UNIVERSITY OF TARTU
DEPARTMENT OF ENGLISH STUDIES

THE COMPARISON OF THE USAGE OF PREFABS IN THE ACADEMIC WRITING OF ESTONIAN EFL LEARNERS AND NATIVE ENGLISH SPEAKERS

BA thesis

LIISI KRAAK
SUPERVISOR: Assoc. Prof. Jane Klavan, PhD

TARTU
2021

ABSTRACT

Languages largely consist of prefabricated expressions (prefabs), more broadly known by the term formulaic language. Accumulating more data in this field of study benefits second- and foreign language acquisition, and more specifically, EFL learners' comprehension and language production in terms of recognising and learning these formulaic patterns. The aim of this thesis is to analyse the usage of prefabs in written academic English between Estonian EFL learners and native English speakers. To achieve this, a corpus-based study was conducted, which utilised the corpus of *Estonian Academic Learner English* (EALE) and the corpus of *British Academic Written English* (BAWE).

The thesis begins with an introduction, which gives an overview of the motivation behind this paper as well as a summary of subsequent chapters. The literature review section defines the core aspects discussed in this thesis such as formulaic language, prefabs and corpus linguistics as well as provides an overview about previous research. The empirical section introduces the methodology, which makes use of the frequency-based approach used in corpus-based studies, followed by the results. Certain prefabs (e.g. *on the other hand*, *in the case of*) are examined further in the form of case studies. This section is followed by the discussion, which expands upon the results and provides options for future studies. The thesis ends with a conclusion.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AFL – Academic Formulas List

BAWE – *British Academic Written English corpus*

BrE – British English

EALE – *Estonian Academic Learner English corpus*

EFL – English as a Foreign Language

L1 – First language

TCELE – *Tartu Corpus of Estonian Learner English*

VESPA-NO – *Varieties of English for Specific Purposes dAtabase*

INTRODUCTION

According to Wray (2002) and Schmitt (2010), certain clusters of words exist in language that are frequently used in everyday life. As this phenomenon has been observed by various scholars, there are multiple different terms to describe it but, in general, these recurring patterns of words are collectively known as formulaic language. While the history of formulaic language can be traced back to as early as the eighteenth-century, the most well-known principle in this field comes from John Sinclair. Sinclair (1991) proposes that language largely consists of preformulated phrases as opposed to smaller units which are put together piece by piece, a phenomenon he calls 'the idiom principle'. This is in line with modern-day understanding of language production.

Formulaic language can be found in both spoken and written language. While it is generally agreed that these formulaic patterns occur more regularly in speech, it also makes up a significant part of written text. According to Erman and Warren (2000), 52% of written language is made up from prefabricated expressions (prefabs). However, studies have found that the estimate can be as low as 32% (Foster 2001) or as high as 80% (Altenberg 1998). The dissimilarity between results is likely due to utilisation of different methods. Needless to say, formulaic language makes up a significant part of spoken and written language which is why it is one of the most prevalent research topics among second- and foreign language acquisition.

Advancement in technology has also been beneficial to this research. Nowadays, language patterns are most frequently analysed via corpus-based studies that utilises corpora containing large numbers of texts. Sorting through each text manually is a tedious task which is why the implementation of computers has made research more efficient as computer-based tools and methods allow much larger numbers of texts to be processed, compiled, and analysed.

Studies using the corpus-based method are conducted in hopes to get a better insight into native speakers' use of language in order to identify characteristics which should be learned to achieve native-like fluency. Additionally, it is used to learn how non-native speakers use these patterns.

Research on the usage of formulaic language has been carried out among several different languages, including Swedish, Norwegian, Russian, Chinese and many others. Yet, there is little done regarding how Estonians interpret formulaic language and even more specifically, how Estonian EFL students use prefabs. A previous corpus-based study was done by Piiri (2020) who examined the usage of formulaic language by native and non-native speakers in academic texts in both spoken and written registers. However, in his discussion, he notes that, due to the source texts used for the entry essays, the analysis of Estonian EFL speakers' use of formulaic language did not provide any clear results (Piiri 2020: 23). Therefore, the current thesis intends to expand upon his work in the form of researching how the usage of prefabs differs between Estonian EFL users and native speakers of English in written academic texts.

The present thesis consists of two chapters. The first chapter is the literature review that aims to give background information on what has been previously researched in the field of formulaic language. The first section of it focuses on establishing what is formulaic language and what terminologies have been used to describe it. In this thesis, the main term used is Erman and Warren's (2000) prefab, the criteria of which is also explained in this section. Besides prefabs, more generalised terms are used alongside 'formulaic language' itself. Further on, the usage and benefits of researching prefabs in various fields is discussed. The second section discusses corpus linguistics and how corpus-based studies are used to identify formulaic

patterns in language. As frequency is an important criterion in corpus-based studies, the benefits and disadvantages of using a frequency-based method are investigated.

The second chapter of this thesis is the empirical analysis in which the first section describes the methodology used in the present thesis. The methodology used in this paper is based on the study *Phraseological teddy bears: frequent lexical bundles in academic writing by Norwegian learners and native speakers of English* (2019) by Hilde Hasselgård in which she examines how the usage of lexical bundles differ between Norwegian learners of English and native English speakers. Similarly, the current thesis investigated the usage of the most frequent four-word bundles (or prefabs) between Estonian EFL learners and native English speakers. In order to draw any definite conclusions, an empirical analysis was carried out using data from two corpora which consist of texts representing novice academic English (in the case of this thesis, texts by undergraduate students). Data for Estonian EFL users was gathered from the corpus of *Estonian Academic Learner English* (EALE) and data for native English speakers was collected from the corpus of *British Academic Written English* (BAWE). This section will further describe each of the corpora, the processing of the aforementioned data as well as the software used to carry out the analysis. Some initial observations are made that are later elaborated on in the discussion section. In the following section, the most frequent prefabs that occur in both corpora are analysed in more detail in the form of case studies. Possible reasons for their frequent usage and distribution among texts are explored. The chapter ends with the analysis of the findings as well as a discussion part where the implications of these findings are further explored. Furthermore, options for future studies are discussed.

In her study, Hasselgård (2019) discusses *phraseological teddy bears* which derive from the term *lexical teddy bears*, first coined by Hasselgren in 1994. Hasselgren (1994: 237)

describes this phenomenon as words which foreign language learners cling to as "[s]tripped of the confidence and ease we take for granted in our first language flow, we regularly clutch for the words we feel safe with: our 'lexical teddy bears'". She further explains that this is likely due to the fact that these words are learned during the early stages of acquiring a new language, widely usable and, most importantly, unlikely to show up as errors, making them 'safe to use' (Hasselgren 1994). Likewise, Hasselgård (2019) describes a similar occurrence with lexical bundles or, in other words, phraseological expressions (hence the name *phraseological teddy bears*). Based on this, it can be hypothesised that similarly to other learners of English, Estonian EFL learners will also exemplify an over-reliance on certain prefabs. Therefore, the initial hypothesis of this thesis is that Estonian EFL students will overuse certain prefabs, whereas there will be a greater variety among native speakers of English. Furthermore, the findings of Hasselgård's (2019) study indicated that Norwegian EFL learners would overuse a small set of prefabs (specifically two), resulting in a sharp frequency decline regarding other bundles. Additionally, text dispersion showed that, although most shared prefabs between native and non-native English speakers had higher frequency rates in the EFL corpus, the most common bundles overall appeared in greater proportion of the texts in L1 English. Whether these results also hold true for Estonian EFL learners will be investigated.

Overall, the aim of the present thesis is to expand the knowledge of the function and distribution of prefabricated expressions in the context of Estonian EFL learners' academic writing to provide a better understanding of this field and valuable data that can be utilised to improve second- and foreign language teaching. Therefore, the main research question is: How does the use of prefabs in academic writing differ between Estonian EFL learners and native English speakers?

1. FORMULAIC LANGUAGE

Formulaic language is a research field that examines the usage of frequently occurring sequences of words. Research on formulaic language goes back to as early as the eighteenth-century. Jespersen (1924) claimed that every language has characteristic formulas and characterised these formulas as a group of words or sentences which are perceived as a single unit. He also argued that the components of a unit cannot be changed. His work laid grounds for future research and theories on the matter. Over the years, formulaic language has been researched by scholars from various disciplines and thus, different terms have been coined to define the same phenomenon. This includes prefabs, chunks, bundles, collocations, fixed expressions, multi-word expressions, recurring utterances and so forth.

Wood (2015) states that formulaic language is used to refer to these terms as a whole. However, he as well as others (Biber and Barbieri 2007; Erman and Warren; Schmitt 2010; Wray 2002), also recognize that it is important to distinguish between these terms as on the surface level they may seem the same but in actuality, their definitions have slight differences, such as minimum length and frequency cut-offs. For example, Biber and Barbieri (2007) differentiate between whether or not they include idiomatic sequences (e.g. expression like *in a nutshell*) and some researchers (e.g. Wray 2002) include all these criteria in an even more complex identification process. Still, most of these terms are used very generally so, it is difficult to pinpoint what the actual criteria are in order to differentiate them. For the most part it seems that differentiating between these terms is mainly important when it comes to methodology as it affects the results. However, theoretically it seems that these terms are often used interchangeably. Therefore, in this thesis, the main terminology used will be prefabs (the

criteria of which will be discussed in section 1.2) as well as other generalised terms used to refer to formulaic language, including 'formulaic language' itself.

