UNIVERZITA PALACKÉHO V OLOMOUCI

Filozofická fakulta

Katedra anglistiky a amerikanistiky

Comparative Study of Lexical Bundles in Medical Research Articles Published in English by Czech Authors

Diplomová práce

Autor: Bc. Barbora Svobodová

Vedoucí práce: Mgr. Michal Kubánek

Olomouc 2021

UNIVERZITA PALACKÉHO V OLOMOUCI

Filozofická fakulta

Katedra anglistiky a amerikanistiky

Comparative Study of Lexical Bundles in Medical Research Articles Published in English by Czech Authors

Srovnávací studie lexikálních svazků v anglicky psaných lékařských odborných článcích publikovaných českými autory

Diplomová práce

Autor: Bc. Barbora Svobodová

Vedoucí práce: Mgr. Michal Kubánek

Olomouc 2021

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

V Olomouci dne 19. srpna 2021 .,……..

Bc. Barbora Svobodová

Poděkování

Ráda bych poděkovala vedoucímu diplomové práce Mgr. Michalovi Kubánkovi za jeho vstřícnost, ochotu a za jeho cenné rady. Taky děkuji své rodině a přátelům za jejich lásku a podporu.

Abbreviations

AM – Acta Medica

BP – Biomedical Papers

CEJPH – Central European Journal of Public Health

CIA – Contrastive Interlanguage Analysis

CZMRAC – Czech Medical Research Article Corpus

ENMRAC – English Medical Research Article Corpus

ELF – English as Lingua Franca

FL – Formulaic language

L1 – First language

L2 – Second language

PMR – Prague Medical Report

PR – Physiological Research

Table of Contents

1 Introduction

One of the first scholars who drew attention to the fact that language is not composed of isolated components but rather of multiword chunks was John Hughlings Jackson in the mid-nineteenth century (Wray 2002, 7). Ever since then, this linguistic phenomenon also known as formulaic language has been widely studied. Wray and Perkins (2000, 1) describe formulaic language as "a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar." Some of the linguists who have investigated the nature of multiword sequences are Pawley and Syder (1983), Wray and Perkins (2000), Wray (2002), Barlow (2011), Paquot and Granger (2012), or Wood (2019), however, there are many more.

A common type of multiword sequences are lexical bundles. According to Hyland (2012, 150), lexical bundles are "statistically the most frequent recurring sequences of words in any collection of texts". They are present in both spoken and written discourses (Biber et al. 1999, Cortes 2004, Biber and Barbieri 2007, Hyland 2012, Hyland and Jiang 2018) and are an essential component in the text production regardless the field of expertise. For example, one of the genres that offers potential for investigation, are medical research articles because there has not been a large volume of studies that focus on lexical bundles in this specific genre. One of the few studies that dealt with this topic was a master's thesis by Mbodj-Diop (2016) that was submitted to Michigan State University. Mbodj-Diop focused on the structures and functions of lexical bundles in the medical research articles published by five international medical journals – *Science, The Lancet, The New England Journal of Medicine, Journal of American Medical Association, The Journal of Clinical Investigation*. Her findings have shown that lexical bundles play an important role in medical writing. In her corpus, she identified all lexical bundle structures as introduced by Biber (1999) and all lexical bundle functions as categorized by Hyland (2008).

The aim of the present study is to identify and analyze four-word lexical bundles in the English medical research articles written by Czech authors and subsequently compare them with the four-word lexical bundles identified in the Mbodj-Diop's (2016) corpus. To be able to do so, a corpus comprised of English medical research articles written by Czech authors will be compiled. The desired size of the corpus will be at least 1,000,000 running words so that its size is comparable

with Mbodj-Diop's (2016) corpus which consisted of 1,177,611 running words. Based on the analysis of the compiled corpus data through AntConc, this thesis will attempt to answer the following questions:

1. What are the most frequent four-word lexical bundles in the corpus of English medical research articles published by Czech authors (CZMRAC)?
2. What are the structures of four-word lexical bundles used in the CZMRAC?
3. What functions do four-word lexical bundles perform in the CZMRAC?
4. How are the identified four-word lexical bundles in the CZMRAC different, if at all, compared to four-word lexical bundles in the English Medical Research Article Corpus (ENMRAC) in their frequency, structures, and functions?

It should be noted that Mbodj-Diop did not differentiate between "nativeness" and "non-nativness" in regard to the authorship of the medical research articles in the ENMRAC. She explains her decision by proposing that if the articles were deemed good enough for publication by the reviewers of the renowned journals, then they could be "considered as representative of the medical research article genre regardless of their authors' first languages" (2016, 19). Therefore, it cannot be claimed that the purpose of this thesis is to compare lexical bundles produced by Czech speakers with lexical bundles produced by Anglo-American authors. Rather, it is a comparison of lexical bundles produced by Czech authors with lexical bundles that were found in the medical research articles published in renowned Anglo-American medical journals. The results of the present study may be beneficial to the authors of the CZMRAs, their reviewers, and the medical journals in the Czech Republic which publish in English as they will provide feedback to their publishing activity. In addition, a byproduct of this study will be a list of the most frequent four-word lexical bundles identified in the CZMRAC and in the ENMRAC which can serve as a part of the teaching material for teachers of English.

The first part of the thesis will provide a theoretical foundation for the subsequent analysis. It will discuss formulaic language (Section 2.1), lexical bundles (Section 2.2), lexical bundles and second language acquisition (Section 2.3), learner corpora (Section 2.4), medical writing (Section 2.5), and Mbodj-Diop's study which served as inspiration for the current study (Section 2.6). The Analytical Part will introduce the corpus design of the CZMRAC (Section 3.1) and will put focus

on presenting the results of this paper and on their discussion (Section 3.2). Finally, conclusions will be drawn in Section 4.

2 Literature Review

The objective of the following part of the thesis is to lay down a theoretical foundation for the analytical part. At first, the phenomenon of formulaic language will be introduced. The following sections will focus on a detailed description of lexical bundles including their structural and functional classification, their relationship to second language acquisition. Next, the concept of learner corpora will be discussed, as well as, some general characteristics of medical writing and the master's thesis, which served as an inspiration for the current thesis.

2.1 Formulaic Language

In her work, Barlow (2011) states that one of the most interesting discoveries in applied linguistics has been the fact that language is typically not produced word by word but is largely formulaic in nature. Despite different figures, the generally accepted rule is that "at least one-third to one-half of language is composed of formulaic elements" (Conklin &Schmitt 2012, 72). In her book, Wray (2002) shares a rather interesting experiment that took place in Britain which can serve as a piece of evidence that language is usually formulaic. In 1993, people were asked what Rice Krispies, a breakfast cereal created by the Kellogg company, were made of. Even though, the name of the cereal contained one of the main ingredients – rice, many people struggled to answer. This illustrates that it is possible to overlook the individual components of a linguistic sequence and to focus more on the meaning of the expression itself, which according to Wray, happens frequently (Wray 2002, 3). This common phenomenon, also called, formulaic language (FL) is defined by Wray and Perkins (2000, 1) as "a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar." Another definition by Wood (2019, 30) suggests that FL is "generally defined as multiword language phenomena which holistically represent a single meaning or function, and are likely mentally stored and used as unanalyzed wholes, as are single words."

According to Wray, one of the first scholars who noticed unexpected levels of fixedness in language was John Hughlings Jackson in the mid-nineteenth century. He studied the ability of aphasic people to fluently express themselves through rhymes, prayer, or routine greetings. Additionally, Wray also identifies Jesperson, who made comments about how difficult it would be

for speakers of a language to know by heart ever single item separately (2002, 7). The field of study that analyzes this phenomenon in detail is known as phraseology. According to Wood (2015), a turning point for phraseology and its development was in 1970s. It was then that lexicographers started collecting information about multiword chunks and conducted research focusing on speech acts and pragmatics. With time, various linguists analyzed FL through different lenses which created some confusion regarding terminology. Wray and Perkins (2000, 3) draw attention to this reality and call it "the multi-faceted nature of formulaic language". In other words, formulaic sequences can exist in various forms. They can be studied according to their form, function, semantic, syntactic and lexical properties, or their relationship with novel (analytic) language. Consequently, this has resulted in over 40 different terms which refer to one or more subtypes of the FL. Table 1 illustrates the variety of terms used in the literature.

Table 1 *Different Terms to Describe Formulaic Sequences by Wray and Perkins (2000, 3)*

Amalgams	Gambits	Preassembled speech
Automatic	Gestalt	Prefabricated routines and patterns
Chunks	Holistic	Ready-made expression
Clichés	Holophrases	Ready-made utterances
Co-ordinate constructions	Idiomatic	Rote
Collocations	Idioms	Routine formulae
Composites	Irregular	Schemata
Conventionalized forms	Lexical(ised) phrases	Semi-preconstructed phrases that constitute single choices
FEISs	Lexicalized sentence stems	Sentence builders
Fixed expressions	Multiword units	Stable and familiar expressions with specialized subsenses
Formulaic language	Non-compositional	Synthetic
Formulaic speech	Non-computational	Unanalyzed chunks of speech
Formulas/formulae	Non-productive	
Fossilized forms	Petrification	
Frozen phrases	Praxons	

After discussing some general information and the complexity of the FL terminology, the next chapters will focus on the characteristics of formulaic sequences which were introduced by Schmitt and Carter (2004, 1-22).

As the first characteristic of the FL, Schmitt and Carter (2004) argue that formulaic sequences are mostly stored in one's mind as holistic units and that they may not have to be learned in all-or-nothing manner. To illustrate this point, idioms can be provided as examples. Their meaning cannot be derived based on the individual components; it must be learned as a holistic unit by heart. Another example can be drawn from the world of phonology. It seems that formulaic sequences are usually uttered more fluently and with a coherent intonation in comparison to sequences that are created in a creative manner. The third example that shows that formulaic sequences are mostly stored as holistic units showcases their cognitive efficacy. For example, a formulaic sequence *by the way* is typically used more often than novel ones such as *It's time for a topic change* if a writer or speaker decides to change a topic. Cognitively, it takes less effort to utilize a formulaic sequence that is already stored in one's mind as a unit. The claim that formulaic sequences are not learned in "all-or-nothing" manner can be supported by an example from first language children acquisition. Schmitt and Carter suggest that in the case of kids, it can sometimes be that "the formulaic sequences are learned over time, with the later stages of acquisition consisting of 'filling in the gaps' in the initial incomplete rendering of the sequence" (4). In other words, it can happen that children can hear an adult utter a formulaic sequence without remembering it completely and then they start using it incompletely until it is learned completely.

Additionally, Schmitt and Carter (2004) claim that even if a large number of formulaic sequences is fixed, they can have slots for enabling flexibility of use as well. An example of a fixed formulaic sequence is *Watch out!* which precisely conveys the message and is instantly recognized. In contrast, the structure '_____ *thinks nothing of* _____' allows for a wide variation of situations which can express something unexpected or exceptional. For instance, *Diane thinks nothing of running 5 miles before breakfast* or *He thinks nothing of driving 100 miles per hour on the freeway*. Nevertheless, it is important to mention that the slots have often semantic constraints. For example, if the structure '_____ *thinks nothing of* _____' was used like in this instance *'She thinks nothing of sleeping 8 hours per night'*, semantically it would not be acceptable because the second slot in such sentence does not capture anything unusual or unexpected. Thus, the second slot must express something unusual because that is the reason why such sentence structure was used in first place.

Another essential characteristic of formulaic sequences is their semantic prosody. As an example, Schmitt and Carter use the word *border*, which without adjacent words, has a relatively wide range of meanings (the edge of something or the state boundary). Nevertheless, once it is used syntagmatically with other words such as *on* as in *bordering on*, its meaning is constrained. Schmitt and Carter propose that out of 100 instances in the British National Corpus, the most common usage was to describe "approaching an undesirable state (of mind)" (Schmitt and Carter 2004, 6).

The last distinguishing feature of formulaic sequences that Schmitt and Carter (2004) describe is the fact that they are tied to particular conditions of use. This means that they serve certain functions. One of the most frequent functions is to maintain social interaction. *Nice weather today*; *Ok, I've got it,* or *How interesting* can exemplify this type of use. Besides social interaction, formulaic sequences also play an important role in discourse organization. They help with topic elaboration with a sequence such as *in other words* or with summarizing and expressing findings with sequences like *in conclusion*. Another function that they can realize is to convey information in a precise manner. Schmitt and Carter present *Cleared to land* as an example from aviation register, which communicates to the pilot very precisely what they are supposed to do and what their rights and responsibilities are. To learn more about the functions of the FL, see Wray and Perkins (2000), who provide an elaborate work on formulaicity as a tool for social interaction through manipulation of others, asserting separate identity, asserting group identity and as shortcutting tool in processing cognitive load.

Before concluding this section dedicated to FL, it is essential to comment on the classification of formulaic sequences. Schmitt and Carter (2004) bring up an important issue which is connected to the multi-faceted nature of formulaicity introduced by Wray and Perkins (2000). They claim that while some formulas are clearly formulaic and recognizable such as idioms, phrasal verbs, or sayings, others can cause confusion. According to Wood (2019, 30), some word clusters such as *and then* or *sooner or later* are very difficult to categorize. Moreover, determining whether a formulaic sequence is a collocation, or an idiom can be problematic as well. However, even if Wood discussed the issues regarding the typology of formulaic sequences, he provided their classification. According to him, they can be categorized based on their structural, semantic or syntactic properties, their pragmatic utility, and their distribution in corpora. To look at his typology in detail, see Wood (2019, 31-36).

Since the aim of this thesis is to investigate a specific type of formulaic sequences - lexical bundles in medical research articles written by Czech authors in English - the next section is dedicated to their more elaborate description.

2.2 Lexical Bundles

According to Biber et al. (1999), the term "lexical bundle" appeared for the first time in Longman Grammar of Spoken and Written English. Numerous researchers have provided a definition of lexical bundles. To name a few, Biber et al. (1999, 989) proposed that "lexical bundles can be regarded as extended collocations: bundles of words that show statistical tendency to co-occur". Biber and Barbieri (2007, 264) described lexical bundles as "the multiword sequences that occur most commonly in a given register ... defined simply as the most frequently recurring sequences of words (e.g., *I don't know if, I just wanted to*). Lexical bundles are usually not structurally complete and not idiomatic in meaning, but they fulfill important discourse functions in both spoken and written texts." Another definition by Hyland (2012, 150) states that lexical bundles are "statistically the most frequent recurring sequences of words in any collection of texts". Additionally, Wood (2015, 45) suggests that lexical bundles are "combinations of three or more words which are identified in a corpus of natural language by means of corpus analysis software programs." The present study follows the definition provided by Biber and Barbieri (2007). Besides defining what lexical bundles are, it should also be noted that researchers use various terms to call them. Some of them are e.g., 'fixed expressions' (Moon 1992), 'multiword lexical units' (Cowie 1992), or 'formulaic sequences' (Wray 2000, 2002). With the help of the provided definitions, this section aims to introduce lexical bundles in more depth and to provide additional observations made by other researchers.

As most definitions emphasize, one of the key characteristics of lexical bundles is their frequency of occurrence. Although the frequency is a deciding factor for their identification, there is no consensus on what the exact threshold of minimum occurrence should be. For example, Biber et. al. (1999) propose that a lexical bundle candidate must occur in at least 5 different texts and at least 10 times per one million words. Cortes (2004) sets stricter criteria suggesting a group of three or more words can be called a lexical bundle if it occurs at least 20 to 40 times per million words. The last example can be provided Hyland (2008) who argues that word clusters must appear 20 times in 10% of the investigated texts. Drawing from the introduced examples, it can be said that the frequency threshold varies based on the researcher, the research requirements, and the corpus size.

The second distinguishing feature of lexical bundles according to Biber and Barbieri (2007) is the fact they are usually not idiomatic. This claim can be supported by Cortes (2004), who suggests that bundles, such as *in the presence of, as a result of*, or *what do you mean*, have a transparent meaning, deduced purely based on the individual words. On the other hand, comprehension of idioms such as *beat around the bush*, cannot be retrieved from the individual words, one needs to know the meaning of the idiom which must be learned by heart. Furthermore, Biber and Barbieri (2007, 270) mention that they do not represent a complete structure unit. This means that they often serve as a bridge between two clauses or two phrases depending on whether they occur during speech or writing. In writing, they bridge two clauses e.g. *I want to know, well that's what I*. In speech, they mostly occur between two phrases e.g. *in the case of, the base of the*.

Wood's (2015) definition also states that lexical bundles are typically identified in a corpus with the help of software programs. Dontcheva-Navratilova (2012, 42) claims that there are two main methods to identify formulaic sequences. The first method uses a search tool which recognizes "lexical co-occurrences of different length, at different cut-off frequency points." In contrast, the second approach investigates a set of fixed expressions that were previously identified in the previous research or study. An example would be a study which assesses "the extent to which foreign language learners have acquired register, discipline and genre-specific target bundles." The present study combines both approaches, as there is a corpus compiled of medical research articles written by Czech speakers, whose results will be compared with a list of lexical bundles extracted by a different researcher.

Along with the aforementioned characteristics of lexical bundles, it is also important to note that they are predominantly found in academic writing. They contribute to well-written academic discourse, because they make language more predictable, and are essential to facilitating pragmatically efficient communication, thus reducing processing time. This happens by guiding readers through the text with chunks, such as *in the next section* or *we can see that*, and by linking ideas with expressions like *is due to* the and *in contrast to*. Furthermore, they play an important role in fluency and they differentiate novices and experts in both spoken and written registers (Hyland and Jiang, 2018, 3-4). Another point, that Hyland and Jiang discussed in their work, was the fact that lexical bundles are not static which means that they vary across genres.

