti (Ed.), *Corpus Linguistics 25 Years on* (pp. 89–105). Radopi.

Wood, D. (2015). *Fundamentals of Formulaic Language: An Introduction.* Bloomsbury Academic.

Resumé

Tato bakalářská práce se zabývá čtyřslovnými lexikálními svazky v lékařských odborných článcích, které jsou psány nerodilými mluvčími anglického jazyka, ve dvou časových obdobích. Tato období jsou rozdělena v závislosti na nástupu asistenčních nástrojích, které zjednodušují psaní v cizím jazyce a mají za účel pomoci nejen nerodilým mluvčím tvořit souvislý text, který svou kvalitou a obsahem co nejvíce odpovídá práci rodilého mluvčího. V úvodní části je stručně představen formulaický jazyk, jeho jednotlivé části jakožto i předmět této práce, čímž jsou lexikální svazky. Současně je vytyčen cíl práce, kterým je tvorba dvou korpusů, ve kterých budou následně identifikované čtyřslovné lexikální svazky, četnost jejich výskytu a jak funkční, tak strukturální klasifikace.

Teoretická část bakalářské práce nejprve detailně popisuje termín formulaický jazyk a následně frazeologii, která se nejčastěji užívá v lékařských odborných článcích. Zohledněn je přístup rodilých i nerodilých mluvčích a oba jsou podloženy již existujícími studiemi. Další kapitola se podrobně věnuje pouze lexikálním svazkům. Jako první je představena jejich obecná funkce v psaném jazyce a co je nejčastěji předmětem studií, které se jim věnují. Stejně tak jsou vymezeny určité problémy, které mohou při jejich identifikaci nastat, například překrývání lexikálních svazků (overlapping of lexical bundles). Jedná se o případy, kdy jsou čtyřslovné lexikální svazky ve skutečnosti součástí více slovného svazku, a tím pádem nemohou být použity k analýze.

Následující podkapitoly definují lexikální svazky na základě jejich funkční a strukturální klasifikace, které byly určeny ve studiích předchozích let. Následující kapitola stručně definuje tzv. žákovské korpusy (learner corpora), které poskytují databázi textů od nerodilých mluvčích. Předposlední kapitola popisuje styl psaní v lékařských odborných článcích a podrobněji i užití lexikálních svazků v těchto publikacích. Opět se opírá o již existující studie. Poslední kapitola definuje asistenční nástroje, k čemu jsou využívány a jejich vliv na psaní nerodilých mluvčích.

Po teoretické části následuje kapitola popisující metodu výzkumu a použité materiály a zároveň vymezuje pracovní hypotézu, tedy že lexikální svazky se budou vyskytovat častěji v období od roku 2011 do roku 2022 právě díky použití asistenčních nástrojů, ale zároveň se bude lišit jejich funkce, vzhledem k tomu, že tyto nástroje nerozpoznávají kontext, a otázky, které je cílem zodpovědět:

1. Jaké jsou nejčastější čtyřslovné lexikální svazky v korpusu v letech 1998-2010 a 2011-2022?
2. Jaká je struktura čtyřslovných lexikálních svazků v korpusu v letech 1998-2010 a 2011-2022?
3. Jaké funkce plní tyto čtyřslovné lexikální svazky v jednotlivých korpusech?
4. Vyskytují se nějaké signifikantní rozdíly v frekvenci, funkci a struktuře v jednotlivých korpusech?

Jak již bylo řečeno, bakalářská práce využívá k výzkumu lékařské odborné články, jejichž autory jsou nerodilí mluvčí. Proto byly zvoleny tři časopisy publikované českými univerzitami, kam přispívají právě samotní studenti: *Acta Medica* (AM) lékařské fakulty Univerzity Karlovy v Hradci Králové, *Biomedical Papers* (BMP) lékařské fakulty Univerzity Palackého Olomouc v Olomouci a *Prague Medical Report* (PMR) 1.lékařské fakulty Univerzity Karlovy. Všechny časopisy poskytují online databázi všech publikovaných článků, které byly staženy a dle příjmení autorů a školy, kterou studují, byly vybrány takové, jejichž autory jsou nerodilí mluvčí. Následně byly převedeny z formátu pdf do formátu txt, aby bylo možné provést další úpravy. Z každého článku byly odstraněny nedůležité údaje, například soubor použité literatury či poděkování. Konečné množství článků bylo 533 v období od roku 1998 do roku 2010 a 412 v období od roku 2011 do roku 2022. Toto rozdělení bylo určeno na základě asistenčních nástrojů, které byly od roku 2011 a dále poměrně běžnou součástí jakéhokoli uživatele internetu.