1.1 Idiom principle

Some scholars (e.g. Wray 2002; Schmitt 2010) define formulaic language as certain clusters of words that occur frequently together and exist in everyday use. This is in accordance with Sinclair's (1991: 109-110) idiom principle, which proposes the idea that a language is put together from larger preformulated phrases as opposed to smaller units which are constructed piece by piece (what he refers to as the open-choice principle). The open-choice principle suggests that practically every part of a sentence can be decided and corresponds with the traditional way of approaching formulaic language. The traditional way will be further discussed in a later part of this section.

Sinclair (1991: 110) believes that both the idiom principle and the open-choice principle are used together in natural speech and writing yet, the idiom principle is what helps to structure language and significantly reduce the number of choices one has to make while producing a 'normal text'. He suggests that this could be due to the fact that humans have a tendency to act according to *economy of effort* (Sinclair 1991:110), meaning, to maximise the output of information (i.e. text or speech) while minimising the effort it takes to produce said information. Sinclair (1991: 110) also points out that this could have developed due to the necessity of real-time conversations. This notion is also accepted by Wray (2002) who writes that formulaic language helps to reduce the effort it takes to process language, thus making it possible to focus on unrelated tasks while still being able to hold a conversation. She finds it unlikely that prefabs lessen the processing effort needed for writing, as text can be rewritten many times, but does believe that they could be helpful to readers (Wray 2002).

According to Erman and Warren (2000), the traditional view is that language production consists mainly of primitives which are then organized according to a number of rules. However, Bolinger (1976, as cited in Erman and Warren 2000) argued against the traditional view as he believed that when constructing sentences, speakers do relatively the same amount of remembering as they do putting sentences together. His view was that due to humans being capable of remembering vast amounts of information, language production would be based on using memorized units of words that can be used to form sentences. This sentiment is also echoed by other scholars. Pawley *et al.* (1983, as cited in Erman and Warren 2000) argue that the traditional approach does not cover idiomaticity nor fluency and Simpson-Vlach and Ellis (2010) add to it by stating that fluency derives from processing one's knowledge of the language automatically (i.e. from memory). Therefore, most scholars seem to favour the new approach which suggests that language mostly consists of prefabricated sequences stored in the memory.

1.2 Prefabs

Prefabricated expressions (or prefabs for short) is one of the terms used in formulaic language. Similarly to other terminology regarding formulaic language, the term 'prefabs' can also have varying definitions depending on the scholar. One of the most clear-cut definitions comes from Erman and Warren (2000: 31) who define it as "/…/a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization." In addition to that, Erman and Warren (2000: 32) establish that prefabs also have to adhere to *restricted exchangeability*, meaning that at least one word in the prefabricated expression cannot be replaced by a synonym without changing its meaning, function and/or idiomaticity. They use *good friends* vs *nice friends* as an example, where changing the word 'good' to its synonym 'nice' causes the expression to lose

its idiomatic meaning. Furthermore, *restricted exchangeability* also restricts some syntactic variability regarding prefabs which is normally possible such as negation (*I guess* cannot be *I don't guess*), loss of auxiliary (*it will do* vs *it does*) as well as reversed order (*up here* but not *here up*).

1.3 Usage of prefabs

Like any aspect of language, there are multiple ways of analysing and evaluating its significance in a particular research field. As such, the usage of formulaic language has been studied for various purposes. In their study, Erman and Warren (2000: 52) highlight three areas in which having better understanding of prefabs would have significant practical implications. First of all, they note its usefulness in machine translation, proposing that building contrastive database of prefabs between language would make machine assisted translations more efficient (2000: 52). However, formulaic language persists as a complex issue regarding machine translation as Corpas Pastor *et al.* (2016) state that multi-word units present major problems due to their "/…/semantic, pragmatic and/or statistical idiosyncracies". While they note that the adaption of neural approaches in machine translation have shown improvements, multi-word sequences remain an issue as they pose a challenge even to human translators, mainly because of the linguistic (but also cultural) differences between languages (Corpas Pastor *et al.* 2016).

Secondly, there are genre studies which examines common core and genre-related prefabs. This is important as prefabs can be used to 'mark a style', meaning that certain sequences of words occur more frequently in specific genres thus, making each genre distinct from one another. Furthermore, using them helps to keep texts genre-appropriate (e.g. the different styles used for academic prose vs sports commentaries). According to various scholars (Biber and Barbieri 2007; Schmitt 2010; Wray 2002), prefabs have several functions in daily

communication such as functional use (e.g. apologizing, giving directions), social interactions (phatic expressions), discourse organization and precise information transfer (e.g. jargon). For example, as academic writing has to uphold a certain style, it limits what words and expressions are deemed acceptable. Thus, prefabs (e.g. *in my opinion* or *on the other hand*) are extremely useful as they help to structure academic texts.

The last noteworthy field within the broader field of formulaic language is language learning. This is arguably the most thoroughly studied area regarding formulaic language as there are a multitude of studies pointing out its usefulness for teaching and learning languages (Erman and Warren 2000; Granger 1998; Simpson-Vlach and Ellis 2010 etc.). More specifically, prefabs have been seen as a valuable, educational tool to help language learners to further their understanding of the English language. Simpson-Vlach and Ellis (2010) claim that understanding these formulaic aspects of language is a key part of fluency and Erman and Warren (2000) speculate that incorporating them into learning strategies could help learners gain a better grasp of a foreign language, similarly to how it is used by the native speakers.

However, while Granger (1998) agrees with this idea she argues that most research has failed to take into consideration how the native language of different learners of English affects learning prefabricated sequences. According to her, English as a foreign language materials tend to be generalised and foreign and second language teaching would benefit from analysing and comparing how these sequences are learnt and used in specific mother tongue groups (Granger 1998). As there is little done regarding how Estonian EFL learners use formulaic language (and more specifically prefabs), the present thesis aims to provide further information on this particular research field.

1.4 The role of corpus linguistics in the study of prefabs

Corpus linguistics is an area of language research that analyses language patterns and usage with the help of corpora. Corpora are principled collections of both written and spoken natural texts. According to Reppen and Simpson-Vlach (2002: 89), natural texts refer to data that has been collected from naturally occurring sources. In the case of written texts, samples are collected from sources such as academic works (e.g. essays) as opposed to surveys or questionnaires, while samples from spoken language are acquired by recording and transcribing speech (Reppen and Simpson-Vlach 2002: 89). As one can imagine, manually going through each text would be quite a tedious task, which is why advancements in technology have provided new and simpler ways of analysing language with the help of computers.

In the present-day context, corpus linguistics and the term ’corpus’ have, for the most part, come to be synonymous with computerised corpora and methods. However, corpus linguistics has been around much longer than that, with one of the earliest corpus studies being conducted by F. W. Kaeding in 1898 (Howatt 2004). Still, corpus linguistics truly found its footing with the advancement of technology. Computer-based tools and methods allow much larger numbers of texts to be processed, compiled, and analysed. Additionally, access to these resources has made the use of corpora more widespread among different linguistic branches as it provides insight into how language is used in various ways (e.g. speech vs written language, formal vs casual etc.) (Reppen and Simpson-Vlach 2010: 89). Based on these advancements, Biber *et al.* (1998: 4) propose that there are four essential characteristics which are associated with corpus-based analysis of language:

1. It is empirical, analysing the actual patterns of use in natural texts.

2. It utilizes a large and principled collection of natural texts, known as a ‘corpus’, as the basis for analysis.
3. It makes extensive use of computers for analysis, using both automatic and interactive techniques.
4. It depends on both quantitative and qualitative analytical techniques.

Corpora are also valuable for researching formulaic language because they help analyse the frequency of word sequences by comparing it to a frequency list created from a large number of texts. There are also different types of corpora (e.g. academic corpora) that help to accumulate more accurate results as they contain texts of specific genres or disciplines. Newer corpora also provide options for even greater organization such as distinguishing between age, gender, first language etc. Moreover, if a certain type of corpus is not available, there are resources (such as the online corpus manager Sketch Engine) that allow the user to compile their own corpus based on their requirements and collected data.

Another important factor in studying formulaic language is frequency. Wood (2015: 20) proposes that sequences of words that are used often are generally seen as formulaic. While these sequences must also fulfil other criteria, frequency is often deemed to be the primary criterion. Biber and Barbieri (2007) also affirm this by claiming that high frequency indicates formulaic status. Additionally, Wood (2015: 20) informs that statistical identification is used as the foundation for the frequency-based approach of analysing formulaic language. This involves setting parameters (e.g. minimum length and minimum frequency cutoffs) before scanning and analysing a corpus to find word sequences which fit the predetermined requirements. The same approach is used in corpus linguistics as corpus-based studies utilise text analysing software to gather data. The main way these software identify formulaic word patterns is by frequency.

However, while the frequency-based analysis is a widely used method for identifying formulaic language, it is also important to keep in mind that drawing conclusions purely based on frequency does not always yield the most accurate results.

Wood (2015: 21) acknowledges that there are some drawbacks to the frequency-based analysis. Firstly, it does not indicate psycholinguistic validity as a study done by Schmitt et al (2004) suggested that remembering complete word sequences, although they were highly frequent in corpora, varied among participants. Thus, memory stores and interprets formulaic sequences differently. Additionally, Wood (2015: 21) points out that frequency can also produce meaningless word combinations and that only large corpora, containing texts from specific registers of language and/or academic disciplines, yield the most accurate results, thus limiting the use of small data sets. There are also some word sequences that occur so infrequently that some data sets or even large corpora do not provide an accurate depiction of their usage (Wood 2015: 21; Schmitt 2010: 67). Furthermore, Schmitt (2010: 67) mentions that corpora are limited by the amount and the types of texts that can be collected, meaning, corpora are usually biased towards language types that are more commonly available (for example, collecting samples of publicly accessible texts (e.g. news articles) vs secret intelligence documents).