Furthermore, Hyland (2008, 4) points out that "[lexical] bundles are not only central to the creation of academic discourse, but they offer an important means for differentiating written texts by discipline". This is a vital piece of knowledge because the present study focuses on lexical bundles in the medical research articles written by Czech authors, and it will be beneficial to investigate to what degree, or if at all, they differ compared to the lexical bundles in the same discipline, thus in the Anglo-American respected medical journals.

The following two sections will introduce structural and functional classification of lexical bundles.

2.2.1 Structural Classification of Lexical Bundles

This section briefly introduces the structural taxonomy of lexical bundles proposed by Biber et. al. (1999, 1014-124):

1) **Noun phrase with of-phrase fragment** (the beginning of the, the shape of the)
2) **Noun phrase with other post-modifier fragments** (the way in which, the extent to which)
3) **Prepositional phrase with embedded of-phrase** (as a result of, in the case of)
4) **Other prepositional phrase** (at the same time, on the other hand)
5) **Anticipatory it + verb / adjective phrase** (it is possible to, it should be noted that)
6) **Passive verb + prepositional phrase fragment** (is shown in figure, is based on the)
7) **Copula be + noun / adjective phrase** (is one of the, is part of the)
8) **(Verb phrase +) that- clause fragment** (has shown that, that there is no)
9) **(Verb / adjective +) to-clause fragment** (are likely to be, has been shown to, to be able to)
10) **Adverbial clause fragment** (as we have seen, if there is a)
11) **Pronoun / noun phrase + be (+...)** (this is not the, there was no significant)
12) **Other expressions** (as well as the, than that of the)

Years later, Biber et al. provided a more elaborate description of structural taxonomy of lexical bundles based on their grammatical properties (2004, 381).

Table 2 *Structural Classification of Lexical Bundles by Biber et al. (2004, 381)*

Structural categories	Sub-categories	Examples
1. Lexical bundles that incorporate verb phrase fragments	A. 1st/2nd person pronoun + VP fragment	*I'm not going to*
	B. 3rd person pronoun + VP fragment	*and this is a*
	C. discourse marker + VP fragment	*I mean I don't*
	D. verb phrase (with non-passive verb)	*have a lot of*
	E. verb phrase (with passive verb)	*is based on the*
	F. yes-no question fragments	*are you going to*
	G. WH- question fragments	*what do you think*
2. Lexical bundles that incorporate dependent clause fragments	A. 1st/2nd person pronoun + dependent clause fragment	*I want you to*
	B. WH-clause fragments	*when we get to*
	C. if-clause fragments	*if we look at*
	D. to-clause fragments	*to be able to*
	E. that-clause fragments	*that this is a*
3. Lexical bundles that incorporate noun phrase and prepositional phrase fragments	A. noun phrase with of-phrase fragment	*One of the things*
	B. noun phrase with other post-modifier fragments	*The way in which*
	C. other noun phrase expressions	*A little bit more*
	D. prepositional phrase expressions	*At the end of*
	E. comparative expressions	*As well as the*

The results of the present study will be categorized based on the taxonomy introduced by Biber et al. (1999).

In addition to structural classification, lexical bundles can also be classified according to their function.

2.2.2 Functional Classification of Lexical Bundles

Before introducing functional classification of lexical bundles, it is necessary to point out that there have been certain discrepancies in the way they are organized. This is due to the fact that many lexical bundles are multifunctional and structurally incomplete. Such differences can be seen in both super-ordinate categories and their sub-categories. An example of a lexical bundle that illustrates these differences is, *at the beginning of*, which can be categorized as having both a time, space and deictic function and/or a discourse organizing function. Differences in the functional distribution of lexical bundles can be attributed to a wide range of corpora sizes and their composition. For example, large corpora focused on hard sciences seem to consist of more referential bundles, whereas corpora compiled of social sciences and humanities texts seem to have a higher proportion of discourse organizing expressions (Dontcheva-Navratilova 2012, 40).

Provided below are three various taxonomies introduced by Biber et al. (2004), Hyland (2008) and Dontcheva-Navratilova (2012) depicting how approaches to terminology in the field of lexical bundles differ.

Table 3 *Functional types of lexical bundles according to Biber et al. (2004, 386-388)*

Functional types	Sub-functions	Examples
1. Stance bundles	A. Epistemic stance	*the fact that the*
(express author's or speaker's	B. Attitudinal/modality stance	
attitudes, feelings, judgements, or	B1) Desire	*I want you to*
commitment concerning the	B2) Obligation/directive	*it is important to*
message)	B3) Intention	*we are going to*
	B4) Ability	*to be able to*

2. Discourse organizers (indicate the general overview of the discourse)	A. Topic introduction B. Topic elaboration/clarification C. Identification/ focus D. Imprecision E. Specification of attributes E1) Quantity specification	*in this chapter we* *on the other hand* *on of the things* *or something like that* *a little bit of*
3. Referential bundles (single out some important features of an identity)	E2) Tangible framing E3) Intangible framing F. Time/Place/Text reference F1) Place reference F2) Time reference F3) Text-deixis F4) Multi-functional reference	*in the form of* *on the basis of* *in the United States* *at the same time* *as shown in table* *in the middle of*

Table 4 *Functional classification of lexical bundles according to Hyland (Hyland 2008, 13-14)*

Functional types	Subfunctions	Examples
1. Research-oriented (help writers structure their activities and experiences of the real world)	A. Location (in time/place) B. Procedure C. Quantification D. Description E. Topic (related to field of study)	*at the beginning of, at the same time* *the use of the, the role of the* *the magnitude of the, a wide range of* *the structure of the, the size of the* *in the Hong Kong, the currency board system*
2. Text-oriented (concerned with the organization of the text and its meaning as a message or argument)	A. Transition signals (contrastive/additive links) B. Resultative signals (inferential/causative relations) C. Structuring signals (organize discourse/direct readers elsewhere in text) D. Framing signals	*on the other hand, in addition to* *as a result of, these results suggest that* *in the present study, in the next section, as shown in figure*

	(specify limiting conditions)	*in the case of, with respect to the, on the basis of, in the presence of*
3. Participant-oriented (focus on the writer or reader of the text)	A. Stance features (writer's attitude/evaluations) B. Engagement features (address reader directly)	*are likely to be, may be due to, it is possible that* *it should be noted that, as can be seen*

Table 5 *Functional classification of lexical bundles introduced by Dontcheva-Navratilova (2012, 40-41)*

Functional types	Subfunctions	Examples
1. Referential bundles (largely overlapping with Hyland's (2008) research-oriented bundles)	A. time/place/text-deixis bundles B. attribute bundles (specifying procedure, quantity or description of reality) C. topic specific bundles	*at the end of the, and the beginning of* *a little bit of, the use of the* *in the curricula of, the interpretation of the*
2. Discourse organizers (correspond to Hyland's (2008) text-oriented lexical bundles)	A. logical relations bundles A1) transition bundles A2) resultative signals B. intratextual reference bundles (organizing smaller and larger stretches of discourse) C. framing bundles (focusing, limiting conditions on arguments)	*on the other hand, in addition to the* *as a result of, it was found that* *in the present study, in the next section* *in the case of, with respect to the*
3. Attitudinal bundles (corresponding to Hyland's (2008) participant-oriented features	A. stance bundles B. interactional bundles	*the fact that the, it is possible to* *it should be noted that, as can be seen*

The introduced classifications clearly prove that different researchers indeed utilize various terms for the same concept. This is especially true with the main functional types of lexical bundles. All three researchers, Biber (2004), Hyland (2008) and Dontcheva-Navratilova (2012), observe that lexical bundles can be divided into three main functional types. Even though, they opted for different terms, except for Dontcheva-Navratilova (2012) who kept Hyland's term "referential bundles" and Biber's term "discourse organizers", pragmatically speaking, all three researchers describe the main function types in a fairly similar way.

Stance bundles, participant-oriented bundles or attitudinal bundles express author's or speaker's attitudes. Hyland (2008) and Dontcheva-Navratilova (2012) emphasize that they also serve as interactional markers between the writer and the reader. When it comes to the sub-functions, Biber (2004) describes them in more detail focusing on four different types of stance which are desire, obligation/directive, intention, and ability. The other two researchers choose a more general approach, especially Dontcheva-Navratilova (2012).

Discourse organizers or text-oriented bundles serve as text builders and help with text structure and organization. Hyland's (2008) and Dontcheva-Navratilova's (2012) subfunctions largely overlap, differing only in their organization and their terminology. Dontcheva-Navratilova (2012) uses an umbrella term logical relations bundles for transition and resultative signals and calls Hyland's (2008) structuring signals intratextual reference bundles. Biber (2004) also introduces subfunctions of topic introduction, topic elaboration/clarification, identification/focus, imprecision, specification of attributes, and quantity specification. These differences are most likely the results of various bodies of texts that researchers explored, as mentioned earlier.

Referential bundles or research-oriented bundles help with structuring writer's activities and experiences of the real world. With these bundles, the pattern repeats itself and Biber (2004) again provides the most elaborate taxonomy of subfunctions. Besides references of place, time, and text which Hyland (2008) and Dontcheva-Navratilova (2012) mention as well, Biber (2004) also introduces tangible and intangible framing. An example of the former subfunction is *in the form of* and of the latter subfunction *on the basis of*.

The most notable difference between Biber et al's (2004) and Hyland's (2008) taxonomies is found in the type of register. Mbodj-Diop (2016, 27) explores this in her thesis and explains that Biber et al (2004) wanted to offer a holistic approach to classifying functions of lexical bundles in all types of registers, but they came to a conclusion that functions of lexical bundles in written and spoken registers differ significantly. As a response to this issue, Hyland (2008) proposed his functional classification that focused purely on research writing. Thus, the present study follows the functional taxonomy introduced by Hyland (2008).

After discussing general characteristics of lexical bundles and their functional and structural classification, the next section will explore specifically the relationship between lexical bundles and learners of English, since it is highly relevant for the present study.

2.3 Lexical Bundles and Second Language Acquisition

Having become lingua franca in the modern world, English has played an essential role in many fields of study, such as engineering, business, law, or medicine. This has certainly been enabled due to globalization and the Internet. Articles, research papers, textbooks, and many other publications can be written in one part of the world and in a few minutes they can be studied worldwide. So that their work can be recognized internationally, many researchers publish not only in their native language, but also in English. Additionally, some of them even produce texts solely in English. Therefore, with a rapidly growing number of publications in English by both native and non-native speakers, the interest of linguists in such documents has been enormous. A specific area of many studies are formulaic expressions including lexical bundles. However, Dontcheva-Navratilova (2012) suggests that most studies of lexical bundles have focused mainly on their identification, structure, or discourse functions and that there are significantly fewer researchers who have investigated lexical bundles, that had been produced by non-native researchers. This section presents some of them.

In their work of spoken language, Pawley and Syder (1983) state that native speakers store thousands of memorized sequences in their mental lexicon. These prefabricated sequences help native speakers with fluency and effortless speech production. This logically leads to a conclusion that in order to acquire a certain language level, it is expected to improve one's lexicon by broadening it with formulaic expressions such as lexical bundles as well. Although, Pawley and Syder claim that it is rare for second language learners to fully accumulate the native repertoire of formulaic language and that it is usually a stumbling block even for advanced language learners. It is more common for them to over-generate grammatical utterances that are not idiomatic. Of course, this statement could be a subject of discussion considering that it was uttered almost 30 years ago, and that second language acquisition has expanded in many aspects.

One of the major proponents of formulaic language and its studies, Alison Wray, conducted a survey in 1999 to illustrate that formulaic sequences are used in different ways by different speakers. Wray (1999) focused on groups of adult native learners, aphasics (people who lost their linguistic ability after a stroke or other brain damage), child first language learners, and both child and adult second language learners. In the survey, she proposes that it cannot be expected to gain

a clear idea of formulaic language production from learners with different abilities, experiences, attitudes, learning styles, interactional agendas. However, she is aware that the results are lumped together because if the studies were subdivided, only a few would be left. Some of the interesting findings were that young children learners seemed to use formulaic language in a similar way to those of first language acquisition, whereas adult learners applied them in two ways. First, learners in the natural environment used them for effective communication without accuracy. Second, classroom learners could analyze them more readily than apply them idiomatically. Another finding regarding classroom learners showed that they tended to over-generate formulaic sequences that were grammatically correct, but unidiomatic. On the other hand, results that were common among all study groups showed that all groups used formulaic sequences to achieve interactional goals (greeting, chastising) and to sustain the interaction even though taught adults used them least frequently. Last, all groups except aphasics, used them to express individual and group identity.

In his article, Allen (2010) focused on learners' use of lexical bundles in written production of science research papers that he later compared with reference corpora comprised of native speaker writing. The areas that Allen investigated were accuracy, grammatical class, and function of lexical bundles. The target group in his article were students from the University of Tokyo. In his conclusions, Allen claimed that the two corpora showed considerable convergence which he contributed the continual process of revising and editing. However, he emphasized the need for pedagogical intervention when it came to noun-phrases constructions because Japanese students tended to overuse them or to use them incorrectly. As an example, he stated bundles including the stem *result of* instead of *results of*. Furthermore, Allen also pointed out that lexical bundles are not easy to categorize and there is inherently a level of subjectivity deciding the function or category of a lexical bundle.

Dontcheva-Navratilova (2012) conducted a study more relevant to Czech learners of English. She investigated the use of lexical bundles in non-native speaker academic discourse with the focus on diploma theses written by Czech students of English. She specifically focused on the extent to which the students use the main functional types of lexical bundles in the fields of linguistics, methodology, and literature. According to Dontcheva-Navratilova. lexical bundles are an essential aspect in language acquisition because: they are necessary for a fluent text discourse production

and comprehension, they represent shared knowledge of a professional discourse community, show communicative competence, and play an important role for differentiating written texts by discipline. In her conclusion, she writes that Czech students of English use lexical bundles less frequently in comparison to expert academic discourse. She concluded that structural inaccuracy in the use of lexical bundles was not frequent, but the distribution of functional categories differed considerably from expert academic writing. She explains this by students' insufficient level of rhetorical skills and by interference from L1 writing conventions. Additionally, she shares important pedagogical implications that unconscious learning i.e., student's exposure to academic reading with a high frequency of lexical bundles, does not lead to their acquisition. This can be explained by their lack of perceptual salience. Thus, she recommends to deliberately focus on lexical bundles acquisition by recognition of lexical bundles in academic discourse, pattern practice, including some contextualization, substitution drills, checking against corpora, and creative use in written performance.

Heng, Kashiha and Tan (2014) conducted a study of lexical bundles in the Malaysian context. Their goal was to identify and analyze the structural and functional characteristics of lexical bundles in university group discussions carried out by proficient ESL students. They concluded that the overall use of lexical bundles by students was frequent mainly because they increased students' ability to engage in the discussion. Structurally, students used phrasal lexical bundles most frequently. Among them were verb-phrase, noun phrase and prepositional phrase fragments. As a result, they suggested that instructors should pay more attention to teaching lexical bundles during lessons. Functionally, the findings showed that students used lexical bundles automatically and applied them unconsciously to show ownership of ideas, to elaborate on a topic or to negotiate directions in the development of the discourse. Another important use was to agree or disagree, thus facilitating a crucial element of any discussion.

To summarize the findings of the provided studies, it can be concluded that correct use of lexical bundles is beneficial for various reasons. One being that students can increase their fluency in discourse production and show that they possess shared knowledge of a professional discourse community. Furthermore, lexical bundles aid in comprehension, enhance students' confidence to speak and show ownership of ideas and both agreement and disagreement. Additionally, lexical bundles increase students' legitimacy in the discipline one writes in and support linguistic inclusion in the community and helps with interaction. Studies have also shown that language

learners tend to over-generate lexical bundles and there seems to be interference from L1 writing conventions. Challenges linked to studying lexical bundles produced by second language learners exist inherently in the fact that most texts undergo numerous revisions and editing which makes it difficult to distinguish with certainty whether studied lexical bundles were indeed produced by non-native speakers. Besides that, Wray (1999) elaborates a crucial point suggesting that different types of speakers apply formulaic sequences in different ways which results in findings that cannot always be lumped together and compared.

Acquiring lexically bundles can be accomplished in a variety of ways. They can be embedded in traditional coursebook instruction, taught explicitly with or without context, or through exposure to authentic native input during active listening or reading. However, according to Hyland and Jiang (2018, 1), the most convenient way to identify lexical bundles and to study them, is through various corpus analysis software, such as WordSmith tools, or AntConc, which retrieve multi-word units according to frequency and distribution criteria specified by the researcher.

2.4. Learner Corpora

It was computer-aided corpus studies that made a significant breakthrough in the study of formulaic language. With the use of computer corpora, researchers were able to investigate large bodies of texts and use a wide range of automated tools to identify and analyze formulaic sequences in learner language. As a result, since late 1980's and early 1990's, learner corpus research has become a highly dynamic and an essential branch of corpus linguistics. Since there are many variables to control, it is a heterogenous field of study. Learners and learning situations vary which means that researchers must set very strict design criteria which deal with characteristics pertaining to a learner (age, gender, mother tongue background etc.) and the task (medium, topic, or timing). Paquot and Granger (2012, 3) define learner corpora as "electronic collections of texts produced by foreign or second language learners." According to Paquot and Granger, learner corpora are powerful tools to study formulaic language because they 1) contain continuous stretches of oral or written discourse rather than decontextualized words, phrases, or sentences and 2) they often include data resulting from pedagogical tasks that allow students to use their own words rather than being requested to produce a particular structure such as in essay writing. Gablasova, Brezina and McEnery (2017, 131) add that corpora "can provide a detailed description of interlanguage at different stages of development and thus offer insights into the sources of variation in the language use and level of proficiency reached by language learners". Granger (2008, 1) describes interlanguage as a "transitional language produced by second or foreign language learners." According to Granger, language learners can be described as students who learn a new language, which is neither their mother tongue nor an institutionalized additional language where they live. In other words, a corpus can be considered a learner corpus if it contains data from students of English in Spain or Sweden, but not if the data were produced by English learners in countries such as India or the Philippines. As unproblematic as the aforementioned definition may seem to be, Granger claims that the situation is more complicated further when English as a Lingua Franca (ELF) is taken into consideration. ELF serves as a means of communication among speakers whose first languages vary and whose proficiency in English is highly advanced. Even though ELF and English in learner corpora are sometimes treated as opposed, Granger views them as two sides of the same coin. The present study follows Granger in that regard and does not significantly differentiate between ELF and learner English.