Následně bylo možné vytvořit dva korpusy, které obsahovaly přes jeden milion slov, a zároveň přejít k analytické části. K tvorbě korpusů byl použit program *AntConc,* který umožňuje identifikaci nejen lexikálních svazků. Celkem bylo identifikováno 311 lexikálních svazků v korpusu 1998-2010 a 238 v korpusu 2011-2022. Z každého korpusu však bylo vybráno pouze prvních sto, z nichž byly některé eliminovány např. kvůli překrývání lexikálních svazků či chybně určenému čtyřslovnému svazku. Konečný počet je tedy 75 a 61. Současně s lexikálními svazky je uvedena i frekvence jejich výskytu. Nejčastěji se v obou korpusech objevoval lexikální svazek *on the other hand* s frekvencí 304 ve starším časovém období a s frekvencí 225 v současném. Následovaly svazky *in the case of* a *at the time of,* které byly ve stejném pořadí v obou korpusech. Tento výsledek zodpověděl první otázku.

Poté byly vyznačeny lexikální svazky, které se objevily v obou korpusech a díky online programu *Corpus Frequency Test Wizard* bylo možné určit, zdali je rozdíl ve frekvenci statisticky signifikantní. Celkem bylo určeno 39 lexikálních svazků, které se shodovaly v obou korpusech, z nichž se prokázalo 18 jako statisticky signifikantní. To odpovídá 46 %.

Další podkapitola se věnuje strukturální klasifikaci, která se opírá o tu, kterou představil ve své studii Biber et al. (2004). V obou korpusech se ukázaly jako nejčastější struktury tvořené předložkovými frázemi s *of* frází. V korpusu 1998-2010 tvořily předložkové fráze s *of* frází 32 %, zatímco v korpusu 2011-2022 to bylo 31 %. Hned na druhém místě, opět v obou korpusech, šlo o předložkové fráze s jinými výrazy. Obecně lze tedy říci, že předložkové fráze tvoří největší část strukturální klasifikace. Příkladem této struktury jsou například svazky *with the use of, with the exception of* nebo *for the development of.*

Následující podkapitola se soustředí na funkční klasifikaci lexikálních svazků, která se shoduje Hylandovou (2008). Kategorie s největším zastoupením lexikálních svazků byla zaměřena na výzkum (Research-oriented) a v korpusu 1998-2010 tvořila 58,66 %, zatímco v korpusu 2011-2022 to bylo 60,65 %, tedy o něco vyšší četnost. Zároveň je ale nutné brát v potaz rozdílný počet lexikálních svazků v obou korpusech. Stále se však nejedná o zásadní rozdíl, obzvlášť když obě skupiny tvořily více než 50 %. Druhá nejčastější kategorie se zaměřuje na text (Text-oriented) a výsledky byly opět velice podobné – 38,66 % a 37,70 %. Nejméně zastoupená byla kategorie zaměřující se na účastníky (Participant-oriented), tedy samotné čtenáře. Vzhledem k povaze odborných lékařských článků není toto zjištění neočekávané; autoři se soustředí na fakta a konkrétní výsledky a čtenářům pouze předkládají tyto informace. Navíc bylo zjištěno, že v obou korpusech se vyskytly lexikální svazky, které nejen že spadaly do více než jedné kategorie, ale zároveň šlo a lexikální svazky, které se vyskytovaly v obou časových obdobích.

Na základě těchto výsledků je možné určit, že nejčastější čtyřslovné lexikální svazky se v prvních třech případech shodovaly v obou korpusech a jedná se o *on the other hand, in the case of* a *at the time of.* Co se týče strukturální klasifikace, oba korpusy poskytly stejné výsledky, a to předložkové fráze. Ani srovnání funkční klasifikace nepřineslo žádné signifikantní rozdíly, oba korpusy byly více než polovinou tvořeny kategorií zaměřující se na výzkum. Přestože se úvodní hypotéza nepotvrdila, výsledky představují zajímavý obraz toho, že ani přes zvýšenou přítomnost a užívání asistenčních nástrojů nebyla zásadně ovlivněna frekvence, struktura a ani funkce užitých lexikálních svazků. Naopak byly výsledky velice podobné a přestože je nelze přímo spojit s asistenčními nástroji, jejich přítomnost je v dnešní době nevyhnutelná a nelze je tedy ani kompletně oddělit.