Although the frequency-based approach has some shortcomings, it is one of the most effective and commonly used methods of identifying formulaic language. Still, based on the reasons above, some intuition is required as computers, although capable of compiling and sorting vast amounts of data, are not equipped to choose what data should be analysed nor how to interpret the findings (Reppen and Simson-Vlach 2010: 90). Therefore, to produce the most accurate results, both qualitative and quantitative methods should be employed.

2. EMPIRICAL ANALYSIS

2.1 Methodology

For the methodology part of the thesis, a study conducted by Hasselgård (2019) was used as a basis for the analysis. In her paper, Hasselgård (2019) compares the usage of lexical bundles in English between students who are native Norwegian speakers and native English speakers. She uses two main corpora for her investigation, *Varieties of English for Specific Purposes dAtabase* (VESPA-NO) and the *British Academic Written English* (BAWE) corpus. Annotations were used to remove any material (e.g. linguistic examples, quotations and bibliographies) from the texts that were not produced by the students. As previously mentioned, this thesis makes use of the methods found in Hasselgård's (2019) paper. However, due to the unavailability of some resources as well as some methods being inapplicable to this thesis, the methodology used in this paper ultimately differs somewhat from Hasselgård's (2019) study. Still, similarly to how Hasselgård (2019) compared the usage of prefabs between Norwegian EFL speakers and native English speakers, this thesis intends to compare the results of the use of prefabs in English between students whose native language is Estonian and native English speakers.

In order to start the analysis, examples of academic texts from both Estonian EFL speakers and English L1 speakers were needed. For Estonian EFL speakers, the corpus of *Estonian academic learner English* (EALE) was used. The corpus consists of the bachelor's theses defended at the Department of English Studies at the University of Tartu. These texts are in the process of being added to the *Tartu Corpus of Estonian Learner English* (TCELE). Currently, the TCELE corpus only consists of university entrance exam essays. As the essays are all written about the same topic and heavily influenced by the source text used for the essay,

the prefabs found in them would be biased. Therefore, the addition of the BA texts would give better insight into Estonian academic learner writing in English. However, before the texts could be analysed, some clean-up was required. Similarly to Hasselgård (2019), all instances of non-student written material were removed. This was done manually by going through each file using Notepad++ (Ho 2021). After the files were cleaned-up, the next step was to import the text files into a text analysis software.

In her study, Hasselgård (2019) uses WordSmith Tools 6 (Scott 2012) to extract recurring word sequences from texts. However, as the program is not free, it was substituted by an equivalent software called AntConc 3.5.8 (2019). AntConc is developed by Laurance Anthony and it is a freeware, multiplatform tool which allows its users to conduct corpus linguistic research by automatically sorting through data from collected text files. The n-grams tool, provided in the software, allows the user to scan through imported text files for 'n' word bundles. 'N' refers to the number of words that make up a bundle (e.g. one word, two words, etc.). This helps to find recurring expressions within the texts. Results can be sorted by frequency or range. Frequency indicates how many times a singular word bundle occurs in the texts altogether. On the other hand, range indicates the number of different texts that contain at least one instance of a certain word bundle.

After importing the cleaned text files into AntConc, the next step was to run the analysis via the n-grams tool. However, some parameters were required to be set beforehand. The first parameter was the n-gram size, which was set to 4 based on Hasselgård's (2019) study. According to her, a bundle size may consist of any number of words (minimum of two), however she references Hyland (2008) who states that four-word bundles "/.../are far more common than five-word strings and offer a clearer range of structures and functions than 3-

word bundles". In addition, two- and three-word bundles often appear within four-word bundles, which is why an n-gram size of four is reasonable for research purposes.

The next parameter to be set was frequency. The frequency cut-off for identifying prefabs is somewhat arbitrary, as different scholars (e.g. Biber and Barbieri 2007; Simpson-Vlach and Ellis 2010) have set it to varying degrees, ranging from 10 to 200 times per million words. However, as the length and the amount of sample texts used in this thesis is much smaller than that of previous studies, it was set to a manageable size of ten. The final parameter was range which was set to five to correspond with Hasselgård's (2019) study. After the parameters were set, an analysis was run via AntConc (see Figure 1.) and the results were copied to an Excel spreadsheet.

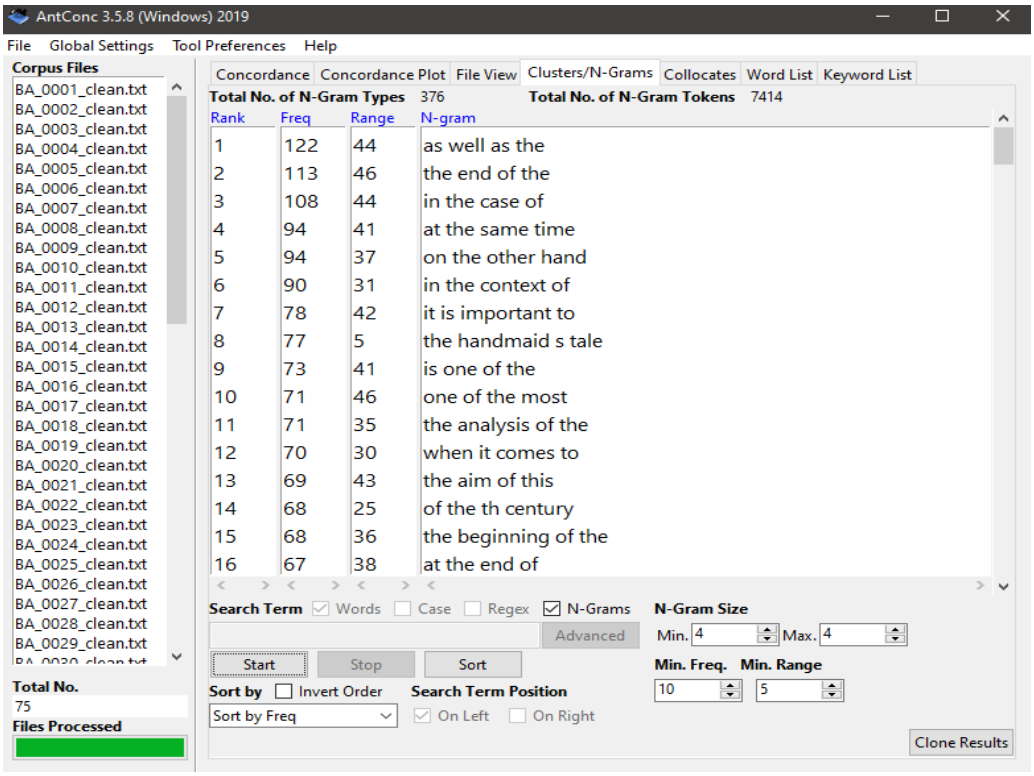

Figure 1. The top prefabs in the EALE corpus according to frequency as displayed in AntConc.

Texts from BAWE were used to collect data about native English speakers' use of prefabs. The BAWE corpus is compiled from academic works written at the universities in the UK. It features text from a variety of disciplines (Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences) and across multiple levels of study (Heuboeck *et al.* 2010). Text files for this corpus are also available to be downloaded online. This option was utilised in the present thesis. A spreadsheet was included with the downloaded text files which made it easier to find the necessary samples for comparison as well as to filter out any other texts which did not meet the requirements.

The following requirements were used to select texts from BAWE in order to make comparisons with the prefabs found from EALE. First of all, the first language of the author was set to English. As the texts used for Estonian EFL learners were bachelor's theses, texts from all bachelor's courses pertaining to the English or Linguistics discipline were chosen since both English and Linguistics fall under the Department of English Studies. Regarding text genres, unfortunately, the BAWE corpus does not include any samples of thesis text files. Therefore, the majority of the text used were essays as they were deemed to be closest to the language style used in the BA theses. There were also a few examples of texts that fell under the literature survey and methodology recount genres. Based on the language style used in these texts, they were also considered appropriate to compare with bachelor's theses. Once the requirements were set, texts files meeting these criteria were copied to a new folder so they would be easier to clean up. The text files were then cleaned up using the exact same methods as the Estonian EFL texts. After cleaning up the text files, they were imported into AntConc and processed (see Figure 2). Results were then copied over to the same spreadsheet as the EFL results.