Most of the learner corpora that is available contains written data. Examples of such corpora are *International Corpus of Learner English* (ICLE) which was one of the first learner corpora containing argumentative essays by learners with numerous mother tongue backgrounds; *Uppsala Student English Corpus*; *Active Learning of English for Science Students*; *Cologne-Hanover Advanced Learner Corpus* (Paquot, Granger 2012, 4). The next few paragraphs are dedicated to the learner corpora typology following Granger (2008).

2.4.1 Commercial vs. Academic

Commercial corpora are typically initiated by major publishing companies. They tend to be larger in size than academic corpora and offer a wider range of mother tongue backgrounds. Two examples of commercial corpora for English are the *Longman Learners' Corpus* and the *Cambridge Learner Corpus*. Academic corpora are more frequent and are compiled in educational settings with usually only one mother tongue background.

2.4.2 Big vs. Small

Big corpora containing millions of words are extremely beneficial due to the representativeness of the data and generalizability of the results. However, small corpora can be of a significant value primarily when tailored to an individual or group profile of learner use.

2.4.3 English vs. non-English

The majority of learner corpora compile English data. Examples of such corpora were provided in the previous paragraphs. However, learner corpora containing languages other than English are on the rise. In addition to English, some of the compiled languages are French, Spanish, German, Swedish, or Norwegian.

2.4.4 Writing vs. Speech

The world of learner corpora is dominated by written discourse and Language for Academic Purposes. Although, this may change in the future because there is a wide variety of computer-mediated communication among language learners. For example, Granger (2008) mentions *Telekorp* which is a large bilingual learner corpus that contains conversations between students of German in America from telecollaborative language courses.

2.4.5 Longitudinal vs. Cross-sectional

Longitudinal corpora cover data from the same learners over time and are rare mainly because they are tremendously time-consuming. Instead, researchers often conduct quasi-longitudinal corpora which are collected at a single point of time but from language learners with different language skills. On the other hand, cross-sectional corpora are significantly more frequent and "contain data gathered from different categories of learners at a single point of time" and cover more than one type of interlanguage (Granger 2008, 3).

2.4.6 Immediate vs. Delayed pedagogical use

Corpora for delayed pedagogical use serve as a description of one specific interlanguage and/or designing tailor-made pedagogical tools. These corpora benefit learners who have the same profile as the students who have gathered the corpus data, such as mother tongue background or same level of proficiency. Users of immediate learner corpora are at the same time their producers. Such data are compiled by teachers during classroom activities.

To apply the introduced typology in a practical way, the learner corpus in the present study is academic, rather small in size (1,074,607 words), cross-sectional with delayed pedagogical use, and is compiled of English medical research articles.

A crucial aspect of learner corpora which should be addressed as well is their appropriate use. Since corpus data are in an electronic form, it is possible to study them quantitatively which is, especially in the case of learner interlanguage, very beneficial. For instance, researchers can investigate co-occurrence and recurrence of learners' vocabulary, as in the case of the present study. Furthermore, Gablasova, Brezina and McEnery (2017) suggest that the majority of learner corpora focused studies are comparative which means that researchers either compare one or more L2 corpora to a L1 corpora or that two or more L2 corpora are compared to each other. Granger (2008) calls these types of studies a Contrastive Interlanguage Analysis (CIA). CIAs enable researchers to study a wide variety of topics such as frequent vocabulary, modals, connectors, or phraseological units and have helped shed light on those that tend to be overused or underused. Nevertheless, Paquot and Granger (2012) warn that corpus-based studies should avoid generalizations of results. Instead, they should interpret results as hypotheses, which should be tested due to the high degree of heterogeneity of data and methods used in the corpus linguistics.

2.5 Medical Writing

Goodman and Edwards (2006, 104) claim that "Sloppy language means sloppy science." To elaborate this statement, they suggest that even if the research itself is meticulous, if authors of medical research articles do not pay close attention to detail in their writing, the reader can wonder whether the same sloppy approach was used in their research as well. In their book, they focus on examples of poor medical writing so that they can use them as a teaching material for those authors who want to avoid recurring mistakes. Unfortunately, Goodman and Edward argue that too many producers and consumers of medical writing are tolerant to clumsy, inaccurate, and/or obscure writing (2006, 3). When confronted about their inadequate writing, authors of medical texts provide five common defences (Goodman and Edwards 2006, 11):

1) Everyday language is inappropriate and not precise enough to describe the results of a medical study.

2) Long words are more scientific.

3) Their writing style follows the convention for writing medical papers.

4) However writers write, editors will alter the writing to suit themselves.

5) Everyone has their own style, and to use this style is better than conforming to a supposedly correct set of rules.

Since number four is a relevant point for this thesis, it will be discussed in greater detail later.

Besides author's defences, Goodman and Edwards (2006, 20-21) state four main reasons for the lack of competency in medical writing – limitations, imitations, obfuscations, and specializations. Regarding *limitations*, the authors suggest that unclear medical writing is largely produced as a result of lack of awareness, failure of education and training, and absence of editorial guidance. Another limitation which Goodman and Edwards describe as *imitating* is manifested when authors use expressions that they copied from another author. They may choose to do so because they want to conform to the writing of their peers or superiors and they might feel that such writing will help them gain respect or to succeed. An example that Goodman and Edwards provide for such imitating is the excessive use of hedging devices which also contributes to *obfuscate* writing. One could think that authors write so vaguely as if they were afraid of legal consequences. The last limitation they detail are *specializations*. Goodman and Edwards quote Hayes (1992) who says

that medical literature contains so much specialized language that it is increasingly obscure to all but the initiates. This happens frequently when young authors are required to publish academic papers without being equipped with formal guidance and without proper editorial process.

After understanding why there is an abundance of "bad" medical writing, it is also important to emphasize some general guidelines offered by Goodman and Edwards (2006, 29) that are essential to follow when producing medical writing. First, quite simply, authors should use familiar words instead of unfamiliar ones. They should try formulating their ideas and knowledge as if they were conversing informally with their colleague. Second, they should create short sentences with shorter words because long sentences can be confusing and result in incorrect grammar. Third, it is helpful to have at least two people read the article, preferably someone who is not a close colleague to ensure that it is as objective as possible. Fourth, they recommend reading books in English.

Goodman and Edwards (2006) are proponents of clear and unambiguous medical writing. Knowing that lexical bundles contribute to the overall comprehension, coherence and discourse organization as mentioned in the earlier chapters, it can be argued that their correct use can be one of the tools that can contribute to elimination of so called "bad" writing, even if Goodman and Edwards do not mention lexical bundles specifically.

2.5.1 Editing Process in Creating Medical Research Articles

Stephen Hardy, CEO of Medical Journal Editors, said: "While attending graduate school at Harvard, I had the chance to work with some of the world's most talented international scholars. However, these scholars often had difficulty preparing manuscripts that fully represented their excellent research. This issue inspired me to create this service (Medical Journal Editors) so that all academic work receives the attention it deserves." This statement supports the comments from the previous section and shows the importance of the editing process in creating medical research articles. Thus, the purpose of the following section is to briefly discuss this topic and to introduce the selected Czech medical journals publishing in English along with their instructions for the authors.

Before introducing the Czech medical journals selected for the present thesis, a few general observations regarding editing in medical papers will be made. Goodman and Edwards (2006)

claim that editors have ultimate power when it comes to deciding whether a paper is accepted or rejected. However, due to time restraints, it is often the case that medical papers are reviewed only by editorial assistants, subeditors, or copyeditors. In some cases, it can even happen that there is no subediting, at all, due to the lack of subeditors or cost-effectiveness. It scarcely happens that subeditors change papers drastically e.g., make them shorter or clearer. Instead, they identify serious grammatical mistakes, correct them, and adapt the paper to the house style. Such changes can be representation of numbers and units, references, illustration and so on.

Goodman and Edwards (2006, 12) believe that "editors of medical and scientific journals should have a feel for language". That is, they should be good writers themselves. Being a well-known specialist and having published a tremendous number of articles does not suffice. Nevertheless, they maintain that this view is unpopular. They even claim that "it is difficult to avoid thinking that the editors are concerned only with filling the pages. To solve this problem, they propose that editors and subeditors should not be afraid to send papers back for rewriting and even more importantly they should not only be professional writers, but they should also be familiar with the medical field. As a result, the scientific aspect of medical papers would be less compromised. Goodman and Edwards also mention editing second-language texts. They consider such work very demanding and time-consuming but imperative for international medical communication.

Following a brief discussion of the editing process of medical papers, the next chapters will introduce five Czech medical journals selected for the corpus compilation in this study. Included with their brief introductions are provided the instructions for the authors that were detailed on their official websites.

Acta Medica (AM) is a medical journal that is published solely in English by the Faculty of Medicine in Hradec Králové. AM was founded in 1958 and is published four times a year online and in print. The character of the journal is multidisciplinary. The readers find in it contributions such as review articles, original articles, brief communications, case reports, and announcements. All submitted papers are reviewed by an editor. The evaluation focuses on novelty, originality, impact, and validity. If a submitted paper does not comply with publication quality, it is rejected. However, the instructions for authors do not specifically mention how they measure quality. If a paper is accepted, it is sent to an external reviewer for peer review. The identity of peer reviewers

is not made known to the authors, but the identity of the authors is known to the reviewers. If a paper needs minor or major changes, there are additional peer reviews to ensure expected quality.

Biomedical Papers (BP) is a medical journal published by the Faculty of Medicine and Dentistry of Palacký University in Olomouc. Annually, there are four issues with reviews and original articles reporting on basic and clinical research in medicine. Submitted manuscripts are reviewed by independent reviewers chosen by the editors. The independent reviewers evaluate whether a manuscript is suitable for publication. As for linguistic requirements, the BP official website states that submitted manuscripts should be written in good and clear English. In addition, the instructions recommend that the article candidates should be read by a native English speaker before submitting.

Central European Journal of Public Health (CEJPH) has been published since 1993 and offers insights on disease protection, environmental impacts on health, and the role of nutrition in health promotion. The submitted papers are reviewed by two independent reviewers and the submission-to-publication process lasts approximately 18 months.

Physiological Research (PR) is an open access journal that is published every two months by The Czech Academy of Sciences. The instructions for the authors state that published articles cover topics such as normal and pathological physiology, biochemistry, biophysics, and pharmacology that report of original work of the authors. U.S. English spelling is preferred and submitted papers are peer-reviewed by at least two peer-reviewers, who know the identity of the author, but the author does not know the identity of the peer-reviewers. In case that two evaluations differ significantly, a third peer-reviewer is asked to provide an additional evaluation.

Prague Medical Report (PMR) is an English biomedical journal published by the First Faculty of Medicine of Charles University. Originally, the journal started as Sborník lékařský in 1885 and since 2004 has been published in English. The PMR is indexed and abstracted in the following overview portals: DOAJ, EBSCO, Index-medicus, MEDLINE, PubMed, Scopus, CNKI, and EuroPub. Readers encounter scientific studies, short communications, case reports, reviews, lectures of great interest and information about activities of the First Faculty of Medicine. British English is required. Submitted manuscripts are evaluated by the editorial board and by at least two peer reviewers. The authors are informed about the result and in case that major changes are

required, the manuscripts are sent back to the authors to revise them and make the necessary corrections. The final versions of the articles are sent for language revisions.

In summary, this part of the thesis focused on medical writing, its editing process and on the introduction of the selected Czech medical journals. The next section of this literature review presents the master's thesis called *Lexical Bundles in Medical Research Articles: Structures and Functions* which served as a theoretical framework for the present thesis and whose results will be partially compared with the results of this thesis.

2.6 Lexical Bundles in Medical Research Articles: Structures and Functions

In 2016, Mbodj-Diop wrote a master's thesis that focused on the structures and functions of lexical bundles in medical research articles (MRAs). She sought to answer the following research questions:

1) What lexical bundles are frequently used in MRAs?

2) What are the structural patterns of lexical bundles in MRAs?

3) What rhetorical functions do such bundles have in MRAs?

To collect the data for her thesis, Mbodj-Diop created a corpus which consisted of 250 MRAs from five international medical journals - *Science, The Lancet, The New England Journal of Medicine, Journal of American Medical Association,* and *The Journal of Clinical Investigation.* While selecting specific articles, she followed three main variables to control when collecting a corpus introduced by Loi (2010) – genre, journals under investigation, and the authorship.

Regarding the variable of genre, Mbodj-Diop focused only on quantitative MRAs that were published in the above-mentioned journals. Originally, she opted for MRAs with three topics diabetes, hypertension, and tropical disease. Later, however, she also added other topics that were selected randomly for the sake of generalizability.

When it comes to reputation of the journals, Mbodj-Diop sought help from a field expert who recommended the five journals, she eventually selected. To ensure that she would investigate the most recent writing in the field of medicine, she selected only articles from most recent ten years at the time she wrote her thesis (2006-2015). She randomly chose five articles from each year. Her goal was to collect a corpus of at least 1,000,000 running words with articles that contained anywhere from 3,000 – 7,000 words.

The last variable that Mbodj-Diop controlled during the article selection was the authorship. She did not consider the issue of "nativeness" or "non-nativeness." She justifies her decision in three ways:

1) Since English is Lingua Franca, it can be difficult to differentiate who is a native speaker and who is not.

2) The matter of "nativness" seems to be irrelevant in the context of a given discourse community.

3) Authors of MRAs may be influenced more by the research training that they received than by the language environment that they grew up in.

Mbodj-Diop's parameter for quality and "nativness" was the fact that the articles were deemed good enough to be published in the internationally recognized journals.

Based on the three criteria, Mbodj-Diop created a corpus that comprised 1,177,611 words. To identify the lexical bundles, she employed the AntConc software. Initially, she set the cutoff parameter at a frequency of 20 times per million words occurring in at least 10% of the articles. Nevertheless, if she had followed this criterion, she would have had to exclude all lexical bundles that occurred at least 20 times per million words because they would have not met the criterion of occurring in 25 texts (10%). Therefore, she decided to lower the cutoff from 10% to 5%. In other words, she investigated only lexical bundles that appeared at least 20 times in her corpus and that were found in at least 12 different texts. Moreover, they had to occur in at least two of the five journals, and they could not be interrupted by any other character. For a structural classification of the identified bundles, Mbodj-Diop followed the taxonomy of 12 categories introduced by Biber et al. (1999). To categorize their functions, she used the classification introduced by Hyland (2008).

Overall, Mbodj-Diop identified 204 four-word lexical bundles. Table 6 shows 50 most frequent lexical bundles in her corpus.

Table 6 *50 Most Frequent Lexical Bundles in MbodjDiop's Corpus (2016, 29)*

Bundles	Frequency	Range	Bundles	Frequency	Range
1. on the basis of	306	122	26. the intention to treat	55	31
2. in the placebo group	235	25	27. end of the study	54	25
3. with the use of	228	57	28. as compared with the	52	32
4. in the control group	199	27	29. as well as the	51	42
5. in the presence of	139	50	30. the results of the	51	39
6. at the time of	137	83	31. did not differ between	50	22

7. the end of the	129	59	32. the use of a	50	35
8. in the absence of	117	67	33. were included in the	50	37
9. in the intervention group	115	13	34. presence or absence of	49	24
10. at the end of	109	49	35. randomly assigned to receive	49	31
11. an increased risk of	83	36	36. the proportion of patients	48	19
12. of patients in the	79	27	37. associated with and increased	47	25
13. years of follow up	74	26	38. in the present study	47	29
14. were randomly assigned to	72	43	39. have been shown to	46	39
15. in the number of	71	39	40. with an increased risk	45	24
16. was associated with a	71	46	41. in accordance with the	44	39
17. in the context of	69	46	42. between the two groups	42	24
18. the total number of	66	41	43. there was no significant	63	41
19. the primary point of	65	23	44. was approved by the	59	55
20. the use of the	65	35	45. as a result of	58	46
21. has been shown to	64	45	46. these data suggest that	58	39
22. did not differ significantly	63	39	47. was defined as a	41	32
23. there were no significant	57	36	48. were more likely to	41	26
24. for the treatment of	56	36	49. are shown in table	39	33
25. the basis of the	56	42	50. in the risk of	38	22

As a next step, Mbodj-Diop classified the identified lexical bundles according to their structure. Her results are illustrated below in Table 7.

Table 7 *Distribution of bundles structures Mbodj-Diop's corpus (2016, 32)*

Structure	Number of bundle types	% of total number of bundles
Prepositional phrase with embedded *of*-phrase fragment	36	31.37
Other prepositional phrase fragments	28	
Noun phrase with *of*-phrase fragment	32	23.03
Noun phrase with other post-modifier fragment	15	
Other expressions	29	14.22
Passive verb + prepositional phrase fragment	26	12.75
(Verb phrase) + *that*-clause fragment	11	05.40
(Verb/adjective) + *to*-clause fragment	09	04.41
Be + noun/adjective phrase	07	0 3.43
Pronoun/noun phrase + *be* (+…).	06	02.94
Adverbial clause fragment	03	01.47
Anticipatory *it* + verb/adjective phrase	02	00.98
Totals	204	100

As evident in Table 7, all lexical bundles structures, as introduced by Biber (1999), were present in Mbod-Diop's corpus. Bundles beginning with prepositional phrases and noun phrases such *as a result of* and *a reduction in the* were more prevalent than other structures. These types of structures accounted for more than half of all the lexical bundle structures in her corpus representing 54.40%. In the results, Mbodj-Diop mentions an interesting observation that is noteworthy, especially in the context of the present thesis. The lexical bundle *an increase in the* appeared in Mbodj-Diop's corpus multiple times. Nevertheless, its common counterpart *a decrease in the* was lacking in the list of the identified lexical bundles. Alternatively, the bundle *a reduction in the* was found. She comments this by emphasizing the importance of novice writers education in discourse construction in their fields of expertise. Furthermore, she suggests that learners frequently opt for alternative word combinations with a similar function, instead of a lexical bundle that is preferred by professionals.