Appendices

Appendix 1: The List of Lexical Bundles from the 1998-2010 Corpus and Their Frequency

No.	Lexical Bundle	Frequency
1.	On the other hand	304
2.	In the case of	270
3.	At the time of	119
4.	In the course of	107
5.	In the Czech Republic	105
6.	As well as in	100
7.	It is necessary to	100
8.	As well as the	99
9.	In the group of	98
10.	On the basis of	96
11.	In the presence of	93
12.	In comparison with the	90
13.	At the age of	87
14.	At the same time	86
15.	In the control group	78
16.	At the end of + (the)	77
17.	It is possible to	72
18.	As a result of	69
19.	Of the dentate gyrus	68
20.	In the present study	66
21.	The results of the	63
22.	An important role in	62
23.	Of the left ventricle	62
24.	In the development of	60
25.	The central nervous system	60
26.	One of the most	58
27.	Is based on the	57
28.	In the number of	56
29.	At the beginning of + (the)	56

30.	At the end of	56
31.	At the level of	53
32.	In the form of	52
33.	Is one of the	52
34.	Are shown in table	50
35.	The course of the	50
36.	The aim of the	50
37.	In combination with atropine	49
38.	To be the most	49
39.	Were found in the	49
40.	And the number of	47
41.	In accordance with the	46
42.	In the treatment of	46
43.	In the area of	45
44.	The influence of the	45
45.	At the department of	44
46.	Hradec Králové Czech Republic	44
47.	Dose of mg kg	43
48.	In the absence of	43
49.	One way anova test	43
50.	The same as in	42
51.	A significant decrease in	39
52.	In our study we	39
53.	In the pathogenesis of	39
54.	It is important to	39
55.	Was found in the	39
56.	Is one of the + (most)	39
57.	For the treatment of	38
58.	Has been shown to	38
59.	With the use of	38
60.	The size of the	37
61.	For the detection of	36
62.	The total number of	36

63.	There were no significant	36
64.	Was used for the	36
65.	With respect to the	36
66.	The fact that the	36
67.	The beginning of the	36
68.	In rats exposed to	35
69.	Was found to be	35
70.	Are summarized in table	34
71.	As a marker of	34
72.	In relation to the	34
73.	As a consequence of	33
74.	Is considered to be	33
75.	The results of our	33

Appendix 2: The List of Lexical Bundles from the 2011-2022Corpus and Their Frequency

No.	**Lexical Bundle**	**Frequency**
1.	On the other hand	225
2.	In the case of	200
3.	At the time of	194
4.	In the Czech Republic	175
5.	In accordance with the	86
6.	Are shown in table	84
7.	At the same time	80
8.	In our study we	79
9.	The results of the	79
10.	In the control group	77
11.	In the presence of	75
12.	In the treatment of	73
13.	As a result of	70
14.	As well as in	69
15.	As well as the	68
16.	In the present study	63

17.	On the basis of	63
18.	Was approved by the	61
19.	At the department of	60
20.	For the treatment of	60
21.	Is one of the + (most)	57
22.	At the end of	56
23.	Are summarized in table	55
24.	Was used for the	55
25.	Mann Whitney U test	54
26.	In the pathogenesis of	52
27.	Of the university hospital	52
28.	It is necessary to	51
29.	In the course of	48
30.	The quality of life	48
31.	The total number of	48
32.	In the development of	47
33.	In the form of	45
34.	Is shown in table	45
35.	With the use of	44
36.	The group of patients	43
37.	Patients with and without	43
38.	Is one of the	42
39.	Of the Czech Republic	42
40.	With the exception of	42
41.	One of the most	41
42.	The time of diagnosis	41
43.	There was a significant	41
44.	No significant difference in	40
45.	The declaration of Helsinki	40
46.	The number of patients	40
47.	Was not statistically significant	40
48.	At the end of + (the)	39
49.	BMI body mass index	39

50.	For the development of	39
51.	In the number of	39
52.	The fact that the	39
53.	There were no significant	39
54.	Was found in the	39
55.	No statistically significant differences	38
56.	Sensitivity and specificity of	38
57.	Was found to be	38
58.	At the age of	38
59.	And the presence of	37
60.	In comparison with the	37
61.	In the diagnosis of	37