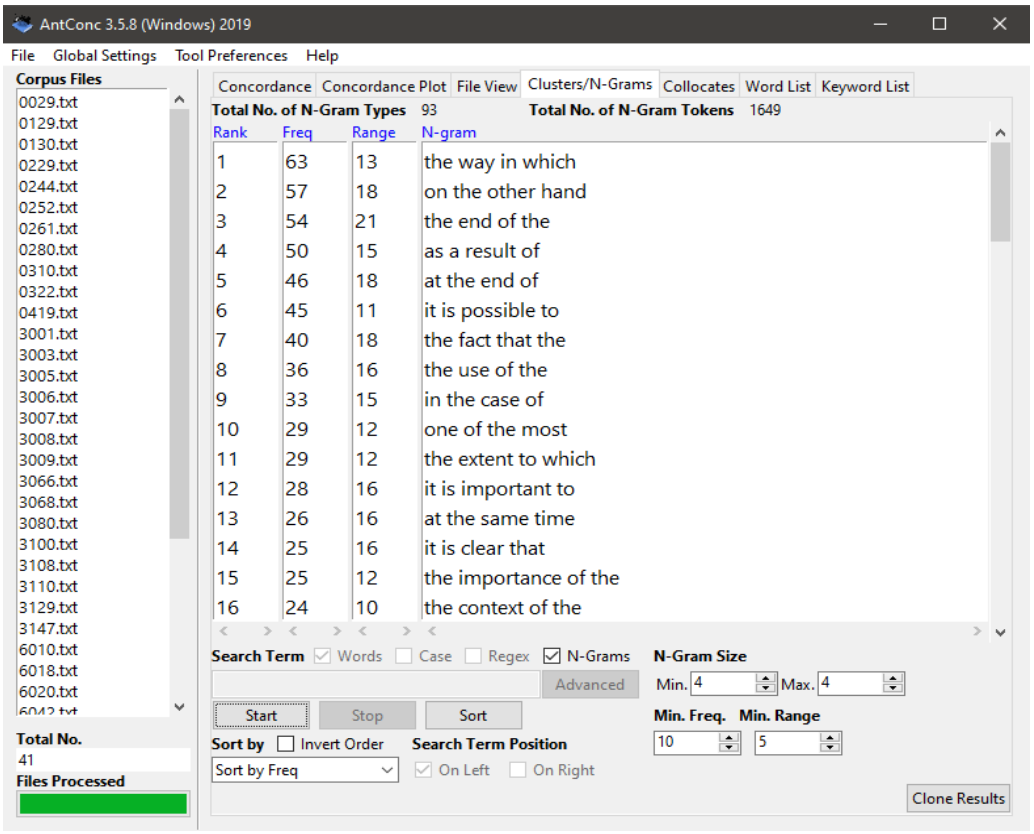

Figure 2. The top prefabs in the BAWE corpus according to frequency as displayed in AntConc.

After importing the results for both English L1 and EFL students into a single spreadsheet, the next step was to start comparing the results. It is important to note that programs have a tendency to pick up proper nouns as prefabs. Therefore, to properly compare the results, incidences of four-word bundles that were considered proper nouns (such as book titles) were removed. Furthermore, as the size and word count of the corpora (given in Table 1) turned out to be too disparate, normalization needed to be undertaken.

Table 1. The two corpora used for the empirical analysis.

Corpora	Number of texts	Words
EALE (EFL)	75	589,633
BAWE (L1, BrE)	41	348,800

The frequency counts for both corpora were normalized to a frequency per 100,000 words. This was done in order to be able to compare the results to Hasselgård's (2019) study. A frequency per 100,000 words was calculated by multiplying the absolute frequency (the number of times a certain prefab occurred in the texts) by 100,000 and then dividing by the total word count of the corpus it appeared in.

Additionally, prefabs were categorised by function to further explore the possible reasons for their frequent usage. Categorisation was done by using the study conducted by Biber *et al.* (2004) as a guideline. In their paper, the functional use of prefabs was divided into three primary categories:

1. Stance expressions (S) – "/…/express attitudes or assessments of certainty/…/" (Biber et al. 2004: 384) (e.g. the fact that the, it is important to etc.)
2. Discourse organizers (D) – "/…/reflect relationships between prior and coming discourse." (ibid.) (e.g. if you look at, as well as the etc.)
3. Referential expressions (R) – "/…/make direct reference to physical or abstract entities, or to the textual context itself/…/" (ibid.) (e.g. is one of the, the nature of the etc.)

Biber *et al.* (2004) note that the primary categories can be further divided into subcategories based on more specific functions and meanings. However, in this thesis, classification was retained to the three broader categories.

2.2 Corpus analysis

After running the texts from both L1 and EFL corpora through AntConc, the total amount of different types of prefabs in the EALE corpus was 376 and in the BAWE corpus, 93.

However, some bundles were removed from EALE corpus as they did not meet the requirements to be considered prefabs. Thus, the total amount of different prefabs in the EALE corpus ended up being 361. The number of total tokens of prefabs in EALE was 7,103 (1,205 per 100,000 words), in BAWE it was 1,649 (472 per 100,000 words). It should be kept in mind that the size of the EALE corpus is almost twice as big as the BAWE corpus. If we look at the type/token distribution of the prefabs in the present sample, the proportion is almost the same - 0.05 (361/7,103) in EALE and 0.06 (93/1,649) in BAWE. In order to conclude anything more substantial about the general differences between the use of different prefabs in academic writing produced by native speakers and Estonian learners of English, a different approach should be taken. Currently, the minimum frequency for a prefab to be included in the study was set to 10.

Out of the total amount of prefabs, the top ten four-word prefabs, according to a frequency per 100,000 words for both corpora were examined in more detail (listed in Table 2). Prefabs that are shared between both corpora are marked in shaded cells. Immediately, it is possible to make some parallels with Hasselgård's (2019) study as two of the most frequently used prefabs in English academic writing *(in the case of* and *on the other hand*), as claimed by Biber *et al.* (1999: 994), also occur in the results of the present thesis. Biber *et al.* (1999: 994) also note that both prefabs are the only four-word bundles that exceed over 100 hits per million words. Similar results can be found in Simpson-Vlach and Ellis's (2010) Academic Formulas List (AFL), in which *in the case of* has a frequency of 135 per million words and *on the other hand* a frequency of 119 per million words, thus making them the most commonly used four-word prefabs in written academic English.

Table 2. Top ten four-word prefabs in EALE and BAWE corpora (frequencies per 100,000 words)

Prefabs in EALE	Function	Freq.	Prefabs in BAWE	Function	Freq.
as well as the	D	20.7	the way in which	R	18.1
the end of the	R	19.1	on the other hand	D	16.3
in the case of	D	18.3	the end of the	R	15.5
at the same time	R	15.9	as a result of	D	14.3
on the other hand	D	15.9	it is possible to	S	12.9
in the context of	R	15.3	the fact that the	S	11.5
it is important to	S	13.2	the use of the	R	10.3
is one of the	R	12.4	in the case of	D	9.5
one of the most	R	12.0	one of the most	R	8.3
the analysis of the	R	12.0	the extent to which	R	8.3

Initially, the seventh spot from EFL students' texts was occupied by 'the handmaid s tale'. However, as it is a proper noun, it was removed from the list. It should be noted that the software made some interesting decisions regarding the way it picked n-gram bundles. For example, the suffix "-'s" was considered as a separate word. Moreover, the suffix 's' was the only bound morpheme that the software considered as a standalone word. It is also interesting to note that the software did not consider numbers (if they were written in Arabic numerals) as standalone words. For example, let us take the initially confusing bundle *the s and s*. At first glance, this bundle was incoherent and therefore, needed to be checked in the context of the texts. AntConc has a feature that allows the user to view all instances of a certain bundle in context by clicking on it. After checking the context, it turned out that the 's' indicated decades (e.g. *the 1970s and the 1980s*), which again, shows that the software considers the suffix 's' as

a standalone word. This begs the question whether bound morphemes should be considered separate from lexical words or not. In this thesis, such bundles were disregarded as they were either at the bottom of the list or did not meet the requirements to be considered prefabs.

There were also instances of overlapping. Originally, both the bundles *the end of the* and *at the end of* appeared quite high on the list. However, after checking the contexts for both bundles, it was deemed that *at the end of* occurred too frequently within *the end of the*, creating the five-word prefab *at the end of th*e. Additionally, there was hardly any variation for the bundle *at the end of* as there were only twelve instances for Estonian EFL students and eighteen for native English speakers where *at the end of* was not proceeded by the definite article *the*. Moreover, in the Estonian EFL students' texts, five out of twelve instances were grammatical errors as, in context, the article *the* should have occurred but was missing. For example, "[a]lthough at the end of novel/…/" (BA_0023) and "/…/at the end of 18th century." (BA_0057). Therefore, due to its infrequency, the bundle *at the end of* was removed. Arguably, *is one of the* and *one of the most* also seem as if they would overlap but, after checking the contexts, there was much greater variation between these prefabs which is why neither of them were removed.

Regarding function, the results between Estonian EFL learners and native English speakers were quite similar (see Table 2). The top ten list of the EALE corpus contained one stance expression, three discourse organizers, and six referential expressions, whereas the BAWE corpus's results were comprised of two stance expressions, three discourse organizers, and five referential expressions. Yet, they differed in distribution. In the EALE corpus, all three discourse organizers resided at the top of the list, while in the BAWE corpus, functions were more evenly distributed. Stance expressions occurred the least out of the three functional

classifications which also corresponds with Hasselgård's (2019) findings. This is likely due to the fact that academic texts in general are written using a neutral tone, meaning, expressing attitudes or assessments is kept to a minimum. In addition, the results of the study done by Biber *et al.* (2004: 396) showed that stance expressions occurred more commonly in spoken discourse as oppose to academic prose.

In regard to frequency, it is noticeable that the most common prefabs in the EALE corpus have an overall higher frequency than the ones in BAWE (see Table 2). This also holds true for the prefabs shared between both corpora with the only exception being *on the other hand* which was more frequently used by native English speakers (further examined in section 2.3.1). It is also interesting to note that frequencies declined faster in the BAWE corpus, suggesting that native English speakers' use of prefabs is more varied than that of Estonian EFL learners. However, as discussed in section 1.2.1, frequency does not always yield the most accurate results which is why the most frequent prefabs in both corpora were also checked by distribution among texts (given in Table 3). Distribution percentage was calculated by dividing the total number of different texts containing at least one instance of a certain prefab by the total number of texts in a corpus and multiplying by 100. Note that the order of the most frequent prefabs differs between Table 2 and Table 3.