As a result of Mbdoj-Diop's functional classification, she created Table 8 which summarizes the distribution of lexical bundle functions in her corpus.

Table 8 *Summary of lexical bundle functions in Mbodj-Diop's corpus (2016, 34)*

	Research-Oriented	Text-Oriented	Participant-oriented	Totals
Number of Bundles	122	72	14	208*
%	58.65	34.62	6.73	100

* total is higher than 204 because of certain bundles appearing in more than one category.

The provided data clearly demonstrate that the most common lexical bundles in the Mbodj-Diops corpus fell into the research-oriented category. This conclusion is consistent with previous studies. Text-oriented bundles also proved to be relevant for medical research articles representing a little more than 34% of the overall lexical bundle functions in her corpus. The participant-oriented bundles showed the least frequency, which can be explained by the nature of the research articles. Mbodj-Diop's findings will be discussed further in the section *Results and Discussion*. Since the aim of the present study is not to investigate rhetorical functions of lexical bundles, the discussion of Mbodj-Diop's results in this area is omitted.

As for pedagogical implications, Mbodj-Diop states that based on previous research e.g., Cortes (2004), explicit teaching of lexical bundles has proven to be insufficient. Instead, she suggests raising student's awareness of the presence of lexical bundles in academic writing. Next, she proposes that lists of lexical bundles from specialized fields can serve as a starting point for teaching materials. Furthermore, she encourages the use of concordance lines to show lexical bundles in different contexts.

3 ANALYTICAL PART

The analytical part will present the methodology of the present study, the key characteristics of the compiled corpus CZMRAC, and the discussion of the results.

3.1 Corpus Design

To be able to analyze the lexical bundles present in the English-written medical research articles produced by Czech authors and to compare them with lexical bundles from Mbodj-Diop's (2016) corpus, it was necessary to build a corpus as well. Thus, the first sections of the analytical part focus on the corpus design. To distinguish between the MbodjDiop's corpus and the corpus of the present study, the terms ENMRAC (English Medical Research Article Corpus) and CZMRAC (Czech Medical Research Article Corpus) will be used.

3.1.1 Article Selection Criteria for the CZMRAC

Since this thesis is inspired by Mbodj-Diop's thesis, it has been designed to follow her corpus parameters as closely as possible. As stated earlier, Mbodj-Diop created a corpus which was comprised of 1,177,611 running words. It consisted of 250 texts from years 2006-2015, with 50 texts from each year, collected from five internationally recognized medical journals. She chose three criteria for her article selection – genre, authorship, and journals under investigation. In this chapter, these three criteria will be utilized to introduce the selected medical research articles in the CZMRAC more closely.

Genre – The texts used for the corpus compilation in the present thesis are research articles from the field of medicine. At first, the ambition was to choose specific topics for the selected texts such as diabetes or dentistry, however, after delving into the medical journals published in English in the Czech Republic, it was clear that if specific topics were selected, there would not be enough research articles to create a corpus of the desired size. Consequently, it was determined that the research articles investigating any medical topic would be best to ensure the compilation of a corpus with at least 1,000,000 running words.

Authorship – The aim of the thesis is to study lexical bundles in English-written medical research articles produced by Czech authors and to compare them with those in the Anglo-American medical journals. Thus, it was critical to ensure the highest likelihood of the investigated texts were indeed produced by Czech speakers. To do so, only research articles from Czech medical journals that also publish articles in English were considered. Three of the journals are published by three major Czech universities – Charles University in Prague, University of Hradec Králové, and Palacký University in Olomouc. These universities have many international students so while selecting the articles, only those with Czech sounding names were downloaded for the corpus. However, even then, there is a chance that during the revision by the journals' editorial board, the texts may have been reviewed by native speakers, but all the selected magazines mention in their instructions for the authors that changes during revisions are only minor. In cases where the quality of English is insufficient, the article is rejected. This suggests that the majority of the articles truly are products of the authors without significant revisions by editors who are native speakers. It may also be that some authors with Czech sounding names are bilingual, but this limitation is almost impossible to control. In contrast, Mbodj-Diop (2016) did not distinguish between "native-like English" and "non-native English". In her opinion, the fact that a certain medical research article has been published in a reputable medical journal ensures sufficient quality.

Journals under Investigation – The investigated research articles were downloaded from five Czech medical journals that publish their articles in English. The selected journals are called *Acta Medica*, *Biomedical Papers*, *Central European Journal of Public Health*, *Physiological Research*, and *Prague Medical Journal*. To match the criterion of research articles covering a period of 10 years, only articles that were published during years 2012 to 2021 were downloaded. However, an exception was made in the case of CEJPH in the year 2021 because the journal did not provide open access for the issues published in this year. To compensate for the missing articles which corresponded to approximately 40,000 words, more articles were downloaded in the previous years. A summary of the collected articles will be provided in the next section.

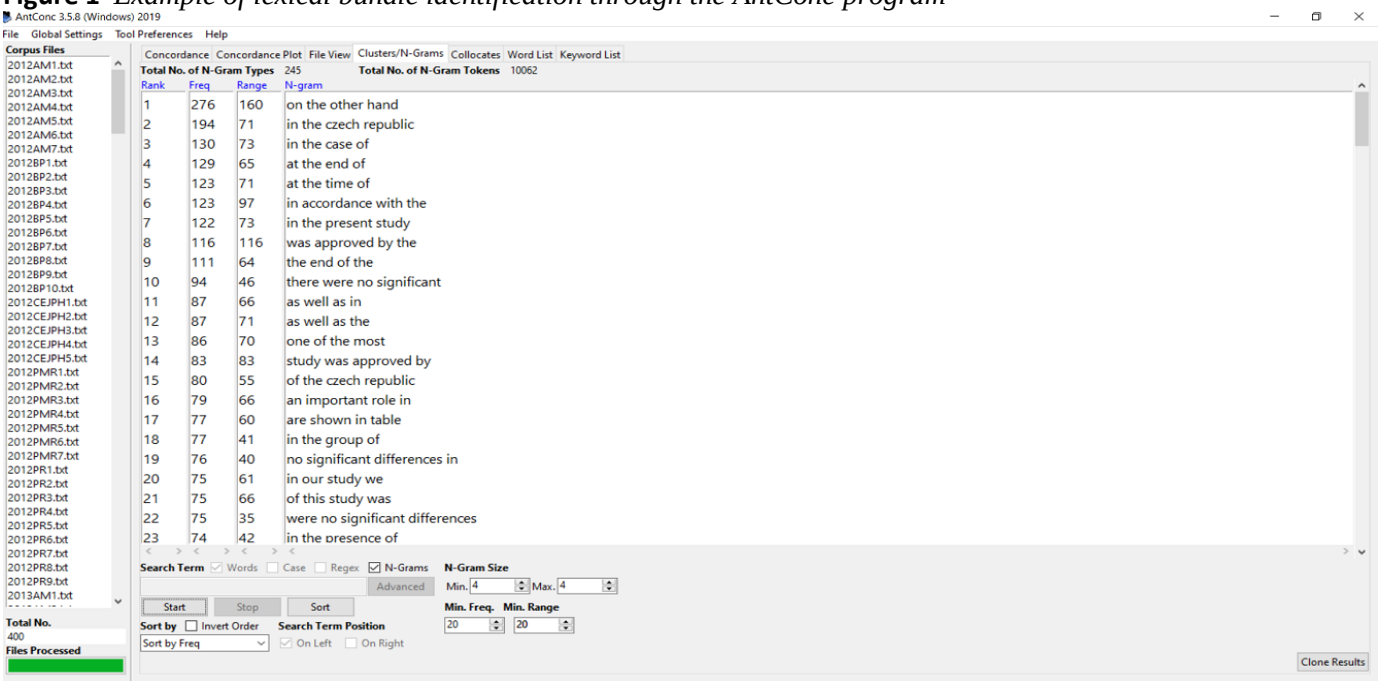
Besides these three criteria, Mbodj-Diop set a word count for her articles ranging from 3,000 to 7,000 words. Since the options for the Czech medical journals are fairly limited compared to the Anglo-American medical journals, this criterion for text selection was not followed. Therefore,

articles of any word count were compiled. The following chapters introduce the CZMRAC characteristics in greater detail.

3.1.2 The CZMRAC Compilation and Its Key Characteristics

For the compilation and analysis of the CZMRAC, a program called AntConc was utilized. This program, which is used especially in corpus linguistics to search and analyze electronic texts, was developed and introduced by Laurence Anthony of Waseda University in Tokyo. Its primary purpose is to identify word clusters, based on selected criteria, which enables to study the patterns of language. It facilitates the study of word lists, concordance, collocates, clusters/N-grams, and keyword lists.

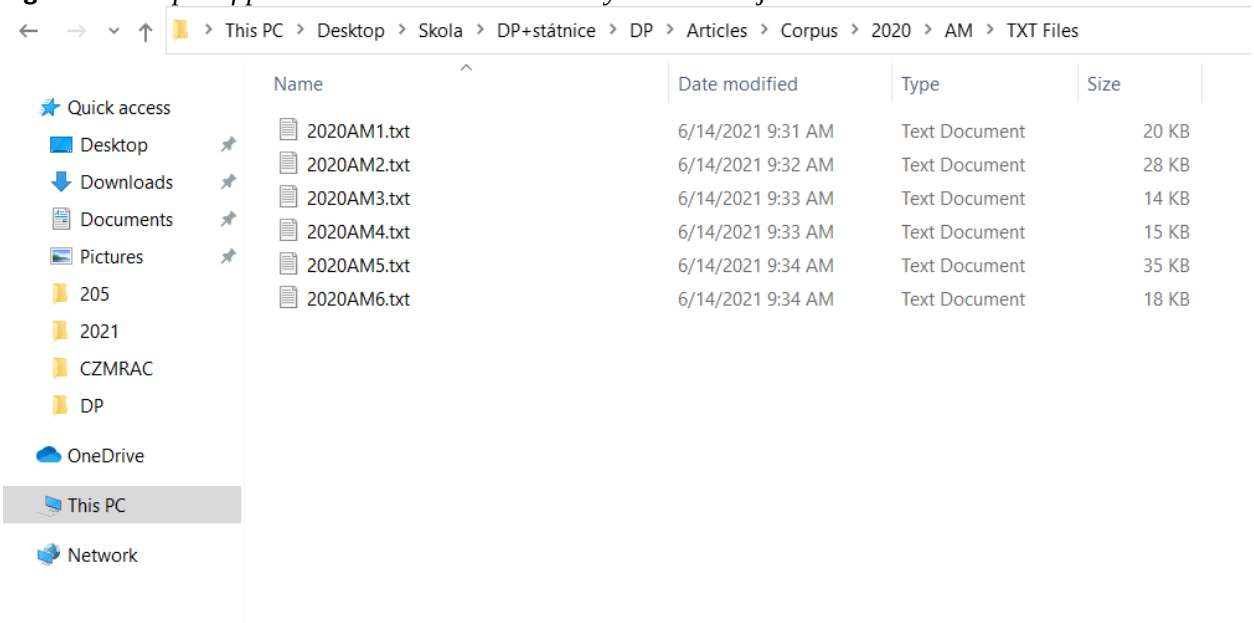
Figure 1 *Example of lexical bundle identification through the AntConc program*



To be able to study the selected medical research articles, it was necessary to clean them and convert them into AntConc program readable files because as Evans (2007, 2) says "… most corpus investigation software will not read the kind of complex embedded formatting associated with common word processing packages like Microsoft Word or pdfs." The process of the text preparation for the CZMRAC is described below.

Initially, all the available research articles based on the above-mentioned criteria were downloaded. When choosing the articles, only those that were called “Original Articles” were downloaded to increase the likelihood they were indeed produced by Czech authors and were not summaries or reviews of other studies already published in English. To organize the texts, a folder was created in which they were saved based on the year of publication. Afterwards, the texts were renamed according to the year of publication, the name of the magazine, and the order of the articles in which they were worked on. For example, an article from 2020, published in Acta Medica that was worked on first, was titled 2020AM1, as can be seen in Figure 2.

Figure 2 *Example of plain texts saved based on the year and the journal*



Next, an Excel file was created consisting of ten tables representing each year. This file served as an overview of how many texts had been downloaded, how many words each article had, and how many words there were in total. As the next step, each PDF article was opened in Microsoft Word, the language of the document was changed into English, and the text which was in two columns was changed into a single column. After that, the documents were cleaned. In each of them, all the metadata were deleted, and the body of the text was kept meaning that the final text started with the section “Introduction” and ended with the section “Conclusion”. Besides that, all the tables, figures, and their descriptions were deleted. When the cleaning process was done, the word count of each article was noted in the Excel file named “Word Count per Year” and then each article was saved as a plain text with UTF-8 coding.

The original plan was to collect 250 research articles, with 50 texts from each year, in the same manner as Mbodj-Diop did. However, after the first 100 documents it was clear that 250 texts would not suffice, because most of them were significantly shorter than the articles in the ENMRAC. As a result, the focus of having 50 articles each year was shifted to setting a goal of compiling at least 100,000 words for each year. This ensured that after having at least 100,000 per year, the final word count from ten years would be 1,000,000 running words and would be comparable with the size of the ENMRAC. Nevertheless, as mentioned in section 3.1.1, year 2021 was underrepresented because one of the medical journals (CEJPH) did not have open access for the articles published in 2021. This caused the word count from that year to be short approximately 40,000 words. To compensate for this shortage, more articles were downloaded for the previous years. Table 9 shows the representation of the research articles in each year and the total word count per journal for each year.

Table 9 *Summary of the collected articles and their wordcount per journal*

Journal	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012	Article count per journal	Word count per journal
AM	2	6	5	6	5	8	8	7	7	7	**61**	**139,979**
BP	13	12	13	12	15	15	10	11	7	10	**118**	**290,269**
CEJPH	-	5	5	5	5	6	5	6	8	5	**50**	**141,206**
PMR	2	2	4	3	5	5	6	2	6	7	**42**	**104,333**
PR	6	15	16	18	15	9	14	16	13	9	**129**	**398,820**
Totals	**23**	**39**	**43**	**44**	**45**	**43**	**43**	**42**	**40**	**38**	**400**	**1,074,607**

As shown in Table 9, the total of all medical research articles collected for the CZMRAC was 400, resulting in 1,074,607 running words. The totals of articles per journal per year differ because each journal publishes articles in different volumes, and some journals have more foreign contributing authors than others. For example, in years 2021, 2020 and 2014 the PMR is represented only by two articles. In the first instance, the number is so low because only two issues out of four have been published in 2021 so far. The other two instances are examples of years where most of the articles were written by foreign authors and thus did not meet the requirements for the article selection. On the other hand, the BM and PR are journals that provided many articles written by

Czech authors. For that reason, the research articles from these two journals represent more than a half of the CZMRAC.

3.2 Results and Discussion

This section of the analytical part will introduce the results along with their analysis.

3.2.1 Identified Lexical Bundles

Taking into consideration the criteria that Mbodj-Diop set for her research, below are provided the criteria for the present thesis:

1) only four-word lexical bundles are considered
2) they must occur at least 20 times in the CZMRAC
3) they must be used in at least 5% articles of the CZMRAC (20 of the 400 articles)
4) they must appear in at least two out the five selected journals
5) they cannot be interrupted by any other character

Based on the introduced criteria, the AntConc program initially extracted from the CZMRAC 245 word clusters. After a closer investigation and following the above-mentioned criteria, there were 41 word clusters that were excluded from the original list of lexical bundle candidates. The following list provides an overview of the eliminated multiword chunks. The numbers in the brackets behind each word cluster represent frequency and range in this order.

Word clusters including "et al" - et al in our (68/47), et al in the (58/49), et al on the (42/37), al in our study (39/31), et al in addition (37/30), et al reported that (36/24), et al it is (35/28), al on the other (33/30), et al however the (30/26), et al found that (29/26), et al it was (28/22), et al in this (26/24), et al showed that (25/25), al in the present (24/21), et al demonstrated that (24/21), et al as well (22/20)

Word clusters that proved to be lexical bundles with more than 4 words - approved by the ethics (55/55), by the ethics committee (55/55), faculty of medicine charles (30/25), of medicine charles university (30/25), to the best of (22/21), the best of our (22/21), institute for clinical and (21/20), for clinical and experimental (21/20)

Word clusters with possessive nouns - to the manufacturer's (55/44), the manufacturer's instructions (38/31), student's t-test (26/24), fisher's exact test (24/23)

Word clusters with proper names - mann whitney u test (45/31), st louis mo usa (32/22), the mann whitney u (26/21)

Word clusters with number slots - and stored at_°C (43/36), a total of_ patients (40/30), stored at_°C until (36/34), at_°C for min (27/21), shown in table_the (27/23), h light dark cycle (24/24), were stored at_°C (22/21)

Word clusters overlapping a sentence boundary - considered statistically significant results (34/34)

Word clusters specific to one journal - the institute of physiology (30/21)

Word clusters with three words - c-reactive protein crp (22/22)

As the list shows, the first group of lexical bundle candidates that was eliminated was the category with "et al". Such clusters often obtained number slots and crossed the sentence boundary (*et al. 2000. On the*). Word sequences including more than four words were the candidates that the AntConc software categorized as two lexical bundles even though it was only one. For example, *approved by the ethics committee* was divided into *approved by the ethics* and *by the ethics committee* with the same frequency and range. Word clusters with possessive nouns were eliminated because they did not qualify as four-word lexical bundles (*to the manufacturer's*). The word clusters with number slots were not used further in the study because Mbodj-Diop did not investigate them either and for the sake of comparability they were removed from the final list of lexical bundles as well. As with word sequences containing "et al", the lexical bundle candidate *considered statistically significant. Results* was removed from the original list as well because the bundle stretched through two sentences. The next eliminated clusters were those that were found only in one journal which turned out to be only one *the institute of physiology*. This lexical bundle was identified only in the research articles that were published in the *Physiological Research* journal. And finally, *C-reactive protein crp* was excluded from the final list because the AntConc software identified the word c-reactive as two words instead of one which turned a three-word bundle into a four-word bundle even if it was not.

After removing the word sequences that did not qualify as lexical bundles as defined above, the final list of 204 four-word lexical bundles remained. The fact that the total of lexical numbers in the CZMRAC was 204 is very interesting because the same total was identified in the ENMRAC

as well. However, it is necessary to note, that this finding is rather a coincidence and cannot be generalized. Nevertheless, it can be argued that four-word lexical bundles are present in both Czech and Anglo-American medical journals.

3.2.2 Most Frequent Four-Word Lexical Bundles in the CZMRAC

The aim of this section is to introduce the most frequent four-word lexical bundles in the CZMRAC and to briefly compare them with the most frequent lexical bundles in the ENMRAC. Table 10 shows the 50 most frequent lexical bundles that were found in the CZMRAC along with the top 50 frequent lexical bundles found in the ENMRAC. The lexical bundles that are matching are in bold letters and marked with green color.

Table 10 *Comparison of the top 50 most frequent lexical bundles in the CZMRAC and ENMRAC*

CZMRAC				ENMRAC			
	LB	F	R		LB	F	R
1.	on the other hand	276	160	1.	on the basis of	306	122
2.	in the Czech Republic	194	71	2.	in the placebo group	235	25
3.	in the case of	130	73	3.	with the use of	228	57
4.	**at the end of**	129	65	4.	**in the control group**	199	27
5.	**at the time of**	123	71	5.	**in the presence of**	139	50
6.	**in accordance with the**	123	97	6.	**at the time of**	137	83
7.	**in the present study**	122	73	7.	**the end of the**	129	59
8.	**was approved by the**	116	116	8.	in the absence of	117	67
9.	**the end of the**	111	64	9.	in the intervention group	115	13
10.	**there were no significant**	94	46	10.	**at the end of**	109	49
11.	as well as in	87	66	11.	an increased risk of	83	36
12.	**as well as the**	87	71	12.	of patients in the	79	27
13.	one of the most	86	70	13.	years of follow up	74	26
14.	study was approved by	83	83	14.	were randomly assigned to	72	43
15.	of the Czech Republic	80	55	15.	in the number of	71	39
16.	an important role in	79	66	16.	was associated with a	71	46
17.	**are shown in table**	77	66	17.	in the context of	69	46
18.	in the group of	77	41	18.	**the total number of**	66	41
19.	no significant differences in	76	40	19.	the primary end point	65	23
20.	in our study we	75	61	20.	the use of the	65	35

21.	of this study was	75	66	21.	has been shown to	64	45
22.	were no significant differences	75	35	22.	did not differ significantly	63	39
23.	**in the presence of**	74	42	23.	**there were no significant**	57	36
24.	is one of the	74	65	24.	for the treatment of	56	36
25.	**the results of the**	70	52	25.	the basis of the	56	42
26.	the study was approved	66	66	26.	the intention to treat	55	31
27.	in the development of	64	48	27.	end of the study	54	25
28.	the aim of this	64	60	28.	as compared with the	52	32
29.	are summarized in table	62	49	29.	**as well as the**	51	42
30.	**in the control group**	62	36	30.	**the results of the**	51	39
31.	the beginning of the	61	35	31.	did not differ between	50	22
32.	the ethics committee of	61	61	32.	the use of a	50	35
33.	this study was to	61	58	33.	were included in the	50	37
34.	at the beginning of	60	35	34.	presence or absence of	49	24
35.	according to the manufacturer's	59	48	35.	randomly assigned to receive	49	31
36.	in the pathogenesis of	59	39	36.	the proportion of patients	48	19
37.	of our study was	59	54	37.	associated with an increased	47	25
38.	at the age of	58	25	38.	**in the present study**	47	29
39.	the other hand the	57	51	39.	have been shown to	46	39
40.	at the same time	56	42	40.	with an increased risk	45	24
41.	ethics committee of the	56	56	41.	**in accordance with the**	44	39
42.	of the present study	56	42	42.	between the two groups	42	24
43.	aim of our study	55	51	43.	there was no significant	63	41
44.	body mass index bmi	51	45	44.	**was approved by the**	59	55
45.	in agreement with the	51	45	45.	as a result of	58	46
46.	**the total number of**	51	32	46.	these data suggest that	58	39
47.	in our study the	50	44	47.	was defined as a	41	32
48.	it is necessary to	49	46	48.	were more likely to	41	26
49.	in the treatment of	48	38	49.	**are shown in table**	39	33
50.	was used for the	48	39	50.	in the risk of	38	22

Out of the first 50 most frequent lexical bundles in the CZMRAC and ENMRAC, there were 13 matching – *at the end of, at the time of, in accordance with the, in the present study, was approved by the, the end of the, there were no significant, as well as the, are shown in table, in the presence*

of, the results of the, in the control group, the total number of. Translating that into percentage, 26% of the 50 most frequent bundles were matching. The most frequent lexical bundle in the CZMRAC was *on the other hand* with the frequency of 276 and range of 160 articles. In the ENMRAC, it was *on the basis of* with frequency of 306 and range of 122 articles. Interestingly, *on the basis of* did not make it on the list of top 50 CZMRAC bundles having appeared only 45 times in 38 articles. The same is true for *on the other hand* which appeared only 30 times in the ENMRAC, unlike in the CZMRAC, where this bundle occurred 276 times. In regard to structural classification, both top bundles start with a prepositional phrase. *On the basis of* belongs to the category "Prepositional phrase with embedded of-phrase fragment" and *on the other hand* falls into the category "Other prepositional phrase fragments". Besides these two prepositional structures, another seven structures were identified in the top 50 bundles in the CZMRAC. This means that out of the 12 structures introduced by Biber et al (1999), the first 50 most frequent bundles included nine of them. The categories that were lacking were "(Verb phrase) + that-clause fragment", "(Verb/adjective) + to-clause fragments", and "Adverbial clause fragment categories". In addition, as the most frequent lexical bundle in the CZMRAC suggests, the most common bundles in the 50 most frequent bundles were those with prepositional phrases representing a little over 50% (26 bundles). A more elaborated description of the structures of all identified lexical bundles in the CZMRAC is provided in section 3.2.4.

As for the functional classification of the top 50 four-word lexical bundles in the CZMRAC, the most frequent bundle, *on the other hand*, served as a transition signal expressing a contrastive relationship. In the EMRAC, *on the basis of* served as a framing signal limiting and specifying conditions. The functional category with the most lexical bundles in the top 50 was the category with bundles describing procedure. In total, there were 14 bundles that belonged to this category which accounted for 28% of the functions in the 50 most frequent lexical bundles in the CZMRAC. The least represented category was the category with the participant-oriented bundles having only one lexical bundle, *it is necessary to*, with an engagement function.

3.2.2.1 Summary

The results provided in this section answer the first research question: "*What are the most frequent bundles in the CZMRAC?*" Table 10 provides a list of them and shows their frequency and their

range. Along with that, the table also offers a comparison with the top 50 four-word lexical bundles in the ENMRAC. Drawing from the structural and functional analyses of the 50 most frequent lexical bundles in the CZMRAC, it can be concluded that the most frequent bundles in the investigated English-written medical research articles produced by Czech authors start with prepositional phrases and function as procedure signals. The next part of the thesis is devoted to the comparison of the frequency distribution of all the matching lexical bundles in the CZMRAC and ENMRAC.

3.2.3 Comparison of the Frequency Distribution of the Matching Bundles

Out of the 204 identified four-word lexical bundles in the CZMRAC, there were 60 that were matching with the bundles in the ENMRAC which is percentagewise 29.41%. To compare the frequency of the matching lexical bundles, a test for statistical significance was performed. The test was run by an online calculator called *Corpus Frequency Wizzard Tool* which compared "the frequency of two samples across two different data sets" (Zagolová 2017, 54). Figure 3 shows the frequency comparison of the most frequent four-word lexical bundle in the CZMRAC, *on the other hand*, which also appeared in the ENMRAC and Figure 4 shows results of this comparison.

Figure 3 *Example of frequency comparison in the CZMRAC and ENMRAC via Corpus Frequency Wizzard Tool*

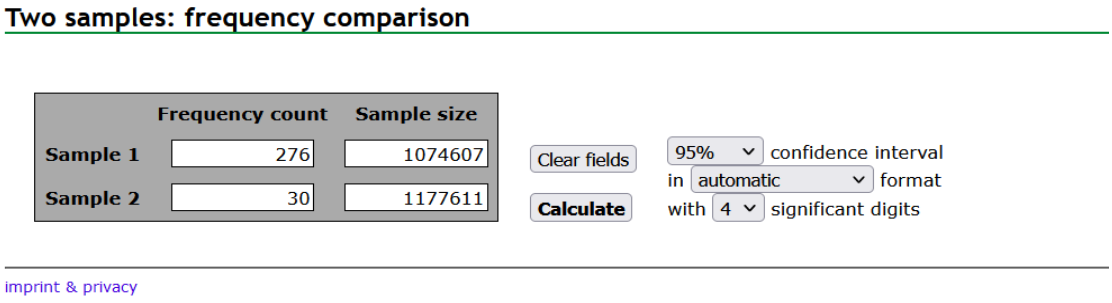


Figure 4 *Example of the corpus frequency test results*

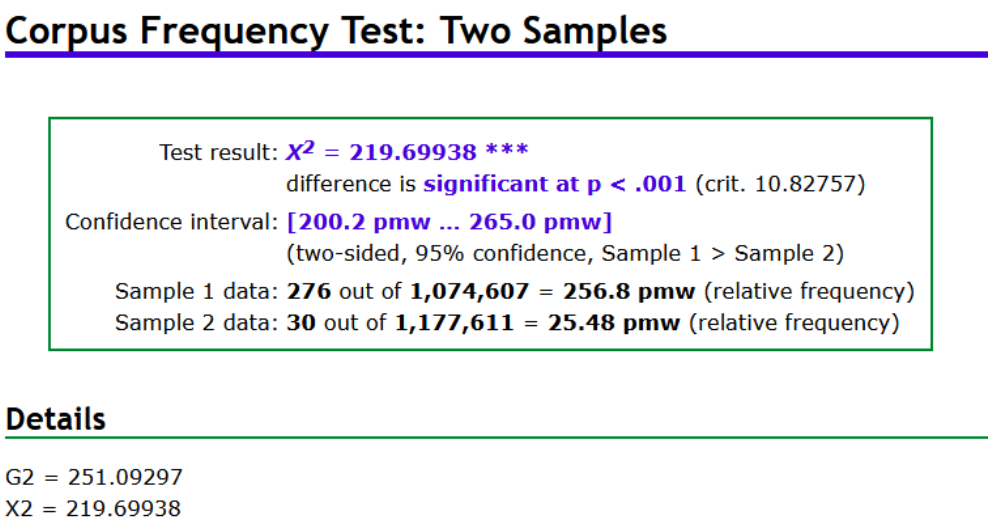


Table 11 summarizes the results of the comparison of the frequency distribution of the matching bundles in the CZMRAC and ENMRAC. The bold letters mark the higher frequency between the two examined corpora and the green color marks the lexical bundles showing a statistically significant difference in their frequency.

Table 11 *Comparison of the frequency distribution of the matching bundles*

No.	Matching bundles	Frequency in the CZMRAC	Frequency in the ENMRAC	Statistically significant	p-value
1.	on the other hand	**276**	30	yes	p < .001
2.	in the case of	**130**	20	yes	p < .001
3.	at the end of	**129**	109	no	
4.	at the time of	123	**137**	no	
5.	in accordance with the	**123**	44	yes	p < .001
6.	in the present study	**122**	47	yes	p < .001
7.	was approved by the	**116**	59	yes	p < .001
8.	the end of the	111	**129**	no	
9.	there were no significant	**94**	63	yes	p < .01
10.	as well as in	**87**	29	yes	p < .001
11.	as well as the	**87**	51	yes	p < .001
12.	one of the most	**86**	27	yes	p < .001
13.	study was approved by	**83**	24	yes	p < .001

14.	an important role in	**79**	20	yes	$p < .001$
15.	are shown in table	**77**	39	yes	$p < .001$
16.	in the presence of	74	**139**	yes	$p < .001$
17.	is one of the	**74**	25	yes	$p < .001$
18.	the results of the	**70**	51	yes	$p < .05$
19.	in the development of	**64**	28	yes	$p < .001$
20.	in the control group	62	**199**	yes	$p < .001$
21.	in the pathogenesis of	**59**	28	yes	$p < .001$
22.	at the same time	**56**	28	yes	$p < .001$
23.	the total number of	51	**66**	no	
24.	in the treatment of	**48**	26	yes	$p < .01$
25.	as a result of	47	**58**	no	
26.	for the treatment of	47	**56**	no	
27.	in the regulation of	**46**	22	yes	$p < .01$
28.	on the basis of	45	**306**	yes	$p < .001$
29.	to the development of	33	28	no	
30.	it is important to	**40**	20	yes	$p < .01$
31.	there was no significant	40	**63**	no	
32.	in the absence of	38	**117**	yes	$p < .001$
33.	no significant difference in	**37**	**37**	match	
34.	were included in the	37	**50**	no	
35.	in the general population	34	**36**	no	
36.	in the context of	33	**69**	yes	$p < .01$
37.	there was a significant	**32**	21	no	
38.	with the exception of	32	**38**	no	
39.	in this study we	31	**35**	no	
40.	were approved by the	31	32	no	
41.	informed consent was obtained	30	**31**	no	
42.	play a role in	**30**	21	no	
43.	studies have shown that	30	**31**	no	
44.	at the level of	**29**	23	no	
45.	has been shown to	29	**64**	yes	$p < .01$
46.	we did not observe	**29**	26	no	
47.	no significant differences between	28	**38**	no	

48.	in the number of	27	**71**	yes	p < .001
49.	it is possible that	27	**38**	no	
50.	there was no difference	**27**	23	no	
51.	studies are needed to	**26**	21	no	
52.	the results of our	**26**	22	no	
53.	were excluded from the	26	**29**	no	
54.	significant differences between the	25	**28**	no	
55.	animal care and use	**24**	**24**	match	
56.	a significant increase in	23	**36**	no	
57.	this is the first	**23**	20	no	
58.	with respect to the	23	**32**	no	
59.	with the use of	22	**228**	yes	p < .001
60.	by the presence of	20	**21**	no	

As the results in Table 11 indicate, out of the 60 matching bundles, the *Corpus Frequency Wizzard Tool* determined that the frequency of 28 lexical bundles was significantly different. These bundles were: *on the other hand, in the case of, in accordance with the, in the present study, was approved by the, there were no significant, as well as in, as well as the, one of the most, study was approved by, an important role, are shown in table, in the presence of, is one of the, there results of the, in the development of, in the control group, in the pathogenesis of, at the same time, in the treatment of, in the regulation of, on the basis of, it is important to, in the absence of, in the context of, has been shown to, in the number of, with the use of.* Out of the 28 matching bundles with a significant statistical difference in frequency, there were eight that were overrepresented in the ENMRAC (*in the presence of, in the control group, on the basis of, in the absence of, in the context of, has been shown to, in the number of, with the use of)* and 20 which were overrepresented in the CZMRAC (*on the other hand, in the case of, in accordance with the, in the present study, was approved by the, there were no significant, as well as in, as well as the, one of the most, study was approved by, an important role in, are shown in table, is one of the, the results of the, in the development of, in the pathogenesis of, at the same time, in the treatment of, in the regulation of, it is important to*.) This shows that even if there are 60 matching bundles in both corpora, their frequency differed. Moreover, most of the significantly different matching bundles were overrepresented in the CZMRAC as shown above.

It should also be noted that the most frequent lexical bundle in the CZMRAC, *on the other hand* (F 276), had a much lower frequency in the ENMRAC (F 30). The same is true for the most frequent four-word lexical bundle in the ENMRAC, *on the basis of,* which appeared 306 times in the Mbodj-Diop's corpus and only 45 times in the CZMRAC. This means that *on the basis of* showed the largest difference in the frequency distribution among the matching bundles.

In addition, it is also important to comment on the matching lexical bundles that showed no statistically significant difference in their distribution. Altogether, there were 32 of them (*at the end of, at the time of, the end of the, the total number of, as a result of, for the treatment of, to the development of, there was no significant, were included in the, in the general population, there was a significant, with the exception of, in this study we, were approved by the, informed consent was obtained, play a role in, studies have shown that, at the level of, we did not observe, no significant differences between, it is possible that, there was no difference, studies are need* to, *the results of our, were excluded from the, significant differences between the, a significant increase in, this is the first, with respect to the, by the presence of, animal care and use, no significant difference in*). Among the matching lexical bundles, there occurred two bundles with the identical frequency. These bundles were *no significant difference in* with frequency of 37 hits and *animal care and use* with frequency of 24 hits. The presented bundles in this paragraph appear to be used in the medical research articles written by Czech authors and in the Anglo-American journals with similar frequency. However, it should be noted that these results cannot be generalized and they are true only for the medical research articles that were in the CZMRAC and ENMRAC.