Table 3. Top ten four-word prefabs in EALE and BAWE corpora (distribution across texts)

Prefabs in EALE	Texts	Prefabs in BAWE	Texts
the end of the	61.3%	the end of the	51.2%
one of the most	61.3%	on the other hand	43.9%
as well as the	58.7%	the fact that the	43.9%
in the case of	58.7%	the use of the	39.0%
it is important to	56.0%	as a result of	36.6%
at the same time	54.7%	in the case of	36.6%
is one of the	54.7%	the way in which	31.7%
on the other hand	49.0%	one of the most	29.3%
the analysis of the	46.7%	the extent to which	29.3%
in the context of	41.3%	it is possible to	26.8%

Based on distribution, prefabs also occurred more frequently across text in the EALE corpus which once more suggests that Estonian EFL students have a tendence to overuse certain prefabs in comparison to native English speakers. Seven out of ten of the most common prefabs in EALE appeared in over half of the text, while only one occurred in the BAWE corpus. Moreover, there is once again a steeper decline in frequency in the BAWE corpus than in the EALE corpus which indicates that the native English speakers' use of prefabs is more diverse than that of Estonian EFL students.

2. 3 Case studies

In her paper, Hasselgård (2019) conducted additional case studies where she further explored some of the most frequently used bundles. Specifically, she looked at three bundles which were overused by learners of English compared to native speakers of English and one which was underused. In this thesis, similar case studies will be conducted as, based on frequency, there were four prefabs which occurred in the top ten list for both sets of data; three of these were overused and one underused by Estonian learners of English.

2.3.1 On the other hand

On the other hand is an idiomatic expression which is "used to introduce a statement that contrasts with a previous statement or presents a different point of view" (Merriam-Webster dictionary n.d.). It is sometimes preceded by the phrase *on the one hand*, however, according to Byrd and Coxhead (2010: 46) it is frequently used as an independent transition and contrast marker. As such, it is one of the most widely used prefabs in academic written English as it helps to structure academic texts. See examples 1a-2b how this prefab is used in the two corpora.

(1a) EALE: "On the other hand, it has been argued by/…/" (BA_0003)

(1b) EALE: "Some women, on the other hand, defected from/…/" (BA_0006)

(2a) BAWE: "On the other hand, signifying an overriding difference/…/" (text 0129)

(2b) BAWE: "The British on the other hand did run into problems/…/" (text 0280)

This is also brought up in Hasselgård's (2019) study as it was the most overused bundle for Norwegian EFL users. In contrast, for Estonian EFL learners it was the only expression occurring in the top ten list for both corpora that was underused. Still, the underuse was quite small (only a difference of 0.4). Moreover, based on distribution the prefab *on the other hand* appeared more frequently in EALE than BAWE texts (5.1% difference). Therefore, although *on the other hand* occurred slightly more frequently in BAWE, it was still used more by Estonian EFL learners based on distribution. A possible explanation for this is that it is due to phraseological teddy bears, a phenomenon where foreign language learners cling to certain prefabs because they are familiar with them and therefore, have deemed them safe to use. However, phraseological teddy bears are not exclusive to non-native speakers. According to

Hasselgren (1994), this phenomenon is also observable in native speakers' use of their L1, however it occurs more often in learner language. Nevertheless, this provides and explanation as to why, by frequency, the prefab *on the other hand* occurred more often in BAWE yet, less by distribution.

Another possible explanation for the discrepancy between results could be that it is due to the use of different corpora. However, this could also be influenced by how much Estonian EFL learners use corresponding expressions in Estonian. In her paper, Hasselgård (2019) explores this idea by comparing the prefabs that were used by Norwegian EFL students in English to their L1 counterparts. This was done using the *English-Norwegian Parallel Corpus* (ENPC) and the KIAP corpus, which both contain research articles in Norwegian. However, as there are currently no equivalent corpora for Estonian academic writing, it was impossible to draw any definitive conclusions on how Estonian EFL students' L1 affects the use of the bundle *on the other hand* as well as subsequent prefabs.

2.3.2 In the case of

Similarly to *on the other hand*, *in the case of* is used as a discourse organizer. It is primarily used to establish a main topic, to introduce additional points or to refer to a specific example, akin to bundles such as *in regard to, in reference to, in the matter of* etc. As such, it is useful for structuring academic texts. Furthermore, it is one of the most commonly used prefabs in academic prose which also explains its high frequency in the results of this thesis. See examples 3a-4b how this prefab is used in the two corpora.

(3a) EALE: "However, in the case of poetry translation, using/…/" (BA_0009)

(3b) EALE: "In the case of Elinor Dashwood, her most/…/" (BA_0064)

(4a) BAWE: “In the case of fast mapping, learning occurs/…/” (text 6067)

(4b) BAWE: “/…/use of language, for example in the case of tag questions.” (text 6120)

Regarding frequency, *in the case of* had the largest frequency difference out of the four prefabs that occurred in the top ten list of both corpora (8.8 higher in EALE). The reason for this discrepancy was difficult to discern. However, a possible reason could be that, as native English speakers have innate fluency, they are able to create more complex grammar structures to illustrate their ideas and continue discourse, whereas Estonian EFL learners are bound to the grammar structures they have learned, thus an over-reliance on certain discourse organizers.

2.3.3 The end of the

The prefab *the end of the* is quite straightforward in it meaning as it literary refers to the end of something. In the case of this thesis, it was the most frequently occurring prefab in both corpora by distribution. A possible explanation for this is that both corpora contain a sufficient amount of texts which discuss literature, drama or film in some way. Evidently, in both corpora over half of the instances of the prefab *the end of the* were proceeded by nouns such as book/novel/film etc. Otherwise, it was mostly used to refer to time (e.g. the end of the 19th century) or linguistics (e.g. the end of the third person plural). This indicates that both corpora show a bias towards certain prefabs (ones which often occur in texts related to English language and literature). Therefore, the high frequency of the prefab *the end of the* is most likely due to the specific data sets used in this thesis. See examples 5a-6b how this prefab is used in the two corpora.

(5a) EALE: “At the end of the novel, Nick is revealed/…/” (BA_0028)

(5b) EALE: "/…/chronologically taking place at the end of the story." (BA_0002)

(6a) BAWE: "The first turn appears at the end of the second stanza/…/" (text 3110)

(6b) BAWE: "By the end of the 16th century/…/" (text 0261)

2.3.4 One of the most

One of the most is used to refer to someone or something that is deemed to be one of the most in a certain grouping. This can be based on facts or depend on the writer's own personal opinion. The prefab is usually followed by an adjective, an adverb + verb or a participle that establishes the category that the noun falls under (e.g. one of the most influential authors). See examples 7a-8b how this prefab is used in the two corpora.

(7a) EALE: "One of the most prominent and memorable scenes/…/" (BA_0007)

(7b) EALE: "Her article is one of the most recent and thorough studies/…/" (BA_0063)

(8a) BAWE: "/…/it starts with one of the most clichéd openings/…/" (text 3066)

(8b) BAWE: "One of the most integral parts of the /…/" (text 6020)

Based on frequency, it was the least regularly occurring prefab that was shared between both corpora. However, in the EALE corpus, *one of the most* was tied with *the end of the* as the most recurring bundle by distribution (in 61.3% of texts) which was significantly more frequent than in BAWE (only in 29.3% of texts). The reason for this disparity remained undetermined. Nonetheless, a possible explanation for its high frequency in both corpora can be attributed to the specificity of the data sets. As *one of the most* is quite ambiguous in its meaning, it can be used in humanities to give general contexts to something or someone. One the other hand, its

usage is likely less common in disciplines which favour directness and accuracy (e.g. formal sciences). However, additional research is needed to test the validity of this claim.

DISCUSSION

According to the results of this thesis and the findings of Hasselgård's (2019) study, it was evident that EFL students overused certain prefabs in comparison to native English speakers. However, in contrast to Hasselgård's (2019) results, the EALE corpus did not have any prefabs that were significantly higher in frequency than the rest. This shows that while Estonian EFL students do overuse certain prefabs, there is still a considerable amount of variation, similarly to native English speakers. The difference in the results of the present thesis and Hasselgård's (2019) study could be because of the divergence of corpora but it could also be because English and Norwegian are closely related languages (both belong in the Germanic languages branch). In her paper, Hasselgård (2019) looks at possible Norwegian equivalents to the prefabs highlighted in the case study and found that the most overused bundles had very similar Norwegian counterparts which is likely why Norwegian EFL students tended to overuse them. A similar analysis could potentially be done with Estonian as well. For example, a similar Estonian phrase to the prefab *on the other hand* is *teisest küljest*. The next step would be then to examine how much do Estonian students use the bundle *teisest küljest* in academic texts written in their native language and if that affects the way they choose corresponding prefabs in English. However, this would require an academic corpus comprised of texts in Estonian which, as of writing this thesis, is not available. Therefore, examining how much are equivalent phrases used in Estonian and whether they influence EFL students' decision making can be a possible research topic in the future.

Regarding distribution, prefabs shared between corpora were more frequent among Estonian EFL texts. This once again contrast with Hasselgård's (2019) results where she concluded that the reason for the smaller distribution rates was because Norwegian EFL users tended to have varying phraseological teddy bears (i.e. prefabs that they overuse). On the other hand, Estonian EFL learners tended to be more uniform in their use of prefabs. The reason for this difference remained unclear based on the results, yet a possible explanation could be that Estonian EFL students are uniformly taught the same prefabs. As early acquisition of prefabs influences how they are used later on, determining what kind of bundles are taught in schools would provide a better understanding of why certain prefabs become overused. However, proving the validity of this explanation would require an additional study which analyses EFL materials used in Estonian schools.