3.2.3.1 Summary

Overall, it can be concluded that a considerable number of the 204 identified four-word lexical bundles in the CZMRAC matches with the lexical bundles found in the ENMRAC accounting for 29.41%, equaling to a total of 60 bundles. The more detailed analysis of the results showed that there were 28 matching lexical bundles that were significantly different in their frequency and 32 matching bundles whose frequency did not differ significantly. Among the second group were also two lexical bundles with the same frequency (*no significant difference in* and *animal care and use*). In addition, most of the matching lexical bundles with a significant difference in frequency were overrepresented in the CZMRAC (20 bundles out of 28).

3.2.4 Structural Classification of the CZMRAC Lexical Bundles

After touching upon the structural classification of the 50 most frequent lexical bundles in the CZMRAC, Table 12 provides a full classification of the identified lexical bundles. The categorization is based on Biber et al. (1999).

Table 12 *Structural classification of the identified bundles in the CZMRAC*

Noun phrase with of-phrase fragment
the end of the, one of the most, the results of the, the aim of this, the beginning of the, the ethics committee of, ethics committee of the, aim of this study, the total number of, the aim of our, on the basis of, committee of the institute, the aim of the, the declaration of Helsinki, the majority of patients, first faculty of medicine, the results of our, a wide range of, the ethical committee of, one of the main, a large number of, best of our knowledge, ethical committee of the, small number of patients, results of our study, aim of our study, group of patients with
Noun phrase with other post-modifier fragment
an important role in, no significant differences in, important role in the, the fact that the, no significant difference in, the Czech Republic in, the relationship between the, a significant decrease in, no significant differences between, significant differences in the, role in the development, significant differences between the, the difference between the, a decrease in the, a significant increase in, significant difference in the
Prepositional phrase with embedded of-phrase fragment
in the case of, at the end of, at the time of, in the group of, in the presence of, in the development of, in the pathogenesis of, at the age of, in the treatment of, for the treatment of, in the regulation of, as a result of, at the department of, to the development of, of the institute of, in the absence of, in the area of, as a marker of, in the majority of, in the context of, with the exception of, for the development of, in the form of, at the level of, in the course of, in the number of, with the results of, for the purpose of, in any of the, with the use of, by the presence of, as a part of, at the beginning of
Other prepositional phrase fragments
on the other hand, in the Czech Republic, in accordance with, in the present study, as well as in, of the Czech Republic, in our study we, of this study was, in the control group, according to the manufacturer's, of our study was, at the same time, of the present study, in agreement with the, in our study the, in comparison with the, by the fact that, in the general population, by the ethical committee, by the local ethics, in this study the, in this study we, in relation to the, of the university hospital, with regard to the, with respect to the, of this study is, in accordance with the
Be + noun/adjective phrase
were no significant differences, is one of the, is in agreement with, was significantly higher in, are in agreement with, is the most common, was not statistically significant, was significantly lower in, is associated with, is in accordance with, was no significant difference

Passive verb + prepositional phrase fragment
was approved by the, are shown in table, are summarized in table, was used for the, included in the study, is shown in table, were included in the, approved by the local, is based on the, be explained by the, performed in accordance with, approved by the ethical, was found to be, are presented in table, were approved by the, was observed in the, were considered statistically different, is considered to be, was found in the, were excluded from the, could be explained by, were found in the, was used as a, were enrolled in the, were found to be, divided into two groups, enrolled in the study, were used for the, are expressed as mean, was obtained from all, was considered statistically different
Anticipatory it + verb/adjective phrase
it is necessary to, it is important to, it is possible to, it is possible that, it should be noted, it has been shown
(Verb phrase) + that-clause fragment
studies have shown that, has been shown that, should be noted that
(Verb/adjective) + to-clause fragment
was to evaluate the, has been shown to, were considered to be, studies are needed to, study was to compare, study was to evaluate, our study was to, present study was to
Adverbial clause fragment
when compared to the
Pronoun/noun phrase + be (+…)
there were no significant, this study was to, there was no significant, there were no differences, there was a significant, there was no difference, this is the first, the present study was
Other expressions
body mass index bmi, study was approved by, the study was approved, play an important role, test was used to, we did not find, the local ethics committee, the present study we, plays an important role, test was used for, statistical analysis was performed, consent was obtained from, did not find any, informed consent was obtained, play a role in, no statistically significant differences, we did not observe, the Czech Republic and, no statistically significant difference, results are expressed as, analyses were performed using, did not observe any, patients were treated with, animal care and use, clinical and experimental medicine, was performed using a, experiments were performed in, protocol was approved by, care and use committee, the Czech Republic the, the other hand the, as well as the

As illustrated in Table 12, all 12 categories of lexical bundle structures were found in the CZMRAC, which demonstrates that all structural categories, as introduced by Biber (1999), play an important role in the production of the medical research articles. This finding agrees with Mbodj-Diop's findings. The most represented categories are those including prepositional and noun phrases, passive verbs, and other expressions. A more detailed summary of structural

classification is shown in Table 13, providing the totals of lexical bundles in each category, their percentage in the CZMRAC, and an overall comparison of the CZMRAC structures and the ENMRAC structures.

Table 13 *Comparison of the lexical bundle structures in the CZMRAC and ENMRAC*

Structure	Number of bundle types		% of total number of bundles	
*The most frequent bundles in the category and their frequency	CZMRAC	ENMRAC	CZMRAC	ENMRAC
Prepositional phrase with embedded of-phrase fragment (*in the case of, F 130**)	33 (1)	36 (1)	29.90%	31.37%
Other prepositional phrase fragments (*on the other hand, F 276*)	28	28		
Noun phrase with of-phrase fragment (*the end of the, F 111*)	27 (2)	32 (2)	21.07%	23.03%
Noun phrase with other post-modifier fragment (*an important role in, F 79*)	16	15		
Other expressions (*study was approved by, F 83*)	32 (3)	29 (3)	15.68%	14.22%
Passive verb + prepositional phrase fragment (*was approved by the, F 116*)	31 (4)	26 (4)	15.19%	12.75%
Be + noun/adjective phrase (*were no significant differences, F 75*)	11 (5)	7 (7)	5.39%	3.43%
(Verb/adjective) + to-clause fragment (*our study was to, F 42*)	8 (6)	9 (6)	3.92%	4.41%
Pronoun/noun phrase + be (+…) (*there were no significant, F 94*)	8 (7)	6 (8)	3.92%	2.94%
Anticipatory it + verb/adjective phrase (*it is necessary to, F 49*)	6 (8)	2 (10)	2.94%	0.98%
(Verb phrase) + that-clause fragment (*studies have shown that, F 30*)	3 (9)	11 (5)	1.47%	5.40%
Adverbial clause fragment (*when compared to the, F 23*)	1 (10)	3 (9)	0.49%	1.47%

On the basis of the information provided in Table 13, the category with the most lexical bundles is the category "Prepositional phrase with embedded of-phrase fragment" which consists of 33 bundles with *in the case of* as the most frequent one. Together with the category "Other

prepositional phrase fragments", these two categories make up 29.90% of all the structures in the CZMRAC which corresponds to 61 bundles. This finding corresponds with that of Mbodj-Diop's whose corpus included 64 four-word lexical bundles falling into the categories starting with prepositional phrases which accounts for 31.37% of all lexical bundle structures in the ENMRAC. The second largely represented group in both corpora was the group consisting of noun phrases (Noun phrase with of-phrase fragment + Noun phrase with other post-modifier fragment). In the CZMRAC, this group was represented by 43 bundles (21.07%). The percentage of the noun-phrase categories in the ENMRAC was 23.03% which is again very close to the percentage in the CZMRAC. Moreover, an interesting observation is that the total of the lexical bundles in the category "Other prepositional phrase segments" is identical in both corpora (28 bundles). Right behind the prepositional phrase and noun phrase categories are categories "Other expressions" and "Passive verb + prepositional phrase fragment" which again corresponds with Mbodj-Diop's findings. In contrast, the category that showed the largest difference in the results was the category "(Verb phrase) + that-clause fragment". In the CZMRAC, there were only three identified bundles that belonged to this category whereas in the ENMRAC there were 11. This finding translates into the biggest percentage gap in the results (3.93%). Moving on to the least represented categories, in the CZMRAC they were "(Verb phrase) + that-clause fragment" with three lexical bundles and "Adverbial clause fragment" with only one bundle. The latter category was one of the two categories that had the least bundles in the ENMRAC as well having three bundles. The second category was the category "Anticipatory it + verb/adjective phrase" which had only two lexical bundles. The rest of the results were fairly consistent with the Mbodj-Diop's results.

3.2.4.1 Summary

In conclusion, the most frequent structures of the four-word lexical bundles in the CZMRAC were those that started with prepositional phrases. Examples of such bundles are as *in the case of* or *on the other hand*. This group of lexical bundles made up 29.90% of all the identified structures and along with the categories starting with noun phrases they represented a little more than half of the structures in the CZMRAC (50.97%). Mbodj-Diop's results showed similar findings. The most frequent structural category of the lexical bundles in the ENMRAC was the category "Prepositional phrase with embedded of phrase" with 31.37%. Moreover, the majority of the structures that Mbodj-Diop identified fell into the categories starting with the prepositional and noun phrases as well accounting for 54.4%. These results suggest that the dominating structures

of four-word lexical bundles in the selected medical research articles were the structures starting with prepositional or noun phrases whether they were written by Czech authors or were published in the prestigious Anglo-American medical journals.

3.2.5 Functional Classification of the CZMRAC Lexical Bundles

As mentioned in section 2.4.2, the functions of the identified bundles in the CZMRAC were going to be determined based on Hyland (2008). However, as in Mbodj-Diop's case, a relatively high occurrence of the statistical lexical bundles called for an additional subcategory "Statistical" which was introduced by Cortes (2004), and which was added to the Hyland's category "Resultative". The functional classification of the identified four-word lexical bundles in the CZMRAC is shown in Table 14.

Table 14 *Functional classification of the identified bundles in the CZMRAC*

Category	Subcategory	Bundles
Research-oriented bundles	Location (time/place)	in the Czech Republic, at the end of, **in the present study**, the end of the, in our study we, the beginning of the, in the control group, at the beginning of, the Czech Republic the, the Czech Republic in, in the general population, of the university hospital, the Czech Republic and, at the department of, in this study the, in this study we, at the level of, in the course of, **at the same time,** in our study the, of the Czech Republic, at the time of
	Procedure (methodology or purpose of research)	study was approved by, an important role in, the study was approved, was used for the, for the treatment of, play an important role, important role in the, consent was obtained from, informed consent was obtained, were used for the, was obtained from all, protocol was approved by, experiments were performed in, by the presence of, were enrolled in the, enrolled in the study, patients were treated with, were excluded from the, was used as a, analyses were performed using, were approved by the, statistical analysis was performed, for the development of, play a role in, was approved by the, in the development of, with the use of, was performed using a, approved by the local, approved by the ethical, performed in accordance with,

		the aim of the, test was used for, for the purpose of, study was to compare, study was to evaluate, divided into two groups, was to evaluate the, present study was to, of this study was, the aim of this, this study was to, aim of this study, of our study was, the aim of our, aim of our study, test was used to, our study was to, were included in the, the present study was, the present study we, according to the manufacturer's, of the present study, included in the study, of this study is, as a part of, **in accordance with the, in agreement with the**
	Quantification + Statistical	there were no significant, no significant differences in, were no significant differences, the total number of, there was no significant, was significantly higher in, no significant difference in, there were no differences, was not statistically significant, no statistically significant differences, a wide range of, in any of the, a decrease in the, a significant increase in, a large number of, one of the main, significant differences between the, was no significant difference, was considered statistically significant, the majority of patients, was significantly lower in, in the number of, a significant decrease in, one of the most, small number of patients, significant differences in the, in the majority of, no statistically significant difference, the difference between the, significant difference in the, there was a significant, is one of the, no significant differences between, there was no difference
	Description (qualities or properties of material)	is the most common, this is the first, the fact that the, by the fact that, as a marker of, plays an important role, is considered to be
	Topic (related to the field of research)	animal care and use, the declaration of Helsinki, by the ethical committee, by the local ethics, the ethical committee of, ethical committee of the, care and use committee, clinical and experimental medicine, the ethics

Text-oriented bundles		committee of, ethics committee of the, body mass index bmi, of the institute of, committee of the institute, first faculty of medicine, the relationship between the, the local ethics committee
	Transition signals (contrastive, additive links)	on the other hand, **at the same time**, in comparison with the, as well as in, as well as the, the other hand the, **in agreement with the, in accordance with the**
	Resultative signals (comparison, inferential, causative relations)	the results of the, as a result of, with the results of, results of our study, were found to be, did no observe any, were found in the, the results of our, we did not observe, is associated with a, did not find any, we did not find, are in agreement with, was found to be, is in accordance with, **in accordance with the**, is in agreement with, **in agreement with the**, were considered to be, were considered statistically significant, to the development of, it has been shown, studies have shown that, has been shown that, has been shown to, results are expressed as, was found in the, did not observe any, when compared to the, are expressed as mean
	Structuring signals	are shown in table, is shown in table, are presented in table, **are summarized in table,** in our previous study, **in the present study**
	Framing signals (limiting conditions)	in the case of, in the treatment of, in the regulation of, in the context of, group of patients with, with regard to the, with respect to the, in relation to the, on the basis of, with the exception of, at the age of, in the presence of, in the pathogenesis of, in the absence of, is based on the, in the area of, was observed in the, role in the development, in the group of
Participant-oriented bundles	Stance features (writer's attitudes and evaluations)	it is possible to, it is possible that, could be explained by, best of our knowledge, be explained by the

	Engagement features (address the readers directly)	it is important to, it is necessary to, studies are needed to, it should be noted, should be noted that

The lexical bundles that are in bold letters have been classified as having more than one function. Altogether, there were four bundles that fulfilled more than one function. Two of them, *in agreement with the* and *in accordance with the* were found to serve three different functions. In the case of *in agreement with*, they were resultative, transition, and procedure and in the case of *in accordance with the* they were procedure, resultative, and transition. The lexical bundles serving more than one function in the CZMRAC are introduced below.

In accordance with the

1) Procedure - *The study was performed* ***in accordance with the*** *Declaration of Helsinki and approved by IKEM Ethical Committee (REC number: MEK – 2442/10/A 11-01-02).* (2015PR9)

2) Resultative signals - *These data are* ***in accordance with the*** *observations of previous studies, where treatment with 7NI did not cause any significant change…* (2015PR1)

3) Transition signal (beginning of a paragraph) - ***In accordance with the*** *theory of successful ageing, many older adults regard themselves as happy and well, even in the presence of disease or disability*. (2013CEJPH5)

In agreement with the

4) Resultative - *This is* ***in agreement with the*** *results of multivariate analysis which showed that longer AFL CL in women was not gender-related.* (2012PR2)

5) Transition signal - *Disc-related diseases of the lumbar vertebral column caused by the repeated vertical effects of whole-body vibration"). **In agreement with the** Recommendation, the dorsopathies can be acknowledged as occupational diseases…* (2016CEJPH3)

6) Procedure - *All experiments were performed* ***in agreement with the*** *Animal Protection Law of the Czech Republic (311/1997) and were approved by the Ethics Committee of the Institute for Clinical and Experimental Medicine.* (2018PR6)

In the present study

7) Structuring signal - ***In the present study,*** *a high prevalence of low T-score values (corresponding to values for osteoporosis) was found in the patients aged over 50 years.* (2015BP2)

8) Location - *Due to a very variable and in part very low staining intensity of CB immunoreactive neurons with unsatisfactory staining of dendritic tree, CB immunostained sections, unsuitable for quantitative analysis, were not used* ***in the present study****.* (2021PR5)

At the same time

9) Location - *To eliminate the influence of circadian SBP variation, the measurements were always done approximately* ***at the same time*** *of day (between 8:00 and 10:00 a.m.).* (2015PR14)

10) Transition signal - ***At the same time****, Lewis rats showed a significantly better survival rate as compared with Wistar rats.* (2014PR15)

After a brief discussion of the lexical bundles multifunctionality in the CZMRAC, Table 14 provides an overview of the distribution of the lexical bundle functions in the CZMRAC. It is necessary to note that the total of all functions exceeds the total of the identified lexical bundles. As illustrated above, it is a result of multifunctionality, because multiple lexical bundles were assigned more than one function resulting in a total of 210 identified functions.

Table 15 *Distribution of the lexical bundle functions in the CZMRAC*

Research-oriented		Text-oriented		Participant-oriented	
Location	22	Transition signals	8	Stance features	5
Procedure	59	Resultative signals	29	Engagement features	5
Quantification	34	Structuring signals	6		
Description	7	Framing signals	19		
Topic	16				
Number of bundles	**138**	**Number of bundles**	**62**	**Number of bundles**	**10**
Total %	**65.71%**	**Total %**	**29.52%**	**Total %**	**4.76%**

Looking at Table 15, it is evident that the predominant functional category of lexical bundles in the CZMRAC is the "Research-oriented category" with 65.71%. This means that the research oriented four-word lexical bundles represent more than a half of the identified functions. This

finding is consistent with the results of the Mbodj-Diop study. As Table 15 shows, the most represented subcategory among the research-oriented bundles is the "Procedure" category with 59 four-word lexical bundles. Some examples of such bundles are *the aim of this, were enrolled in the*, or *were approved by*. It is not surprising that this subcategory had the most bundles because a significant part of research articles consists of methodology and procedure description. The next largely represented subcategories were "Quantification" followed by "Location", and "Topic". On the other hand, the subcategory with the lowest total of bundles was "Description" (7 lexical bundles).

The lexical bundles in the "Text-oriented category" made up 29.52% of the identified functions. The most frequent subcategory in this group of functions was "Resultative signals" with 29 lexical bundles which seconds the fact that lexical bundles are necessary for the analysis of the results. The other subcategories were represented as well. The framing signals were the second most represented subcategory with 19 lexical bundles, followed by the transition signals consisting of 8 bundles, and by the structuring signals with 6 bundles, which was the least represented subcategory.