Based on the case studies, it appeared that referential bundles tended to be more biased towards the source texts that discourse organizers. Discourse organizers are used to structure texts and bridge a connection between prior and coming discourse which is why they are widely usable and unlikely to be biased towards any particular type of genre. In addition, they have a rigid structure and usage which means that once they are learnt they are unlikely to be used incorrectly. These aspects of discourse organizers are similar to the characteristics of phraseological teddy bears which is likely why this function type is frequently overused by EFL learners. That said, referential bundles still occurred the most out of the three function types. This is in line with Biber *et al.*'s (2004: 398) findings which showed that academic prose mostly consists of referential bundles. This is reasonable since, as stated before, discourse organizers have a distinct usage and placement in texts which means they can not be used as often throughout a single text as referential bundles and stance bundles.

Another finding of this thesis was that although the sample text used for native and non-native English speakers were from different genres (thesis vs essays), the results of the top ten four-word bundles shows that there are some prefabs which frequently occur across both types of academic texts. As such, these prefabs would be beneficial to teach to EFL students as they have a wide range of usage. That said, in order to validate what kind of prefabs should be taught to EFL learners, results should also be compared to expert academic writing as it utilises more diverse grammar structures and is less likely to contain errors, thus making it a better learning target. Additionally, the topic of this thesis can be expanded upon by using the same methods to analyse and compare texts from different disciplines (e.g. arts and humanities vs natural sciences etc.) to see if there are any significant changes in the prefabs used. Doing so provides a more accurate depiction of how prefabs are used in general, as opposed to a single discipline. Additionally, combining all the results of the different disciplines together would highlight the most frequently used prefabs in novice academic English.

The prefab *as well as the* is also worthy of interest as it was the most frequently occurring prefab in the EALE corpus. Although three-word bundles were not the focus of this thesis, it is interesting to note that the three-word cluster *as well as* was also number one by frequency out of its respective bundle size. Moreover, both of these prefabs showed significant over usage by Estonian EFL learners compared to native English speakers. *As well as the* had an over usage of 14.7 and *as well as* a massive difference of 50.3 between corpora. In contexts, both of these prefabs were used similarly to how the conjunction *nii… kui ka* is used in Estonian which is a possible reason for their popularity among Estonian EFL learners. Another possible explanation as to why *as well as* was vastly overused by Estonian EFL learners is that in Estonian there are two equivalent words for the conjunction *and* in the forms of *ja* and *nagu*. Although it is not a

rule per se, in Estonian it is good practice not to use *ja* twice in a row (unlike in English where the usage of *and* is more relaxed). Thus, a second *ja* is often substituted with *nagu*. In English, the closest construction to fill the function of *nagu* is *as well as* which is likely why it is overused by Estonian EFL learners. However, it should be noted that *and* and *as well as* are not equivalents of each other which also opens up the possibility that, to some extent, Estonian EFL learners use *as well as* incorrectly. Still, in order to be able to make any definitive conclusions on this matter, further research is needed to be done. Therefore, the usage of *as well as* and/or *as well as the* could be a possible topic for a future study.

CONCLUSION

Based on a multitude of studies, it is generally believed that language is made up from prefabricated expressions that are collectively known as formulaic language. Formulaic language has been researched in order to obtain a better understanding of the qualities in a native speaker's use of their language which gives it its native-like fluency and how non-native speakers learn and utilise these features. While there are plenty of studies done regarding the use of prefabs in English by learners with varying mother tongues (and of native English speakers themselves), there is virtually no data on how Estonian EFL learners use these patterns. Therefore, the thesis at hand intended to provide insights into this matter. In the beginning of the thesis, an initial hypothesis was proposed based on Hasselgård's (2019) results, which was that due to the phenomena known as phraseological teddy bears, Estonian EFL users were more likely to exhibit an over-reliance on certain prefabs than native English speakers. Additionally, whether Estonian EFL learners would exhibit similar trends to the Norwegian EFL learners was also investigated.

The main question of this thesis was: How does the use of prefabs in academic writing differ between Estonian EFL learners and native English speakers? Based on results of the thesis at hand, 361 types of different prefabs occurred in the EALE corpus and 93 in the BAWE corpus. The number of total tokens of prefabs in EALE was 7,103 (1,205 per 100,000 words), while in BAWE it was 1,649 (472 per 100,000 words). By examining the type/token distribution of the prefabs in the sample, the proportion size turned out to be quite similar (0.05 in EALE and 0.06 in BAWE), yet no other conclusions were able to be drawn based on these results. Thus, concluding anything more substantial about the general differences between the use of different prefabs in academic writing produced by native speakers and Estonian learners of English would require a different approach to the one used in the present thesis.

Still, it was evident that Estonian EFL users tended to overuse specific prefabs in comparison to native English speakers as the frequencies in the top ten list of the EALE corpus were overall higher than those in the BAWE corpus. Nevertheless, compared to Hasselgård's (2019) study, there were no prefabs which were significantly more overused than the others which indicates that Estonian EFL learners' use of prefabs still contained substantial amount of variation, yet less than that of native English speakers. However, the present thesis only gives an overview of the type and frequency of prefabs that occurred in the BA theses of the Department of English Studies and not how prefabs are generally used among Estonian EFL users in the context of written academic English. Therefore, the topic of this thesis would benefit from further studies done regarding various disciplines. Combining the results of different disciplines together would highlight the most commonly used prefabs in novice academic English which, by also comparing it to expert academic writing, can determine the most

valuable prefabs to teach to EFL learners to improve their comprehension and language production skills.

The distribution of prefabs among texts also differed from Hasselgård's (2019) results. Estonian EFL learners tended to collectively overuse similar prefabs, while Norwegian EFL learners' usage was more individualistic. The reason for this difference was unclear, however, a possible cause could be that the Estonian school curriculum has a uniform way of teaching English. Still, to test the validity of this claim would require research into EFL materials used in the Estonian education system.

Additional case studies were carried out to further examine bundles that were shared between the top ten list of both corpora. Possible reasons for their overuse were discussed. However, due to a lack of a corpus containing texts of written academic Estonian, possible influences of Estonian EFL learners' L1 in their choice of using certain prefabs remained undetermined. Thus, the research on the use of formulaic language among Estonian EFL learners would benefit from a corpus consisting of academic texts written in Estonian, to be able to make comparisons between the two languages. Furthermore, one of the case studies also demonstrated the issue with solely using a frequency-based approach in corpus-based analyses. Based on frequency, the prefab *on the other hand* was used more by native English speakers. Yet, by also examining its distribution among texts, it was determined that the prefab was used more by Estonian EFL users, therefore, indicating the importance of analysing prefabs beyond frequency alone.

In regard to the functional use of prefabs, discourse organizers had some of the highest frequencies in the EALE corpus (first, third and fifth highest prefabs by frequency). Still,

referential bundles were the most frequent function type among texts. They also proved to be used more in terms of distribution (also exhibited in Biber *et al.* (2004) findings), yet as examined in the case studies, this may have been due to the specificity of the data sets used in the analysis of this thesis. Nevertheless, prefabs used as discourse organizers at least proved to be overused by Estonian EFL learners in comparison native English speakers.

Based on the sample used in this thesis, it can be concluded that Estonian EFL users tend to overuse certain prefabs and functions of prefabs compared to native English speakers. However, there is a possibility that the reason for these differences was due to the use of different academic texts (BA theses vs essays) as the standards for academic texts might be quite different between countries. Therefore, the present thesis would benefit from being able to compare the use of prefabs in the BA theses of Estonian EFL learners and the BA theses written by native English speakers, as it would provide more data to complement the results of this paper.

Overall, the current thesis serves as a starting point into how Estonian EFL users utilise prefabs in written academic English. Although the present thesis only focused on prefabs within the context of disciplines pretraining to the Department of English Studies, the topic of this paper can be further expanded on by research questions and shortcomings drawn from the findings. These findings alongside subsequent results of future studies can provide a clearer understanding of how prefabs are used among Estonian EFL learners and subsequently, how this information can be used to improve second- and foreign language acquisition.

REFERENCES

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word combinations. In Annette Cowie (ed.), *Phraseology: Theory, Analysis and Applications*, 101–122. Oxford: Oxford University Press.

Biber, Douglas and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes*. 26:3 263 –286.

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

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad and Viviana Cortes. 2004. 'If you look at...': Lexical bundles in university teaching and textbooks. *Applied Linguistics.* 25: 371–405.

Bolinger, Dwight. 1976. Meaning and Memory. *Forum Linguisticum*. 1: 1-14.

Byrd, Pat and Averil Coxhead. 2010. On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*. 5: 31-64.

Corpas Pastor, Gloria, Johanna Monti, Violeta Seretan and Ruslan Mitkob (eds.). 2016. *Workshop Proceedings Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2015)*. Geneva: Editions Tradulex.

Erman, Britt and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text*. 20: 1, 29-61.

Foster, Pauline. 2001. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In Martin Bygate, Peter Skehan, and Merrill Swain (eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing*, 75-93. Harlow: Longman.

Granger, Sylviane. 1998. PREFABRICATED PATTERNS IN ADVANCED EFL WRITING: COLLOCATIONS AND LEXICAL PHRASES. In Anthony P. Cowie (ed.) *Phraseology: Theory, Analysis and Applications*, 145-160. Oxford: Clarendon Press.

Hasselgård, Hilde. 2019. Phraseological teddy bears: frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In M. Mahlberg and V. Wiegand (eds), *Corpus Linguistics, Context and Culture*, 339-362. Berlin: De Gruyter.

Hasselgren, Angela. 1994. Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*. 4: 2.

Heuboeck, Alois, Jasper Holmes and Hilary Nesi. 2010. The BAWE Corpus Manual. Available at https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/, accessed April 26, 2021.

Ho, Don. 2020. Notepad++. Available at https://notepad-plus-plus.org, accessed April 12, 2021.

Howatt, Anthony P.R. 2004. *A History of English Language Teaching* (second edition). Oxford: Oxford University Press.

Jespersen, Otto. 1924. *The philosophy of grammar*. London: Allen and Unwin.

Laurence, Anthony. 2019. AntConc (Version 3.5.8). [Windows]. Tokyo, Japan: Waseda University. Available at https://www.laurenceanthony.net/software, accessed April 12, 2021.

Merriam-Webster. n.d. On the other hand. Available at https://www.merriam-webster.com/dictionary/on%20the%20other%20hand, accessed May 8, 2021.

Pawley, Andrew and Frances Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards and Richard W. Schmidt (eds.) *Language and Communication* 7:1, 191-226. London: Longman.

Piiri, Andreas. 2020. A corpus based study of formulaic language use by native and non-native speakers. Available at http://hdl.handle.net/10062/69933, accessed April 17, 2021.

Reppen, Randi and Rita Simpson-Vlach. 2010. Corpus Linguistics in: Schmitt, Norbert (ed.) *An Introduction to Applied Linguistics* (2nd ed.). London: Hodder & Stoughton.

Schmitt, Norbert, Sarah Grandage, and Svenja Adolphs. 2004. Are corpus-derived recurrent clusters psycholinguistically valid. In Norbert Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use*, 127-151. Amsterdam: John Benjamins Publishing.

Schmitt, Norbert. 2010. *Researching Vocabulary: A Vocabulary Research Manual*. London: Palgrave Macmillan.

Simpson-Vlach, Rita and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*. 31: 4, 487-512

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Wood, David. 2015. *Fundamentals of formulaic language: An introduction*. London and New York: Bloomsbury.

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

APPENDICES

Appendix 1

The list of different types of prefabs that occurred in the EALE corpus sorted by frequency.

Rank	Freq	Range	Prefab	Rank	Freq	Range	Prefab
1	122	44	as well as the	42	31	5	the year of the
2	113	46	the end of the	43	30	23	can be seen as
3	108	44	in the case of	44	30	18	the rest of the
4	94	41	at the same time	45	29	22	gives an overview of
5	94	37	on the other hand	46	29	11	in front of the
6	90	31	in the context of	47	29	19	it can be said
7	78	42	it is important to	48	29	18	it can be seen
8	73	41	is one of the	49	29	18	the use of the
9	71	46	one of the most	50	28	13	by the fact that
10	71	35	the analysis of the	51	28	18	can be said that
11	70	30	when it comes to	52	28	19	in the form of
12	69	43	the aim of this	53	28	18	the case of the
13	68	25	of the th century	54	28	23	the first part of
14	68	36	the beginning of the	55	27	16	i am going to
15	63	29	can be seen in	56	27	6	of the source text
16	55	35	an overview of the	57	27	17	the majority of the
17	52	28	part of the thesis	58	27	26	the thesis consists of
18	48	30	the context of the	59	26	17	for the first time
19	47	29	it is possible to	60	26	20	in this case the
20	46	26	in addition to the	61	25	18	as a result of
21	44	26	at the beginning of	62	25	15	end of the novel
22	44	31	of this thesis is	63	24	20	a part of the
23	44	27	the fact that the	64	24	15	can be found in
24	42	21	in the th century	65	24	19	it is clear that
25	41	27	on the basis of	66	24	16	of the novel and
26	39	16	the results of the	67	24	17	the second part of
27	38	27	aim of this thesis	68	23	15	can be used to
28	38	17	the meaning of the	69	23	21	first part of the
29	38	30	this thesis is to	70	23	18	for example in the
30	37	24	with the help of	71	23	13	in the empirical part
31	36	23	be seen in the	72	23	9	in the united states
32	36	25	one of the main	73	23	15	part of this thesis
33	36	15	we can see that	74	23	7	the order of the
34	34	18	in a way that	75	23	10	used to refer to
35	34	26	the aim of the	76	22	19	as one of the
36	34	24	to the fact that	77	22	17	as well as a
37	33	17	the empirical part of	78	22	13	for the sake of
38	32	8	in the estonian translation	79	22	16	in the novel the
39	31	20	as can be seen	80	22	14	it is possible that
40	31	9	in the target language	81	22	15	one of the reasons
41	31	16	of the novel the	82	22	18	that there is a

Rank	Freq	Range	Prefab	Rank	Freq	Range	Prefab
83	21	14	beginning of the novel	132	17	10	in the use of
84	21	15	for the purpose of	133	17	16	of the most important
85	21	17	is based on the	134	17	12	points out that the
86	21	17	is to find out	135	17	7	the early th century
87	21	12	of the present thesis	136	17	13	the structure of the
88	21	10	the history of the	137	17	16	thesis focuses on the
89	21	9	the translation of the	138	17	14	was one of the
90	21	6	the world of the	139	16	11	in the middle of
91	20	8	in the estonian language	140	16	13	is important to note
92	20	12	in the first place	141	16	14	is the fact that
93	20	13	in the sense that	142	16	10	of the novel as
94	20	13	is considered to be	143	16	11	of the use of
95	20	15	it is difficult to	144	16	12	seems to be the
96	20	18	of the thesis is	145	16	14	the basis of the
97	20	16	one of the first	146	16	9	the fact that she
98	20	14	second part of the	147	16	13	the importance of the
99	20	10	the time of the	148	16	5	the most frequently used
100	20	5	the use of this	149	16	11	this can be seen
101	19	17	does not have a	150	16	13	will focus on the
102	19	13	empirical part of the	151	15	6	at the university of
103	19	11	end of the book	152	15	8	can also be used
104	19	19	findings of the thesis	153	15	11	can be considered as
105	19	16	in order to make	154	15	10	can be seen that
106	19	15	of the novel is	155	15	10	for the purposes of
107	19	16	the author of the	156	15	14	in the beginning of
108	19	16	the first chapter of	157	15	9	in the novel is
109	19	11	the length of the	158	15	6	in the original text
110	19	7	to make sense of	159	15	11	it is necessary to
111	19	11	to refer to the	160	15	10	it is not possible
112	18	12	at the time of	161	15	15	main findings of the
113	18	12	be said that the	162	15	6	native speakers of english
114	18	15	can be seen from	163	15	15	of the thesis introduction
115	18	13	despite the fact that	164	15	13	of the thesis the
116	18	12	does not seem to	165	15	13	on the analysis of
117	18	13	due to the fact	166	15	11	on the one hand
118	18	12	for the analysis of	167	15	10	the focus of the
119	18	12	important to note that	168	15	15	the main findings of
120	18	12	in order to find	169	15	11	the middle of the
121	18	14	in order to understand	170	15	10	the novel and the
122	18	12	in the literature review	171	15	12	the second half of
123	18	9	in the present thesis	172	15	6	the source text and
124	18	11	on the topic of	173	15	14	thesis is to analyse
125	17	10	a short overview of	174	15	9	to the use of
126	17	14	aim of the thesis	175	15	7	with the use of
127	17	12	as well as in	176	15	5	written and spoken language
128	17	13	can also be seen	177	14	7	a corpus based study
129	17	16	first chapter of the	178	14	12	an important role in
130	17	13	in the first chapter	179	14	14	and the conclusion the
131	17	8	in the light of	180	14	11	but at the same

Rank	Freq	Range	Prefab	Rank	Freq	Range	Prefab
181	14	11	by the end of	231	13	9	to do with the
182	14	12	chapter of the thesis	232	13	6	to refer to a
183	14	9	could be said that	233	13	6	turns out to be
184	14	11	does not mean that	234	12	7	are used in the
185	14	8	does not want to	235	12	10	as a way of
186	14	7	during the th century	236	12	8	as a way to
187	14	8	in my thesis i	237	12	10	be seen as a
188	14	9	it could be said	238	12	5	can see that the
189	14	10	it is evident that	239	12	8	characters of the novel
190	14	7	it is used to	240	12	5	context of the novel
191	14	10	it should be noted	241	12	9	could be interpreted as
192	14	10	of this paper is	242	12	11	for a long time
193	14	13	of this thesis the	243	12	10	in order to get
194	14	10	should be noted that	244	12	7	in the english language
195	14	11	the development of the	245	12	7	in the real world
196	14	6	the estonian translation of	246	12	7	is not possible to
197	14	9	the role of the	247	12	7	it is interesting to
198	14	10	the story of the	248	12	10	it is not a
199	14	11	there is also a	249	12	9	it seems that the
200	14	11	this thesis focuses on	250	12	11	not seem to be
201	14	9	through the eyes of	251	12	10	of the novel in
202	14	11	to find out the	252	12	9	of the reasons why
203	14	12	to look at the	253	12	11	of the thesis will
204	14	8	was first published in	254	12	6	out of all the
205	13	11	a brief overview of	255	12	5	the characters of the
206	13	13	and a conclusion the	256	12	8	the course of the
207	13	9	aspects of the novel	257	12	8	the creation of the
208	13	6	be explained by the	258	12	9	the differences between the
209	13	12	be seen from the	259	12	11	the fact that they
210	13	5	both positive and negative	260	12	10	the focus of this
211	13	9	do not have a	261	12	10	the novel in the
212	13	12	give an overview of	262	12	10	the other hand the
213	13	11	in addition to that	263	12	10	the present thesis is
214	13	11	in the second part	264	12	10	the purpose of this
215	13	8	is also important to	265	12	11	the thesis ends with
216	13	12	is to analyse the	266	12	10	thesis is divided into
217	13	11	is used as a	267	12	9	this means that the
218	13	9	it can be concluded	268	12	10	was used in the
219	13	8	it is also important	269	12	8	when looking at the
220	13	10	it would have been	270	11	9	a member of the
221	13	9	provides an overview of	271	11	9	also points out that
222	13	8	she is unable to	272	11	10	an analysis of the
223	13	9	that he does not	273	11	7	be used as a
224	13	11	that there is no	274	11	7	can be concluded that
225	13	12	the purpose of the	275	11	9	can be used as
226	13	10	the relationship between the	276	11	9	considered to be a
227	13	12	this is not the	277	11	7	empirical part of this
228	13	10	this paper is to	278	11	9	from the perspective of
229	13	9	throughout the novel the	279	11	5	i would like to
230	13	12	to be able to	280	11	7	in order to see