The third category of "Participant-oriented bundles" consisted of the least lexical bundles. There were only 10 lexical bundles which belonged to this category being equally divided between the two subcategories "Stance features" and "Engagement Features". The former subcategory is represented by bundles *it is possible to, it is possible that, could be explained by, best of our knowledge, be explained by the* and the latter by *it is important to, it is necessary to, studies are needed to, it should be noted, should be noted that*. These results once again indicate that writer's attitudes and evaluations are not common in medical writing as well as it is rare for the authors to address the readers directly.

To compare the distribution of lexical bundles functions in the CZMRAC and ENMRAC, Table 16 is provided below.

Table 16 *Comparison of the lexical bundle functions in the CZMRAC and ENMRAC*

	Research-oriented		Text-oriented		Participant-oriented	
	CZMRAC	ENMRAC	CZMRAC	ENMRAC	CZMRAC	ENMRAC
Number of bundles	138	122	62	72	10	14
Total %	65.71%	58.65%	29.52%	34.62%	4.76%	6.73%

Based on Table 16, it can be concluded that the distribution of the lexical bundle functions in the CZMRAC and ENMRAC does not differ significantly. In both corpora, it was the research-oriented lexical bundles that were used predominantly. In the CZMRAC, they accounted for 65.71% and in the ENMRAC for 58.65% which equals to a difference of 15 bundles. The category of the text-oriented bundles consisted of considerably less bundles in both corpora, but still made up a significant percentage of the overall functional distribution in both of them. In the CZMRAC, they were represented by 29.52% and in the ENMRAC by 34.62%. The participant-oriented bundles were found to be the least frequent bundles which was true for both, the CZMRAC and ENMRAC. Ten bundles were identified in the CZMRAC accounting for 4.76% of the total distribution and 14 bundles were found in the ENMRAC which translated into 6.73%.

Summary 3.2.5.1

The purpose of this section was to analyze the functions of the four-lexical bundles in the CZMRAC and to compare them with the lexical bundle functions in the ENMRAC. As illustrated in Tables 13, 15, and 16, the majority of four-word lexical bundles in the CZMRAC served as research-oriented bundles. They accounted for 65.71% with 138 lexical bundles. These bundles were mostly used to describe the procedure and the methodology of the research. The second most represented functional category was the "Text-oriented category" which made up 29.52% of the identified functions. The participant-oriented lexical bundles were least represented in the CZMRAC with only 10 bundles. In regard to comparing the distribution of lexical bundle functions in the CZMRAC with the ENMRAC results, the data propose that the overall distribution of the lexical bundle functions is very similar. Both corpora show that the most frequent functions in the selected English-written medical research articles are research-oriented. Furthermore, the research-oriented lexical bundles in both the CZMRAC and the ENMRAC accounted for more than half of the identified functions. In the ENMRAC, they accounted for 58.65% and in the CZMRAC, they made up 65.71% of the overall functional distribution. In both cases, the text-oriented bundles ended up taking a second position with 29.52% in the CZMRAC and 34.62% in the ENMRAC and the participant-oriented bundles demonstrated the lowest presence. Based on these results, it can be concluded that the use and distribution of lexical bundle functions in the

medical research articles published by Czech authors appears to be consistent with that of the Anglo-American medical journals. Both corpora comprised of a whole variety of lexical bundles that served various functions. As a matter of fact, all the functions introduced by Hyland (2008) were identified in both corpora even if some were less prevalent than others. Moreover, both corpora consisted of lexical bundles that served more than one function. Some examples of such bundles are *at the same time* (location and transition) or *in the present study* (structuring signal and location).

4 Conclusion

The aim of the present thesis was to identify and analyze four-word lexical bundles in English medical research articles published by Czech authors and to compare their frequency, structures, and functions with four-word lexical bundles that had been identified by Mbodj-Diop (2016) in prestigious Anglo-American medical journals. The objective of the analysis was to answer the following research questions:

1. What are the most frequent four-word lexical bundles in the corpus of English medical research articles published by Czech authors (CZMRAC)?
2. What are the structures of four-word lexical bundles used in the CZMRAC?
3. What functions do four-word lexical bundles perform in the CZMRAC?
4. How are the identified four-word lexical bundles in the CZMRAC different, if at all, compared to four-word lexical bundles in the English Medical Research Article Corpus (ENMRAC) in their frequency, structures, and functions?

The first part of the analysis focused on the frequency of four-word lexical bundles in the CZMRAC. Interestingly, the CZMRAC consisted of the identical total of four-word lexical bundles as the ENMRAC, equivalent to 204 bundles. To answer the first research question regarding the most frequent four-word lexical bundles in the CZMRAC, the top 50 CZMRAC lexical bundles were introduced. The most frequent lexical bundle was, *on the other hand,* with frequency of 276 hits and range of 160 articles. Functionally, *on the other hand* served as a contrastive signal and structurally belonged to the category "Other prepositional phrase fragments". A little over half of the top 50 lexical bundles started with prepositional phrases. Moreover, except for structural categories "(Verb phrase) + that-clause fragment", "(Verb/adjective) + to-clause fragments", and "Adverbial clause fragment categories", all other structures, as introduced by Biber (1999), were identified in the 50 most frequent lexical bundles. The functional analysis that followed Hyland's (2008) classification, has revealed that the majority of the most frequently used lexical bundles in the CZMRAC were used to describe the procedure. On the other hand, the most frequent lexical bundle in the ENMRAC, was *on the basis of*. It occurred 306 times in 122 different medical research articles. It should be noted that *on the basis of* did not make it on the list of top 50 CZMRAC bundles. The same is true for *on the other hand*

which appeared only 30 times in the ENMRAC, unlike in the CZMRAC, where it occurred 276 times. Finally, it is important to mention that there were 13 lexical bundles that were matching in the top 50 bundles in the examined corpora. They were – *at the end of, at the time of, in accordance with the, in the present study, was approved by the, the end of the, there were no significant, as well as the, are shown in table, in the presence of, the results of the, in the control group, the total number of.*

Next, a comparison of the frequency distribution of the matching four-word lexical bundles was conducted. Out of 204 identified bundles, there was a total of 60 matching bundles. Percentagewise, the matching bundles accounted for 29.41% of all the identified bundles. Having compared their frequency using an online calculator *Corpus Frequency Wizzard Tool*, it was determined that 28 matching lexical bundles proved to be significantly different in their frequency and 32 of the matching bundles showed no significant difference. In addition, among the lexical bundles with no significant difference, there were two bundles with the identical frequency – *no significant difference in* with frequency of 37 hits and *animal care and use* with frequency of 24 hits. Moreover, most of the lexical bundles with a significant difference in frequency were overrepresented in the CZMRAC. To summarize results regarding frequency, it can be claimed that four-word lexical bundles were frequently used in the analyzed medical research articles, whether they were published by Czech authors or in prestigious Anglo-American journals. In the present study, the total of lexical bundles turned out to be identical. Thus, it is necessary to stress that such results cannot be generalized. Further research is needed to investigate, whether this was only a coincidence, or if Czech medical journals and Anglo-American journals indeed utilize four-word bundles in such a similar manner. Furthermore, it has also been shown that even if there are matching four-word lexical bundles in both corpora, their frequency differs, sometimes significantly and sometimes insignificantly.

The structural classification of all 204 CZMRAC four-word lexical bundles has revealed that all 12 lexical bundle structures, as identified by Biber (1999), were present in the CZMRAC. The most frequent structures proved to be those starting with prepositional phrases, accounting for 29.90%. The second largely represented structural category was composed of lexical bundles starting with noun phrases, resulting in 21.07% of all identified structures. In total, these four categories – "Prepositional phrase with embedded of-phrase", "Noun phrase with of-phrase fragment", "Other prepositional phrase fragments", "Noun phrase with other post-modifier

fragment" – accounted by a narrow margin for over half of the lexical bundle structures with 50.97%. In contrast, the least represented structures in the CZMRAC were "(Verb phrase) + that-clause fragments" with *when compared to the* being the only lexical bundle in this category and "Adverbial clause fragments" with bundles such as *studies have shown that*. When comparing the structural classification of the four-word lexical bundles in the CZMRAC and ENMRAC, it has been concluded that most of the results agree. As in the case of CZMRAC, the ENMRAC most frequent lexical bundle structures started with prepositional and noun phrases. These two categories made up more than half of the identified structures in the ENMRAC. On the other hand, the least frequent structures were "Adverbial clause fragments" and "Anticipatory it + verb/adjective phrases". The category that showed the most significant contrast of 8 lexical bundles in the distribution of structures, was the category of "(Verb phrase) + that-clause fragment". In the CZMRAC, there were only three bundles and in the ENMRAC, there were 11. In conclusion, these findings have illustrated the importance of all lexical bundle structures, as introduces by Biber (1999) in medical research articles. This has been found to be true whether they were published by Czech authors or in prestigious Anglo-American journals. It can also be concluded that there were not any significant differences between the distribution of the individual structural categories. The present study supports Mbodj-Diop's findings that the predominant lexical bundle structures in medical research articles start with prepositional and noun phrases.

In addition, the functional classification of the lexical bundles identified in the CZMRAC have largely overlapped with Mbodj-Diop's (2016) findings. Both corpora consisted mainly of research-oriented lexical bundles. These functions accounted in both corpora for more than half of the identified functions. In the CZMRAC, the overall percentage of research-oriented bundles reached 65.71%. In the ENMRAC, the percentage was slightly lower, yet still significant, accounting for 58.65%. Text-oriented bundles were largely represented as well, in the CZMRAC they were represented by 29.52% and in the ENMRAC by 34.62%. The least lexical bundles were classified as having participant-oriented functions. In the CZMRAC, there were only 10 lexical bundles performing these functions and in the ENMRAC, there were 14. Percentagewise, they made up 4.76%, in the case of CZMRAC and 6.73% in the case of ENMRAC. Once again, the results of the functional classification have strongly supported Mbodj-Diop's findings and have shown that all functional categories and subcategories, as introduce by Hyland (2008), occur in medical research articles.

The data provided in this conclusion, make it evident that the examined lists of four-world lexical bundles extracted from the CZMRAC and ENMRAC have shown a considerable convergency in the results. Even though, some of the results slightly differed in their percentage or numbers, there were found no significant differences in the overall frequency, structural, and functional distribution. These findings could be contributed to education, revision by editors and subeditors, and/or to Czech authors themselves, who could be familiar with the genre of medical research articles to such extent that it is natural for them to produce them. However, these results are valid solely for this comparative study and cannot be generalized. Furthermore, the present thesis has not dealt with the correct or colloquial use of lexical bundles by Czech authors. Therefore, further research could focus on this area more closely and investigate whether Czech authors not only use four-word lexical bundles, but also whether they use the lexical bundle functions and structures correctly.

The contribution of this study lies in its feedback to Czech authors who publish research medical articles in English, and to their editors and subeditors. It is equally important to mention that this study could be a helpful material for English teachers and students. The investigated lists of four-word lexical bundles can serve as examples of frequently used lexical bundles and can encourage students of English to use them in spoken or written discourse in order to increase their language proficiency.

5 Appendices

Appendix A: List of the 204 CZMRAC four-word lexical bundles

No.	Frequency	Range	Lexical bundle
1.	276	160	on the other hand
2.	194	71	in the czech republic
3.	130	73	in the case of
4.	129	65	at the end of
5.	123	71	at the time of
6.	123	97	in accordance with the
7.	122	73	in the present study
8.	116	116	was approved by the
9.	111	64	the end of the
10.	94	46	there were no significant
11.	87	66	as well as in
12.	87	71	as well as the
13.	86	70	one of the most
14.	83	83	study was approved by
15.	80	55	of the czech republic
16.	79	66	an important role in
17.	77	60	are shown in table
18.	77	41	in the group of
19.	76	40	no significant differences in
20.	75	61	in our study we
21.	75	66	of this study was
22.	75	35	were no significant differences
23.	74	42	in the presence of
24.	74	65	is one of the
25.	70	52	the results of the
26.	66	66	the study was approved
27.	64	48	in the development of
28.	64	60	the aim of this
29.	62	49	are summarized in table
30.	62	36	in the control group
31.	61	35	the beginning of the
32.	61	61	the ethics committee of
33.	61	58	this study was to
34.	60	35	at the beginning of
35.	59	48	according to the manufacturer
36.	59	39	in the pathogenesis of
37.	59	54	of our study was
38.	58	25	at the age of
39.	57	51	the other hand the
40.	56	42	at the same time
41.	56	56	ethics committee of the
42.	56	42	of the present study
43.	55	51	aim of this study

44.	51	45	body mass index bmi
45.	51	45	in agreement with the
46.	51	32	the total number of
47.	50	44	in our study the
48.	49	46	it is necessary to
49.	48	38	in the treatment of
50.	48	39	was used for the
51.	47	39	as a result of
52.	47	33	for the treatment of
53.	46	33	in the regulation of
54.	46	40	play an important role
55.	46	46	the aim of our
56.	45	39	included in the study
57.	45	36	is in agreement with
58.	45	38	on the basis of
59.	45	31	significant differences in the
60.	44	38	at the department of
61.	44	33	to the development of
62.	43	39	important role in the
63.	43	29	in comparison with the
64.	42	42	aim of our study
65.	42	41	our study was to
66.	40	33	it is important to
67.	40	35	of the institute of
68.	40	38	test was used to
69.	40	35	there was no significant
70.	40	28	we did not find
71.	39	34	the fact that the
72.	39	31	was significantly higher in
73.	38	33	are in agreement with
74.	38	27	in the absence of
75.	38	24	in the area of
76.	38	27	is shown in table
77.	38	28	the czech republic the
78.	37	28	as a marker of
79.	37	31	it has been shown
80.	37	28	no significant difference in
81.	37	32	were included in the
82.	36	36	approved by the local
83.	36	36	committee of the institute
84.	36	34	in our previous study
85.	36	31	is based on the
86.	36	27	it is possible to
87.	36	23	the czech republic in
88.	36	36	the local ethics committee
89.	36	30	the present study we
90.	36	30	there were no differences
91.	35	31	the present study was
92.	34	32	be explained by the
93.	34	31	by the fact that
94.	34	20	in the general population

95.	34	27	in the majority of
96.	34	33	plays an important role
97.	34	29	test was used for
98.	34	34	the aim of the
99.	34	34	the declaration of helsinki
100.	33	29	in the context of
101.	33	29	is the most common
102.	33	33	performed in accordance with
103.	33	24	the relationship between the
104.	32	31	approved by the ethical
105.	32	32	by the local ethics
106.	32	24	there was a significant
107.	32	26	was found to be
108.	32	26	with the exception of
109.	31	22	a significant decrease in
110.	31	27	are presented in table
111.	31	30	by the ethical committee
112.	31	29	in this study the
113.	31	25	in this study we
114.	31	31	statistical analysis was performed
115.	31	27	was not statistically significant
116.	31	31	were approved by the
117.	30	30	consent was obtained from
118.	30	21	did not find any
119.	30	26	for the development of
120.	30	24	in relation to the
121.	30	27	in the form of
122.	30	29	informed consent was obtained
123.	30	22	play a role in
124.	30	25	studies have shown that
125.	30	27	was observed in the
126.	29	24	at the level of
127.	29	26	has been shown that
128.	29	24	has been shown to
129.	29	20	no statistically significant differences
130.	29	26	of the university hospital
131.	29	21	the majority of patients
132.	29	24	we did not observe
133.	29	29	were considered statistically significant
134.	28	20	in the course of
135.	28	21	no significant differences between
136.	28	27	present study was to
137.	28	21	the czech republic and
138.	28	20	was significantly lower in
139.	27	23	first faculty of medicine
140.	27	23	in the number of
141.	27	23	is associated with a
142.	27	26	is considered to be
143.	27	22	is in accordance with
144.	27	23	it is possible that
145.	27	22	no statistically significant difference

146.	27	25	results are expressed as
147.	27	22	significant difference in the
148.	27	22	there was no difference
149.	27	26	was considered statistically significant
150.	27	25	was found in the
151.	27	26	with the results of
152.	26	26	analyses were performed using
153.	26	23	role in the development
154.	26	20	studies are needed to
155.	26	26	the results of our
156.	26	22	were excluded from the
157.	25	21	a wide range of
158.	25	22	could be explained by
159.	25	22	did not observe any
160.	25	24	for the purpose of
161.	25	21	group of patients with
162.	25	23	in any of the
163.	25	23	it should be noted
164.	25	20	patients were treated with
165.	25	22	significant differences between the
166.	25	24	the ethical committee of
167.	25	23	was no significant difference
168.	25	22	were found in the
169.	24	23	animal care and use
170.	24	22	one of the main
171.	24	24	the difference between the
172.	24	21	was used as a
173.	24	20	were enrolled in the
174.	24	21	were found to be
175.	23	23	a large number of
176.	23	21	a significant increase in
177.	23	22	best of our knowledge
178.	23	21	divided into two groups
179.	23	21	enrolled in the study
180.	23	22	ethical committee of the
181.	23	20	study was to compare
182.	23	22	study was to evaluate
183.	23	22	this is the first
184.	23	21	were used for the
185.	23	23	when compared to the
186.	23	22	with regard to the
187.	23	20	with respect to the
188.	22	21	a decrease in the
189.	22	21	care and use committee
190.	22	21	clinical and experimental medicine
191.	22	20	should be noted that
192.	22	20	small number of patients
193.	22	21	was performed using a
194.	22	22	with the use of
195.	21	21	are expressed as mean
196.	21	20	of this study is

197.	21	20	results of our study
198.	21	20	was to evaluate the
199.	21	20	were considered to be
200.	20	20	as a part of
201.	20	20	by the presence of
202.	20	20	experiments were performed in
203.	20	20	protocol was approved by
204.	20	20	was obtained from all

6 Shrnutí

Cílem této práce je identifikovat a analyzovat 4-slovné lexikální svazky v korpusu sestaveném z anglicky psaných lékařských odborných článků (CZMRAC), které byly publikovány českými autory a posléze je srovnat s lexikálními svazky, jež byly představeny v magisterské práci od Mbodj-Diopové (2016), která zkoumala lexikálními svazky v prestižních lékařských anglo-amerických časopisech (ENMRAC). Účelem komparativní analýzy je zodpovědět následující výzkumné otázky:

1) Které 4-slovné lexikální svazky se objevují v CZMRACu nejčastěji?

2) Které strukturální typy 4-slovných lexikálních svazků jsou využívány v CZMRACu?

3) Jaké funkce plní 4-slovné lexikální svazky identifikované v CZMRACu?

4) Jak se liší, pokud vůbec, 4-slovné lexikální svazky identifikované v CZMRACu ve srovnání se 4-slovnými lexikálními svazky, které byly identifikovány v ENMRACu z hlediska frekvence, strukturálních typů a funkcí?

Teoretická část si klade za cíl vymezit základní pojmy a koncepty, na jejichž základě jsou zpracovány výsledky této práce. Nejprve je představen lingvistický jev, který je znám pod termínem formulaický jazyk. Další části se zaměřují na lexikální svazky a jejich obecnou charakteristiku a na studie, které se zabývají lexikálními svazky a jejich rolí při osvojování jazyka. Další sekce je zaměřena na korpusy obsahující texty nerodilých mluvčí a jejich typologii. Předposlední část teoretické části představuje obecnou charakteristiku tvorby lékařských textů a rovněž krátce popisuje proces jejich korektury. Poslední sekce se soustředí na představení výchozí studie, které sloužila jako inspirace pro tuto práci.

Analytická část podrobně popisuje metodologii práce a samotnou kompilaci CZMRACu, který se skládá ze 400 anglických lékařských odborných článků publikovaných v těchto odborných časopisech – *Acta Medica, Biomedical Papers, Central European Journal for Public Health, Physiological Research* a *Prague Medical Report* a obsahuje 1 074 607 slov. Po vymezení definice lexikálních svazků pro danou práci jsou představeny výsledky společně s jejich komentáři. Následující odstavce stručně popisují výsledky této práce a komparativní analýzy.

Nejčastější 4-slovné lexikální svazky v CZMRACu

K identifikaci lexikálních svazků byl použit program s názvem *AntConc*, který umožňuje studovat sekvence slov na základě zadaných parametrů. Program *AntConc* celkem vyextrahoval 245 4-slovných lexikálních svazků. Na základě analýzy jednotlivých svazků byl vytvořen finální seznam 4-slovných lexikálních svazků, který obsahoval 204 svazků. Zajímavostí této práce je, že celkový počet identifikovaných svazků se rovná celkovému počtu svazků, které ve své práci identifikovala Mbodj-Diopová (2016). Dále bylo představeno 50 nejčastějších svazků, které se vyskytovaly v CZMRACu. Svazek s nejvyšší frekvencí byl svazek *on the other hand*. Tato formule se v korpusu objevila celkem 306krát v 160 různých článcích. Nejčastějším svazkem v ENMRACu byl svazek *on the basis of,* který se v něm vyskytl celkem 306krát v 122 různých článcích. Podrobnější analýza 50 nejfrekventovanějších svazků v CZMRACu ukázala, že kromě tří strukturálních typů – „(Verb phrase) + that-clause fragment", „(Verb/adjective) + to-clause fragments" a „Adverbial clause fragment categories" se mezi nejčastějšími svazky v CZMRACu objevily všechny ostatní strukturální typy, které představil Biber (1999). Analýza strukturálních typů také odhalila, že více jak polovina nejčastějších svazků začínala předložkovými frázemi jako např. *in the case of*. Na základě funkční analýzy lze říci, že většina nejfrekventovanějších svazků sloužila k popisu výzkumu (procedure). Jako poslední je nutné uvést, že mezi 50 svazky s nejvyšší frekvencí bylo celkem 13 svazků, které se mezi dvěma zkoumanými korpusy shodovaly – *at the end of, at the time of, in accordance with the, in the present study, was approved by the, the end of the, there were no significant, as well as the, are shown in table, in the presence of, the results of the, in the control group, the total number of*.

Na základě komparativní analýzy 204 identifikovaných 4-slovných svazků v CZMRACu byl vytvořen seznam 60 svazků, které se shodovaly se svazky ze seznamu Mbodj-Diopové (2016). Tyto svazky byly posléze podrobeny testu tzv. statistické signifikance za využití online programu *Corpus Frequency Wizzard Tool*, který na základě dvou uvedených frekvencí rozlišuje míru statistické signifikance dvou vzorků ze dvou korpusů. Tento program vyhodnotil, že 28 lexikálních svazků prokázalo signifikantní rozdíl ve frekvenci a 32 svazků se objevovalo ve zkoumaných korpusech bez signifikantního rozdílu. Mezi svazky bez signifikantního rozdílu se také objevily dva svazky, které měly totožnou frekvenci – *no significant difference in* a *animal care and use.*

Klasifikace strukturálních typů 4-slovných svazků v CZMRACu

Klasifikace 204 identifikovaných 4-slovných strukturálních typů potvrdila výsledky Mbodj-Diopové (2016). Výsledky obou prací ukazují, že autoři anglicky psaných lékařských článků využívají všechny strukturální typy lexikálních svazků, které představil Biber (1999), nezávisle na tom, jestli byly napsány českými autory, nebo byly publikovány v anglo-amerických lékařských časopisech. Závěry obou studií dokládají, že nejčastějšími strukturálními typy lexikálních svazků v lékařských odborných článcích jsou struktury, které začínají jmennými a předložkovými fázemi. Procentuální zastoupení těchto strukturálních typů v CZMRACu tvořilo 50,97 %. Naopak nejméně zastoupené kategorie v CZMRACu byly „(Verb phrase) + that-clause fragments" (pouze jeden lexikální svazek *when compared to the*) a "Adverbial clause fragments", která obsahovala svazky jako např. *studies have shown that*.

Funkční klasifikace 4-slovných svazků v CZMRACu

Všechny identifikované lexikální svazky byly klasifikovány na základě funkční typologie, kterou představil Hyland (2008). Stejně jako u výsledků analýzy strukturálních typů se výsledky funkční analýzy značně shodovaly s výsledky Mbodj-Diopové (2016). V obou korpusech byla nejvíce zastoupena kategorie lexikálních svazků, které se zaměřují na výzkum (*research-oriented*). Procentuálně tvořily 65,71 % všech identifikovaných funkcí. V ENMRACu byly tyto svazky zastoupeny o něco méně, avšak stále pokryly 58,65 % celkové distribuce funkcí. Druhou nejvíce zastoupenou funkční skupinou v obou korpusech byla kategorie svazků, které se zaměřovaly na samotný text (*text-oriented*). V CZMRACu to bylo celkem 29,52 % a v ENMRACu 34,62 %. Lexikální svazky, které se orientují na participanta (*participant-oriented*) se vyskytovaly v obou korpusech nejméně. V CZMRACu jich bylo identifikováno pouze 10, což je percentuálně 4,76 % a v ENMRACu 14, což odpovídá 6,73 %.

Výsledky uvedené v tomto shrnutí ukazují, že se výsledná data této práce a práce Mbodj-Diopové (2016) výrazně shodují. Z hlediska frekvence lze říci, že zajímavostí této komparativní analýzy je skutečnost, že je celkový počet identifikovaných 4-slovných lexikálních svazků zcela totožný s celkovým počtem identifikovaných svazků v práci od Mbodj-Diopové. Zajímavým zjištěním je také to, že se ve zkoumaných korpusech celkem vyskytlo 60 totožných lexikálních svazků, což odpovídá 29,41 %. Tyto výsledky nasvědčují, že 4-slovné lexikální svazky jsou využívány

v lékařských odborných článcích nezávisle na tom, jestli byly publikovány českými autory, nebo jestli byly zveřejněny v prestižních anglo-amerických lékařských časopisech. Dále je potřeba také upozornit na to, že i když se značné procento lexikálních svazků v obou korpusech shoduje, jejich frekvence se často liší, někdy signifikantně a někdy bez signifikantních rozdílů. Co se týká strukturálních typů a funkcí lexikálních svazků, je nutno poznamenat, že nedošlo k žádným překvapivým závěrům. Oba korpusy prokazují poměrně velké zastoupení jmenných a předložkových frází (*Noun-phrase* a *Prepositional-phrase fragments*) a signifikantní zastoupení svazků, které popisují samotný výzkum (*research-oriented*). Na závěr lze říci, že i když se procentuální zastoupení jednotlivých kategorií a subkategorií mírně liší, rozdíly jsou zanedbatelné. Celková distribuce strukturálních typů lexikálních svazků a jejich funkcí se výrazně podobá. Tyto závěry mohou být výsledkem obdrženého vzdělání, kvalitních jazykových korektur pod vedením editorů a subeditorů, nebo i zkušeností samotných českých autorů, kteří jsou natolik zvyklí na kontakt s odbornou lékařskou literaturou v anglickém jazyce, že jim tvorba lékařských odborných článků, alespoň z hlediska lexikálních svazků, nečiní problémy. Nicméně je důležité upozornit na skutečnost, že výsledky komparativní analýzy jsou validní pouze pro zkoumané korpusy v rámci této práce, a tudíž je nelze generalizovat. Další výzkum by mohl nahlédnout především do problematiky správného a uzuálního využití lexikálních svazků v lékařských textech v podání českých autorů píšících anglicky, jelikož tato tématika nebyla předmětem této studie.

7 List of Figures and Tables

8 Works Cited

Allen, David. 2010. “Lexical Bundles in Learner Writing: An Analysis of Formulaic Language in the Aless Learner Corpus.” Komaba Journal of English Education. https://www.academia.edu/4614255/Lexical_Bundles_in_Learner_Writing_An_Analysis_of_Formulaic_Language_in_the_ALESS_Learner_Corpus.

Barlow, Michael. 2011. “Corpus Linguistics and Theoretical Linguistics.” International Journal of Corpus Linguistics. John Benjamins, January 1, 2011. https://www.jbe-platform.com/content/journals/10.1075/ijcl.16.1.02bar.

Biber, Douglas, and Federica Barbieri. 2007. “Lexical Bundles in University Spoken and Written Registers.” English for Specific Purposes, 2007. https://www.academia.edu/376249/Lexical_Bundles_In_University_Spoken_and_Written_Registers.

Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004.“If You Look at ...: Lexical Bundles in University Teaching and Textbooks.” OUP Academic. Oxford University Press, September 1, 2004. https://academic.oup.com/applij/article-abstract/25/3/371/179465.

Biber, Douglas et al. 1999. “Longman Grammar of Spoken and Written English.” Google Books. Longman, 1999. https://books.google.com/books/about/Longman_Grammar_of_Spoken_and_Written_En.html?id=vjomAQAAMAAJ.

Conklin, Kathy, and Norbert Schmitt. 2012. “The Processing of Formulaic LANGUAGE: Annual Review of Applied Linguistics.” Cambridge Core. Cambridge University Press, December 5, 2012. https://www.cambridge.org/core/journals/annual-review-of-applied-linguistics/article/processing-of-formulaic-language/8728FEDCAB7A9A280B6EF1CD899B94B1.

Cortes, Viviana. 2004.“Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology.” English for Specific Purposes. Pergamon, January 29, 2004. https://www.sciencedirect.com/science/article/pii/S0889490603000851.

Cowie, A.P. 1992. “Multiword Lexical Units and Communicative Language Teaching,” 1992. https://link.springer.com/chapter/10.1007%2F978-1-349-12396-4_1.

Dontcheva-Navratilova, Olga. 2012. Lexical bundles in academic texts by non-native speakers, January 1, 2012. https://digilib.phil.muni.cz/handle/11222.digilib/126942.

Evans, David. 2007.“Corpus Building and Investigation for the Humanities.” http://docplayer.net/29259380-Corpus-building-and-investigation-for-the-humanities.html, 2007. https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/Intro/Unit2.pdf.

Gablasova, Dana, Vaclav Brezina, and Tony McEnery. (2017) “Exploring Learner Language through Corpora: Comparing and Interpreting Corpus Frequency Information.” Wiley Online Library. John Wiley & Sons, Ltd, March 15, 2017. https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12226.

Goodman, Neville W., and Martin B. Edwards. (2006) *Medical Writing: A Prescription for Clarity*. *Www.cambridge.org/9780521858571*. New York, United States of America: Cambridge University Press, 2006. www.cambridge.org/9780521858571.

Granger, Sylviane. (2008) “Learner Corpora.” Research Gate, 2008. https://www.researchgate.net/publication/273480731_Learner_Corpora.

Hardy, Stephen. (n.d.) “About Us.” Medical Journal Editors (MJE), 2021. http://www.medicaljournaleditors.com/about-us.

Hayes, Donald P. (1992). “The Growing Inaccessibility of Science.” Nature News. Nature Publishing Group, 1992. https://www.nature.com/articles/356739a0.

Heng, C.S., H. Kashiha, and H. Tan. (2014). “Lexical Bundles: Facilitating University ‘Talk’ in Group ...” Research Gate, 2014. https://www.researchgate.net/publication/271336404_Lexical_Bundles_Facilitating_University_Talk_in_Group_Discussions.

Hyland, K. (2008a). “As Can Be Seen: Lexical Bundles and Disciplinary Variation ...” Research Gate, 2008. https://www.researchgate.net/publication/222407470_As_can_be_seen_Lexical_bundles_and_disciplinary_variation.

Hyland, Ken, and Feng Jiang. (2018). “Academic Lexical Bundles How Are They Changing?” Research Gate, 2018. https://www.researchgate.net/profile/Kevin-Jiang-2/publication/329937930_Academic_lexical_bundles_How_are_they_changing/links/5c26bfc692851c22a34c40ea/Academic-lexical-bundles-How-are-they-changing.pdf.

Hyland, Ken. (2012). “Bundles in Academic Discourse.” Annual Review of Applied Linguistics, 2012. https://www.academia.edu/22895067/Bundles_in_Academic_Discourse.

Loi, Chek Kim. (2010). “Research Article Introductions in Chinese and English: A Comparative Genre-Based Study.” Journal of English for Academic Purposes. Pergamon, October 20, 2010. https://www.sciencedirect.com/science/article/pii/S1475158510000573.

Mbodj-Diop, N. B. (2016) “Lexical Bundles in Medical Research Articles: Structures and Functions.” d.lib.msu.edu, 2016. https://d.lib.msu.edu/etd/3841/datastream/OBJ/view.

Moon, R. (1992). “Textual Aspects of Fixed Expressions in Learners’ Dictionaries.” Springer, 1992. https://link.springer.com/chapter/10.1007/978-1-349-12396-4_2.

Paquot, M., and S. Granger. (2012). “Formulaic Language in Learner Corpora.” Research Gate, 2012. https://www.researchgate.net/publication/259704872_Formulaic_Language_in_Learner_Corpora.

Pawley, A., and F. H. Syder. (1983). *Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency*. London: J. C. Richards & R. W. Schmidt, 1983.

Schmitt, Norbert, and Ronald Carter. (2004). “Formulaic Sequences in Action: An Introduction.” John Benjamins Publishing Catalog. John Benjamins Publishing Company, 2004. https://benjamins.com/catalog/lllt.9.02sch.

Wood, David. 2015. “Fundamentals of Formulaic Language: An Introduction. London .” Research Gate, 2015. https://www.researchgate.net/publication/316522342_David_Wood_Fundamentals_of_formulaic_language_An_introduction_London_Bloomsbury_2015_Pp_vii_198.

Wood, David. (2019) “Classifying and Identifying Formulaic Language.” Routledge Handbooks, 2019. https://www.routledgehandbooks.com .

Wray, A., and M. R. Perkins. (2000). “The Functions of Formulaic Language: An Integrated Model.” citeseerx.ist.psu.edu, 2000. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.558.6248&rep=rep1&type=pdf.

Wray, Alison. (2002). “Formulaic Language and the Lexicon.” assets.cambridge.org, 2002. https://assets.cambridge.org/97805217/73096/frontmatter/9780521773096_frontmatter.pdf.

Wray, Alison. (1999). “Formulaic Language in Learners and Native Speakers.” www.cardiff.ac.uk, 1999. https://www.cardiff.ac.uk/__data/assets/pdf_file/0009/152919/wraylt1999.pdf.

Zagolová, Pavlína. (2017). “Czech as Source and Target Language in a Comparable Corpus of Journalistic Texts.” Thesis, Palacký University, 2017.

9 Abstract

This paper presents a comparative study which deals with four-word lexical bundles in medical research articles. The objective of the comparative study is to identify and analyze four-word lexical bundles in the corpus comprised of English medical research articles written by Czech authors (CZMRAC) and subsequently compare them with four-word lexical bundle identified in prestigious Anglo-American medical journals by Mbodj-Diop (2016). The analysis focuses on the most frequent four-word lexical bundles in the CZMRAC, their structural and functional classification, and their comparison with lexical bundles identified in Anglo-American medical journals.

Key Words

Corpus, comparative study, formulaic language, functional classification, medical writing, medical research articles, learner corpora, lexical bundles, second language acquisition, structural classification

Anotace

Cílem této práce je identifikovat a analyzovat 4-slovné lexikální svazky v korpusu sestaveném z anglicky psaných lékařských odborných článků, které byly publikovány českými autory (CZMRAC), a posléze je srovnat s výsledky magisterské práci od Mbodj-Diopové (2016), která zkoumala lexikální svazky v prestižních lékařských anglo-amerických časopisech (ENMRAC). Komparativní analýza se zaměřuje na nejfrekventovanější 4-slovné lexikální svazky v CZMRACu, na klasifikaci strukturálních typů identifikovaných svazků a na jejich funkce.

Klíčová slova

Formulaický jazyk, korpus, korpus s texty nerodilých mluvčí, komparativní studie, lexikální svazky, klasifikace strukturálních typů lexikálních svazků, funkční klasifikace lexikálních svazků, lékařské odborné články