Rank	Freq	Range	Prefab	Rank	Freq	Range	Prefab
281	11	6	in the eyes of	322	10	5	in the book the
282	11	9	in the novel as	323	10	8	in the end of
283	11	8	in the process of	324	10	9	in the novel that
284	11	10	in the same way	325	10	9	in this case is
285	11	8	in the second half	326	10	8	is a part of
286	11	10	is not the only	327	10	7	is an example of
287	11	9	it comes to the	328	10	6	is seen as a
288	11	10	it is easy to	329	10	7	is the use of
289	11	10	of the most common	330	10	8	it can also be
290	11	11	of this thesis was	331	10	7	it can be assumed
291	11	8	on the use of	332	10	5	it is true that
292	11	8	out to be the	333	10	6	nothing to do with
293	11	8	part of the novel	334	10	6	of the novel when
294	11	9	second half of the	335	10	6	of this study is
295	11	11	seems to be a	336	10	9	short overview of the
296	11	10	that it is not	337	10	9	taking into account the
297	11	8	that she does not	338	10	7	that he is a
298	11	7	the protagonist of the	339	10	9	that it is a
299	11	8	the reason for this	340	10	6	the context of this
300	11	9	the th century and	341	10	9	the events of the
301	11	6	the use of language	342	10	5	the eyes of the
302	11	5	to take care of	343	10	10	the fact that it
303	11	6	used in this thesis	344	10	10	the findings of the
304	10	6	a closer look at	345	10	10	the novel as well
305	10	6	a good example of	346	10	10	the novel does not
306	10	8	aim of this paper	347	10	7	the original and the
307	10	9	analysis is based on	348	10	7	the role of a
308	10	10	and at the same	349	10	9	the same time the
309	10	8	and the fact that	350	10	9	the thesis is divided
310	10	6	as pointed out by	351	10	6	the title of the
311	10	8	but it does not	352	10	10	thesis consists of four
312	10	7	can be assumed that	353	10	10	thesis ends with a
313	10	10	chapters and a conclusion	354	10	7	thesis will focus on
314	10	10	consists of four parts	355	10	7	this part of the
315	10	10	ends with a conclusion	356	10	10	to be the most
316	10	5	estonian translation of the	357	10	9	to find out what
317	10	9	i will focus on	358	10	6	to make the reader
318	10	8	in contrast to the	359	10	6	to the analysis of
319	10	6	in order to do	360	10	7	towards the end of
320	10	9	in terms of the	361	10	5	which can be seen
321	10	9	in the analysis of				

Appendix 2

The list of different types of prefabs that occurred in the BAWE corpus sorted by frequency.

Rank	Freq	Range	Prefab	Rank	Freq	Range	Prefab
1	63	13	the way in which	48	13	6	despite the fact that
2	57	18	on the other hand	49	13	7	in order to make
3	54	21	the end of the	50	13	6	in this essay i
4	50	15	as a result of	51	13	9	it is necessary to
5	46	18	at the end of	52	13	8	it is possible that
6	45	11	it is possible to	53	12	9	by the use of
7	40	18	the fact that the	54	12	7	due to the fact
8	36	16	the use of the	55	12	5	is that of the
9	33	15	in the case of	56	12	8	of the nineteenth century
10	29	12	one of the most	57	12	6	of the united states
11	29	12	the extent to which	58	12	6	one of the first
12	28	16	it is important to	59	12	6	that it is not
13	26	16	at the same time	60	12	5	the use of language
14	25	16	it is clear that	61	12	8	this is due to
15	25	12	the importance of the	62	12	8	to such an extent
16	24	10	the context of the	63	11	5	as can be seen
17	23	10	the ways in which	64	11	9	as one of the
18	21	13	as well as the	65	11	8	be seen as a
19	21	12	through the use of	66	11	9	can be seen as
20	20	8	can be found in	67	11	5	end of the novel
21	20	10	for example in the	68	11	8	for the first time
22	20	15	in the form of	69	11	7	in terms of the
23	20	7	it could be argued	70	11	7	in the use of
24	19	11	an example of this	71	11	8	one of the main
25	19	14	at the beginning of	72	11	8	that it is the
26	19	12	can be seen in	73	11	7	the image of the
27	19	13	to the fact that	74	11	7	the majority of the
28	18	9	a result of the	75	11	9	the structure of the
29	18	7	could be argued that	76	11	9	this can be seen
30	18	13	the beginning of the	77	11	6	with the use of
31	17	11	is an example of	78	10	6	allows the reader to
32	17	11	that there is no	79	10	9	and as a result
33	16	15	that there is a	80	10	7	and the use of
34	16	8	the nature of the	81	10	7	in contrast to the
35	16	11	the rest of the	82	10	5	in this case the
36	16	8	way in which the	83	10	6	in this way the
37	15	9	as a means of	84	10	7	is due to the
38	15	11	in an attempt to	85	10	7	on the part of
39	15	9	the meaning of the	86	10	6	the death of the
40	15	11	to be able to	87	10	8	the idea of the
41	14	11	by the fact that	88	10	9	the power of the
42	14	7	example of this is	89	10	6	the success of the
43	14	9	in the context of	90	10	7	they were able to
44	14	11	in the same way	91	10	5	to the way in
45	14	10	the role of the	92	10	8	towards the end of
46	13	11	be seen in the	93	10	5	we are able to
47	13	8	by the end of				

RESÜMEE

TARTU ÜLIKOOL
ANGLISTIKA OSAKOND

Liisi Kraak

The Comparison of the Usage of Prefabs in the Academic Writing of Estonian EFL Learners and Native English Speakers.
***Prefab*ide kasutuse võrdlus Eesti inglise keelt võõrkeelena (EFL) õppijate ja inglise keelt emakeelena kõnelejate teadustekstides.**

bakalaureusetöö
2021
Lehekülgede arv: 46

Annotatsioon:

Käesolev bakalaureusetöö uurib *prefab*ide kasutust Eesti inglise keelt võõrkeelena (EFL) õppijate ja inglise keelt emakeelena kõnelejate teadustekstides. Töö eesmärk on korpusuuringu meetodil tuvastada, kuidas erineb *prefab*ide kasutus Eesti EFL õppijate ja inglise keelt emakeelena kõnelevate autorite poolt kirjutatud teadustekstides. Töö jäljendab Hasselgårdi (2019) uuringu meetodit, mille järgi otsitakse nii õppijakorpusest kui ka inglise keelt emakeelena kõnelejate korpusest kõige sagedamini esinevaid 4-sõnalisi *prefab*e ning võrreldakse nende kasutamist. Leitud tulemusi võrreldakse Hasselgårdi (2019) järeldustega.

Töö jaguneb kahte peatükki: kirjanduse ülevaade ning empiiriline analüüs. Kirjanduse ülevaates kirjeldatakse terminoloogiat ja antakse ülevaade varasematest uuringutest sellel teemal. Empiirilise analüüsi peatükk tutvustab lähemalt analüüsis kasutatud korpuseid (EALE ja BAWE), metodoloogiat ning tulemusi. Teatud *prefab*e uuritakse lähemalt juhtumiuuringutena. Peatükk lõppeb tulemuste aruteluga.

Kokku oli EALE korpuses 361 ja BAWE korpuse 93 erinevat 4-sõnalist *prefab*i. Nendest uuriti lähemalt mõlema korpuse kümmet kõige sagedamini kasutatavat *prefab*i 100 000 sõna kohta. *Prefabid*e sagedus kui ka jaotus erinevates tekstides oli EALE korpuses üleüldiselt märgatavalt kõrgem kui BAWE korpuses, mis viitab sellele, et Eesti EFL õppijad kasutavad neid liiast. *Prefabid*e funktsiooni poolest oli nii Eesti EFL õppijate kui ka inglise keelt emakeelena rääkijate kasutus sarnane - kõige sagedamini esinev funktsiooni tüüp oli *refrencial expressions* (viitavad väljendid). Lisaks esinesid mõlema korpuse esikümnes neli sama tüüpi *prefab*i, millest Eesti EFL õppijate poolt oli kolm ülekasutatud ning üks alakasutatud. Juhtumiuuringutest tuli välja, et kõik neli *prefab*i olid siiski ülekasutatud Eesti EFL õppijate poolt. Lisaks, võis järeldada, et teatud *prefab*id esinesid sagedamini kasutatutud valimi spetsiifilisuse tõttu. Võrreldes saadud tulemusi Hasselgårdi (2019) tulemustega võis järeldada, et Eesti EFL õppijate *prefab*ide kasutus oli sageduse poolest mitmekesisem kui Norra EFL õppijate oma, kuid jaotuse poolest ühtlasem.

Märksõnad:
Inglise keel ja keeleteadus, akadeemiline inglise keel, õppijakeel, korpusuuring.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Liisi Kraak

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

The Comparison of the Usage of Prefabs in the Academic Writing of Estonian EFL Learners and Native English Speakers,

mille juhendaja on Jane Klavan,

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Liisi Kraak
25.05.2021